University of Reading

Evaluating the consistency of ensemble forecasts

A thesis submitted for the degree of Doctor of Philosophy

Department of Geography & Environmental Science

David Shaw Richardson

September 2024

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

David Richardson

Abstract

Ensemble forecasts play an essential role in providing early warnings to mitigate the impact of hazardous weather events. However, there are still many areas where ensemble information is not fully exploited. One key issue limiting the uptake of ensemble forecasts is the jumpiness that can sometimes occur between successive forecasts. Ensemble forecasts show the range of future weather scenarios that can occur, allowing users to make appropriate risk-based decisions. Occasionally a new forecast seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the previous ensemble. Such inconsistencies can cause users to lose confidence in the forecasting system.

This thesis aims to improve the diagnosis and understanding of these ensemble forecast inconsistencies. First, a methodology is developed to quantify the consistency between a sequence of ensemble forecasts valid for a given time, taking account of the full ensemble distribution. This enables a quantitative evaluation of the consistency or jumpiness between successive forecasts, providing insights into the relationship between jumpiness, skill and spread.

The thesis also provides practical guidance to address user concerns about ensemble jumpiness. It provides specific guidance that will enable users to make better use of the available operational ensemble tropical cyclone track and genesis forecasts. The thesis shows that evaluation of forecast consistency is complementary to the current focus on skill and ensemble spread, and that an integrated approach using both skill and consistency measures can be beneficial in evaluation of ensemble forecast performance.

Implementation of this approach at NWP centres will ensure that users have the necessary information and guidance to mitigate the impact of run-to-run jumpiness and will provide feedback to model developers on model weaknesses, complementing existing evaluation tools. The research in this thesis will help improve the utilisation of ensemble forecasts to provide early warnings of significant weather hazards, contributing to the UN Early Warnings for All initiative.

Acknowledgments

I would first and foremost like to thank my supervisors, Hannah Cloke and John Methven. Their support, advice and guidance throughout my PhD was essential in steering me through the last six years. I am extremely grateful to my line manager Florian Pappenberger and ECMWF Director-General Florence Rabier for enabling me to embark on this PhD while working at ECMWF and who helped me to fit the PhD work around my ECMWF duties. Special thanks to Florian for his constant support, encouragement and many insightful discussions on ensemble forecasting in general and jumpiness in particular. Helen Titley and Ervin Zsótér are owed a special mention for being such helpful PhD buddies at various stages of my journey.

I would also like to thank my friends and colleagues in the Evaluation Section, and across ECMWF, who have helped to develop my thinking on ensemble forecasting over the years. Particular thanks are due to Linus Magnusson and Sharan Majumdar for encouraging me to explore the jumpiness of ensemble tropical cyclone forecasts and for sharing their expertise on tropical cyclones. I have also enjoyed being part of the Water@Reading research group at the University of Reading; thanks for many interesting discussion on a wide range of subjects.

I am also thankful to friends and family who have been a constant support through the PhD. I am incredibly grateful to my family: to my devoted parents, Alex and Norma Richardson, my wonderful wife Christine, daughter Natalie and son Aidan who have had to suffer the consequences of me being busy and stressed with work and the PhD and who have always supported me enthusiastically every step of the way, regardless.

A huge thank you goes to you all!

Contents

List of acronyms

AC	Anomaly correlation
AEW	African Easterly Wave
AIFS	Artificial Intelligence/Integrated Forecasting System
AIFS-ENS	AIFS ensemble forecast
AT	Along-track
BLO	Scandinavian Blocking
BS	Brier score
BSS	Brier skill score
CAPE	Convective available potential energy
CDF	Cumulative distribution function
CRPS	Continuous Ranked Probability Score
CRPSS	Continuous Ranked Probability Skill Score
CT	Cross-track
CTRL	Ensemble control forecast
DI	Divergence Index
ECMWF	European Centre for Medium-Range Weather Forecasts
EFI	Extreme Forecast Index
EM	Ensemble mean
ENS	ECMWF global ensemble
EOF	Empirical Orthogonal Functions
EPS	Ensemble prediction system
ERA5	ECMWF Reanalysis dataset (latest version)
ERA-Interim	ECMWF Reanalysis dataset (predecessor to ERA5)
EW4All	Early Warning for All
FCS	Forecast convergence score
FFI	Flip-flop Index
GEFS	NCEP global ensemble
GDPFS	Global Data Processing and Forecasting System
HPC	High-performance computing
HRES	ECMWF high-resolution deterministic forecast
IBTrACS	International Best Track Archive for Climate Stewardship
IFS	ECMWF Integrated Forecasting System
IWTC	International Workshop on Tropical Cyclones

KS	Kolmogorov-Smirnov test
LC-EPSV	WMO Lead Centre for EPS Verification
MAE	Mean absolute error
ML	Machine learning
MOGREPS-G	Met Office global ensemble
MOS	Model output statistics
MSE	Mean square error
MSLP	Mean sea level pressure
MWU	Mann-Whitney U test
NAO	North Atlantic Oscillation
NCEP	National Centers for Environmental Prediction
NHC	National Hurricane Center
NOAA	National Oceanic and Atmospheric Administration
NW	North west
NWP	Numerical weather prediction
PIT	Probability Integral Transform
PoP	Probability of precipitation
RMSE	Root mean square error
RSMC	Regional Specialized Meteorological Centre
ROC	Relative operating characteristic
ROCA	ROC area
S2S	Sub-seasonal to seasonal
TC	Tropical cyclone
TCP	WMO Tropical Cyclone Programme
TC-PFP	Tropical Cyclone Probabilistic Forecast Products project
TCWC	Tropical Cyclone Warning Centre
TIGGE	The International Grand Global Ensemble
TS	Tropical storm
TT	Tropical transition
UBS	User Brier score
UN	United Nations
US	United States of America
UTC	Coordinated Universal Time
WMO	World Meteorological Organization
WWRP	World Weather Research Programme

Chapter 1 Introduction

1.1 Motivation and aim

The chaotic nature of the atmosphere means that numerical weather prediction (NWP) forecasts are sensitive to small changes in their initial conditions. Operational NWP centres address this by running a number of forecasts from similar starting conditions. The resulting ensemble of forecasts shows the range of future atmospheric states consistent with the known uncertainties in the initial conditions (Leutbecher and Palmer 2008; Swinbank et al. 2016).

The skill of weather forecasts has increased steadily over the years (Bauer et al. 2015; Haiden et al. 2019). This is a result of improvements to the forecast models, an increase in the number of observations, and improvements to the data assimilation which processes the observations to make the initial conditions for the forecast. Bauer et al. (2015) refer to this steady progress as the "quiet revolution" of NWP. Ensemble forecast skill has also improved through better quantification of initial condition uncertainties and representation of model uncertainties (Buizza and Richardson 2017; Buizza et al. 2008; Leutbecher et al. 2017; Swinbank et al. 2016).

Despite the increase in skill and undoubted progress in the use of ensemble forecasts, there are still many areas where the ensemble information is not fully exploited. One of the key issues limiting the uptake of ensemble forecasts is run-to-run jumpiness, when the latest ensemble run seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the earlier forecast. For example, the ensemble forecast made on Monday may predict that the coming weekend will be mild and wet over the UK, with perhaps some uncertainty about the location, timing and amount of rain. With the benefit of additional observations in the initial conditions, the updated forecast made on Tuesday should reduce these uncertainties giving more clarity about where and when the rain will occur. If instead, Tuesday's forecast changes to predict a cold dry weekend, this introduces an inconsistency with the previous forecast: if this cold weather could occur at the weekend then Monday's forecast should have included the possibility among the solutions in the ensemble. Similarly, if Monday's ensemble predicts that a tropical cyclone will make landfall at the weekend somewhere along the coast of Texas, Tuesday's forecast will be consistent if it narrows down the at-risk areas of the Texas coast; but the Tuesday forecast would be inconsistent if it changed to predict landfall in Florida which was not identified as at risk in Monday's forecast. These differences between successive forecasts valid for the same time are referred to as run-to-run consistency or jumpiness. Concerns over such inconsistencies are regularly raised in feedback from users of ECMWF forecasts (Hewson 2021, 2020). They present a significant challenge to forecast centres and can cause users to lose confidence in the forecasting system (Hewson 2020; Pappenberger et al. 2011b; McLay

2011; Elsberry and Dobos 1990; Magnusson et al. 2021; Dunion et al. 2023). A major challenge in ensemble forecasting is to understand why these inconsistencies occur and how ensemble forecast systems need to be improved to address the underlying causes (Dunion et al. 2023).

Additional significant outstanding challenges for ensemble forecasting include the prediction of high-impact weather and regime transitions (ECMWF 2015; Brunet et al. 2023). In order to evaluate progress and make relevant choices to maximise the performance of forecasts, relevant diagnostics and verification tools are needed. Verification and diagnostic tools have been essential in monitoring progress and identifying weaknesses in the forecasting system, helping to guide the direction of research to improve the forecasts. As the modelling systems evolve, new verification and diagnostic approaches will be needed (ECMWF 2015; Ebert et al. 2013, 2018; Dorninger et al. 2018).

One major challenge in the evaluation of forecast performance for high impact weather is the availability of observations. Differences in reporting practices between regions, differences in spatial or temporal scale between observation and model, lack of common definition between observed and modelled quantities (e.g. thunderstorm or tropical cyclone), measurement errors, or complete lack of observations (data sparse regions) all have significant impact on the ability to properly assess forecast quality and identify underlying model weaknesses.

The consistency between successive ensemble forecasts valid for the same time is referred to as run-to-run consistency or jumpiness. Evaluation of the run-to-run consistency of ensemble forecasts will directly address the concerns over ensemble jumpiness raised by forecasts users. This evaluation does not depend on the availability of observations and is therefore not affected by the observational issues described above. Quantifying the level of jumpiness in an ensemble system provides valuable information to the forecast user, while identifying the circumstances in which jumpiness occurs is an important step towards addressing the underlying cause. This may complement the established verification methods and help to identify weaknesses in the ensemble prediction system.

The evaluation of run-to-run ensemble consistency has so far received relatively little attention in the literature. Therefore the main aim of this PhD thesis is:

Aim: to carry out research to improve the use and understanding of ensemble forecasts through the evaluation of run-to-run consistency together with existing verification methods.

1.2 Objectives and research questions

1.2.1 *Evaluation of **the run-to-run consistency in ensemble forecasts***

The first step was to develop and demonstrate an appropriate score to evaluate the run-to-run jumpiness of ensemble forecasts.

Previous work on forecast consistency has focused on deterministic forecasts, for example in the context of model output statistics (Ruth et al. 2009), comparing automated with manual forecasts (Griffiths et al. 2019), comparing deterministic rainfall forecasts from different models (Ehret 2010) and in forecasts of river flow (Pappenberger et al. 2011a). The only study to date on the jumpiness in the ECMWF ensemble focused on the deterministic ensemble mean forecast (Zsoter et al. 2009). None of the methods used in these studies are directly applicable to assess the consistency of a sequence of ensemble forecasts taking account of the full ensemble distribution. There is therefore a need to develop a measure of forecast consistency that accounts for all aspects of the ensemble empirical distribution.

This raised the first key research question: *How can we identify run-to-run consistency in a sequence of ensemble forecasts?*

This question was the motivation for the first objective of this PhD research:

Objective 1: Develop a suitable index to measure the run-to-run consistency in a sequence of ensemble forecasts and demonstrate how this can identify important cases of high ensemble forecast jumpiness.

This objective was addressed in the study presented in Chapter 3 entitled "Evaluation of the consistency of ECMWF ensemble forecasts" and published in Geophysical Research Letters in 2020.

1.2.2 *Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks*

The forecasting of tropical cyclone (TC) tracks is one area where the flow-dependent probabilistic information in operational ensembles is not fully exploited (Titley et al. 2019; Dunion et al. 2023; Conroy et al. 2023) and run-to-run jumpiness in ensemble track forecasts has caused difficulties for forecast users (Magnusson et al. 2021).

Hurricane Laura (2020) was one case which had unusually large inconsistency from run to run in the ECMWF tracks, causing problems for forecasters at the National Hurricane Center (NHC) who were trying to assess the areas at risk as along the US Gulf coast (Magnusson et al. 2021). As well as being unusually jumpy compared to ECMWF forecasts for other TCs, the jumpiness of the ECMWF tracks for Laura was not seen in the track forecasts from other global ensemble systems.

This raised the second key research question: *How does run-to-run consistency vary between ensemble forecasts from different centres, and do these differences shed light on the causes of jumpiness?*, which motivated the second objective:

Objective 2: Evaluate and compare the jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks from three operational centres, identify any common factors and provide guidance to users.

This objective was addressed in the study presented in Chapter 4 entitled “Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks” and published in Weather and Forecasting in 2024.

1.2.3 *Jumpiness of TC genesis*

Prediction of TC genesis is a major scientific challenge and the processes involved are not well understood (Tang et al. 2020; Rajasree et al. 2023). Although ensemble forecasts have been shown to have some skill in predicting TC genesis (Komaromi and Majumdar 2014, 2015; Majumdar and Torn 2014; Yamaguchi and Koide 2017; Yamaguchi et al. 2015), the use of ensemble probabilistic information in TC genesis forecast products for operational TC centres is still limited (Hon et al. 2023).

One of the key issues limiting the uptake of ensemble forecasts is the run-to-run jumpiness that can occur in some situations (Dunion et al. 2023; Magnusson et al. 2021). Another factor is the lack of routine evaluation of the products provided by the global centres: although ECMWF regularly publishes verification results for ensemble forecasts of the track and intensity of existing TCs (Haiden et al. 2023), it does not routinely evaluate genesis forecasts, so users do not have a clear picture of ensemble performance in predicting genesis (Magnusson et al. 2021).

A major difficulty in the evaluation of TC genesis forecasts is the lack of a common definition across different models and observation datasets. Reporting practices differ across regions, while differences in feature identification between different TC trackers can have a significant impact on the number of TCs identified in ensemble forecasts covering the period of an event that is subsequently observed (Conroy et al. 2023) therefore affecting the forecast probability of the genesis event. The absence of a generally agreed best practice for the definition and evaluation of TC genesis has been identified as an important area for the international TC community to address (Dunion et al. 2023).

The lack of routine operational verification, poor understanding of occasional run-to-run jumpiness and the representativeness issues surrounding evaluation of TC genesis against observations led to the third key research question: *Can an integrated approach using both skill and consistency measures be beneficial in evaluation of ensemble forecast performance for weather hazards with significant forecasting challenges and significant observational representativeness or uncertainty issues?*

This motivated the third objective:

Objective 3: Evaluate the skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis, provide guidance to users and identify factors affecting forecast performance.

This objective was addressed in the study presented in Chapter 5 entitled “Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis” which was submitted to Weather and Forecasting in 2024.

1.3 Structure of the thesis

The structure of the thesis (Figure 1.1) is formed around the three main papers produced during the PhD, which are re-formatted into the format of the thesis in order to preserve the literary unity of the thesis as a whole. Table 1-1 summarises the research questions and the corresponding objectives introduced in section 1.2, together with the title of the chapter containing the associated research study.

Chapter 2 provides a literature review to substantiate the motivation for carrying out the research on ensemble forecast consistency and introduce the main datasets used in the thesis. Chapters 3, 4, and 5 contain the three main research outputs of this thesis, addressing in turn the three objectives set in section 1.2. Chapter 6 contains a summary of additional papers and reports that I have contributed to during this PhD and that are connected to the work of the thesis. Chapter 7 discusses the key findings and limitations of the research conducted during the thesis and provides recommendations for next steps to take forward the outcomes of the thesis. Finally, conclusions to the thesis are presented in Chapter 8. A list of references is provided at the end of the thesis, followed by appendices containing the published versions of Chapters 3 and 4 and the submitted version of Chapter 5, together with other important co-authored papers and reports.

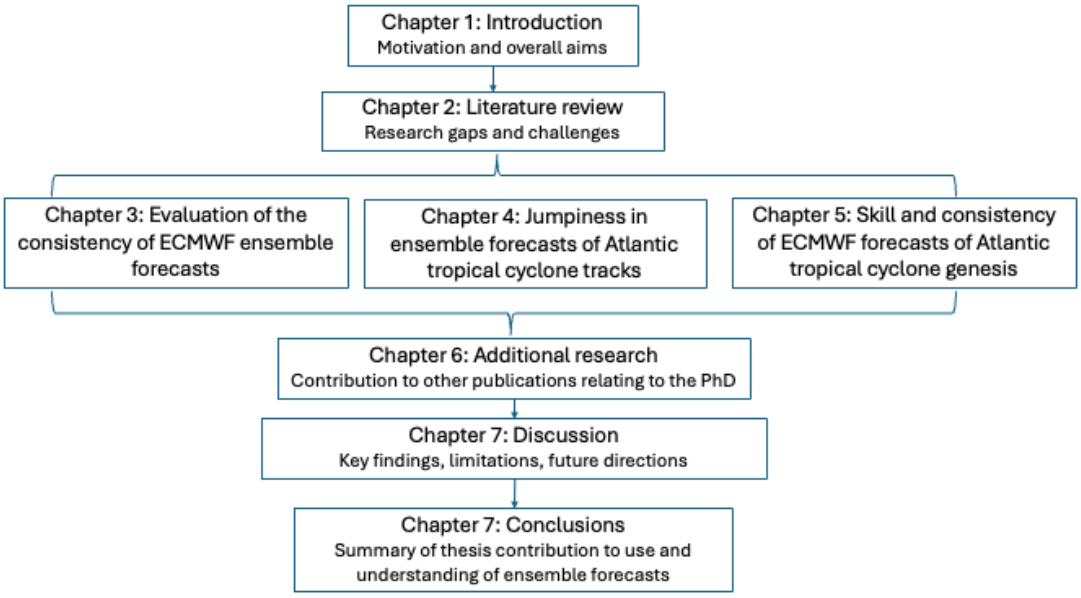


Figure 1.1. Schematic structure of the PhD.

Chapter contents		
Chapter 2: Literature review.		
Research Question	**Objective**	**Chapter details**
How can we identify run-to-run consistency in a sequence of ensemble forecasts?	Develop a suitable index to measure the run-to-run consistency in a sequence of ensemble forecasts and demonstrate how this can identify important cases of high ensemble forecast jumpiness.	**Chapter 3**: Evaluation of the consistency of ECMWF ensemble forecasts
How does run-to-run consistency vary between ensemble forecasts from different centres, and do these differences shed light on the causes of jumpiness?	Evaluate and compare the jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks from three operational centres, identify any common factors and provide guidance to users.	**Chapter 4**: Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks
What factors affect the jumpiness and skill of TC genesis forecasts? Can an integrated approach using both skill and consistency measures be beneficial in evaluation of ensemble forecast performance for weather hazards with significant forecasting challenges and significant observational representativeness or uncertainty issues?	Evaluate the skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis, provide guidance to users and identify factors affecting forecast performance	**Chapter 5**: Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

Table 1-1. Summary of research questions and the associated objectives and chapters.

Chapter 2 Literature review

2.1 Ensemble forecasting

Ensemble predictions have been produced regularly at the European Centre for Medium-Range Weather Forecasts (ECMWF) since December 1992 (Molteni et al. 1996) and at the National Centers for Environmental Prediction (NCEP) (Toth and Kalnay 1993). Today, global ensemble prediction systems (EPS) are run at many operational NWP centres to provide forecasts for the medium range (up to two weeks ahead).

Typically these EPS comprise a control forecast initialized from the operational analysis, plus additional integrations initialized from perturbations to the control analysis (Buizza et al. 1998). The size and resolution of the ensemble is a compromise dependent on available high-performance computing (HPC) resources. The operational EPS usually have 20-50 members and are run at a lower spatial resolution than the single high-resolution deterministic forecast that is also run as part of the operational forecast suite. The EPS complements the deterministic forecast by the provision of information about the probability distribution of future weather, based on uncertainty in the initial analysis.

In 2023 a major milestone in global NWP was achieved when ECMWF upgraded the resolution of the ensemble to match that of the deterministic forecast and so effectively make a unified purely ensemble-based forecasting system (Haiden et al. 2023).

As well as benefitting from improvements to the data assimilation and NWP model, ensemble forecast skill has also improved as the forecast uncertainties have been better quantified through improvements to the initial condition perturbations and representation of model uncertainties (Buizza and Richardson 2017; Buizza et al. 2008; Leutbecher et al. 2017; Swinbank et al. 2016; Zhou et al. 2022; Inverarity et al. 2023)

Ensemble forecasts generate a large amount of data, and the ability to extract and communicate the relevant information for each user's decision-making process is essential to the successful use of the ensemble. Several products have been developed to address different user requirements. Clustering products group similar ensemble members together depending on defined flow characteristics appropriate for different applications (Ferranti et al. 2015, 2018; Neal et al. 2016, 2024; Richardson et al. 2020a; Grams et al. 2017). The Extreme Forecast Index (EFI) was designed to alert users to potentially extreme weather (Lalaurette 2003; Zsoter et al. 2015). Products to automatically identify and track tropical cyclones (Conroy et al. 2023; Heming et al. 2019; Titley et al. 2019; Magnusson et al. 2021) and extratropical cyclones (Hewson and Titley 2010) can be used to identify areas at risk from

potentially damaging storms. This thesis investigates run-to-run ensemble forecast consistency using some of these operational products designed to focus on the large-scale flow (Chapter 3) and high impact weather (tropical cyclones tracks in Chapter 4 and tropical cyclone genesis in Chapter 5).

The main focus of this thesis is on the global medium-range ensembles from the European Centre for Medium-Range Weather Forecasts (ECMWF). The ECMWF forecasting system including the ensemble (ENS) is upgraded about once a year, with improvements introduced to the data assimilation, model physics and dynamics, and ensemble configuration (Haiden et al. 2019, 2021, 2022). Relevant information about the operational configurations used in each part of this thesis is provided in the data sections of Chapter 3, 4 and 5. In Chapter 4, the ECMWF ensemble TC track forecasts were compared with those from two other global centres: the US National Centers for Environmental Prediction (NCEP) and the UK Met Office. The forecast data from the other two centres was retrieved from the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016).

2.2 Run-to-run consistency (jumpiness)

Global NWP centres typically produce a new forecast every 6-12 hours, with the latest forecast starting from initial conditions (the analysis) that takes account of the most recent observations received at the centre. In a sequence of consecutive forecasts valid for the same time, the more recent forecasts will on average be more skilful as they benefit from the additional observations in the initial conditions. However, because of the uncertainties in the initial conditions and the potential rapid growth of small initial differences (see section 2.3.5), there can be large differences between consecutive forecasts, especially at longer forecast lead times, and sometimes a more recent single deterministic forecast will be significantly less skilful than the preceding forecast.

Ensemble forecasts are designed to explicitly account for uncertainties in the initial conditions, and one of the of the expected benefits of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts (Zsoter et al. 2009; Buizza 2008a). Ensemble forecasts show the range of weather scenarios that can occur, allowing users to make appropriate risk-based decisions. An ensemble forecast made two weeks in advance will show a range of possible outcomes. New observations included in the initial conditions for subsequent forecasts will eliminate some of these scenarios and the forecast will become more certain. However, occasionally a new forecast seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the earlier forecast; in this way, the new forecast is inconsistent with the previous forecast.

As an illustration of consistent and inconsistent forecasts, Figure 2.1 shows two hypothetical sequences of forecasts made on consecutive days (labelled as Monday, Tuesday, Wednesday). Each

panel shows the area at risk from a tropical cyclone (TC) as it is forecast to move across the Caribbean and the Gulf of Mexico, from an initial position south of Hispaniola. The ensemble forecast provides a set of TC tracks that will spread out depending on the uncertainty in the forecast. The blue shading shows the areas at risk, based on the ensemble information.

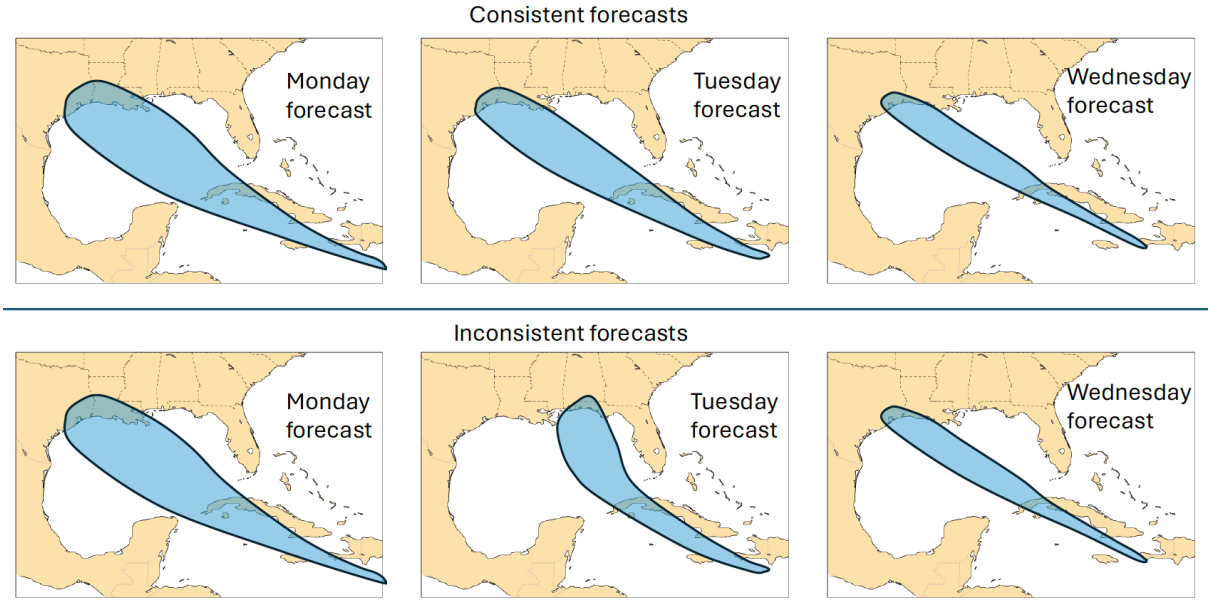


Figure 2.1. Hypothetical examples of sequences of three consistent (top) and inconsistent (bottom) ensemble forecasts initialised on three consecutive days (Monday, Tuesday, Wednesday). In each panel, the blue shading indicates the area at risk from a tropical cyclone that is forecast to move north-westwards over the Caribbean and Gulf of Mexico in the coming days.

The top row is an example of a consistent sequence. The forecast initialised on Monday shows that the TC, initially south of Hispaniola, will move north-west affecting Jamaica and Cuba before continuing towards the US Gulf coast. The exact track is uncertain, and the TC is forecast to make landfall along the coast of Texas or Louisiana. Tuesday's forecast (shown in the middle panel) and then Wednesday's forecast become steadily more certain: the range of solutions in the ensemble is reduced, benefiting from the additional observations received since Monday and therefore the area at risk is better defined, by Wednesday's forecast concentrating the likely area of US landfall around the Texas/Louisiana border.

The bottom row shows a contrasting example of an inconsistent sequence of ensemble forecasts. In this case, Tuesday's forecast is very different from Monday's in terms of US landfall: in Tuesday's forecast, the risk of landfall in the US is concentrated on the eastern Gulf coast. This area was not identified as at risk in Monday's forecast, and therefore the forecasts from Monday and Tuesday are inconsistent with each other. Monday's forecast should identify all possible areas at risk and the following forecast should refine this, eliminating some possibilities, but not introducing new solutions that are not already present in Monday's forecast. Finally in this example, Wednesday's forecast jumps

back towards Monday's solutions, with landfall predicted on the western Gulf coast. The forecasts from Monday and Wednesday are consistent with each other, but neither is consistent with the forecast from Tuesday. The sequence as a whole (Monday, Tuesday, Wednesday) is inconsistent and an example of a flip-flop (Zsoter et al. 2009) or windshield-wiper effect (Broad et al. 2007), where the whole ensemble jumps back and forth between two different solutions.

Concerns over such inconsistent behaviour in the ECMWF ensemble are regularly raised in feedback from users of ECMWF forecasts (Hewson 2021, 2020). In some cases the ensemble distribution as a whole can flip-flop over several consecutive forecasts (Magnusson et al. 2021). Contradictory messages from such jumpiness present a significant challenge to forecasters and decision-makers and can cause users to lose confidence in the forecasting system (Dunion et al. 2023; Elsberry and Dobos 1990; Magnusson et al. 2021; McLay 2011; Pappenberger et al. 2011b)

As users receive the latest NWP output, they must decide how to revise their weather-dependent decisions to take account of the new forecast information. In the example in Figure 2.1, how should the forecaster respond to the large jump in predicted landfall between Monday's and Tuesday's forecast? Should they downgrade any warnings for the western Gulf coast and focus the warning area on the newly identified eastern coast? Or do they keep warnings in place for the west in anticipation of a possible jump back in the next forecast, while extending the warning to also cover the area further east?

Forecasters will need to strike a balance between closely following the changed model guidance and taking a more conservative approach of making a smaller change to mitigate the potential need to make a change in the opposite direction later, that is to avoid a so-called windshield-wiper effect (Broad et al. 2007). This balance is likely to be dependent on the user and their specific application. A survey of Warning Coordination Meteorologists in the US National Weather Service found that forecast inconsistency was a key problem and affected credibility (Sherman-Morris et al. 2018). TC forecasters have similar concerns, and consequently tend to adjust the predictions with the aim to minimise jumpiness in the operational forecasts (Broad et al. 2007).

McLay (2011) applied a simple dynamic decision model to sequences of more or less jumpy ensemble forecasts and demonstrated that the jumpier sequences led to the greatest overall expense incurred by the decision-maker. This provides some qualitative support to the perception that jumpy forecasts are more challenging for users to manage.

To make informed decisions, users need to have the relevant information about the forecast behaviour: how often do jumps occur? When are they likely to happen? Why are ensemble forecasts sometimes inconsistent? How large can these differences be? Information quantifying the consistency

between successive probabilistic forecasts can be important to inform optimal decision making, such as whether to act now or wait for the next forecast (Regnier and Harr 2006; Jewson et al., 2021, 2022). Both noted that such information is not readily available to users.

Evaluation of the run-to-run consistency of NWP forecasts has received little attention in the literature. Ruth et al. (2009) introduced the forecast convergence score (FCS) to measure the number of significant jumps in a sequence of model output statistics (MOS) forecasts valid at the same time. The FCS counts the number of jumps over a threshold and also includes a measure of the sum of forecast differences through the sequence. The FCS can be applied to point forecasts of temperature and probability of precipitation (PoP), but not to ensemble distributions as a whole. Ruth et al. (2009) use the FCS to assess whether MOS forecasts have the same overall level of consistency as the forecasts issued by Weather Forecast Offices. More recently, Griffiths et al. (2019) introduced a flip-flop index (FFI) to compare the jumpiness of automated and manual forecasts for locations in Australia. Like the FCS, the FFI computes the sum of the absolute differences between a sequence of forecasts valid for the same time. However, it does not explicitly count large jumps. To avoid penalizing trends, the difference between the maximum and minimum values in the sequence is subtracted. Like the FCS, the FFI is applied to temperature and PoP forecasts, with the aim to compare the overall consistency of automated and manual systems.

Elsberry and Dobos (1990) investigated the consistency of TC guidance for the Western North Pacific by using the difference in cross-track errors between successive forecasts. Fowler et al. (2015) assessed the consistency of Atlantic TC track forecasts by counting how often in a sequence of forecasts the predicted position changes from one side to the other of a fixed reference track, for example the observed track. However, they caution that biased forecasts may appear to be consistent since successive forecasts may jump considerably without crossing the observed track. Both Elsberry and Dobos (1990) and Fowler et al (2015) recommended that evaluation of forecast consistency be included alongside the routine evaluation of forecast accuracy.

Forecast consistency has also been considered in the context of comparing deterministic rainfall forecasts from different models (Ehret 2010) and in forecasts of river flow (Pappenberger et al. 2011a).

In the only study to date of the run-to-run consistency of the ECMWF ensemble, Zsoter et al. (2009) investigated the consistency of successive ensemble-mean forecasts of the large-scale flow over Europe valid for the same time. They define an inconsistency index as the difference between two fields over a given area, divided by their average standard deviation over the area. They consider cases of large jumps (inconsistency greater than a chosen threshold) and focus on sequences of jumps of opposite sign (flip-flops). Using this methodology, they showed that ensemble-mean forecasts are

more consistent than the corresponding ensemble control forecasts. Zsoter et al. (2009) conclude by noting that to further investigate the benefit of ensemble forecasts compared to single forecast, an index for probabilistic forecasts will need to be developed.

None of the above methods are directly applicable to assess the consistency of a sequence of ensemble forecasts taking account of the full ensemble distribution. There is therefore a need to develop a measure of forecast consistency that accounts for all aspects of the ensemble empirical distribution.

2.3 Evaluation of ensemble forecasts

2.3.1 *Introduction*

Reasons for the evaluation or verification of forecasts are usually grouped into three categories, referred to as administrative, scientific and economic (Jolliffe and Stephenson 2011; Wilks 2020; Brier and Allen 1951).

Administrative verification includes the routine monitoring of forecast performance for quality assurance and monitoring the long-term trend in forecast improvements. Examples include the headline scores used at ECMWF (Haiden et al. 2023) and the WMO standard verification for global NWP and global ensemble NWP (WMO 2023).

The aim of scientific verification is to understand the strengths and weaknesses of the forecasting system and to provide feedback to model developers to guide research on future improvements to the model. As well as standard scores, this evaluation can involve process diagnostics (Day et al. 2020), error tracking (Magnusson 2017a; Grams et al. 2018), ensemble sensitivity analysis (Torn and Hakim 2009) and relaxation experiments (Jung 2011; Jung et al. 2010). The methodologies can be used to address overall systematic errors of the forecasting system or to target particular flow-dependent aspects of the performance. ECMWF's comprehensive annual verification report (Haiden et al. 2023, 2022, 2021) illustrates the wide range of measures used to evaluate the quality of the ECMWF forecasting system and to assess the changes in performance resulting from the introduction of new model cycles.

Economic verification focuses on the user, with the objective of providing the user with the information they need to gain the maximum benefit from the available forecast information. The benefit or value of a given forecasting system can be very different for different users depending on their decision-making requirements. There is a compromise to be made – ideally the verification scores are simple and easy to understand, while also providing the user with relevant guidance on how the forecasts will benefit their decision-making process (Richardson 2011).

2.3.2 *WMO standard verification for EPS*

The World Meteorological Organisation (WMO) has defined a standard set of verification scores for EPS forecasts produced by global NWP centres (WMO 2023). This provides users with consistent verification information on the ensemble products produced by the different centres and enables the centres to compare EPS forecast quality with each other using a consistent evaluation framework. Such intercomparisons help the producing centres to better understand their models' performance and can guide developments to improve predictions.

The WMO standard scores have been chosen to provide key information appropriate to monitor the quality of state-of-the-art EPS while at the same time being straightforward to implement in a consistent way at each producing centre using where relevant a common climatology and set of verifying observations.

The WMO standardized verification for EPS comprises three required components:

- Verification of the ensemble mean (EM) as a deterministic forecast using root-mean-squared error (RMSE) and anomaly correlation
- Evaluation of ensemble spread measured as the standard deviation across ensemble members
- Verification of the ensemble as a probabilistic forecast using the Continuous Ranked Probability Score (CRPS)

In addition, producing centres are encouraged to exchange verification results for EPS probability forecasts of specified dichotomous (yes/no) events. These results are exchanged as tables and the WMO Lead Centre for EPS verification (LC-EPSV) uses these to produce several verification scores including Brier score (BS) and relative operating characteristic (ROC).

The scores are applied to a defined set of forecast variables (including 500hPa height, 850 hPa temperature, near-surface temperature and wind, and precipitation) to give an overview of the EPS performance in forecasting both the large-scale weather systems and the main weather elements.

Scores are regularly exchanged between the participating producing centres and are collected and displayed by the WMO LC-EPSV (see https://epsv.kishou.go.jp/EPSv/).

The definitions for the WMO standard scores are provided on the web sites of the WMO Lead Centres for Deterministic and EPS verification:

- https://confluence.ecmwf.int/display/WLD/Score+definitions+and+requirements
- https://epsv.kishou.go.jp/EPSv/

They are presented in Table 2-1 using the notation used in this thesis for ease of reference. In Chapters 4 and 5 the CRPS is presented in its kernel representation (Gneiting and Raftery 2007); this is mathematically equivalent to the WMO specification (discrete formulation, Hersbach (2000)), but the kernel representation shows more clearly the relationship between the CRPS and the mean absolute error (MAE) as well as the link to the divergence measures used in this thesis to quantify the jumpiness between consecutive ensemble forecasts.

Score	**definition**
Mean square error (MSE)	$\frac{1}{W}\sum_{n=1}^{N} w_n(f_n - y_n)^2$
Root mean square error (RMSE)	$\sqrt{\frac{1}{W}\sum_{n=1}^{N} w_n(f_n - y_n)^2}$
Mean absolute error (MAE)	$\frac{1}{W}\sum_{n=1}^{N} w_n\lvert f_n - y_n\rvert$
Anomaly correlation (AC)	$\frac{\sum_{n=1}^{N} w_n(f_n' - \langle f_n'\rangle)(y_n' - \langle y_n'\rangle)}{\sqrt{\sum_{n=1}^{N} w_n(f_n' - \langle f_n'\rangle)^2 \sum_{n=1}^{N} w_n(y_n' - \langle y_n'\rangle)^2}}$
Ensemble mean (EM)	$\overline{f_n} = \frac{1}{M}\sum_{m=1}^{M} f_{n,m}$
Ensemble standard deviation (spread)	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(f_{n,m} - \overline{f_n}\right)^2}$
Brier score (BS)	$\mathrm{b} = \frac{1}{N}\sum_{i=1}^{N}(p_n - o_n)^2$
Brier score for climate forecast	$\mathrm{b_c} = \frac{1}{N}\sum_{i=1}^{N}(s - o_n)^2$
Brier skill score (BSS)	$B = \frac{b_c - b}{b_c}$
Continuous ranked probability score (CRPS)	$CRPS = \int_{-\infty}^{\infty}\left(P_n(x) - O_n(x)\right)^2 dx$

Continuous ranked probability score (CRPS) for climate forecast	$CRPS(clim) = \int_{-\infty}^{\infty} \big(S(x) - O_n(x)\big)^2 dx$
Continuous ranked probability skill score (CRPSS)	$CRPSS = \frac{CRPS(clim) - CRPS(f)}{CRPS(clim)}$
Relative operating characteristic (ROC)	Plot of hit rate H against false alarm rate F for range of decision thresholds (typically probability thresholds) for a given binary (yes/no) event $H = p(\text{event forecast}\vert\text{event observed})$ $F = p(\text{event forecast}\vert\text{event not observed})$
ROC area (ROCA)	Area under the ROC curve

Table 2-1 Verification scores and related measures used or referred to in this thesis.

For verification over a set of N cases (or N grid points), y_n is the verifying value (observation or analysis), f_n is a single deterministic forecast, and is $f_{n,m}, m = 1, \ldots M$ Is an ensemble of M members. The climatological mean value of the variable is c, anomalies from the climate are represented as $f_n' = f_n - c$ and $y_n' = y_n - c$, and angle brackets $\langle \quad \rangle$ indicate the mean over the verification sample $\langle f_n' \rangle = \frac{1}{W}\sum_{n=1}^{N} f_n'$.

The weights w_n are defined as:

- Verification against analyses over a grid of N grid points: $w_n = \cos\theta_n$, cosine of latitude at grid point n
- Verification against observations or against analysis for single location (over set of N cases): $w_n = \frac{1}{N}$

with the sum of weights $W = \sum_1^N w_n$.

For verification of probabilistic forecasts for a given binary (yes/no) event over a set of N cases, $p_n$is the forecast probability and o_n is the verification, where $o_n = 1$ if the event occurs and $o_n = 0$ otherwise. The climatological probability (base rate) of the event is s. For a continuous variable x, the forecast probability cumulative distribution function (CDF) is $P_n(x)$, the verification expressed as a CDF is $O_n(x)$, where $O_n(x) = 1$ if the verifying value is greater than x and zero otherwise, and the climate CDF is $S(x)$.

2.3.3 *ECMWF verification for EPS*

ECMWF uses a wide range of measures to assess the quality of its medium-range ensemble (ENS). These include the WMO standard verification scores described in the previous section.

ECMWF has selected a set of 8 headline scores to monitor the long-term trend in forecast performance (Haiden et al. 2023). Four of these are for the ENS. Three ENS scores are based on the WMO standard verification CRPS. Two of these aim to monitor the overall ENS performance; they are expressed as skill scores relative to climatology and presented as the lead time at which the skill score reaches a given value, chosen to focus the verification on a particular forecast range appropriate to the forecast variable:

- the lead time at which the Continuous Ranked Probability Skill Score (CRPSS) reaches 25% for 500hPa geopotential over the extra-tropical northern hemisphere
- the lead time at which the CRPSS reaches 10% for 24-h total precipitation over the extra-tropics (both hemispheres combined)

The third is chosen to focus on specific cases of large errors, which can be masked in the overall scores:

- The proportion of large errors (CRPS>5°C) in the ENS probabilistic forecasts of near-surface temperature over the extra-tropics

The large errors in near-surface temperature tend to occur in particular weather situations of calm air in winter (stable boundary-layer) where the near-surface temperature is especially sensitive to differences in cloud cover, near-surface wind speed and snow on the ground (Haiden et al. 2018).

The fourth headline score for the ENS focuses on high-impact weather and measures the skill of the Extreme Forecast Index (EFI) for near-surface wind speed over Europe using the ROC area.

In addition to these ENS-oriented scores, the headline scores include tropical cyclone position error at forecast day 3 for the high-resolution deterministic forecast (HRES).

ECMWF maintains a comprehensive suite of verification measures to give a comprehensive evaluation of the forecasting system including the HRES and ENS. Many of these are provided for users to reference on the ECMWF web site (https://www.ecmwf.int/en/forecasts/quality-our-forecasts). In addition, ECMWF prepares an annual review of forecast performance including a wide range of verification results (Haiden et al. 2023).

2.3.4 *Observation uncertainty and representativeness*

NWP forecasts can be verified against corresponding model analyses or against observations, or both. The apparent skill of the forecasts can sometimes vary considerably depending on the choice of verifying data (Feldmann et al. 2019; Mittermaier 2012; Pinson and Hagedorn 2012).

Verification against the model analysis ensures an equivalence between forecast and verifying data, although any model errors affecting both forecast and analysis will not be apparent in the verification results. On the other hand, verification against observations is affected by representativeness issues – point observations (e.g. from a single synoptic observing station) are not directly representative of the grid scale of the model because of the local variations that can occur within the area covered by a single model grid box. For example, if a model is run with a 20 km grid spacing, the forecast rainfall at each model grid point represents the average rainfall over a 400 km^2 region. For a small-scale event, such as a summer thunderstorm, there can be large differences between the rainfall measured at an individual location and the average rainfall over the larger area represented by the model. When NWP forecasts are verified against point observations, the mismatch between the spatial and temporal scales of the forecast and observation can be a significant contribution to the overall error (Tustison et al. 2001; Bowler 2006; Mittermaier 2008, 2014; Mittermaier and Stephenson 2015). It is important to take account of this scale mismatch in forecast verification, especially in the verification of extreme events (Goeber et al, 2008).

Representativeness can account for a substantial proportion of the apparent under-dispersion of the ensemble forecast. A number of standard measures that assess the calibration or reliability of an EPS, such as rank histograms (PIT diagrams) and reliability diagrams, as well as scores such as BS, CRPS and ROC are sensitive to representativeness issues and care needs to be taken in interpretation of the results if the observation uncertainty is not properly accounted for (Saetra et al. 2004; Candille and Talagrand 2008; Yamaguchi et al. 2016; Rodwell et al. 2016, 2018; Ferro 2017; Bowler 2008).

Another aspect of representativeness is when the model and observation do not directly represent the same variable, for example the verification of thunderstorm forecasts using lightning data (Marsigli et al. 2021) or in comparing model TCs with subjectively-based reports of TC position and intensity (Conroy et al. 2023; Dunion et al. 2023). As the resolution and accuracy of ensemble forecasts increases, there is an increasing focus on the prediction of high-impact weather. Intense high-impact events (e.g. hail, lightning, tornados) are often very localised and not well captured or even measured by conventional observing networks. This has led to an increased interest in exploiting non-conventional observations, including citizen observations, in ensemble verification (Marsigli et al. 2021; Tsonevsky et al. 2018). The lack of direct physical correspondence between model variable and observation, as well as differences in spatial and temporal scales between forecasts and observations,

all contribute to representativeness issues that will need to be addressed to properly evaluate the ensemble performance (Marsigli et al. 2021; Janjić et al. 2018; Casati et al. 2022).

2.3.5 *Error growth and predictability*

The relative contributions of initial condition uncertainty and deficiencies in the forecast model can be investigated using a simple error growth model first introduced by Lorenz (1982) and extended by (Dalcher and Kalnay 1987). The original model (Lorenz 1982) was designed so that small initial errors would grow exponentially at first, representing the chaotic nature of the atmosphere (Lorenz 1963), while at longer ranges the forecast error would on average be the same as that from a randomly chosen atmospheric state, representing the loss of predictability at long forecast lead times. Dalcher and Kalnay (1987) introduced an additional linear error growth term to represent model error. This results in a three-parameter model, representing the doubling time of small errors, the asymptotic level at which errors saturate, and the linear growth of error associated with model deficiencies. As well as providing a way to assess the relative importance of model and initial condition errors (Simmons and Hollingsworth 2002; Magnusson and Källén 2013), the error-growth model can also be used to investigate the potential predictability limit, i.e. the forecast range at which even an almost perfect model started from near-perfect initial conditions would lose skill due to the intrinsic chaotic nature of the atmosphere. This was the aim of the original study of Lorenz (1982) and has been revisited most recently by Zhang et al. (2019). These studies have focused on the deterministic predictability limit, the forecast range at which an individual forecast becomes indistinguishable from a random selection from the climate distribution. Zhang et al. (2019) conclude that reducing the current initial condition uncertainty by an order of magnitude could extend the deterministic limit for daily weather forecasts by up to 5 days.

Froude et al. (2013) consider the predictability limit for both the ECMWF high-resolution deterministic forecast (HRES) and the ensemble mean of the corresponding ECMWF ensemble forecasts (ENS). They show that the error growth is lower for the ensemble mean than for the HRES and that the shape of the error growth curves are different, with the error growth of the ensemble mean starting to decrease around day 10, rather than continuing to grow as for the HRES. They note that a possible reason for this is that the ensemble mean starts to lose predictive skill at this time range.

These studies consider only the average growth rate over a large sample of cases (typically a season) and hence do not consider any flow-dependent aspects of error growth. Zhang et al. (2019) note in their conclusions that determining the limit for probabilistic prediction and for weather regimes is beyond the scope of their work and that future studies that consider the interaction between tropical and midlatitude systems would also be valuable.

It is well known that forecast skill varies from day to day, and an alternative approach to evaluation of forecast performance is to focus more specifically on cases with large forecast error, often referred to as forecast busts.

Rodwell et al. (2013) define a forecast bust as a case where the day 6 forecast for 500 hPa geopotential height over Europe has an RMSE larger than 60m and anomaly correlation of less than 40%, hence ensuring that the bust cases involve errors in both magnitude and pattern (phase). They show that these busts are often linked with the prediction of blocking over Europe and tend to occur in spring in forecasts where the initial conditions have a trough over the Rockies and high convective available potential energy (CAPE).

Also investigating forecast busts over Europe, Lillo and Parsons (2017) found that such bust situations are in general associated with changes in the large-scale flow pattern over the North Atlantic (regime transitions; Ferranti et al. 2015) resulting from the initiation and growth of Rossby wave trains extending from North America towards Europe. They found the most frequent busts in autumn at times with tropical storms recurving in the central Atlantic. The adverse impact of extra-tropical transition of tropical storms on predictability over Europe has also been shown by Jones et al. (2003), Keller et al. (2019) and Brannan and Chagnon (2020).

These studies focus on deterministic forecast busts. For ensemble forecasts in these situations, the reduced predictability should be reflected in increased ensemble spread (Rodwell et al. 2013; Magnusson 2017; Leutbecher and Palmer 2008). Cases where the ensemble spread is small compared to the ensemble mean error could be characterised as ensemble busts. Although suggested by Rodwell et al. (2013), this has not yet been systematically investigated. In a study of flow-dependent mid-latitude predictability, Sánchez et al. (2020) characterised cases where the forecast error grows much faster than the ensemble spread as “predictability barriers” and associated such cases with strong diabatic influences on tropospheric advection. Alternatively, an ensemble bust could be defined as a case where the ensemble error exceeds a specified threshold. Haiden et al. (2019) consider the ECMWF ensemble to have a large error in 2m temperature if the CRPS exceeds 5K. They show the frequency of such large errors has steadily decreased as the forecasting system has improved over the last 20 years. However, there has not yet been any study of the specific situations associated with these cases, or on whether there is any common deficiency in ensemble spread.

Evaluation of the run-to-run consistency of ensemble forecasts is an alternative approach to investigating flow-dependent aspects of forecast performance and may complement the approaches proposed above. The research presented in Chapter 3-5 demonstrates the use of this new approach in identifying and investigation the causes of inconsistent ensemble behaviour.

2.3.6 *Ensemble reliability – the spread-error relationship*

A fundamental requirement for ensemble forecasts is that the spread of the ensemble represents the uncertainty in the forecasting system (due to model errors and imperfect knowledge of the initial conditions). The ensemble spread can be measured for example by the standard deviation across the ensemble members. Averaged over a large number of cases, the spread should be equal to the average error of the ensemble-mean. This is known as the spread-error relationship (Leutbecher and Palmer 2008) and is one basic method for assessing the reliability of an ensemble forecasting system. In practice a small adjustment needs to be added to account for the finite size of operational ensemble forecasts, and it can be important also to account for any mean bias of the forecast model as well as for uncertainty in the verifying observation or analysis (Saetra et al. 2004; Yamaguchi et al. 2016; Rodwell et al. 2016).

Rodwell et al. (2016) provide a detailed derivation and application of the spread-error relationship, taking account of all the above aspects. The resulting "reliability budget" separates the mean-squared difference between ensemble mean and verifying observation into three components representing bias, spread (ensemble variance) and observation error, together with a residual term that represents the reliability deficiency in the ensemble. They demonstrate how regional deficiencies in reliability can be related to both observation error and representation of model uncertainty. These results were based on the application of the reliability budget over a full season, and hence did not take account of any flow-dependent aspects of the ensemble uncertainty. Rodwell et al. (2016) conclude that it would be useful to investigate the spread-error relationship in different flow situations, such as European blocking (Ferranti et al. 2015), or cases of mesoscale convective systems over north America (Rodwell et al. 2013) which have been shown to be important for medium-range predictability.

In a first step towards this more flow-dependent evaluation, Rodwell et al. (2018) applied the same reliability budget approach to a sample of cases where a particular flow pattern over North America was present in the initial conditions. This pattern, a trough over the Rockies and high convective available potential energy (CAPE) over eastern North America, has been shown to lead to increased uncertainty in the forecasts 6 days later over Europe (Rodwell et al. 2013). Rodwell et al. (2018) show that in these situations the uncertainty growth rate is too small in the jet stream region over North America, possibly related to systematic errors in the height that the convection reaches in the areas of high CAPE.

Investigation of the evolution of ensemble spread (uncertainty growth) in other flow types including large-scale weather regimes may give insights into relevant model deficiencies. This may help to identify avenues for model developments that will improve the prediction of transitions between

large-scale weather regimes, currently a major challenge for operational ensemble forecasting systems (ECMWF 2015; Ferranti et al. 2018; Vitart et al. 2017).

The standard spread-error relationship (Rodwell et al. 2016; Leutbecher and Palmer 2008) and error-growth model used to investigate predictability limits for deterministic forecasts (Dalcher and Kalnay 1987) both use root-mean-square metrics of error and dispersion (standard deviation). In Chapter 4, the link between consistency, spread and error is explored using new diagnostic measure of forecast consistency developed in Chapter 3 and corresponding CRPS error score and the related measure of ensemble spread based on the mean absolute difference between ensemble members. This ensures the evaluation considers the full ensemble distribution (Gneiting and Raftery 2007; Thorarinsdottir et al. 2013).

2.4 Tropical cyclones

2.4.1 *introduction*

Tropical cyclones (TCs) are among the most damaging natural hazards. Providing early warnings is essential for mitigating their impacts and protecting life and property. This is coordinated globally under the WMO Tropical Cyclone Programme (WMO/TCP); there are five regional bodies responsible for the different ocean basins where TCs occur (WMO 2017). Although WMO/TCP has worked towards standardising terminology and practices globally there are still differences between the different regions. TCs are categorized according to their intensity based on maximum sustained surface wind speed associated with the TC (WMO 2017). Weaker systems with wind speeds less than 17 ms^{-1} (34 kt) are generally referred to as tropical depressions; more developed systems with wind speeds 18-32 ms^{-1} (34-63 kt) are tropical storms; while hurricanes or typhoons (depending on the region) are the strongest TCs with winds at least 33 ms^{-1} (64 kt). Precise definitions for the different regions are given in the official WMO Regional Operational Plans (https://community.wmo.int/en/tropical-cyclone-operational-plans).

Each year, there are 80-90 TCs that reach at least tropical storm strength (17 ms^{-1}) globally, with the greatest number occurring in the Western North Pacific (Schreck et al. 2014; Frank and Young 2007; WMO 2017). In the North Atlantic basin there are 8-15 tropical storms annually (Schreck et al. 2014). Weaker TCs that fail to reach tropical storm strength are often not included in evaluation studies since the reporting practices are more subjective and particularly variable between regions (WMO 2017).

The process of forming a new TC is referred to as TC genesis or tropical cyclogenesis. TCs develop from precursor disturbances to the background tropical atmospheric state. There are different mechanisms in the different regions (basins) and within a given basin, although the processes involved are still not fully understood (Rajasree et al. 2023; Tang et al. 2020; Emanuel 2022).

The focus of the research in this thesis (Chapters 4 and 5) is on the Atlantic basin. The main precursor in the Atlantic is African Easterly Waves (AEWs) which accounts for around 60% of TC genesis events (Landsea 1993; Russell et al. 2017). A further 10% of TC genesis events are indirectly related to AEW (Russell et al. 2017). However, there is little correlation between the number or intensity of AEWs and the number of TCs in a season (Avila et al. 2000; Russell et al. 2017). The factors that influence why some AEWs lead to TC genesis and others do not (so-called developing and non-developing waves) is an area of continuing research (Núñez Ocasio et al. 2021, 2020; Lawton et al. 2022; Feng et al. 2023). Other mechanisms also lead to the development of TC in the Atlantic, including the tropical transition pathways (McTaggart-Cowan et al. 2013, 2008).

2.4.2 *Ensemble forecast products for TCs*

Global NWP centres produce specialized forecast products for tropical cyclones. This is an automated post-processing of the NWP outputs that identifies TCs in the forecasts and tracks their movement as the forecast evolves. The outputs from the tracker are files containing the predicted location (latitude and longitude), maximum sustained wind speed and minimum MSLP, typically at 6 h steps through the forecast. Tracks are computed for existing TCs (those that are already officially reported at the beginning of the forecast) for the high-resolution deterministic forecast and for each ensemble member. Some centres also generate tracks for TCs that do not exist at the initial time but develop during the forecast. Centres are encouraged to distribute the forecast tracks in real time and to archive them in the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016).

Additional forecast products can be constructed based on these forecast tracks. For example, ECMWF strike probability maps show the probability that a given TC will pass within 120 km of any given location within the next 240 hours. This is one way to summarise the ensemble track information for a given existing TC. ECMWF also generates TC activity maps that shows the probability for an active TC (either pre-existing or one that develops during the forecast) to pass within 300 km of any location within a 48 h window at different forecast lead times.

Each centre runs its own tropical cyclone tracker (Conroy et al. 2023). In Chapter 4, I use the TC tracks for ECMWF (Magnusson et al. 2021), NCEP (Marchok 2021), and the Met Office (Heming 2017) for existing TCs from the TIGGE archive. In Chapter 5, I use the ECMWF ensemble tracks for TCs that develop during the forecast to investigate the performance in predicting TC genesis.

Differences in feature identification between different TC trackers can have a significant impact on the number of TCs identified by a forecast model (Conroy et al. 2023) and there is currently no generally agreed best practice for the definition and evaluation of TC genesis (Dunion et al. 2023). This remains a challenge for the international community to address.

2.4.3 *Verification of ensemble TC products*

There has been significant progress in forecasting TC tracks due to improvements in observing systems, data assimilation and NWP modelling (Landsea and Cangialosi 2018; Yamaguchi et al. 2017). Official forecasts of tropical cyclone (TC) tracks are typically based on guidance from Numerical Weather Prediction (NWP) models (Conroy et al. 2023). NWP ensemble forecasts are increasingly being used. Although their use in official forecasts is often limited to the ensemble mean (EM) track, there is increasing evidence of the benefits of using more of the ensemble probabilistic information (Titley et al. 2019, 2020; Kawabata and Yamaguchi 2020; Leonardo and Colle 2017).

Guidelines for TC verification were developed by the WMO Joint Working Group on Forecast Verification Research (WMO, 2013). The evaluation of operational ensemble TC track forecasts includes EM track errors, ensemble spread and verification of strike probability (e.g. Cangialosi 2022, Haiden et al. 2022, Titley et al. 2020, Heming et al. 2019, Leonardo and Colle 2017). The probabilistic performance of ensemble TC forecasts is usually assessed based on the skill of strike probability forecasts as measured by the Brier skill score (Yamaguchi et al. 2012; Titley et al. 2020; Leonardo and Colle 2017). ROC (Haiden et al. 2023) and reliability diagrams (Haiden et al. 2023; Leonardo and Colle 2017) are also used in some studies.

The TC position error (great circle distance from forecast to observed TC location, in km or often in mi) can be broken down into cross-track (CT) and along-track (AT) error, where the observed TC track provides a reference orientation. CT errors are more associated with the location of landfall, while AT errors are more related to speed issues. The causes of CT and AT errors can be different and they are often investigated independently (Leonardo and Colle 2021, 2020).

The research presented in Chapter 4 investigates the run-to-run consistency of ensemble forecasts of Atlantic TC tracks, focusing on the CT position. It compares the forecast jumpiness with spread and error of the TC tracks. An innovation in Chapter 4 is the use of CRPS as a measure of the ensemble TC track error to account for the full ensemble distribution, rather than using the error of the ensemble mean.

Although ECMWF regularly publishes verification results for ensemble forecasts of the track and intensity of existing TCs (Haiden et al. 2023), it does not routinely evaluate genesis forecasts, so users do not have a clear picture of ENS performance (Magnusson et al. 2021). Previous studies have shown that ensemble forecasts do have skill in predicting TC genesis (Komaromi and Majumdar 2014, 2015; Majumdar and Torn 2014; Yamaguchi and Koide 2017; Yamaguchi et al. 2015). In Chapter 5, the run-to-run consistency and skill of ECMWF ensemble forecasts for TC genesis are investigated; the skill is evaluated using the Brier skill score following the approach of (Yamaguchi et al. 2015).

2.4.4 *Best track data*

Official reports of TC tracks and intensity are prepared by the World Meteorological Organization (WMO) Regional Specialized Meteorological Centres (RSMCs) and Tropical Cyclone Warning Centres (TCWCs). For the Atlantic basin, the responsible centre is the US National Hurricane Center (NHC; RSMC Miami). As well as providing track information in real time, the RSMCs and TCWCs also carry out post-event analysis to make an official record of the history of the tropical cyclone over its lifecycle, known as the best track. This is done subjectively by the forecasters using all available observational data, some of which may not have been available in real time when the original reports were made (Landsea and Franklin 2013). Best track data includes the location and intensity of the tropical cyclone every 6 hours. Intensity is defined as the maximum wind associated with the tropical cyclone at 10m above the surface, and conventionally reported in knots for the Atlantic basin (1kt = 0.5 m/s). Central pressure of the TC is also included. The best track reports from all RSMCs and TCWCs are collated and archived in the International Best Track Archive for Climate Stewardship (IBTrACS, (Knapp et al. 2018, 2010)). The IBTrACS data is used in the studies in Chapters 4 and 5.

2.4.5 *Uncertainties in best track data*

In situ observations of TC structure are limited and to a large extent the best track relies on information derived from satellite data. TC intensity can be estimated from an analysis of satellite cloud images (visible and infra-red) using the semi-subjective Dvorak technique (Dvorak 1984; Velden et al. 2006) or the automated Advanced Dvorak Technique (Olander and Velden 2007). Both methods analyse organized cloud patterns to derive an index of TC intensity, then use look-up tables to associate the derived intensity index to a maximum surface wind and MSLP for the TC.

In the Atlantic basin, aircraft data are also available for some TCs, with reconnaissance missions being flown by U.S. Air Force Reserve's 53rd Weather Reconnaissance Squadron C-130s and aircraft from the National Oceanic and Atmospheric Administration (NOAA) Aircraft Operations Center. These aircraft observations are mainly limited to the western parts of the basin (west of 60°W) and are available for approximately 30% of best track report times (Rappaport et al. 2009; Torn and Snyder 2012).

Uncertainties in the best track information were investigated by Torn and Snyder (2012) and Landsea and Franklin (2013). Position uncertainties were found to be larger for weaker TCs, while uncertainties in intensity were greater for stronger TCs. Both studies found similar results, with average position errors around 35 n mi (65 km) for TS without aircraft data and 22 n mi (40 km) with aircraft. Maximum wind speed uncertainties were 8-10 kt.

In Chapter 4, the best tracks are used as a frame of reference to define cross-track and along-track directions, and in Chapter 5 the best track position and intensity at the time and location of TS genesis are used as reference points for investigating the ensemble jumpiness. An important benefit of evaluating ensemble jumpiness is that it is a property of the ensemble system and does not depend on the observations. Therefore the uncertainties in the best track estimates do not directly affect the results on TC consistency that are presented in Chapters 4 and 5.

However, the research in this thesis does also compare the forecast jumpiness to forecast skill for position error (Chapter 4) and genesis (Chapter 5). The best track uncertainties for position are relatively small compared to the medium-range forecast position errors and will have limited impact on the forecast skill results (Landsea and Franklin 2013; Torn and Snyder 2012). However, the uncertainty in intensity is significant and should be accounted for in verification of intensity forecasts; at present this is not done in operational verification and is an area where further research is needed. In Chapter 5 the best track intensity is used to identify the observed genesis – the first point on the observed track with maximum wind greater than 17 m/s (34 kt). In comparing the ensemble forecast tracks, I allow a tolerance of 500 km and 24h in position and timing of the observed genesis. In this way I account for the uncertainty in when the best track reaches TS strength.

2.4.6 *Challenges and issues*

Although the skill of ensemble probabilistic forecasts is increasingly recognised, the pull-through into operations has been limited. In a report to the 9th WMO International Workshop on Tropical Cyclones (IWTC-9), Titley et al. (2019) identified a number of challenges and issues that limited the use of ensemble probabilistic forecast information. They identified several areas for research to enable forecasters to make better use of ensemble forecasts, including: improving ensemble forecast skill, user-oriented verification, best practice, communicating uncertainty information.

One outcome of IWTC-9 was the formation of the WMO/WWRP Tropical Cyclone-Probabilistic Forecast Products (TC-PFP) project, which is also endorsed as a WMO Seamless GDPFS Pilot Project (Dunion et al. 2023). Additional challenges and issues limiting the use of ensemble forecasts identified by this project include run-to-run jumpiness in the ensemble (which can lead to reduced forecaster confidence), lack of verification for TC genesis, and it was recommended that research should be done to address these issues (Dunion et al. 2023).

These issues were also recognised by ECMWF as requiring research to develop guidance to users of the ECMWF forecasts, in particular to address the lack of routine verification for TC genesis and the run-to-run jumpiness in ensemble forecasts, especially for TCs (Magnusson et al. 2021).

These research gaps are addressed in Chapter 4 which addressed jumpiness for TC tracks and Chapter 5 which evaluates the skill and jumpiness of ECMWF ENS genesis forecasts. The aims of the research are to identify jumpy cases, investigate common factors that may help to understand the ensemble model weaknesses and to establish a baseline evaluation of the ability of the current ECMWF ENS to predict TC genesis in the Atlantic.

2.5 Summary

This chapter has reviewed the current state and recent progress in ensemble forecasting including for TC track and genesis, reviewed recent progress in different aspects of forecast evaluation and documented previous work on forecast consistency. The chapter provides additional background information supporting the motivation of the aim and main objectives of the thesis and identifying relevant research gaps and introduces the datasets used in the thesis. The following three chapters contain the three main papers produced during the PhD and details of the datasets and methods used in each study are presented in the relevant section of each chapter.

Chapter 3 Evaluation of the consistency of ECMWF ensemble forecasts

The first objective of the PhD was to develop a suitable index to measure the run-to-run consistency in a sequence of ensemble forecasts and demonstrate how this can identify important cases of high ensemble forecast jumpiness. The paper addressing this objective was published in Geophysical Research Letters with the following reference:

Richardson, D.S., Cloke, H.L. and Pappenberger, F. (2020) 'Evaluation of the Consistency of ECMWF Ensemble Forecasts', *Geophysical Research Letters*, 47(11), p. e2020GL087934. Available at: https://doi.org/10.1029/2020GL087934.

The contributions of the authors of this paper are as follows: D.R. designed the study with advice from H.C. and F.P., produced the datasets, carried out the analysis, and led the writing of the manuscript. All authors assisted with writing the manuscript. Overall, 90% of the writing was undertaken by D.R.

The published article can be found in Appendix A1.

Key Points:

- A new divergence index is introduced to measure inconsistency (jumpiness) in a sequence of ensemble forecasts
- The ECMWF ensemble has occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7-9 days lead
- To understand the causes of jumpiness it is important to consider the time evolution of each ensemble (eg using phase space trajectories)

Abstract. An expected benefit of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts. Inconsistent (jumpy) forecasts can cause users to lose confidence in the forecasting system. We present a first systematic, objective evaluation of the consistency of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble using a measure of forecast divergence that takes account of the full ensemble distribution. Focusing on forecasts of the North Atlantic Oscillation and European Blocking regimes up to two weeks ahead, we identify occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7-9 days lead. However, care is needed in the interpretation of ensemble jumpiness. An apparent clear flip-flop in a single index may

hide a more complex predictability issue which may be better understood by examining the ensemble evolution in phase space.

Plain Language Summary. Ensemble forecasts show the range of weather scenarios that can occur, allowing users to make appropriate risk-based decisions. An ensemble forecast made two weeks in advance will show a range of possible outcomes. New observations included in subsequent forecasts will eliminate some of these scenarios and the forecast will become more certain. Occasionally a new forecast seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the earlier forecast. Such inconsistencies can cause users to lose confidence in the forecasting system. We present a new method to assess the consistency of ensemble forecasts of large-scale weather patterns over Europe made by the European Centre for Medium-Range Weather Forecasts. We show that a careful analysis of each forecast is needed to understand how and why these jumps occur. Understanding and reducing the occurrence of inconsistent ensemble forecasts will increase user confidence and improve decision-making.

3.1 Introduction

The chaotic nature of the atmosphere means that numerical weather prediction (NWP) forecasts are sensitive to small changes in their initial conditions. Operational NWP centres address this by running a number of forecasts from similar starting conditions. The resulting ensemble of forecasts shows the range of future atmospheric states consistent with the known uncertainties in the initial conditions (Leutbecher and Palmer 2008; Swinbank et al. 2016). One of the expected benefits of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts (Zsoter et al. 2009; Buizza 2008a). Inconsistent (or jumpy) forecasts are difficult to handle and can cause users to lose confidence in the forecasting system (Pappenberger et al. 2011b; Hewson 2020). However, this aspect of ensemble forecasts has received little attention in the literature.

The inconsistency between successive ensemble-mean forecasts valid for the same time was investigated by Zsoter et al. (2009). They define an inconsistency index as the difference between two fields over a given area, divided by their average standard deviation over the area. They consider cases of large jumps (inconsistency greater than a chosen threshold) and focus on sequences of jumps of opposite sign (flip-flops). Using this methodology, they showed that ensemble-mean forecasts are more consistent than the corresponding ensemble control forecasts. Zsoter et al. (2009) conclude by noting that to further investigate the benefit of ensemble forecasts compared to single forecast, an index for probabilistic forecasts will need to be developed. Forecast consistency has also been considered in the context of model output statistics (Ruth et al. 2009), comparing automated with

manual forecasts (Griffiths et al. 2019), comparing deterministic rainfall forecasts from different models (Ehret 2010) and in forecasts of river flow (Pappenberger et al. 2011b).

None of the above methods are directly applicable to assess the consistency of a sequence of ensemble forecasts taking account of the full ensemble distribution. In this work, for the first time, we investigate the consistency of the European Centre for Medium-Range Forecasts (ECMWF) ensemble (ENS) using a measure of forecast divergence that accounts for all aspects of the ensemble empirical distribution.

We focus on two key characteristics of the large-scale flow over the European-Atlantic region: the North Atlantic Oscillation (NAO) and Scandinavian Blocking (BLO). Predicting transitions between such large-scale weather regimes two weeks or more ahead is a significant scientific challenge and at the frontier of numerical weather prediction (ECMWF 2015). These transitions are associated with large-scale changes in temperature and winds over Europe (Ferranti et al. 2018; Yiou and Nogaj 2004) and hence have significant societal impacts, for example on health (Charlton-Perez et al. 2019) and on energy production (Grams et al. 2017). We consider the full 15-day forecast range of the operational ENS.

The data and indices used are introduced in section 3.2. Methods, including the definition of the forecast divergence are described in section 3.3. We then evaluate the inconsistency of the ENS forecasts for NAO and BLO and compare the jumpiness of the ENS with that of the ensemble mean (EM) and control forecasts in section 3.4. We present concluding remarks and avenues for future work in section 3.5.

3.2 Data

We study the time evolution of the NAO and BLO patterns that are associated with high-impact temperature anomalies over Europe (Ferranti et al. 2018). Following the approach of Ferranti et al. (2018), we use a 2-dimensional phase space based on the two leading Empirical Orthogonal Functions (EOFs) of mid-tropospheric flow computed over the Euro-Atlantic region. The EOFs are computed using daily geopotential height at 500 hPa computed for the Euro-Atlantic region (30°N to 88.5°N, 80°W to 40°E) from 29 years of extended winter periods (October to March) of ECMWF ERA-Interim data (Dee et al. 2011; Berrisford et al. 2011). For the EOF computation a 5-day running mean was used and the mean seasonal cycle was removed. The first EOF represents the positive phase of the NAO (NAO+): a negative anomaly over Iceland and positive anomaly to the south (Cassou 2008). The second EOF has a positive anomaly (high pressure) over Scandinavia, and a low to the east over the Atlantic, representing the flow pattern associated with blocking events over northern Europe (Ferranti et al. 2015). We refer to Ferranti et al. (2018) for further details.

We study the consistency of the operational ECMWF ensemble forecasts (Buizza and Richardson 2017; Ben Bouallègue et al. 2019) of the large-scale flow over the North-Atlantic Europe region for DJF 2016-2019, ie 1 Dec 2015 to 28 Feb 2019, a total of 361 cases. All forecasts verifying at 00 UTC between 1 December and 28/29 February are included in the evaluation. The ENS comprises 50 perturbed members and 1 control member. The forecasts are valid for lead times of 1 to 15 days (at 24-hour intervals). The 500 hPa fields of each ENS forecast are extracted on a 1x1 degree grid and projected onto the two EOFs. The projections describe the magnitude of the NAO and BLO in each forecast, calculated relative to the climatological standard deviation. Following Ferranti et al (2018), cases with projections greater than one standard deviation are considered large amplitude events.

3.3 Methods

We consider a sequence of ensemble forecasts valid for the same time t_v and started from initial conditions between 1 and L days before, $f(t_v, i), i = 1, \dots L$. Each ensemble consists of M members, $f_m(t_v, i), m = 1, \dots M$. We consider NAO and BLO separately, so f_m are univariate and real-valued.

To measure the difference between two ensembles f and g with M and N members respectively, we use the divergence function given by

$$d(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|f_i - g_j| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_i - g_j| \qquad (3\text{-}1)$$

d is the divergence function associated with the Continuous Ranked Probability Score (CRPS), which is widely used to measure of the quality of ensemble forecasts (Gneiting and Raftery 2007). If either M or N is equal to one, then d reduces to the CRPS, while if both are one, d is simply the absolute distance $|f - g|$. This means that d can also be used to measure the difference between two ensemble-mean or control forecasts. d shares the important property of propriety with CRPS (Gneiting and Raftery 2007) and as shown by Thorarinsdottir et al. (2013) these properties make d a particularly suitable choice.

The difference between two ensemble forecasts initialized on consecutive days and valid for the same time is

$$D(t_v, i) = d\big(f(t_v, i), f(t_v, i-1)\big), i = 1, \dots, L \qquad (3\text{-}2)$$

where $f(t_v, 0)$ is the set of initial perturbed ensemble members at time t_v.

To measure the overall divergence (or inconsistency) between the sequence of forecasts valid for a given time we sum the divergence between successive pairs of forecasts. To focus on the jumpiness

within the sequence rather than a general trend across lead times (or a single large jump representing a one-time change in predictability), we subtract the difference between the first and last forecast of the sequence, and define the Divergence Index (DI) for a given case as:

$$DI(t_v) = \frac{1}{L-1}\left(\left(\sum_{l=1}^{L} D(t_v, l)\right) - d\big(f(t_v, L), f(t_v, 0)\big)\right) \tag{3-3}$$

The divergence index is calculated for the ENS and also for the ensemble control forecast (CTRL) and the ensemble mean (EM). We refer to DI(ENS), DI(CTRL) and DI(EM) respectively. In this study, all ensemble forecasts have M=50 members (control not included) and we consider forecasts up to lead time of L=15 days.

As noted above, for a single forecast such as CTRL and EM, the divergence is equal to the absolute difference. For these forecasts, DI is similar (though not identical) to the Flip-Flop Index of Griffiths et al. (2019).

3.4 Results

Figure 3.1 (upper panel) shows the DI(ENS) for NAO (solid) and BLO (dashed) for each day of the last four winters (December-February 2015-16 to 2018-19; the vertical dotted lines indicate the start of each season). Positive values indicate higher inconsistency. There is similar variability in DI for both regimes (standard deviation of 0.027 for NAO and 0.028 for BLO). However, the peaks of high/low consistency occur at different times. Winter 2018-19 was more inconsistent than usual for blocking, while forecasts for NAO were not unusually inconsistent in this season. Overall, there is no strong correlation between inconsistency in forecasts of blocking and NAO (correlation =-0.1 over the full set of cases).

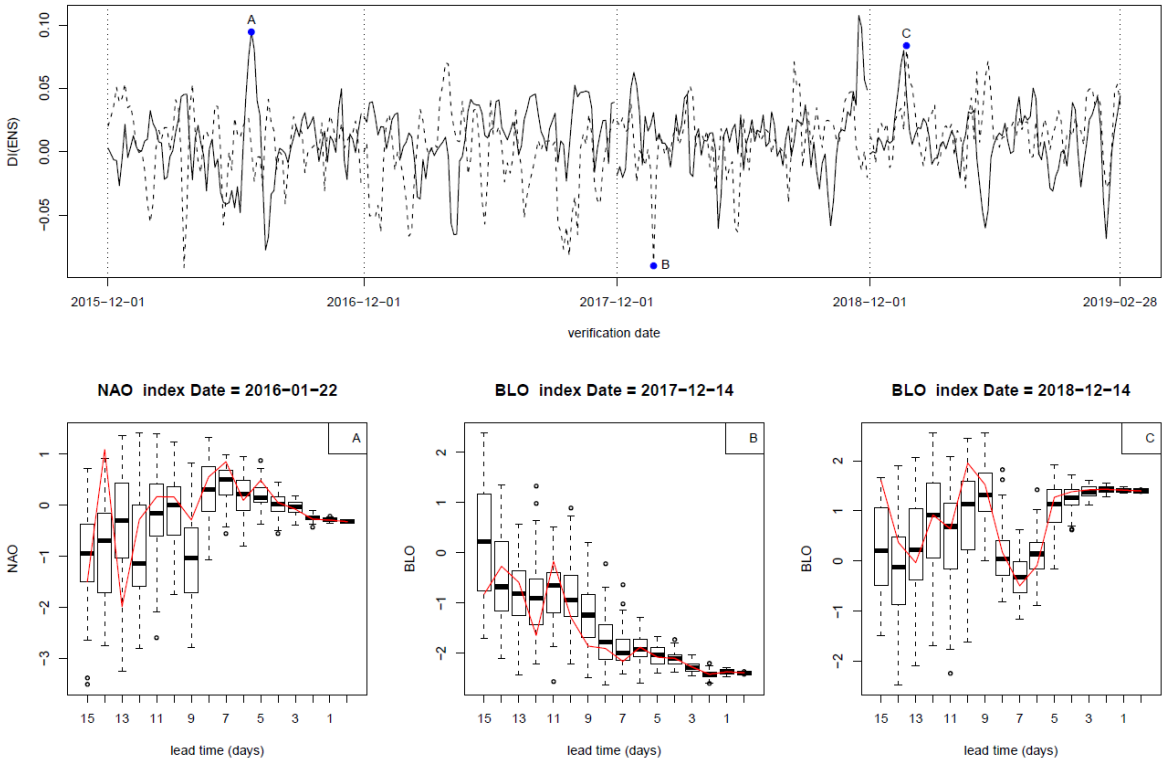


Figure 3.1 Consistency in ENS forecasts of the NAO and BLO regimes. Upper panel: timeseries of overall consistency DI(ENS) of 1-15 day forecasts verifying during winters (December-February) 2015-2019. Positive values indicate lower consistency. Lower panels show examples of both consistent and inconsistent cases for each regime. Each example shows the distribution of ENS forecasts verifying for a given date with lead time of 1 to 15 days; box and whiskers show min, max and 25, 50 and 75 percentiles of the ENS distribution (50 perturbed members); red line shows the ENS control.

To illustrate the different levels of consistency associated with the high and low DI, three example cases are shown in the lower panel (labelled A, B, C on top panel). B (centre) shows an example of a case with very good consistency in forecasting the BLO regime. The plot shows the amplitude of blocking for 14 December 2017 predicted by forecasts initialized between 30 November and 13 December. The 15-day ENS forecast has a broad distribution (large spread), similar to the climate distribution. Subsequent forecasts show smaller spread and a consistent shift of the ENS towards negative BLO.

C (right) shows a contrasting case with poor consistency in forecasting blocking. The plot shows the amplitude of blocking for 14 December 2018 predicted by forecasts initialized on 30 November to 13 December. The longest-range forecasts are similar to the climate distribution; there is a trend over the following days showing an increasing probability for blocking. However, there is then an abrupt change in the forecast to a strong signal for neutral conditions, followed by an equally abrupt change back to blocking. This is the most inconsistent BLO case of this whole period.

A (left) is a case of large inconsistency for the NAO. This occurs at the end of an extended period of strong NAO- (and associated cold weather over NW Europe). The forecasting challenge in this case is to identify when this cold event will end. The longest-range forecasts show large uncertainty, but with

probability of around 50% for a return to near-normal conditions (NAO magnitude <1). The forecasts from 11 January onwards show much higher probability for the end of the NAO- event, with the exception of the forecast from 13 January which again gives a higher probability for the cold spell to continue beyond 21 January.

These cases of large inconsistency illustrate the challenge for users – in both there is an apparent increase in certainty for a change in weather type (regime). But this is thrown into doubt by a large change in a subsequent forecast. The following jump back is also difficult for the user to manage - can it be trusted, or will the following forecast jump again? While such cases are uncommon in the ENS (Figure 3.1, top), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

The consistency of ENS is compared with that of the control forecast and of the EM in Figure 3.2 for NAO (results for BLO are similar). Overall, DI is much larger for the EM (mean DI 0.14) and especially CTRL (0.42) than for ENS (0.01), reflecting how the full ENS distribution does mitigate the jumpiness seen in the deterministic forecasts. The cases with large DI(ENS) also tend to have large DI(EM), and vice versa. The examples of inconsistent ENS forecasts in Figure 3.1 are typical – there is a substantial shift of the whole ENS distribution, which is reflected in both DI(EM) and DI(ENS). For more consistent cases, the correlation is less strong. When the whole ENS distribution is very consistent, the EM must also be consistent. However, when the EM is consistent there may still be variation in the ENS distribution as a whole (for example changes in spread) that can lead to larger DI(ENS).

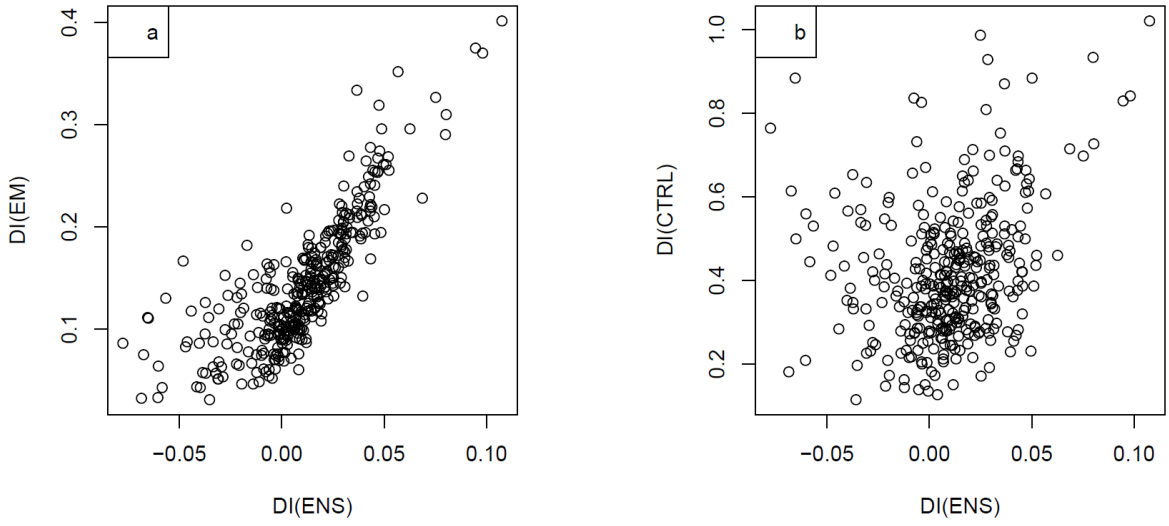


Figure 3.2. Comparison of consistency of ENS, CTRL and EM for NAO. Each panel shows a scatter plot of DI(ENS) on the x-axis against a) DI(EM) and b) DI(CTRL). 361 cases verifying during winters (December-February) 2015-2019.

There is much less correlation between DI(ENS) and DI(CTRL). The most inconsistent cases for ENS tend to be associated with a substantial shift in the whole ENS distribution and the control also shows

large inconsistency as expected. However, there are also cases with large DI(CTRL) but small DI(ENS) – large jumps in CTRL are not reflected in the ENS as a whole, as seen in the examples. This is an important result that demonstrates that jumpiness in the ENS is not simply a consequence of a corresponding jumpiness in the CTRL.

Figure 3.3 shows the distribution of magnitude of the individual jumps ($|D(t_v, i)|$, absolute value of difference between forecasts started 1 day apart) at each lead time for both ENS and CTRL. The two inconsistent cases A and C from Figure 3.1 are highlighted. As well as having large overall DI(ENS), both cases have some of the largest individual ENS jumps between consecutive forecasts at any lead time. As for DI, the magnitude of the individual jumps is much larger for CTRL than for ENS.

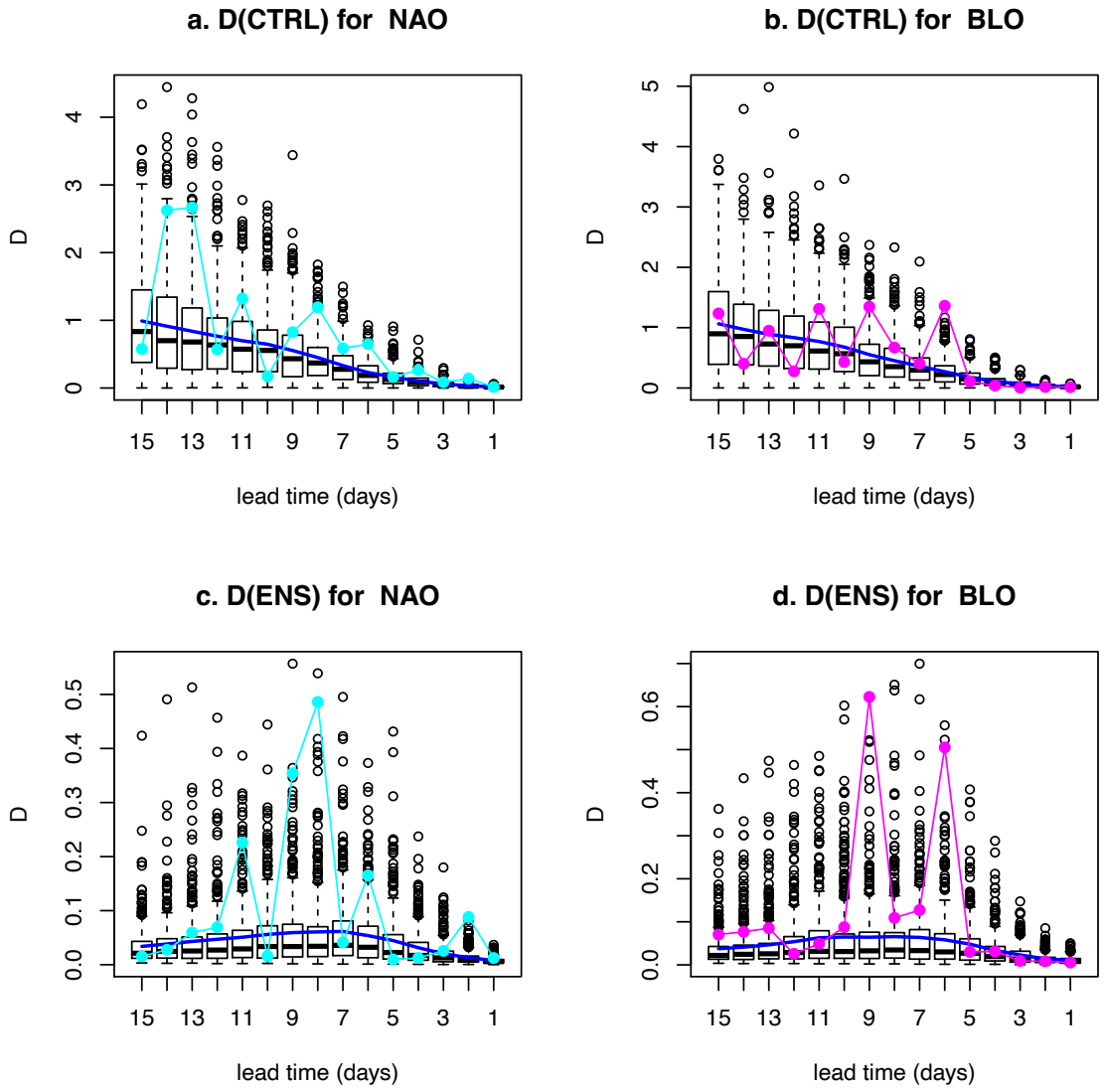


Figure 3.3. Distribution of jumps ($|D(t_v, i)|$) at each forecast lead time (i days) for CTRL (top) and ENS (bottom) for the NAO (left) and BLO (right) regimes. box and whiskers show 25, 50 and 75 percentiles of the ENS distribution, with outliers shown by open circles; thick blue lines show the mean value. The values for the sequence of forecasts verifying on 22 January 2016 for NAO (cyan) and 14 December 2018 (magenta) correspond to the two examples of inconsistent forecasts shown in Figure 3.1.

Figure 3.3 highlights another important difference between the jumpiness of the ENS and CTRL. For CTRL, $|D(t_v, i)|$ increases with lead time, with the mean jump approaching 1 by day 15. However, for ENS the largest mean value and most extreme jumps tend to occur at around 7-9 days lead. At longer lead times, as memory of the initial conditions is lost, the limit of predictability is reached and each forecast behaves like a random draw from the climate distribution. This means that at long lead the difference between two control forecasts will be on average the same as the difference between two randomly selected states from the climate (see text S1 for details). In contrast, at this range, two ENS forecasts will represent two statistically indistinguishable samples from the same climate distribution. Any difference between them will only be due to sampling and for a sufficiently large ensemble $D(t_v, i)$ will be small.

We have seen that DI can identify cases of high inconsistency in the ENS. A more detailed investigation of such cases is merited to understand what aspects of the ensemble forecast configuration lead to such behaviour. The high-DI cases A and C (Figure 3.1) both occur in situations of transitions between large-scale regimes. A compact way to visualize these transitions is in a phase-space plot which can be used to examine how the magnitude of both BLO and NAO evolve through the forecast for each ensemble member (Ferranti et al. 2018). Following this approach for high-DI cases also brings some new insight into the jumpiness itself.

To illustrate this, we consider the BLO case of 14 December 2018 (C in Figure 3.1) and examine the phase-space trajectories of the relevant forecasts. We compare the forecasts started on 5 and 9 December (which both predict a positive BLO pattern) with the contrasting forecast from 7 December which has largest probability for a negative BLO to occur (Figure 3.4a). Figure 3.4b (and Figure 3.S1) shows the phase-space evolution of the forecasts from 5, 7 and 9 December 2018. The forecast from 9 December follows the observed trajectory with only a few members moving too quickly away from the block. The forecast from 7 December also follows the observed trajectory for the first 4-5 days of the forecast, but then most members fail to maintain the blocking and evolve too quickly towards the more mobile NAO+ pattern, leading to the poor 7-day forecast for BLO (Figure 3.4a, cyan). The forecast from 5 December does not follow the observed trajectory so well from 9 December onwards: most ENS members move too quickly into a strong blocking, and NAO-. Although this forecast gives a strong indication of blocking for 14 December (day 9 forecast, Figure 3.4a, blue), the evolution leading to this is clearly inconsistent with the observed development. While Figure 3.4a suggests that the forecast from 7 December has lost the signal that was present in earlier forecasts, the analysis of the phase space trajectories shows that the situation was more complex. In fact, the forecast from 7 December better captured the observed evolution up to 11 December, with significantly smaller ENS spread. Neither the 5 December nor the 7 December forecast captured the observed trajectory after this time.

It was only the later forecasts, from 9 December onwards that correctly predicted the observed evolution.

This shows us that care is needed in the interpretation of the ensemble jumpiness. An apparent clear flip-flop in a single index may hide a more complex predictability issue. When investigating the cause of a case of high DI, it is important to frame the analysis in the right context, as shown by Figure 3.4. From a diagnostic point of view, Figure 3.4a raises the question: why does the forecast from 7 December lose the signal that was present in the earlier forecast from 5 December? In contrast, looking at the wider context of Figure 3.4b raises the question: what mechanism caused the two successive changes in predictability, first to avoid the too strong NAO-/BLO (5 December forecast), and secondly to maintain the block and not move too quickly to NAO+ (7 December forecast). Error tracking (Magnusson 2017a; Grams et al. 2018) shows that both these errors can be traced back to the initial mishandling of developing trough-ridge patterns over eastern North America (Figures 3.S2 and 3.S3).

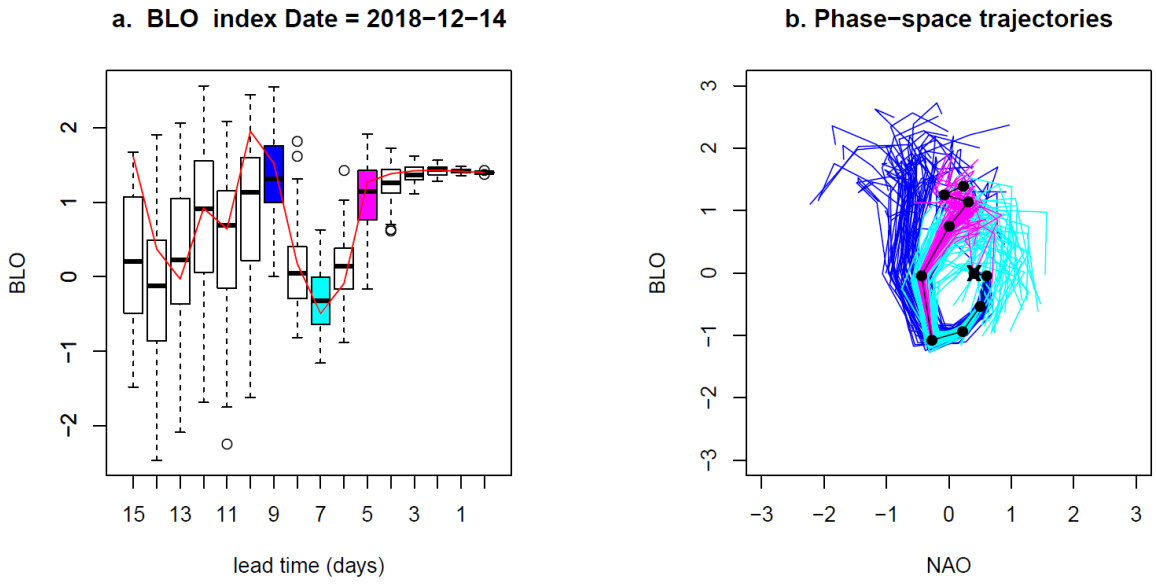


Figure 3.4. Phase space trajectories of ENS forecasts initialized on 5, 7, 9 December 2018. a: amplitude of blocking for 14 December 2018 predicted by forecasts from different initial times up to 15 days ahead with forecasts from 5, 7 and 9 December highlighted; box and whiskers show min, max and 25, 50 and 75 percentiles of the ENS distribution, with outliers shown by open circles; red line shows the ENS control. b: phase space trajectories of ENS forecasts initialized on 5, 7, 9 December 2018 (blue, cyan, magenta respectively) and verifying analysis trajectory (black; analysis position on 5 December marked by x, subsequent days marked by dots).

3.5 Conclusions

Predicting transitions between large-scale weather regimes two weeks ahead is a significant forecasting challenge. Occasionally, successive ensemble forecasts can give contradictory indications about the probability for a change in weather type. Such jumpiness or “flip-flopping” is difficult for

users to manage since the forecast does not give a consistent message for decision making. While such cases are uncommon (Figure 3.1), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

For the first time, we have carried out a systematic, objective evaluation of the consistency of ECMWF ensemble forecasts that takes account of the full ensemble distribution. This extends the earlier work of Zsoter et al. (2009) who focused specifically on flip-flops of the ensemble mean.

We investigated the ENS consistency for two key flow patterns for Europe, NAO and blocking. We used a measure of the divergence between two ensembles started at different times but valid for the same time. This allowed us to quantify both individual jumps and the overall consistency of a sequence of ENS forecasts valid for a given time. Our main conclusions are:

- In general, the peaks of high and low consistency occur at different times for NAO and BLO; there is no strong correlation between inconsistency for NAO and BLO (Figure 3.1).
- DI for the ENS is on average much lower than for EM and especially for CTRL (Figure 3.2) demonstrating benefit of the ensemble in mitigating the jumpiness of the deterministic forecasts by representing the range of possible scenarios.
- The largest individual jumps for ENS tend to be days 7-9, while for the CTRL the magnitude of individual jumps continues to increase throughout the forecast (Figure 3.3). This is associated with the different asymptotic behaviour of the (deterministic) CTRL forecast and the ENS at long forecast lead.
- Care is needed in the interpretation of the ensemble jumpiness. What looks at first sight to be a clear case of flip-flopping in a single index (BLO or NAO) may be a more complex predictability issue. This may be better understood by examining the phase-space evolution of both components together (Figure 3.4).

In this work, we assessed the consistency of the univariate forecast of NAO and BLO separately. However, we also showed how it is important to consider the ensemble trajectories in the 2-dimensional phase space to properly understand the reason for apparent jumpiness. It will therefore be valuable to extend the divergence and DI methodology to the multivariate situation so that the consistency of NAO and BLO can be evaluated together. This will also enable investigation of the consistency of other aspects of ensemble performance such as for tropical cyclone tracks.

The divergence index (DI) allows us to identify important cases of high ensemble forecast inconsistency, and to routinely monitor the occurrence of such cases. Careful diagnosis of these cases will help to identify the causes of the inconsistency and hence to address the relevant aspects of

ensemble configuration and modelling. Reducing the occurrence of inconsistent (or jumpy) ensemble forecasts will increase user confidence and improve decision-making.

3.6 Supporting information

3.6.1 *Introduction*

This supporting information provides details on the asymptotic limit for jumps of the CRTL forecast and four figures with explanatory text detailing the evolution of the errors in the forecasts from 5, 7, 9 December 2018.

Text S1 explains the theoretical limit for the magnitude of individual jumps in the CTRL forecast.

Text S2 describes the evolution of the errors in the ensemble mean forecasts from 5, 7, 9 December 2018 which are shown in Figures 3.S2-S4. Figure 3.S1 shows the magnitude of the BLO and NAO projections for all ensemble members of these forecasts at 24-hour intervals. This is the same information as shown in Figure 3.4b, but with each lead time shown separately for 9-14 December for extra clarity.

3.6.2 *Text S1*

Figure 3.3 shows the distribution of magnitude of the individual jumps $|D(t_v, i)|$ at each lead time for both ENS and CTRL. For CTRL, For CTRL, $|D(t_v, i)|$ increases with lead time, with the mean value approaching 1 by day 15. Here we consider the asymptotic limit for this mean value.

At long lead times, each forecast behaves like a random draw from the climate distribution, i.e. two control forecasts $f(t_v, i)$ and $f(t_v, i-1)$ will be uncorrelated for sufficiently large i. The average distance (divergence) between two such random states, f and g, is

$$\overline{d_r} = \overline{|f - g|}$$

where the overbar denotes the average of all cases, the subscript r indicates this is for random selection of states, and recall that for the deterministic forecast the divergence d is the absolute distance $|f - g|$.

If the climatology is normally distributed then we can compute $\overline{d_r}$ analytically

$$\overline{d_r} = \frac{2}{\sqrt{\pi}}\sigma \approx 1.13\sigma$$

where σ is the climate standard deviation. In our study the NAO and BLO projections are already normalized by the climate standard deviation, so that $\sigma = 1$. Hence we could expect the mean curves for CTRL shown in Figure 3.3 to tend to 1.13 in the long-range as predictability is lost. The fact

that this limit has not quite been reached by day 15 is suggests there is some predictability still at this range.

The above relies on the assumption that the climate distribution is normal. If we make no assumption about the climatological distribution of the projections, then we cannot make an analytic value for d_r but we can derive an upper limit.

The mean absolute distance $\overline{d_r}$ is related to the mean squared distance $\overline{d_r}^2$:

$$\overline{\left(d_r - \overline{d_r}\right)^2} = \overline{d_r}^2 + \overline{d_r}^2 - 2\overline{d_r}^2 = \overline{d_r}^2 - \overline{d_r}^2$$

The mean-squared difference between two random states can be written as

$$\overline{d_r}^2 = \overline{(f-g)^2} = \overline{f^2} + \overline{g^2} - 2\overline{fg} = 2\sigma^2$$

since by definition the random states f and g are uncorrelated, $\overline{fg} = 0$, and $\overline{f^2} = \overline{g^2} = \sigma^2$.

This is the standard result that asymptotically the control forecast error will be equal to twice the climatological variance.

Hence, from the above two equations we see that

$$\overline{d_r}^2 = \overline{d_r}^2 - \overline{\left(d_r - \overline{d_r}\right)^2} = 2\sigma^2 - \overline{\left(d_r - \overline{d_r}\right)^2}$$

This shows that the mean absolute distance $\overline{d_r}$ is never greater than the root mean squared distance $\sqrt{\overline{d_r}^2}$. As noted above, in this study $\sigma = 1$, and so the upper limit for the asymptotic (long-lead) limit for the mean curves for CTRL in Figure 3.3 is $\sqrt{2}$.

3.6.3 *Text S2*

The forecast from 5 December develops too strong BLO and NAO- for 11-12 December (Figure 3.S1). This can be seen clearly in the ensemble mean forecast (Figure 3.S2) – there is a large error in the ridge over the E Atlantic, the axis of the ridge is too far west and there is a too strong north-westwards extension over Iceland and towards Greenland (consistent with the NAO- signal). The troughs either side of this main ridge are too deep, giving overall a much too strong omega block (enhanced omega pattern). The error pattern established by 11 December remains through the rest of the forecast. This error pattern can be traced back through the forecast to an over amplification of the trough-ridge structure over eastern North America the western Atlantic in the 72-hour forecast: enhanced ridge over the Hudson Bay, slightly extended trough off the eastern seaboard and small overdevelopment of the ridge in the mid-Atlantic.

The forecast from 7 December captures much better the initial trough ridge structure (on 9 December) and does not overextend the meridional pattern, resulting in much lower errors over N Atlantic/Europe on 11-12 December (Figure 3.S3). However, upstream errors in a following trough-ridge pattern (also originating with positive error over the Hudson Bay and negative errors over the east coast) amplify downstream. In this case though the interaction with the pre-existing ridge appears to speed up the anticyclonic wave breaking and the high pressure moves further downstream. This results in especially large error over western Europe.

It is worth noting that the forecast from 5 December also has very similar error structure that develops in this second trough-ridge pattern (compare the centres of the error positive and negative over the west Atlantic on 12 December in Figures 3.S2 and 3.S3). However, the much extended and higher amplitude pre-existing block in the central Atlantic appears to limit the impact of this second error pattern.

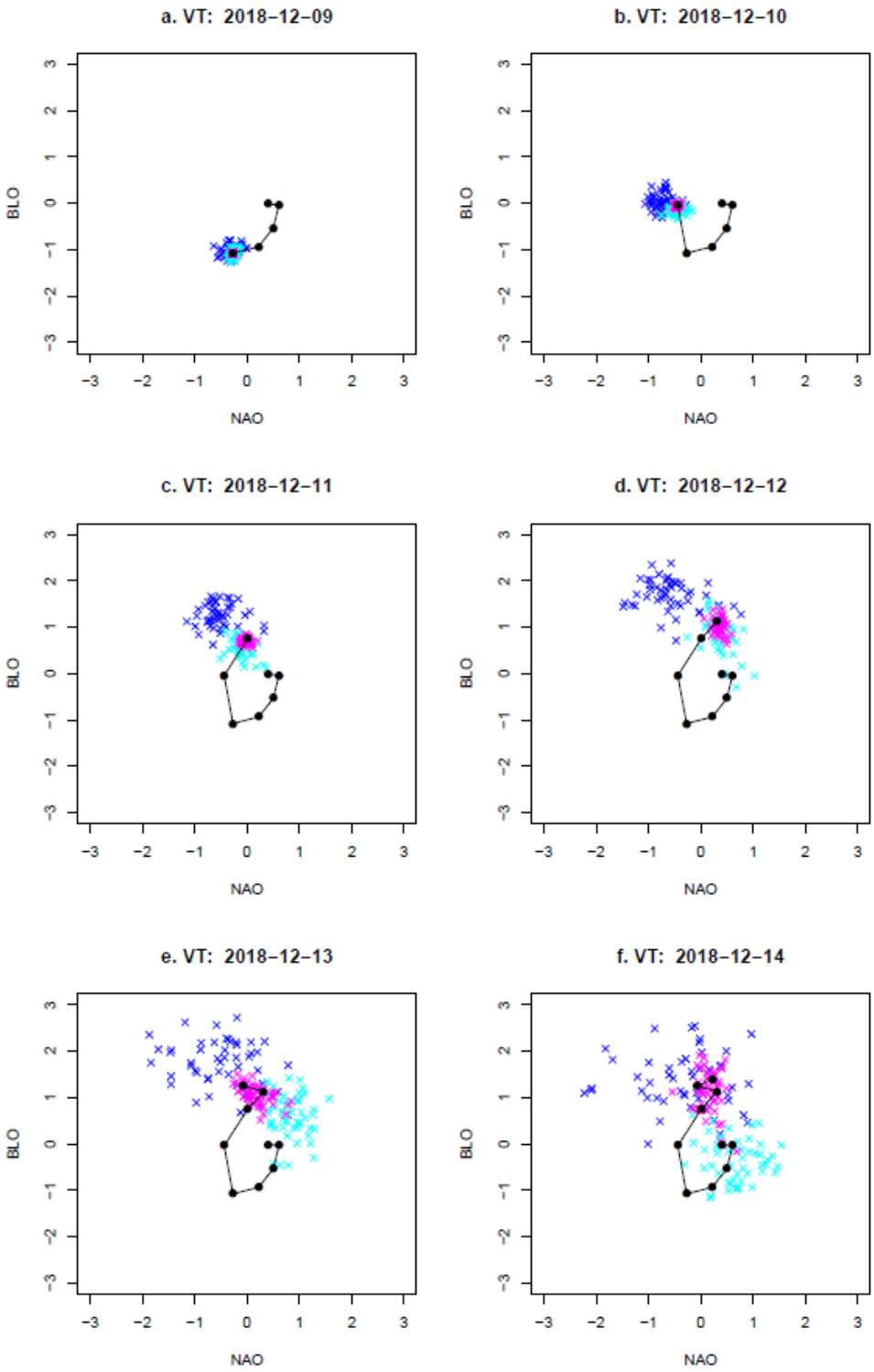


Figure 3.S1. Phase space plots of ENS forecasts initialized on 5, 7, 9 December 2018 (blue, cyan, magenta respectively) verifying on 9-14 December (panels a to f). In each panel the verifying analysis trajectory is shown at 24-hour intervals from 5 December to the verifying date (black line).

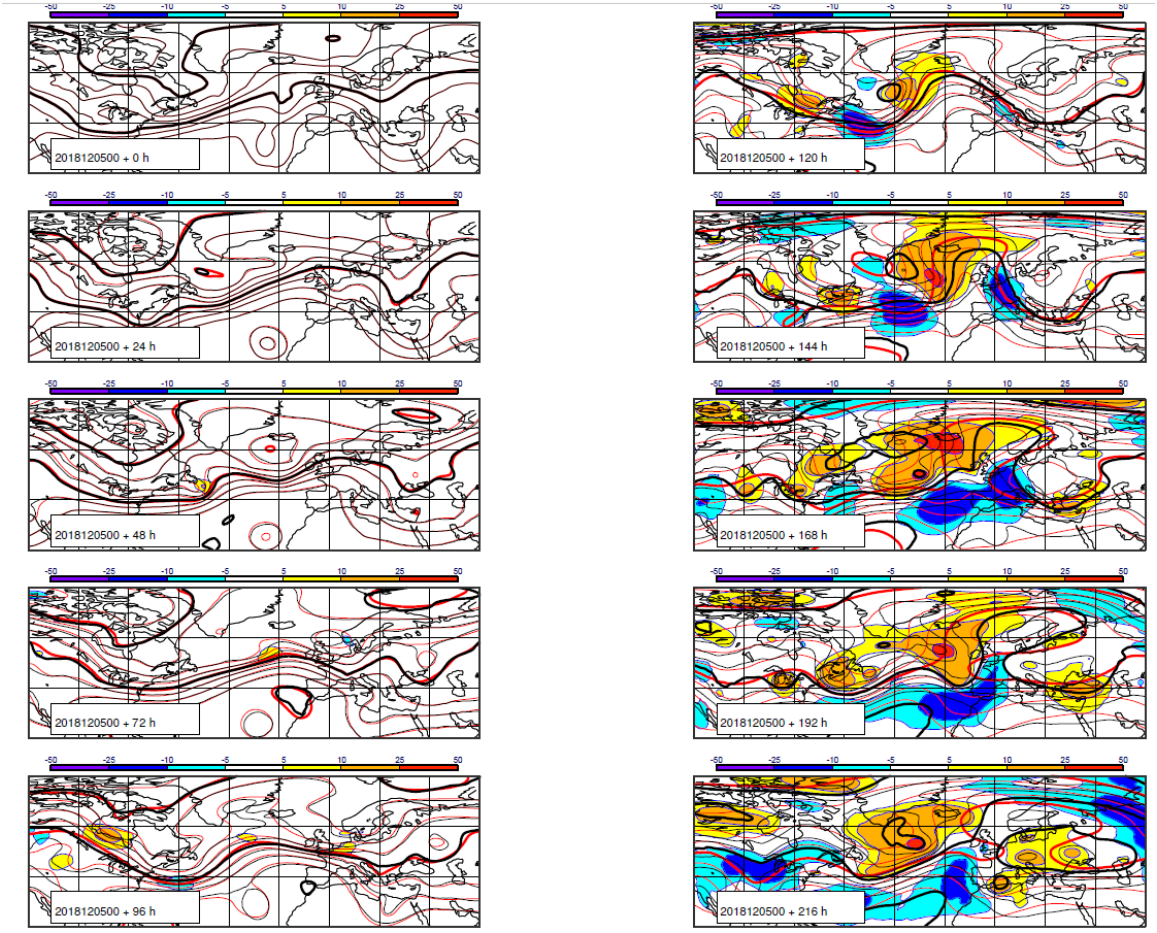


Figure 3.S2. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 9 December 2018 together with the verifying analysis (black), at 24-hour intervals.

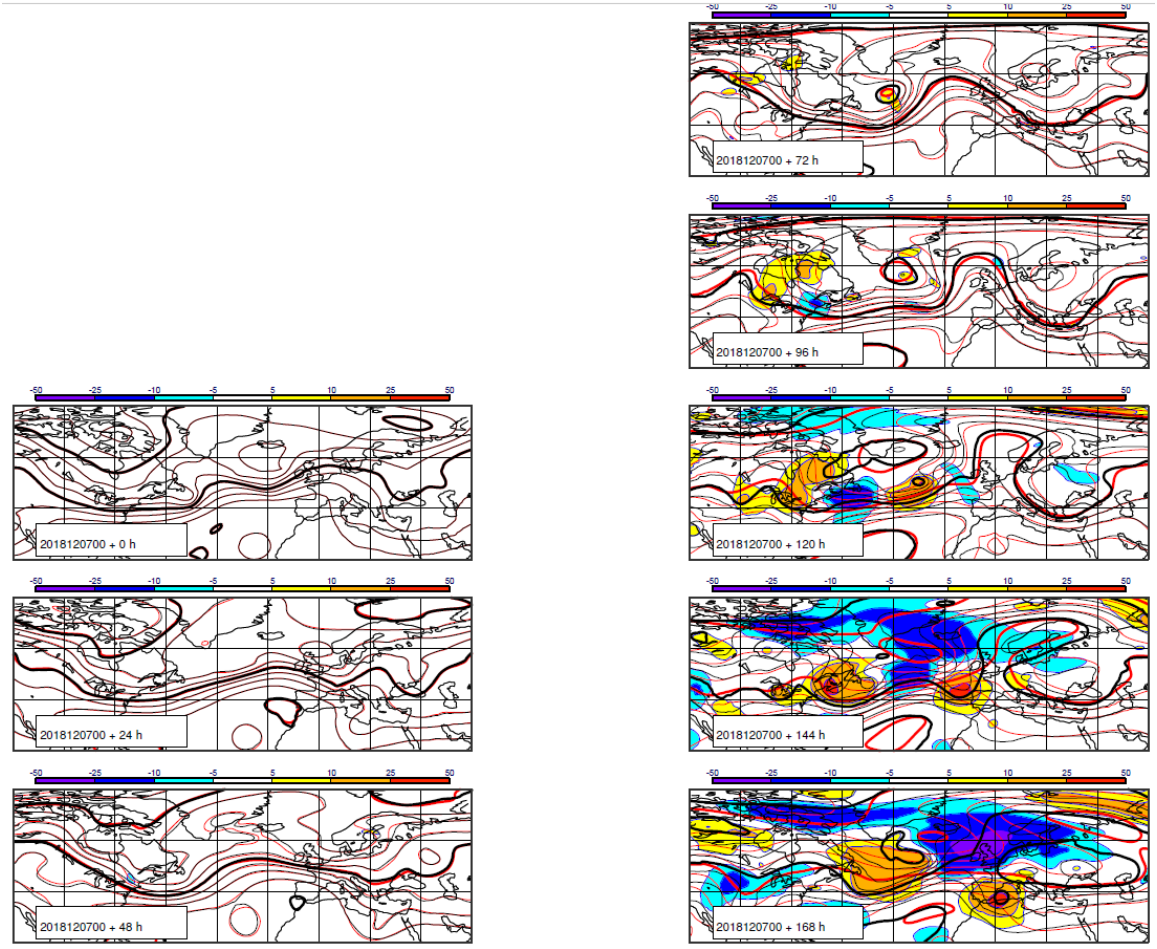


Figure 3.S3. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 7 December 2018 together with the verifying analysis (black), at 24-hour intervals.

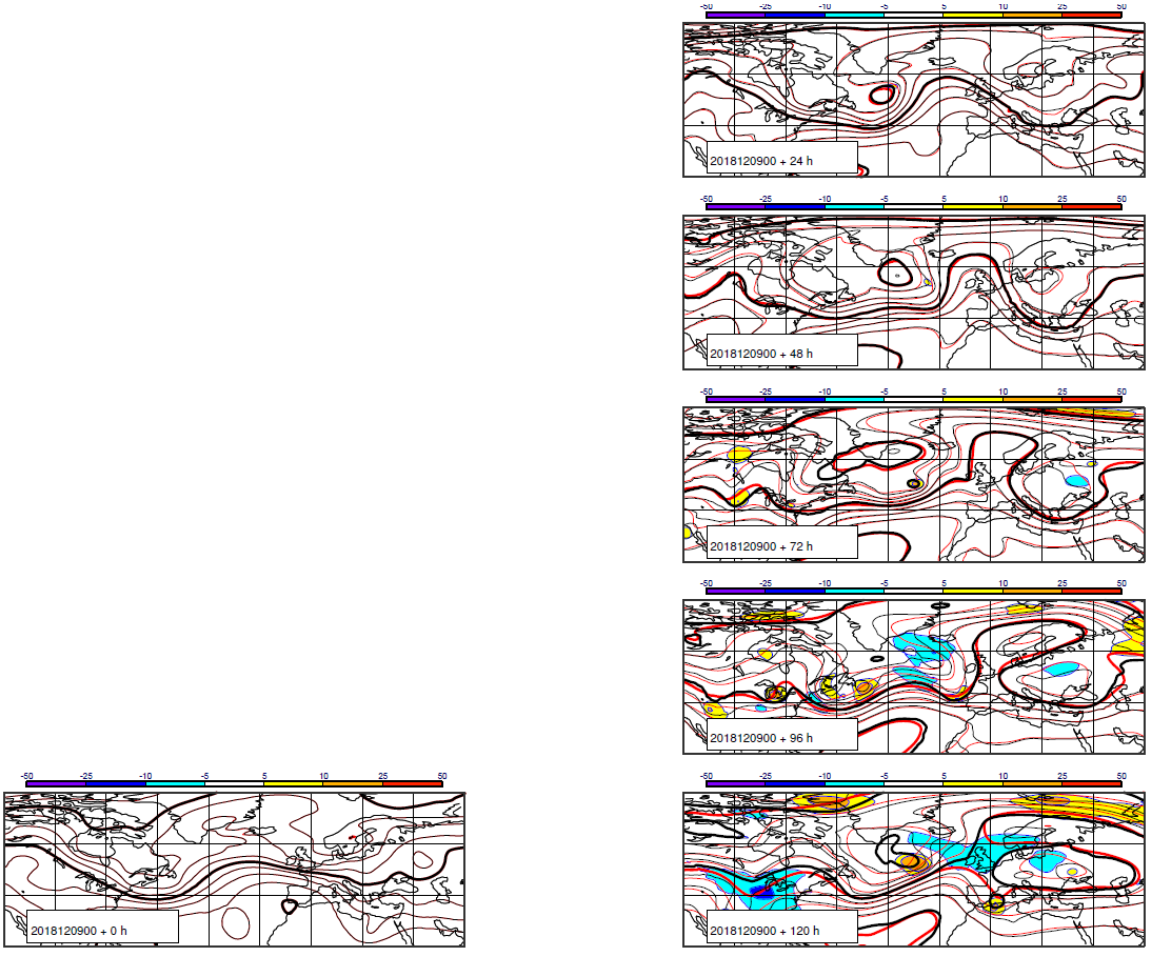


Figure 3.S4. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 9 December 2018 together with the verifying analysis (black), at 24-hour intervals.

Chapter 4 Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks

The second objective of the PhD was to evaluate and compare the jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks from three operational centres, identify any common factors and provide guidance to users. The paper addressing this objective was published in Weather and Forecasting with the following reference:

Richardson, D.S., Cloke, H.L., Methven, J.A. and Pappenberger, F. (2024) 'Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks', *Weather and Forecasting*, 39(1), pp. 203–215. Available at: https://doi.org/10.1175/WAF-D-23-0113.1.

The contributions of the authors of this paper are as follows: D.R. designed the study with advice from H.C., J.M. and F.P., obtained the datasets, carried out the analysis, and led the writing of the manuscript. All authors assisted with writing the manuscript. Overall, 90% of the writing was undertaken by D.R.

The published article can be found in Appendix A2.

Abstract. We investigate the run-to-run consistency (jumpiness) of ensemble forecasts of tropical cyclone tracks from three global centres: ECMWF, the Met Office and NCEP. We use a divergence function to quantify the change in cross-track position between consecutive ensemble forecasts initialized at 12-hour intervals. Results for the 2019-2021 North Atlantic hurricane season show that the jumpiness varied substantially between cases and centres, with no common cause across the different ensemble systems. Recent upgrades to the Met Office and NCEP ensembles reduced their overall jumpiness to match that of the ECMWF ensemble. The average divergence over the set of cases provides an objective measure of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80-90 km in the cross-track direction between a forecast for 120 hours ahead and the updated forecast made 12 hours later for the same valid time. This quantitative information can support users' decision making, for example in deciding whether to act now or wait for the next forecast. We did not find any link between jumpiness and skill, indicating that users should not rely on the consistency between successive forecasts as a measure of confidence. Instead, we suggest that users should use ensemble spread and probabilistic information to assess forecast uncertainty, and consider multi-model combinations to reduce the effects of jumpiness.

Plain Language Summary. Forecasting the tracks of tropical cyclones is essential to mitigate their impacts on society. Numerical weather prediction models provide valuable guidance, but occasionally there is a large jump in the predicted track from one run to the next. This jumpiness complicates the creation and communication of consistent forecast advisories and early warnings. In this work we aim to better understand forecast jumpiness and we provide practical information to forecasters to help them better use the model guidance. We show that the jumpiest cases are different for different modelling centres, that recent model upgrades have reduced forecast jumpiness, and that there is not a strong link between jumpiness and forecast skill.

4.1 Introduction

Official forecasts of tropical cyclone (TC) tracks are typically based on guidance from Numerical Weather Prediction (NWP) models (Conroy et al. 2023). NWP ensemble forecasts are increasingly being used. Although their use in official forecasts is often limited to the ensemble mean (EM) track, there is increasing evidence of the benefits of using more of the ensemble probabilistic information (Titley et al. 2019, 2020; Kawabata and Yamaguchi 2020; Leonardo and Colle 2017). One benefit of using ensembles is the increased consistency between consecutive forecasts (Buizza 2008b; Zsoter et al. 2009). There are nevertheless occasions where an ensemble is unexpectedly jumpy with the predicted TC locations flip-flopping over several consecutive forecasts (Magnusson et al. 2021). Such cases can be difficult to interpret, complicating the creation of consistent forecast advisories and early warning communications. Understanding the frequency and reasons for these cases as well as information about the overall levels of consistency in operational ensemble forecasts can help forecasters to better use the available ensemble track data.

As new forecast information arrives (usually every 6-12 hours for global NWP models), forecasters need to decide how to revise their forecasts to take account of the new forecast information. National Hurricane Center (NHC) Tropical Cyclone Advisories often discuss the change in forecast track due to updated guidance, making adjustments to the path depending on the new information. There is a balance to be struck between closely following the changed model guidance and taking a more conservative approach of making a smaller change to minimise the potential need to make a change in the opposite direction later, that is to avoid a so-called windshield-wiper effect (Broad et al. 2007). Contradictory messages from such jumpiness can cause difficulties for decision-makers and reduce users' confidence in the forecasts (Hewson 2020; Pappenberger et al. 2011b; McLay 2011; Elsberry and Dobos 1990). Information quantifying the consistency between successive probabilistic forecasts can be important to inform optimal decision making, such as whether to act now or wait for the next

forecast (Regnier and Harr 2006; Jewson et al. 2022, 2021). Both noted that such information is not readily available to users.

Evaluation of operational ensemble TC track forecasts includes EM track errors, ensemble spread and strike probability (e.g. Cangialosi 2022; Haiden et al. 2022; Titley et al. 2020; Heming et al. 2019; Leonardo and Colle 2017). However, few authors have addressed the jumpiness of TC track forecasts. Elsberry and Dobos (1990) investigate consistency of TC guidance for the Western North Pacific by using the difference in cross-track errors between successive forecasts. Fowler et al. (2015) assess consistency of Atlantic TC track forecasts by counting forecast crossovers – how often in a sequence of forecasts the predicted position changes from one side to the other of a fixed reference track, for example the observed track. However, they caution that biased forecasts may appear to be consistent since successive forecasts may jump considerably without crossing the observed track. Both Elsberry and Dobos (1990) and Fowler et al. (2015) recommend the regular evaluation of forecast consistency in addition to the standard assessments of forecast accuracy.

More generally, there has been limited investigation of forecast jumpiness, especially for ensemble forecasts. Zsoter et al. (2009) considered flip-flops in sequences of forecasts all valid for a given time and showed that EM forecasts are more consistent than the corresponding ensemble control forecasts. Griffiths et al. (2019) introduced a flip-flop index to compare the consistency of automated and manual forecasts, while Ruth et al. (2009) assessed how model output statistics improved forecast consistency. Forecast consistency has been considered for rainfall (Ehret 2010) and river flow (Pappenberger et al. 2011b).

These previous studies were mainly focused on deterministic forecasts (either single runs or EM) and the methods are not directly applicable to assess the jumpiness in sequence of ensemble forecasts taking account of the full ensemble distribution. Recently, Richardson et al. (2020b) introduced a measure of forecast jumpiness based on forecast divergence that accounts for all aspects of the ensemble empirical distribution. They used this to investigate jumpiness of ensemble forecasts for the large-scale flow over the Euro-Atlantic region.

In the present study we apply the forecast jumpiness measure introduced by Richardson et al. (2020b) to ensemble forecasts of Atlantic TCs, focusing on the run-to-run consistency in the cross-track direction which is most important in determining the location of TC landfall. The aim is to provide forecasters and model developers with information about the jumpiness of ensemble TC forecasts. This will help forecasters and decision-makers better understand the expected changes between successive forecasts. We address the following questions:

- How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centres?
- Is there a common cause of 'jumpy' cases – are the ensembles from different centres particularly jumpy for the same TC cases and if so what is the reason?
- Have recent ensemble model upgrades had a noticeable effect on the forecast consistency?
- What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness – what information is practically useful? Is there any useful link between jumpiness and skill?

We investigate these questions using ensemble forecast data from three global NWP centres. The data used in this study and the methods to assess forecast jumpiness are introduced in sections 4.2 and 4.3. Results are presented in section 4.4. We start with a case study to illustrate the issues of ensemble TC track jumpiness. Then we look at the overall jumpiness over the 2019, 2020 and 2021 Atlantic hurricane seasons. Finally, we consider the relationship between jumpiness, error and spread. We conclude with a summary, recommendations for forecasters and avenues for future work in section 4.5.

4.2 Data

In this study we investigate the run-to-run consistency of ensemble tropical cyclone track forecasts from three global centres: the European Centre for Medium-Range Weather Forecasts (ECMWF), the US National Centers for Environmental Prediction (NCEP) and the UK Met Office. Each centre runs its own tropical cyclone tracker (Conroy et al. 2023) and the resulting track forecasts are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the TIGGE forecast tracks for all available dates from the Atlantic basin for 2019, 2020 and 2021 for forecasts initialized at 00 and 12 UTC from the ECMWF ensemble (ENS, 51 members integrated on ~18 km grid), NCEP ensemble (GEFS, 21 members, ~34 km grid until 22 September 2020; 31 members, ~25 km grid from 23 September 2020 onwards), and Met Office ensemble (MOGREPS-G, 36 members, ~20 km grid). A given TC is not always tracked in every ensemble member (for example because the system dissipates in that member or the forecast intensity is below the threshold used in the tracking algorithm) and we exclude cases where a centre has fewer than 10 members that track the TC at each forecast step.

We use the observed TC positions from International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al. 2018, 2010). We concentrate our analysis on named Atlantic tropical cyclones and for each cyclone include all 00 and 12 UTC verification times when the observed system is at least tropical storm strength (winds at least 34 kt) and the system is reported as tropical in IBTrACS (Titley

et al. 2020; Goerss 2000). For each of these verification times we consider all available TIGGE forecasts. These include forecasts initialized when the TC is still a tropical depression. However, TIGGE forecast tracks are only generated for existing TCs, so longer lead-time forecasts are not always available for verification times close to when the TC is first analysed as a tropical storm. This means that overall there are fewer forecasts for longer lead times than for shorter lead times in our sample.

We make a homogeneous sample by only including a case if the ensemble data is available from each of the three centres. This ensures that we are comparing the different centres over the same set of cases. The total number of cases decreases with forecast lead time from 356 for 12-hour forecasts to 91 for 120-hour forecasts. To maintain a reasonable sample we restrict the study to forecasts of 120 hours or less.

Our focus is on the changes between successive forecasts for a given verification time. We therefore need to set a minimum number of consecutive initial times over which we can assess these changes. For a given verification time t_v, we require a minimum of 6 consecutive forecasts, initialized at $(t_v - 12\ \text{hours})$, $(t_v - 24\ \text{hours})$, up to $(t_v - 72\ \text{hours})$, all valid for t_v. To ensure homogeneity, the same cases must be available from all three centers. With these conditions, the total number of available cases to assess the run-to-run jumpiness is 139 over the three-year period.

Each NWP centre has made upgrades to their operational ensemble system during the 2019-21 period used in this study. A major upgrade to the GEFS was implemented on 23 September 2020, including the introduction of a new forecast model and an increase in the number of ensemble members from 20 to 30 (Zhou et al. 2022). This upgrade brought significant improvements to the ensemble performance, including for tropical cyclone forecasts. The MOGREPS-G ensemble was upgraded on 4 December 2019, including a major change to the generation of the ensemble perturbations (Inverarity et al. 2023) and revised model physics (Walters et al. 2019). This upgrade improved TC track errors (Met Office 2019).

Upgrades to the ECMWF ENS in June 2019 (Haiden et al. 2019), June 2020 (Haiden et al. 2021) and May 2021 (Rodwell et al. 2021) were neutral in terms of TC track performance, although the latter two brought improvements to intensity forecasts (Rodwell et al. 2021; Bidlot et al. 2020). A later upgrade in October 2021 did also improve TC track forecasts (Haiden et al. 2022); however, there was only one Atlantic TC in 2021 after this date. Overall, the ECMWF ensemble track forecast performance can be considered relatively stable over the period of this study. We therefore use the ENS as a reference against which to evaluate the impact of the upgrades of the other centres on ensemble jumpiness.

4.3 Methods

For each tropical cyclone, the observed track provides a convenient frame of reference. We consider jumpiness in a sequence of forecasts in terms of changes in the predicted cross-track location (Elsberry and Dobos 1990). A positive cross-track position indicates that the forecast is to the right of observed track (facing the observed direction of travel). We also consider the links between jumpiness, ensemble error and spread. All scores – error, spread and jumpiness – are computed in terms of the cross-track distance and are defined below.

We measure the cross-track error of the ensemble forecasts using the Continuous Ranked Probability Score (CRPS). The CRPS is widely used for evaluation of ensemble forecasts. It is a so-called proper score: if the 'true' forecast probability distribution is F, a proper score ensures that the best expected score will be achieved using the forecast F rather than any other forecast distribution $G \neq F$. Hence forecasters are rewarded for honest forecasts reflecting their true beliefs. As a proper score, CRPS discourages hedging (Gneiting and Raftery 2007) and rewards both reliability and resolution (Hersbach 2000).

For an ensemble of M members $f_i, i = 1, \ldots M$ the CRPS is given in its kernel representation by

$$\mathrm{CRPS}(f) = \frac{1}{M}\sum_{i=1}^{M}|f_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j| \tag{4-1}$$

where y is the verifying observation (Gneiting and Raftery 2007). The first term is the mean of the absolute error of the individual ensemble members and the second term is the mean of the distances between the different ensemble members which accounts for the ensemble spread.

The ensemble mean forecast is given by

$$\bar{f} = \frac{1}{M}\sum_{i=1}^{M} f_i \tag{4-2}$$

For a single deterministic forecast, the CRPS is equal to the mean absolute error, so the error of the ensemble mean is

$$\mathrm{CRPS}(\bar{f}) = |\bar{f} - y| \tag{4-3}$$

To allow us to compare the mean spread and error over the sample of cases, we use a measure of ensemble spread that is also based on the mean absolute difference. The spread measure which corresponds to the mean absolute error of the ensemble mean is the mean absolute deviation of ensemble members from the ensemble mean:

$$s = \frac{1}{M}\sum_{i=1}^{M}\left|f_i - \bar{f}\right| \tag{4-4}$$

On average over a large sample of cases the ensemble mean error (Eq. 4-3) and spread (Eq. 4-4) should be equal for a well-tuned ensemble system.

To measure the 'jump' from one forecast to the next we follow Richardson et al. (2020b) and use the divergence function d associated with the CRPS. For two ensembles f and g with M and N members respectively, d is given by

$$d(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|f_i - g_j\right| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|f_i - f_j\right| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_i - g_j\right| \tag{4-5}$$

The first term measures the distance between the two ensembles f and g, while the second and third terms reflect the variability (spread) in each ensemble, f and g, respectively. Comparing Eq. 4-5 to Eq. 4-1 shows that the divergence reduces to the CRPS if either M or N is equal to one. If both M and N are one then d is the absolute distance $|f - g|$. The divergence d takes account of both location and spread differences between f and g and, like the CRPS, d is a proper score (Gneiting and Raftery 2007; Thorarinsdottir et al. 2013) which discourages hedging.

Consider a given verification time t_v: an ensemble forecast f valid for this time and initialized h hours before is written $f(t_v, h)$ and individual ensemble members are $f_i(t_v, h)$. In this study $f_i(t_v, h)$ represents the distance (in km) in the cross-track direction from the observed TC location at verification time t_v. The difference between two consecutive ensemble forecasts initialized at time $(t_v - h)$ and $(t_v - (h - 12))$ and valid for the same time t_v is

$$D(t_v, h) = d\big(f(t_v, h), f(t_v, h - 12)\big) \tag{4-6}$$

where d is the divergence function (Eq. 4-5).

To measure the overall divergence between the sequence of L forecasts valid for a given time we use the mean divergence between successive pairs of forecasts:

$$\overline{D(t_v)} = \frac{1}{L-1}\left(\sum_{l=2}^{L} D(t_v, 12l)\right) \tag{4-7}$$

Larger values of $\overline{D}$ indicate greater change (in position, spread or both) between successive forecasts in the sequence. However it does not necessarily indicate jumpiness in the sense of flip-flopping back and forth between different solutions. For example, if in the initial ensemble forecast all members are far to the right of the observed position and subsequent forecasts become progressively closer to the observed location, this will result in large $\overline{D}$. To distinguish between 'trend' cases and 'flip-flop' cases,

we use the difference between the first and last forecasts of the sequence to represent this overall change (trend). Subtracting this difference from $\overline{D}$ gives the Divergence Index (DI) introduced by Richardson et al. (2020b) which highlights jumpiness (flip-flops) in the sequence:

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1} d\big(f(t_v, 12L), f(t_v, 12)\big) \tag{4-8}$$

In this way, DI will be less sensitive than $\overline{D}$ to trends caused by bias or to cases with single large jumps (resulting for example from a sudden increase in predictability). This means that the larger values of DI will be more closely related to flip-flops in the sequence of forecasts.

Our focus is on the performance of the ensemble forecast distribution and both D and DI are computed using all available ensemble members. However, because the ensemble mean (EM) track is also often used in operational forecasting we also compute the same measures for the ensemble mean. Note that for tropical cyclone tracks, the ensemble mean refers to the Euclidean mean position of the tracks from the individual ensemble members and not to a track calculated from the ensemble mean spatial fields.

The statistical significance of differences between the different centres' distributions of $\overline{D}$ and DI are assessed using the Kolmogorov-Smirnov (KS) and Mann-Whitney U (MWU) tests (Wilks 2020). Both tests are non-parametric statistical methods to compare the empirical cumulative distributions of two samples. The MWU test is mainly sensitive to differences in location (e.g. differences in the median), while the KS test is sensitive to differences in both location and shape of the distributions.

4.4 Results

We start with an example to illustrate the issues of jumpiness and sampling. Then we look at the overall jumpiness over 2019, 2020 and 2021 seasons. Finally, we consider the relationship between jumpiness, error and spread.

4.4.1 *Example: Hurricane Laura, August 2020*

Hurricane Laura formed initially as a tropical storm in the western tropical Atlantic on 20 August 2020 and affected several Caribbean countries. After travelling across the Caribbean, it reached hurricane strength on 25 August as it entered the Gulf of Mexico. It made landfall in Louisiana at 06 UTC on 27 August. Here we focus on the ECMWF ensemble (ENS) forecasts for 00 UTC on 27 August, just before the Louisiana landfall. Figure 4.1 shows the ENS tracks for Laura from forecasts initialized every 12 hours between 21 and 25 August. The earliest forecasts, from 20 August 12 UTC (not shown) and 21 August 00 UTC were almost all to the north-east (right-hand side) of the observed track throughout the forecast, and predicted landfall most likely along the central and eastern Gulf coast. From 21

August 12 UTC, the forecasts showed a higher probability for landfall further west, although with a large uncertainty as shown by the distribution of the tracks from the individual ensemble members. Between 22 August 00 UTC and 24 August 00 UTC, successive forecasts exhibited a “flip-flop” behaviour, alternating between the western or more central Gulf coast as the most likely landfall location. Finally, from 24 August 12UTC onwards, the forecasts more consistently indicated the western solution as most likely and it turned out that the observed track was at the eastern (right-hand) end of the range of predicted locations.

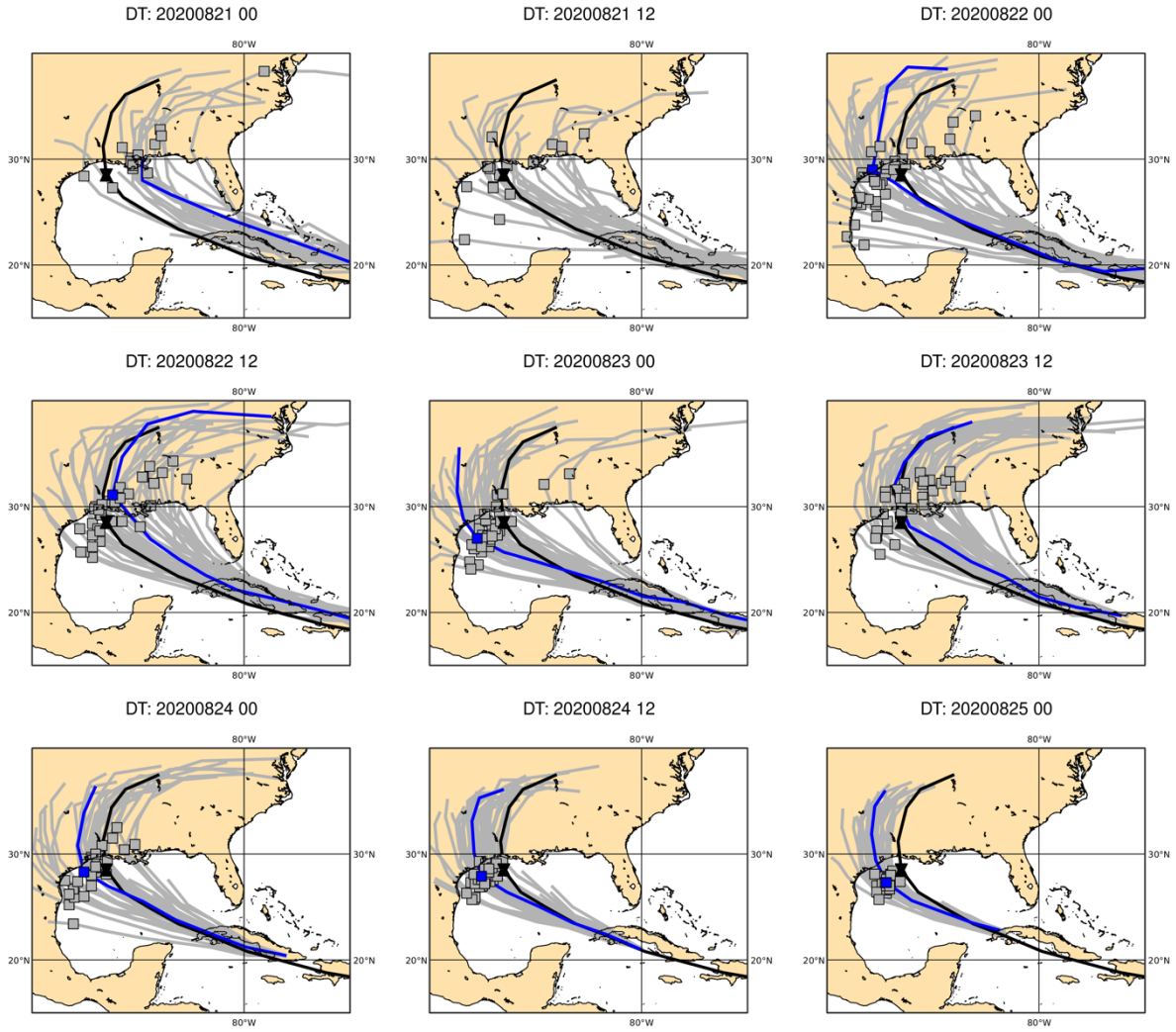


Figure 4.1. Hurricane Laura: ECMWF ensemble forecast tracks (blue: control; grey: perturbed members) and observed track (black). Forecast start dates (DT) from 00 UTC on 21 August to 00 UTC on 25 August 2020. Coloured symbols show forecast and observed (hourglass) position at 00 UTC 27 August.

We can summarize the variations in successive forecasts for a fixed valid time in a box-and-whisker meteogram (Figure 4.2). This shows the distribution of the position in the cross-track direction for all ensemble members valid for 00 UTC on 27 August, from forecasts initialized every 12 hours between 20 August 12 UTC (the first available forecast) and 26 August 12 UTC. Each ENS forecast has one control forecast and 50 perturbed members. However, the number of members that successfully track Laura

until 27 August is substantially below this, especially for the earlier forecasts. Figure 4.2 clearly shows the jumpiness of the ENS forecasts. The earlier forecasts are mainly to the right of the observed track (too far east), while the shorter-range forecasts are too far west (left of observed track). Intermediate forecasts flip-flop between left and right of the observed position. For each lead time (except the 48-hour forecast from 00 UTC 25 August), the observed track does lie within the ensemble distribution. However, the jumpiness (lack of consistency) between successive forecasts poses a challenge for forecasters trying to assess the most likely location of landfall.

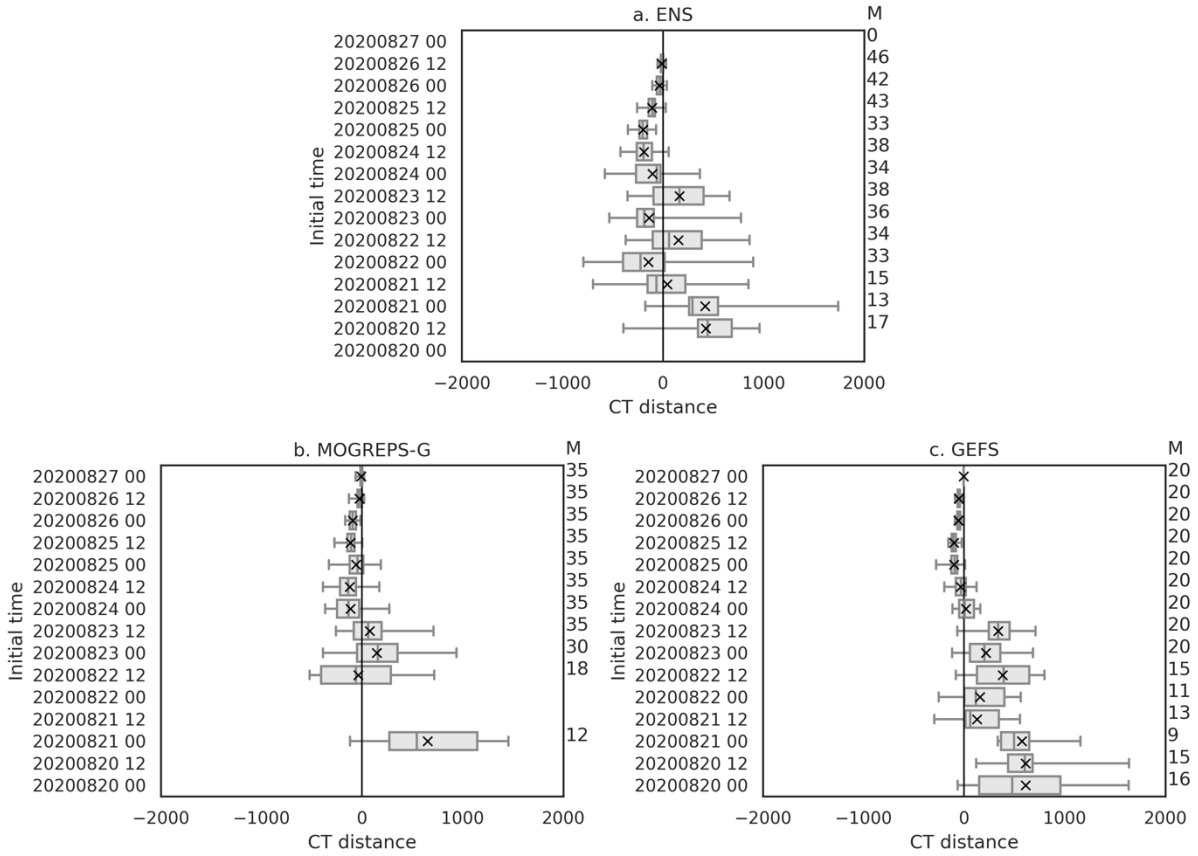


Figure 4.2. Jumpiness of ensemble forecasts for hurricane Laura, valid at 00 UTC on 27 August 2020. Each box plot summarizes the distribution of the cross-track (CT) errors (error at right-angles to the observed direction of travel; negative values indicate left-of-track error) for one ensemble forecast; distance measured in km. Forecasts started every 12 hours from 12 UTC 20 August; y-axis shows the forecast initial time. Box and whisker show the min, max and 25, 50 and 75 percentiles of the ensemble distribution (number of members shown to right of plot). Ensemble mean shown as X. a: ECMWF ENS, b: Met Office MOGREPS-G, c: NCEP GEFS.

This was a particularly jumpy case for the ENS (Magnusson et al. 2021) which merits further investigation. Comparing with other ensemble forecasts may help to identify possible causes. For example, if all centres display the same flip-flop behaviour it might suggest a common cause, such as changes in available observational data between the different analysis times.

Figure 4.2b and Figure 4.2c show the corresponding cross-track position forecasts for the MOGREPS-G and GEFS ensembles. Note that the MOGREPS-G ensemble data is missing from the TIGGE archive

for forecast start times 12 UTC 21 August and 00 UTC 22 August. There are some similarities between all three centres: a general right bias for earlier forecasts (initialized 00 21 August and earlier), with a substantial proportion of members not able to track Laura as far as the verification time of 00 UTC 27 August. Short-range forecasts for all centres are slightly left of the observed position. However, neither MOGREPS-G nor GEFS shows the same degree of flip-flop behaviour as ENS.

The MOGREPS-G forecasts are the most consistent from 12 UTC 22 August onwards, with relatively small changes between successive forecasts. The GEFS forecasts maintain the initial right-hand bias for several successive forecasts, with a notable jump between 00 and 12 UTC on 21 August. There is a second noticeable jump between 12 UTC on 23 and 00 UTC on 24 August, after which the GEFS forecasts are generally close to the observed position, although with a small left bias. It is also worth noting that both MOGREPS-G and GEFS track Laura in all members for forecasts initialized from 12 UTC 23 August onwards, while the ECMWF ensemble does not, even for the shorter ranges. The three centres use different tracking algorithms, and this suggests differences in the sensitivity and robustness of the different trackers (Conroy et al. 2023).

This example was chosen to illustrate jumpiness in the ECMWF ENS, and in particular the flip-flops between successive forecasts. Comparison with the other centres shows that this was not a feature common to all centres. The ENS jumpiness may be related to possible issues with the data assimilation or initial perturbations, but further work is needed to investigate this (Magnusson et al. 2021). Alternatively, this could be just a chance occurrence due to the limited number of ensemble members. For each of the initial times before 25 August, 20-30% of the ENS members did not track Laura as far as the verification time of 00 UTC on 27 August. In some cases, especially for initial times on 24 and 25 August, the ECMWF tracker misassigned some of the later forecast steps to hurricane Marco. However, this does not account for the majority of the missing tracks. These may be related to difficulties in initializing the cyclone due to the land interactions as Laura passed Puerto Rico, Hispaniola, and Cuba, while at earlier initial times, Laura was a relatively weak tropical storm and there was relatively large uncertainty in the initial analysed position (Magnusson et al. 2021). We have recomputed the results including the corrected misassigned tracks and confirmed that this does not affect any of our conclusions.

How typical is this Laura case? To investigate how often such jumpy cases occur and whether jumpiness tends to occur for the same or different cases in different ensemble systems, the following sections consider the run-to-run consistency over all Atlantic tropical cyclones from 2019-2021.

4.4.2 *Ensemble jumpiness 2019-2021*

To summarize the run-to-run inconsistency for a single case, we use the mean divergence $\overline{D}$ and divergence index (DI), both computed over all forecasts verifying at a given time for a given tropical cyclone. $\overline{D}$ measures the overall change in each sequence of forecasts, while DI accounts for the trend over the sequence and highlights any flip-flop behavior.

Figure 4.3 shows the distribution of $\overline{D}$ and DI over all available cases for Atlantic tropical cyclones from 2019-2021 for the ENS, MOGREPS-G and GEFS ensembles. For $\overline{D}$, ENS has the lowest median value and smallest inter-quartile range, while the distribution for GEFS is noticeably broader than for the other centers. The difference between the distributions of GEFS and the other centers are statistically significant at the 1% level for both the KS and MWU tests. Although much closer to each other, the difference between ENS and MOGREPS-G distributions is significant at the 5% level for MWU test (but not significant for KS). For DI, GEFS also has the broadest distribution and ENS has the narrowest distribution. The difference between MOGREPS-G and GEFS is not statistically significant. ENS is significantly different from both MOGREPS-G and GEFS at the 5% level.

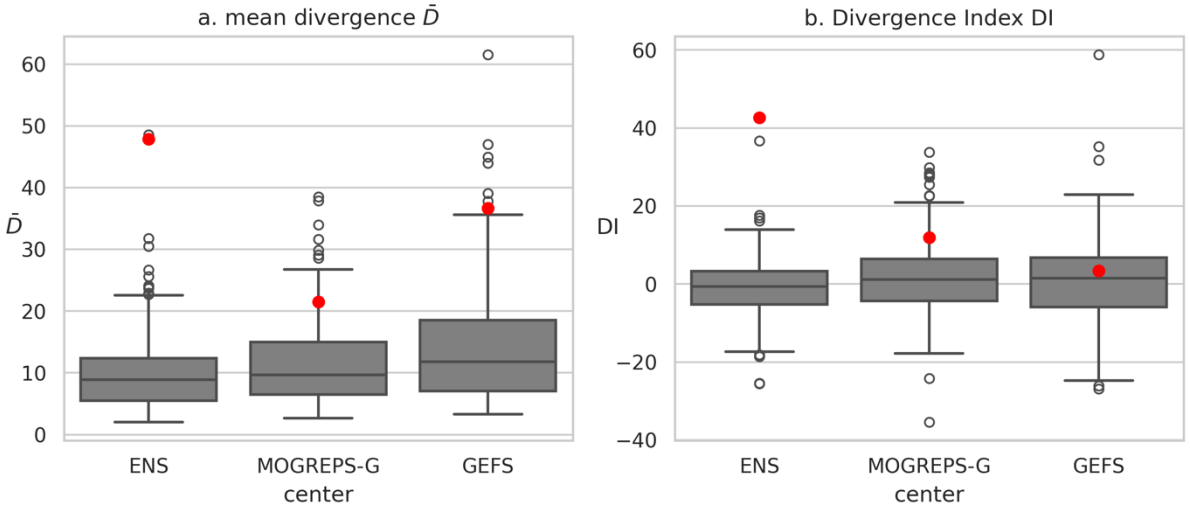


Figure 4.3. Run-to-run inconsistency (jumpiness) of ensemble forecasts for Atlantic tropical cyclone tracks (2019-21). Box plots show the distribution over all cases for the two divergence-based measures, a) mean divergence ($\overline{D}$) and b) Divergence index (DI). Box plots show the interquartile range and the median; the whiskers indicate the minimum and maximum values that are within 1.5 times the interquartile range; any more extreme points are shown with open circles as outliers. For both $\overline{D}$ and DI larger positive values indicate the most inconsistent cases. The points for the example case of hurricane Laura shown in Figure 4.1 and Figure 4.2 (verification time 27 August 2020, 00 UTC) are marked as red filled circles.

In general, a larger ensemble should give a more robust representation of the predicted distribution while a smaller ensemble will be more susceptible to sampling uncertainties and therefore may be expected to jump more from run to run. The above results are therefore consistent with the GEFS ensemble having fewer members than the other centres, especially before the upgrade to 31 members in September 2020. However, other factors can also influence the run-to-run consistency of

the ensemble. For example, a lack of spread due to under-representation of either initial condition or model uncertainties would also tend to make the ensemble more jumpy. The impact of the upgrade is considered in the next sub-section.

High positive values indicate the most inconsistent cases for both $\overline{D}$ and DI. For each centre, points that are more than 1.5 times the inter-quartile range above the upper quartile are classed as outliers (marked with open circles in Figure 4.3). The example case for hurricane Laura discussed in the previous section is highlighted – this is an extreme outlier for ENS for both measures, highlighting the unusually large jumpiness for this case.

For MOGREPS-G and GEFS, this case was not an outlier for DI, consistent with the absence of flip-flops that characterized the ENS forecasts. Although not the most extreme case, this case was an outlier for GEFS using the $\overline{D}$ measure. This was due to the large right bias in the earlier GEFS forecasts. This example illustrates the difference between $\overline{D}$ and DI: ENS had several flip-flops between successive forecasts, while changes between GEFS forecasts were more associated with a trend away from the initial right bias. Both centres had large mean divergence $\overline{D}$, but the underlying cause was different. MOGREPS-G was more consistent than the other centres.

We have seen that while Laura was an example of extreme jumpiness for ENS, this was not such an extreme case for the other centres, especially for DI. Scatter plots of $\overline{D}$ and DI for pairs of centres (Figure 4.4) show that this is a typical example. For each pair of centres, the number of cases that are outliers (high positive values, the most inconsistent cases) for either one centre or both centres are indicated in the figure. The dashed lines in the figures indicate the threshold used for the outliers (1.5 times the inter-quartile range above the upper quartile). The jumpiest cases (high positive DI) for one centre are in general not extremes for the other centres. For DI, none of the other ENS outliers are also outliers for either of the other centres. The results are similar for the outliers from MOGREPS-G and GEFS. There is only one case which is an outlier for more than one centre, MOGREPS-G and GEFS, but that case is not an outlier for ENS. For $\overline{D}$, the highlighted Laura case is unusual in that it has high $\overline{D}$ for both ENS and GEFS, although the cause is different for each center as discussed above. However, more typically the cases of high $\overline{D}$ for one center are not exceptional for the other centers. In the scatter plots, the outliers with high $\overline{D}$ tend to lie away from the diagonal so that there are substantially more cases in the upper-left and lower-right quadrants than in the upper-right.

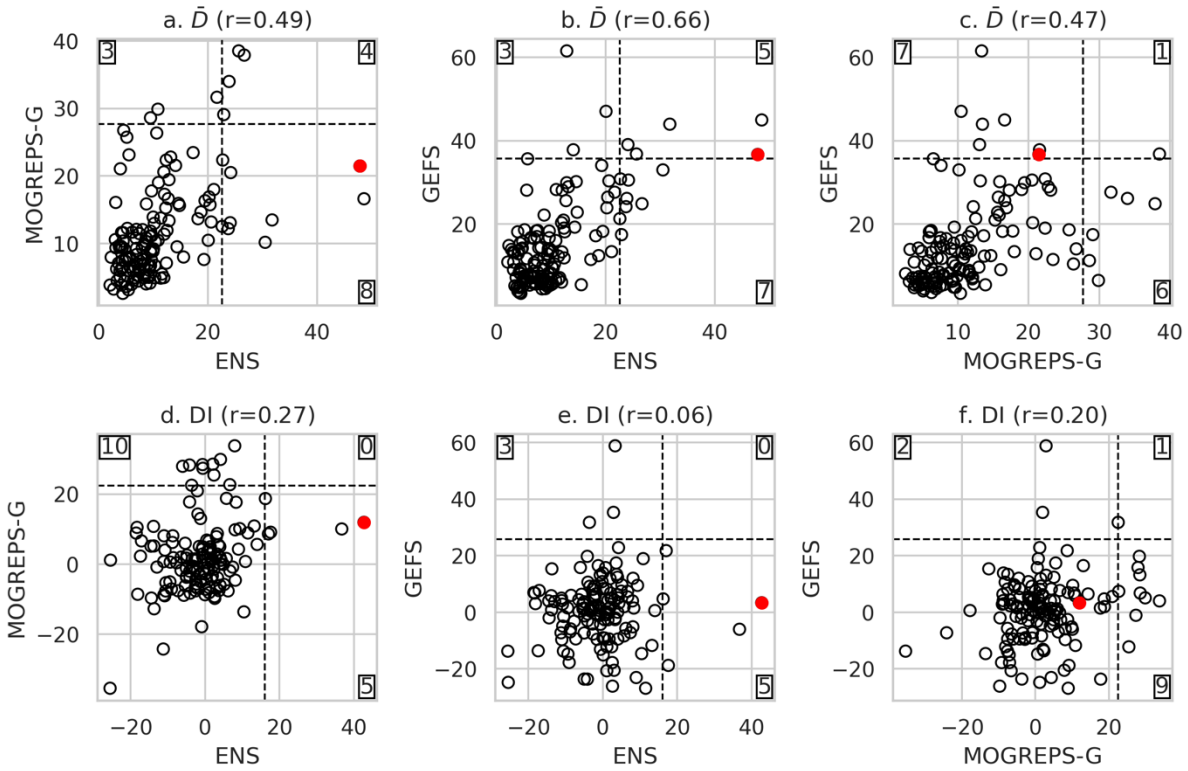


Figure 4.4. Comparison of jumpiness between different centres' ensemble forecasts for Atlantic tropical cyclone tracks (2019-21). Scatter plots show the distribution of the two divergence-based measures, mean divergence ($\bar{D}$, top row) and Divergence index (DI, bottom row) over all cases for pairs of centres. For both $\bar{D}$ and DI larger positive values indicate the most inconsistent cases. Dashed lines mark the threshold for the most inconsistent outliers (1.5 times the inter-quartile range above the upper quartile). In each panel, the number of cases that are outliers for both centres or just one of the centres is indicated in the corresponding quadrant. The points for the example case of hurricane Laura shown in Figure 4.1 and Figure 4.2 (verification time 27 August 2020, 00 UTC) are marked as red filled circles.

These results suggest that the ensemble jumpiness is not strongly linked to the atmospheric situation or to the availability of observations. Rather, they suggest that individual model deficiencies or sampling uncertainties are more likely causes for the jumpiness. Sampling uncertainties will lead to run-to-run jumpiness if the ensemble is not large enough to fully represent the distribution of possible outcomes; a larger ensemble would better sample this underlying distribution and improve consistency from run to run. Alternatively, an ensemble may fail to properly represent the range of possible outcomes because the perturbations to initial conditions are not adequate or because the uncertainties in the model formulation are not sufficiently represented. Either of these will result in the ensemble spread being too small and may lead to jumpy behaviour.

4.4.3 *The effect of recent NWP system upgrades on ensemble jumpiness*

The results of the previous section showed that overall GEFS was more jumpy than the other centres. The GEFS upgrade in September 2020 was the most substantial upgrade of any of the centres during the study period, including a new forecast model, changes to the ensemble perturbations and an increase in the number of ensemble members. It brought a substantial improvement in the spread of

tropical cyclone track forecasts (Zhou et al. 2022). Here we consider the impact of the upgrade on the jumpiness of ensemble track forecasts.

We separate our sample into two subsets initialized before (64 cases) and after (75 cases) the GEFS upgrade. In Figure 4.5 we compare the empirical cumulative distribution of the mean divergence $\overline{D}$ for the three centres before (Figure 4.5a) and after (Figure 4.5b) the upgrade. Overall, $\overline{D}$ is significantly lower after the upgrade (comparing Figure 4.5a and Figure 4.5b). However, this applies also to the results from the other centres, suggesting that the difference is at least partly due to the differences between the observed samples. To mitigate this sampling effect, we focus on the difference between the GEFS ensemble and the other centres for the two subsets of cases.

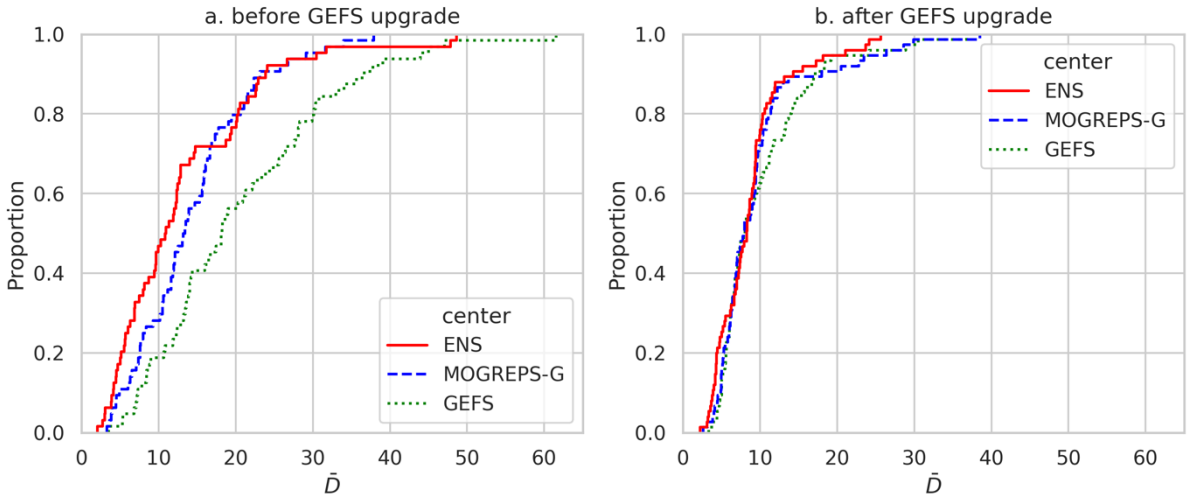


Figure 4.5. Effect of GEFS v12 cycle upgrade, 23 September 2020. Empirical cumulative distribution function of $\overline{D}$ for subsamples of cases (a) before and (b) after the upgrade.

Before the upgrade, the GEFS had substantially more cases with high values of $\overline{D}$ compared to ENS and MOGREPS (Figure 4.5a). The difference in distribution compared to the other centres is highly significant at well below the 1% level for both KS and MWU tests. Differences in the distributions for ENS and MOGREPS-G are not statistically significant. After the upgrade, the GEFS distribution was much closer to those of the other centres (Figure 4.5b) and there were no statistically significant differences between the distributions of any of the centres. These results show that the upgrade to the GEFS did make a significant difference to the consistency in terms of mean divergence $\overline{D}$. As for the full sample, differences in the distributions of DI are smaller (not shown); the only statistically significant difference between GEFS and either of the other centers is with ENS before the GEFS upgrade.

The GEFS upgrade brought a substantial improvement in the spread of tropical cyclone track forecasts. This was considerably under-dispersive in the previous version and the upgrade resulted in a much better spread-error relationship, due to the upgrade to the stochastic model perturbations (Zhou et

al. 2022). The change in $\overline{D}$ is consistent with this increase in spread for the GEFS system. In general, a larger spread will give a broader distribution of tropical cyclone positions and the change between the set of positions for successive forecasts would tend to be less than for a less dispersive ensemble. For the same reason, the improved spread might also be expected to affect DI. Although there was some indication of this in our results (the ENS and GEFS distributions were closer and not significantly different after the upgrade), it was not such a clear change as for $\overline{D}$.

It is possible that additional factors as well as the increased spread also helped to improve $\overline{D}$. For example, a reduction in cross-track bias in the longer-lead forecasts would help to reduce $\overline{D}$, but would not tend to affect DI. Leonardo and Colle (2021) showed that the GEFS had larger cross-track errors than ENS in a large sample of Atlantic tropical cyclones for 2008-2016. We were not able to identify any significant changes in the GEFS bias after the upgrade in our sample of cases. While the change in ensemble spread was large enough to identify in our sample, it may be that other differences require larger samples. Leonardo and Colle (2021) also noted that large year-to-year variability made it difficult to identify any changes due to model upgrades.

The MOGREPS-G upgrade in December 2019 also improved TC track errors and spread (Met Office 2019; Titley et al. 2020). Taking the same approach as above we found that for the subset of cases before the MOGREPS-G upgrade there was a significant difference between the ENS and MOGREPS-G distributions for both $\overline{D}$ and DI (with the MOGREPS-G having overall higher jumpiness). After the upgrade there was no significant difference between the two centres. See Figure 4.S1 in the supplemental material (section 4.6).

We conclude that the recent upgrades to the MOGREPS-G and GEFS systems both improved the run-to-run consistency of the ensemble track forecasts, and that since these upgrades the overall jumpiness is similar for the three ensemble systems.

4.4.4 *Comparison of error, spread and divergence*

We now compare the mean scores over all cases for the three different aspects of ensemble performance: error, spread and divergence. The upper panel of Figure 4.6 shows the ensemble error (CRPS, left), divergence (D, centre) and spread (s, right) at lead times out to five days ahead for the three centres. The vertical bars indicate the bootstrapped 95% confidence intervals for each centre's scores. Overall, the three centres have similar performance and most differences between scores are not statistically significant.

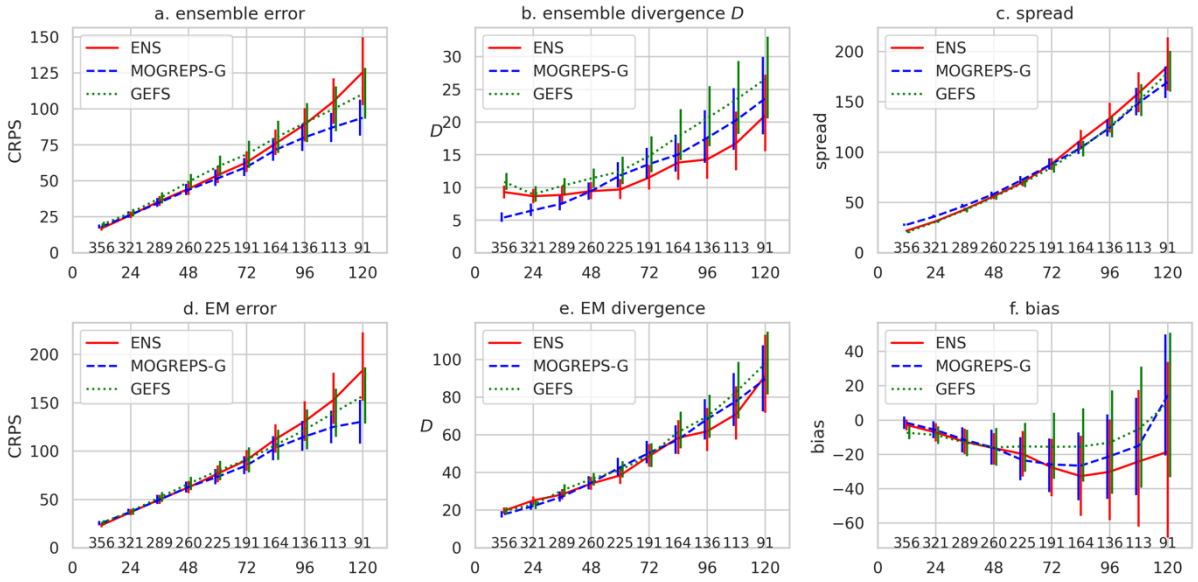


Figure 4.6. Error, spread and divergence for forecast lead time from 12 to 120 hours. Scores for the full ensemble are shown on the upper row; corresponding error and divergence for the ensemble means are shown below. a, d: CRPS error; b, e: divergence; c: ensemble spread; f: bias. Vertical bars indicate 95% confidence intervals. Mean scores over all available cases for each forecast lead time: number of cases indicated above x-axis.

The larger divergences in the short range for ENS and GEFS (Figure 4.6b) are consistent with the lower spread (Figure 4.6c) at these time steps for these centres. MOGREPS-G has larger initial spread (maybe partly due to the time-lagging of the initial conditions of the MOGREPS-G system), and this will tend to reduce the difference (divergence) between consecutive forecasts as seen in Figure 4.6b.

For each centre, the mean ensemble divergence (Figure 4.6b) is approximately equal to the mean difference in CRPS between consecutive forecasts (difference between successive points on the curves in Figure 4.6a). The agreement is particularly strong at short range for all centres, and for ENS at all forecast ranges. In other words, on average the divergence gives an indication of the expected change in error for the next forecast. However, this does not apply in individual cases.

Table 4-1 shows the Pearson correlation between divergence and CRPS across all available cases for each forecast lead time. For comparison, the correlation between ensemble spread and CRPS is also shown. Corresponding scatter plots are shown in Figures 4.S2-4.S5 in the supplemental material (section 4.6). The association between divergence and error is in general substantially weaker than the link between spread and error. These results are consistent with previous studies that show the benefit of using spread as a measure of forecast uncertainty (Majumdar and Finocchio 2010; Titley et al. 2019; Yamaguchi et al. 2009; Kawabata and Yamaguchi 2020). However, the low correlation for divergence suggests that it does not provide useful case-to-case guidance: there is no indication that users should expect less jumpy cases to be more skilful.

Step (h)	ENS	MOGREPS-G	GEFS
72	0.18 (0.45)	0.22 (0.38)	0.07 (0.29)
84	0.25 (0.56)	0.32 (0.47)	0.05 (0.27)
96	0.19 (0.58)	0.36 (0.47)	-0.01 (0.32)
108	0.29 (0.67)	0.42 (0.41)	0.19 (0.44)

Table 4-1. Correlation between divergence and error. Each row shows the correlation between the CRPS error at a given forecast lead time h and the divergence D between h-hour and (h+12)-hour forecasts. For comparison the correlation between the CRPS and the ensemble spread for the h-hour forecasts is shown in brackets.

Table 4-2 shows the Pearson correlation over all cases between the two overall measures, $\overline{D}$ and DI, and the corresponding mean error over all forecast lead times $\overline{\mathrm{CRPS}}$. Although for $\overline{D}$ the correlation is somewhat higher than for the individual forecast steps (Table 4-1), the corresponding scatter plots show large variations in error for cases of both low and high $\overline{D}$. This again suggests that users should be cautious in individual cases – a consistent case with relatively low jumpiness may still have large overall error.

Centre	$\overline{D}$ v $\overline{\mathrm{CRPS}}$	DI v $\overline{\mathrm{CRPS}}$
ENS	0.54	-0.30
MOGREPS-G	0.56	-0.01
GEFS	0.67	-0.30

Table 4-2. Correlation between overall jumpiness and error ($\overline{CRPS}$).

We can do the same analysis for the ensemble-mean forecasts, which are often used in operational TC forecasting (Figure 4.6d,e; lower panel). Again, the divergence gives useful additional information for forecast users. For example, for ENS the ensemble mean cross-track error is around 175 km for 120-hour forecasts (Figure 4.6d), and the ensemble spread is similar (showing that the ensemble system is overall well-tuned; Figure 4.6c). The mean expected change in cross-track EM position between T+120 and T+108 is ~80 km (Figure 4.6e). This is similar for all three centres.

The forecast systematic error (bias) is shown in Figure 4.6f. Overall, each centre has a negative bias, that is the forecast positions tend to be to the left of the observed position. However, there is large uncertainty as indicated by the large confidence intervals shown on the plot. Magnusson et al. (2021) show that the ENS tends to have a left-of-track bias for northward-moving TCs, but a right-of track bias for westward moving systems and this situation-dependent variation in bias may partly explain the large confidence intervals at longer lead times. As for the other scores, the confidence intervals indicate that there is no significant difference between the biases of the different centres. Comparing

Figure 4.6d and Figure 4.6f shows that for all centres the bias is relatively small compared to the total error.

4.5 Conclusions

We have carried out an investigation of the jumpiness or run-to-run consistency of ensemble forecasts of tropical cyclone tracks. We used ensemble forecasts from the TIGGE tropical cyclone track archive for three global centres: ECMWF (ENS), Met Office (MOGREPS-G) and NCEP (GEFS). The forecasts were compared to the observed tracks for all named tropical cyclones from the IBTrACS archive for the Atlantic basin for 2019, 2020 and 2021.

We looked at the change in the distribution of cross-track position (relative to the observed track) for tropical cyclones in consecutive ensemble forecasts initialized at 12-hours intervals. This was quantified using the divergence function D associated with the CRPS error score following (Richardson et al. 2020b). The overall jumpiness of a sequence of forecasts all verifying at the same time was summarized using the mean divergence $\overline{D}$ and the Divergence Index DI.

We present our conclusions in the framework of the questions posed in the introduction.

4.5.1 *How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centres?*

The distribution of DI was similar for each centre, showing substantial variation between centres with a few significant outliers. There was no strong agreement between the centres on which cases were most jumpy. The case shown for Hurricane Laura was a typical example: this was the most extreme case of jumpiness (largest DI) for the ECMWF ENS, showing a clear flip-flopping of the ensemble between being left and right of the observed track in successive forecasts. This behaviour was not apparent in either the MOGREPS or GEFS ensembles. This case also illustrated the difference between the two summary measures $\overline{D}$ and DI. Earlier GEFS forecasts were substantially to the right of the observed track and this right-of-track bias decreased in later forecasts. The large trend over successive forecasts is indicated in the relatively high mean divergence. However, the absence of the flip-flop behaviour seen in the ECMWF ENS results in the DI being close to the overall median value. Using the combination of both $\overline{D}$ and DI can help to distinguish these different behaviours in a sequence of forecasts.

4.5.2 *Is there a common cause of 'jumpy' cases – are the ensembles from different centres particularly jumpy for the same cases and if so what is the reason?*

The jumpiest cases were different for each centre for both $\overline{D}$ and DI, indicating that there is not a common cause of jumpiness across the different ensemble systems. This suggests that the ensemble

jumpiness is not strongly related to the prevailing atmospheric conditions or to the available observations.

Outliers for the different centres may be due more to specific issues in the data assimilation, models or ensemble configurations. Recent studies highlight both continuing progress and ongoing challenges in each of these areas (e.g. Magnusson et al. 2019, 2021). However, a deeper analysis of outliers would require a substantially larger sample than we have used and is beyond the scope of the present work. Leonardo and Colle (2021) used 9 years (2008–16) of Atlantic TC data to investigate the causes of large cross-track errors in the GEFS and ENS. However, we have also seen that recent upgrades to ensemble systems have led to a significant reduction in the ensemble jumpiness and therefore including a longer sample of earlier years may not be representative of the current ensemble capabilities.

Another possible reason for the occasional cases of large jumpiness is sampling uncertainty due to finite ensemble size. This would be consistent with outliers occurring at different times for the different centres. Richardson (2001) showed how even a well-tuned ensemble will appear unreliable if it has insufficient members and that the required number of ensemble members depends on both the underlying distribution and the needs of the users. Leutbecher (2019) and Craig et al. (2022) have demonstrated substantial sensitivity to ensemble size in studies using large ensembles of 200 members and 1000 members respectively. Kondo and Miyoshi (2019) suggest that up to 1,000 ensemble members are necessary to represent important aspects of some forecast distributions. The impact of ensemble size on forecast jumpiness has not been investigated and is a topic for future work.

4.5.3 *Have recent ensemble model upgrades had a noticeable effect on the forecast jumpiness?*

In this study we used a three-year period to provide a sufficient number of cases to assess. During this period upgrades to both the MOGREPS-G and GEFS ensembles resulted in substantial improvements to their predictions of TC tracks. Using the ECMWF ENS as a reference, we found that both these upgrades significantly reduced the jumpiness of the ensembles. Before the upgrades the ENS was significantly less jumpy than the other centres. However, after the upgrades there was no significant difference between the centres. Both upgrades increased the spread of the ensembles, and the improved jumpiness is consistent with this change. These results suggest that it is the overall level of ensemble spread that is important and that differences in initialization and perturbation methodology between the current systems are not a major factor in determining the overall level of ensemble jumpiness.

The more recent upgrade to the ENS at the end of 2021 improved TC track errors by 10% but had little impact on the overall spread (Haiden et al. 2022). This improved the statistical reliability of the TC

track. The impact on jumpiness of this upgrade has not been assessed but can be done once a sufficient sample of cases is available.

4.5.4 *What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness – what information is practically useful? Is there any useful link between jumpiness and skill?*

The divergence D gives an indication of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80-90 km in the cross-track direction between a forecast for 120 hours ahead and the 108-hour forecast for the same time made 12 hours later. The expected change between a 72-hour and 60-hour forecast is around 50 km. These expected changes were similar for all three centres. Corresponding values for the expected divergence for the full ensemble distributions are 20-25 km and 10-15 km respectively. These results address the user requirements identified for example by Regnier and Harr (2006) and Jewson et al. (2022) to provide objective measures of the expected change from run to run so that users can take account of this in their decision making.

We did not find any strong link between either $\overline{D}$ or DI and error (CRPS). This indicates that users should not rely on the jumpiness or consistency between successive forecasts as measure of confidence in the forecasts. This is consistent with the work of Zsoter et al. (2009) who found only a weak link between jumpiness and error in ensemble forecasts for Europe. In contrast, ensemble spread and the ensemble probabilistic information (e.g. strike probabilities) have been shown to provide useful situation-dependent guidance on forecast uncertainty (Majumdar and Finocchio 2010; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020).

Although we note that the effect of more recent system upgrades has not yet been evaluated, users should expect generally similar levels of jumpiness in the three ensemble systems considered in this study. The jumpiest cases will tend to be different for the different centres, likely to be a result of sampling uncertainties or specific deficiencies in the individual ensemble configurations.

One practical approach for users to adopt to address both these potential sources of jumpiness would be to combine the ensemble forecasts from the different centres into multi-model ensembles. Such multi-model combinations have already been shown to improve probabilistic TC track prediction (Yamaguchi et al. 2012; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020). Another option would be to use lagged ensembles, combining consecutive forecasts from one centre. By construction this will reduce jumpiness and this is already used in the MOGREPS-G system to increase ensemble size. Although our aim in this study was to evaluate and compare the jumpiness in the individual systems, the effect of multi-model combinations on ensemble jumpiness is an area for future work.

4.6 Supplementary material

This supplementary material provides 5 additional figures to complement the results shown in the paper.

Figure 4.S1 shows the impact of the MOGREPS-G upgrade in December 2019 on the mean divergence $\overline{D}$ and Divergence Index DI. Before the upgrade, the differences between the empirical cumulative distributions for ENS and MOGREPS-G are statistically significant at the 5% level ($p < 0.03$) for $\overline{D}$ and at the 1% level ($p < 0.005$) for DI using both the Kolmogorov-Smirnov and Mann-Whitney U tests. After the upgrade there was no significant difference between ENS and MOGREPS-G.

Figures 4.S2-4.S5 show scatter plots of divergence against error and of spread against error for lead times of 72, 84, 96 and 108 hours. These figures complement the correlations shown in Table 4-1.

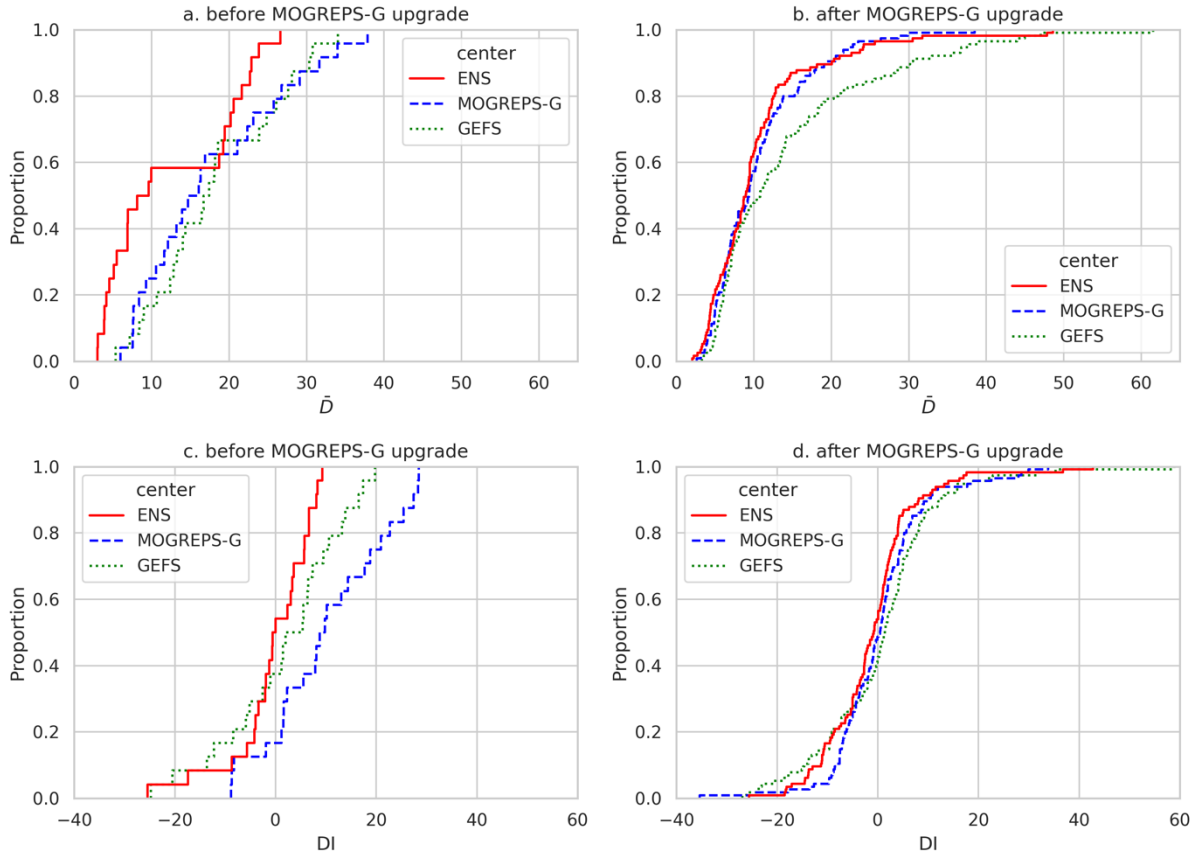


Figure 4.S1 Effect of MOGREPS-G cycle upgrade, 4 December 2019. Top row: empirical cumulative distribution function of $\overline{D}$ for subsamples of cases (a) before and (b) after the upgrade. Bottom row: empirical cumulative distribution function of DI for subsamples of cases (c) before and (d) after the upgrade.

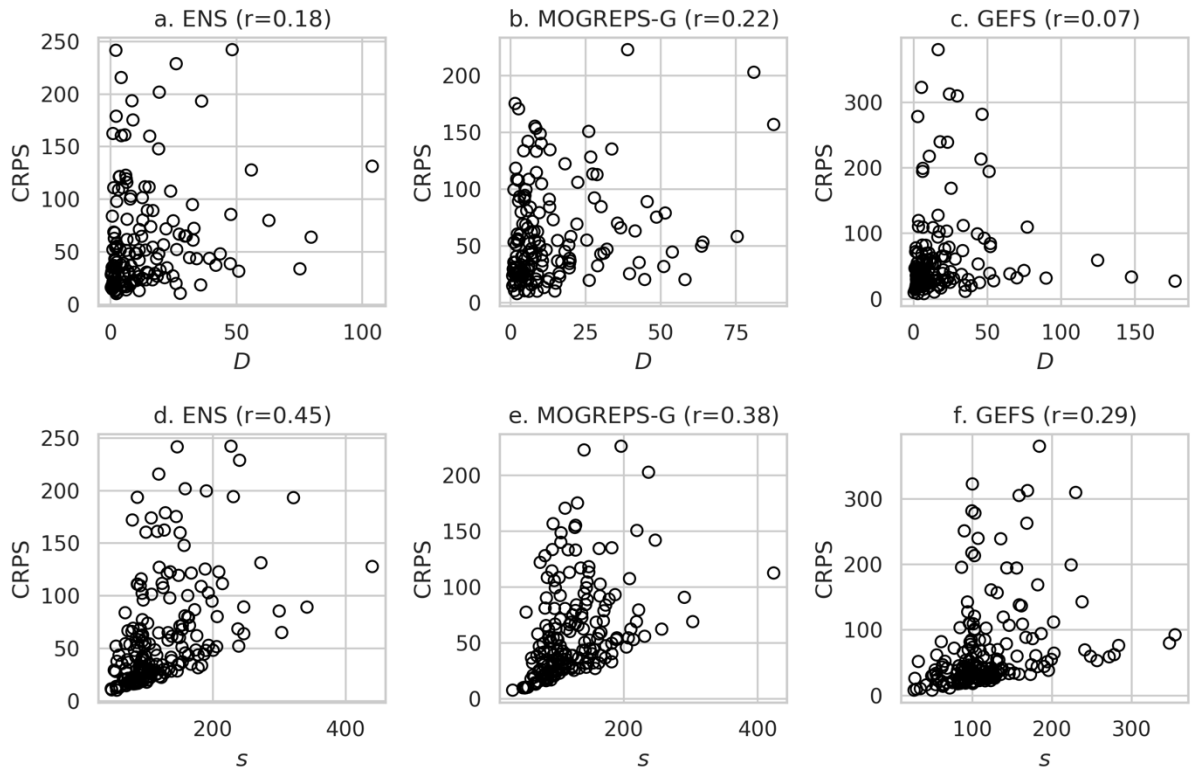


Figure 4.S2. Correlation between divergence and error (top row) and between spread and error (bottom row). On the top row, each panel shows a scatter plot over all cases of the CRPS error at a forecast lead time of 72 hours and the divergence D between 72-hour and 84-hour forecasts for a: ENS, b: MOGREPS-G, and c: GEFS . For comparison the correlation between the CRPS and the ensemble spread s for the 72-hour forecast is shown in the panel below. The Pearson correlation coefficient r for each sample is shown in the title of each panel.

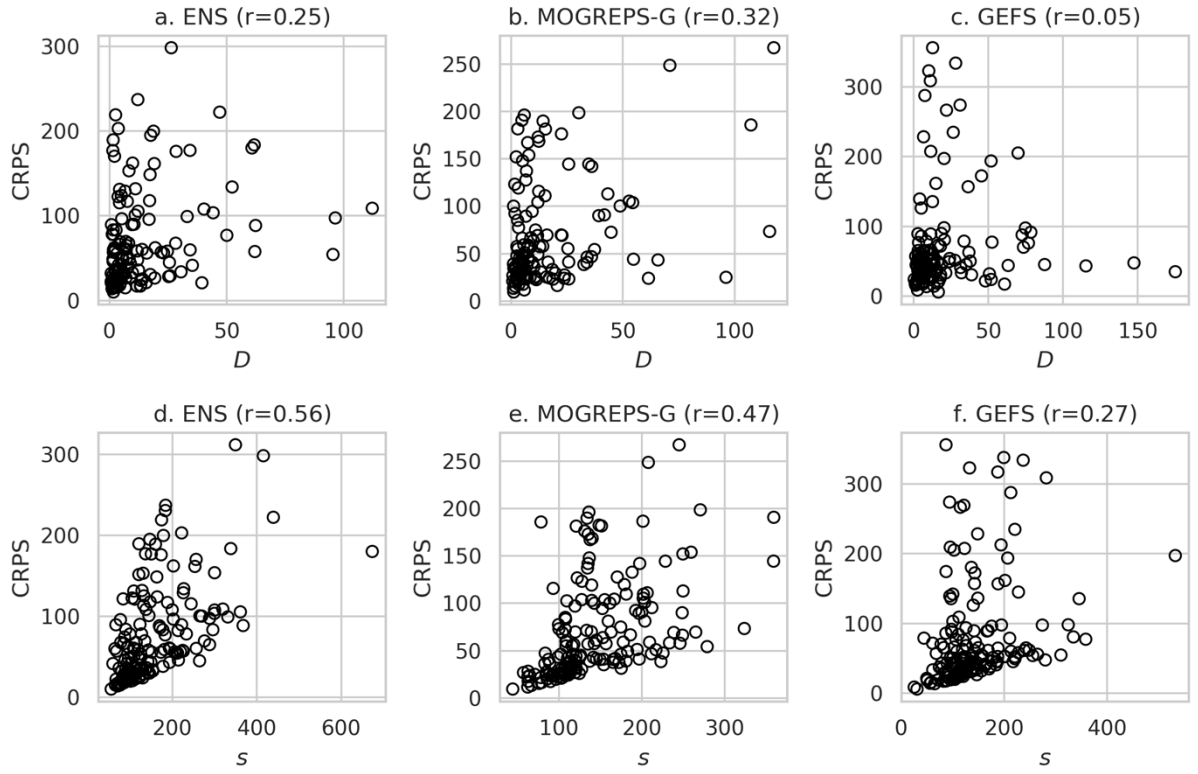


Figure 4.S3. As Figure 4.S2 for 84-hour forecast.

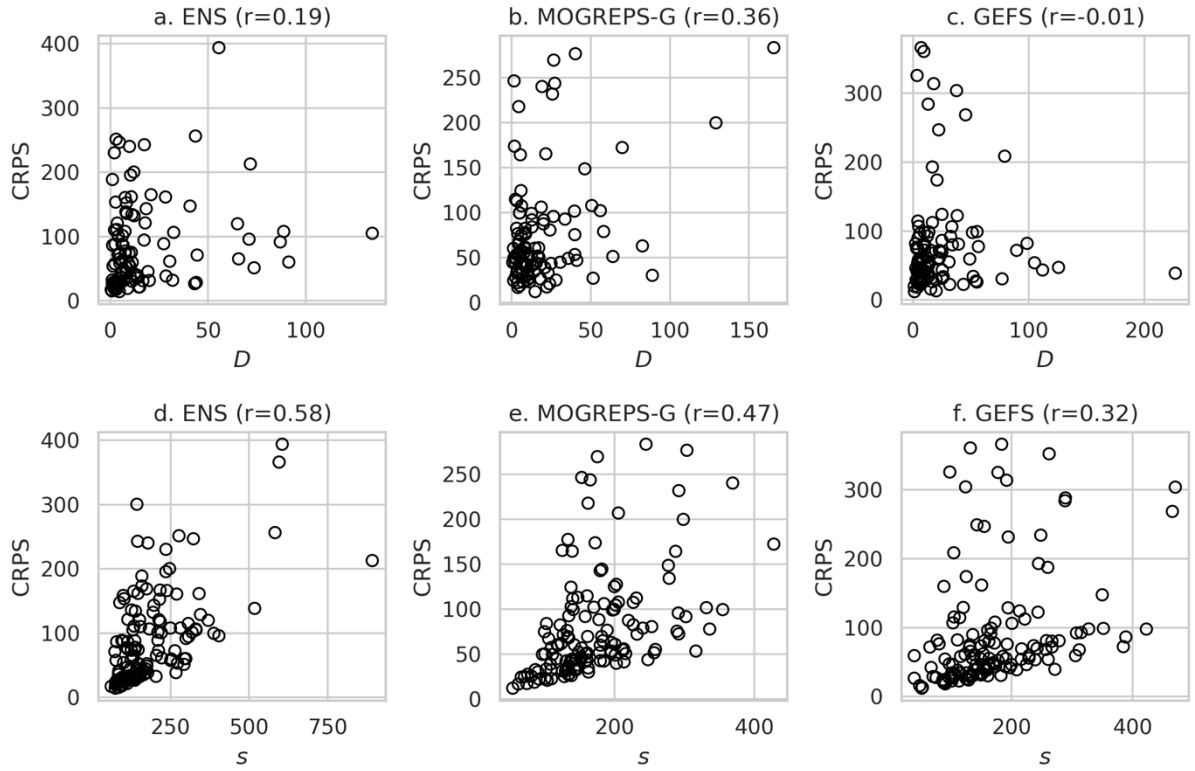


Figure 4.S4 As Figure 4.S2 for 96-hour forecast.

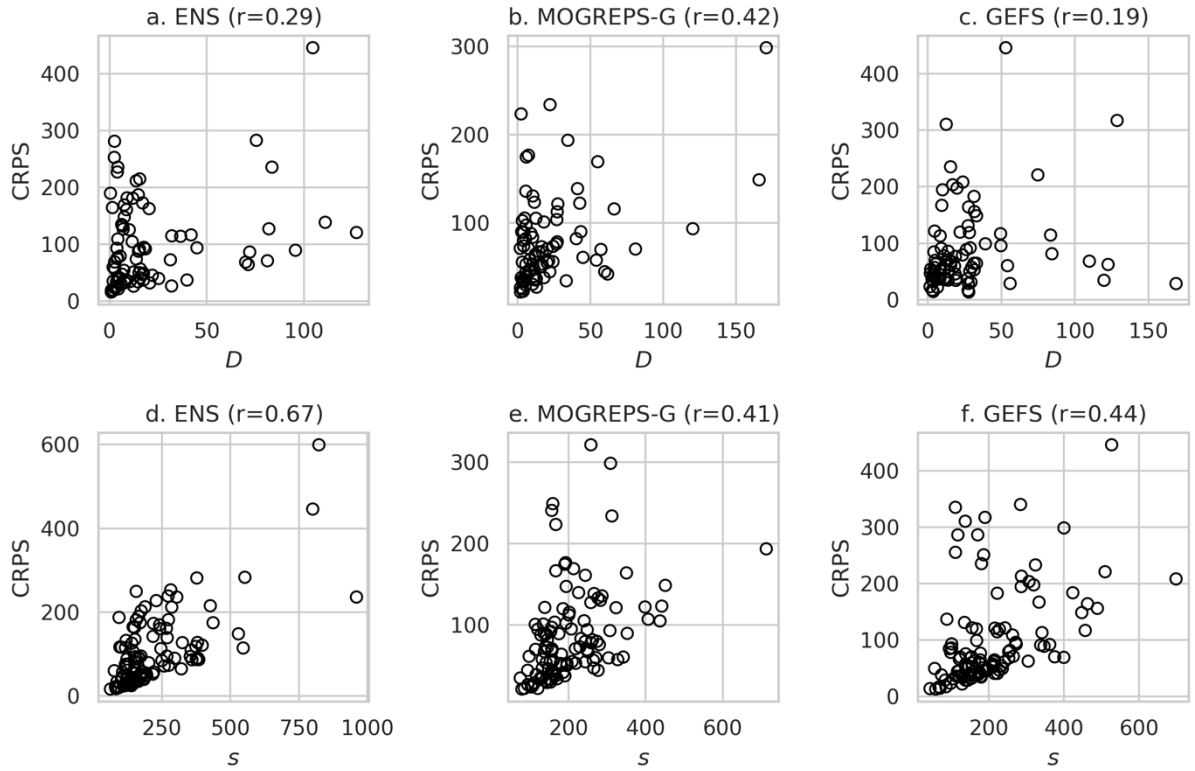


Figure 4.S5. As Figure 4.S2 for 108-hour forecast.

Chapter 5 Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

The third objective of the PhD was to evaluate the skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis, provide guidance to users and identify factors affecting forecast performance. The paper addressing this objective has been submitted to Weather and Forecasting with the following reference:

Richardson, D.S., Cloke, H.L., Magnusson, L., Majumdar., S. J., Methven, J.A. and Pappenberger, F. (2024) 'Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis', *Weather and Forecasting* (submitted June 2024; re-submitted November 2024 following review)

The contributions of the authors of this paper are as follows: D.R. designed the study with advice from L.M. and S.M., obtained the datasets, carried out the analysis, and led the writing of the manuscript. All authors assisted with writing the manuscript. Overall, 90% of the writing was undertaken by D.R.

The submitted article can be found in Appendix A3.

Abstract. We evaluate the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting tropical cyclone genesis in the Atlantic basin. Focusing on the probabilistic performance of the ENS, we assess how far in advance the ENS can predict genesis, quantify the consistency (jumpiness) from run to run and investigate what factors influence the skill and consistency. We find that first indications of genesis are picked up at least 7 days ahead in 50% of the observed cases, although strong signals often only appear less than 3 days before genesis. There are significant regional differences, with observed genesis events predicted 2-3 days earlier in the eastern Atlantic than in other areas. The genesis probabilities can be jumpy from run to run and the jumpiest cases are in the more skilful regions (central and eastern Atlantic) and for situations where the initial signal for genesis appears at longer lead time. In the eastern Atlantic, there is a tendency for the ENS tracks to reach tropical storm strength earlier and further east than observed; this model bias can affect both skill and jumpiness of the genesis forecasts. Our results provide guidance to forecasters on how to use and interpret the ENS predictions. Areas for future work include the link between early intensification in the eastern Atlantic and African easterly wave activity, the relationship between skill and the TC development pathways, and the impact of systematic analysis differences between 0000 UTC and 1200 UTC on forecast intensity.

Significance statement. Forecasting where and when tropical cyclones will appear increases the lead time at which decision makers can begin to take preparatory mitigating action. Numerical weather

prediction models can provide important guidance, but sometimes are not consistent from one run to the next. We evaluate the skill and consistency of a state-of-the-art global model in predicting the formation of tropical cyclones up to ten days ahead and provide guidance to forecasters on how to use and interpret the model predictions. We show that the formation of tropical cyclones can be predicted 2-3 days earlier in the eastern Atlantic than in the western Atlantic and identify some of the factors influencing both skill and consistency.

5.1 Introduction

Following significant progress in forecasting tropical cyclone (TC) tracks (Landsea and Cangialosi 2018) and intensity (Cangialosi et al. 2020), there is increasing focus on predicting TC genesis (Hon et al. 2023). For the Atlantic basin, the US National Hurricane Center (NHC) Tropical Weather Outlook provides forecasts of TC genesis for 2 and 7 days ahead (Hon et al. 2023). By providing information about the likely development of TCs before they have formed, skilful genesis forecasts can effectively increase the lead time at which decision makers can begin to take preparatory mitigating action.

Numerical weather prediction (NWP) forecasts including ensemble forecasts are used in operational genesis forecasts (Titley et al. 2019; Hon et al. 2023), often in combination with statistical methods (Halperin et al. 2017). Use and verification of NWP genesis forecasts has focused on deterministic aspects, assessing hits and false alarms using standard contingency-table measures such as hit rate or probability of detection, success ratio, and the threat score or critical success index (Wilks 2020). These have been applied to the high-resolution global forecasts from different centres (Halperin et al. 2016, 2013; Liang et al. 2021) to ensemble mean forecasts (Li et al. 2016; Wang et al. 2018) and to individual ensemble members (Zhang et al. 2022).

Recently there has been increasing development of probabilistic TC genesis forecast products for operational centres (Hon et al. 2023). For example, Halperin et al. (2017) developed a statistical–dynamical tool to generate TC genesis probabilities using logistic regression models applied to the outputs from several high-resolution global NWP models. A consensus probability is also provided when more than one model predicts a genesis event. Verification using Brier scores and reliability diagrams showed that these provide useful guidance (Halperin et al. 2017), and the products are regularly used in the NHC (Hon et al. 2023). The use of probabilistic information from the ensembles is more limited, although ensemble forecasts have been shown to have skill in predicting TC genesis (Komaromi and Majumdar 2014, 2015; Majumdar and Torn 2014; Yamaguchi and Koide 2017; Yamaguchi et al. 2015).

One of the key issues limiting the uptake of ensemble TC forecasts is the run-to-run jumpiness that can occur in some situations (Dunion et al. 2023; Magnusson et al. 2021). Large jumps in the predicted

probability of TC genesis between successive ensemble forecasts present a significant challenge to forecast centres and lessen users' confidence in the prediction system (McLay 2008; Elsberry and Dobos 1990; Hewson 2020; Dunion et al. 2023; Pappenberger et al. 2011b). Although approaches such as multi-model combinations or lagged ensembles can help mitigate such jumpiness, it is important to identify and understand the underlying causes of such jumpy behaviour. Quantifying the level of jumpiness in an ensemble system provides valuable information to the forecast user. This can be important for example in helping the user to decide between acting now or waiting for the next forecast (Regnier and Harr 2006; Jewson et al. 2022, 2021). Identifying the circumstances in which jumpiness occurs is an important step towards addressing the underlying cause – is it related to model or analysis uncertainty (lack of spread in the ensemble perturbations) or model bias, or is it an indication of insufficient ensemble size to give a reliable uncertainty estimate? Jumpiness of TC track forecasts has been investigated for the western North Pacific (Elsberry and Dobos 1990) and the Atlantic (Fowler et al. 2015; Richardson et al. 2024). However there has been no corresponding assessment of TC genesis forecasts. In this study we conduct a first assessment of the jumpiness of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (ENS) forecasts for TC genesis.

Another factor limiting the use of ensemble TC genesis forecasts is the lack of routine evaluation of the products provided by the global centres. Although ECMWF regularly publishes verification results for ensemble forecasts of the track and intensity of existing TCs (Haiden et al. 2023), it does not routinely evaluate genesis forecasts, so users do not have a clear picture of ENS performance (Magnusson et al. 2021).

These knowledge gaps are addressed in this study which evaluates the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting TC genesis in the Atlantic basin. We address the following questions:

- How far in advance can the ENS forecast TC genesis the Atlantic basin?
- How consistent from run to run are the forecasts of the observed genesis events?
- What are the factors that influence the skill and consistency of the ENS genesis forecasts and what future work will help to improve these forecasts?

In each case, we focus on the probabilistic performance of the ENS. The data we use in this study and the methods we apply to identify genesis events are described in section 5.2, with verification scores and consistency measures introduced in section 5.3. Results are presented in section 5.4, addressing each of the three key questions in turn. We conclude with a summary and discussion of directions for future work in section 5.5.

5.2 Data

We investigate the ability of the ECMWF ensemble (ENS) to predict the genesis of tropical cyclones over the Atlantic. ENS comprises 50 perturbed members integrated on ~18km grid until 27 June 2023 and thereafter on ~9km grid. The ECMWF tropical cyclone tracker (Magnusson et al. 2021) identifies and tracks both existing TCs and those that develop during the forecast. The tracker is applied to all ensemble members. These operational forecast tracks are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the operational forecast tracks for ENS forecasts initialized at 0000 and 1200 UTC from May to December 2019-2023 and consider forecast lead times from one to ten days ahead.

We evaluate the forecasts against the observed TC data from the International Best Track Archive for Climate Stewardship (IBTrACS, (Knapp et al. 2018, 2010). We extract the observed positions and maximum winds from all named Atlantic tropical storms (i.e. tropical cyclones that reach tropical storm strength during their life cycle). We focus our evaluation on the first time the observed system is reported as a tropical system of at least tropical storm strength (winds at least 34 kt; 1kt ~ 0.51 m s^{-1}), which we define as the genesis time for the tropical storm (TS) (Magnusson et al. 2021; Zhang et al. 2022). To ensure a consistent set of forecast lead times throughout the evaluation, we limit the verification times to also be 0000 and 1200 UTC and so the observed genesis time is the first 0000 or 1200 UTC time with wind >17 m s^{-1}. There were 98 observed tropical storms in the Atlantic basin during the 5-year study period. However, TS Imelda (2019) was a TS for less than 12 h and was not included in the verification, therefore we used 97 observed TS genesis in this work.

To investigate how well and how consistently the ENS can forecast the observed TS genesis events, we compute the probability of TS genesis or TS activity at the observed genesis time and location for each of the 97 observed TS.

For a given verification time t_v, we refer to an ensemble forecast f valid for this time and initialized h hours earlier as $f(t_v, h)$ and write the individual ensemble members as $f_m(t_v, h)$. Given the inherent limitations of predictability as well as uncertainties in both forecast and observations (Landsea and Franklin 2013; Torn and Snyder 2012), we do not expect the forecast to predict genesis at exactly the time and location of the reported observed genesis event. Therefore, we define tolerances in both space and time. Several different choices have been used in previous studies (Halperin et al. 2016, 2013; Zhang et al. 2022; Magnusson et al. 2021; Yamaguchi et al. 2015). For each observed TS genesis event, we use the following procedure where t_v represents the observed genesis time:

- For the ENS forecast $f(t_v, h)$ we count how many members m have TC tracks that pass within 500 km of the observed genesis location at any time between $t_v - 24$ h and $t_v + 24$ h. We

define the proportion of members m/M as the forecast probability of TC activity at the observed genesis event. This gives the probability for TC but does not address the intensity or the location of genesis in the forecast. We refer to this set of forecast probabilities as FATC.

- To address the intensity, we select the subset of the forecast tracks that have maximum wind greater than a given threshold. We use 17 m s^{-1} for a direct comparison with the observed intensity, but also consider lower thresholds (e.g., 15 m s^{-1}) to account for potential differences in intensity in the forecasts. We refer to these forecast activity probabilities as FA17 and FA15, respectively.
- Finally, to address the timing of the genesis we again subset the forecast tracks to keep only those that have forecast genesis within 24 h and 500 km of the observed genesis event. We define the forecast genesis event as the first point on the track with wind greater than 17 m s^{-1} and refer to this set of forecast probabilities as FG17.

Table 5-1 summarizes the different sets of forecast probabilities that we consider in this study and the naming convention that we use.

For a broader perspective, to consider the overall forecast probabilities of TC genesis and to include assessment of false alarms, we also conduct some evaluation on a regular 1°x1° latitude-longitude grid. At each grid point, the forecast TS genesis probability is defined as the proportion of ENS members that predict a TS genesis event to occur within 500km of that grid point (center of the 1°x1° box) and between 24h and 216h ahead. Similarly, we define TS genesis to occur if there is an observed TS genesis event within 500 km of the grid point and within the same 192 h (8-day) time window.

Identifier	set	description
FG17	Forecast TS genesis 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and the first time that wind is >17 m s^{-1} along this track is within this time/location tolerance
FA17	Forecast TS activity 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind >17 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind >17 m s^{-1} may have occurred more than 24 h before t_v) and more than 500 km from the given location.
FA15	Forecast TC activity 15 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind >15 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind >15 m s^{-1} may have occurred more than 24 h before t_v) and more than 500 km from the given location. Accounts for overall lower intensity in forecasts
FATC	Forecast TC activity	Forecast TC track passes within 500 km and 24 h of given location (forecast wind may not reach TS strength)

Table 5-1. Different forecast sets considered in this study. Identifier used to refer to each set of forecast probabilities.

5.3 Verification and consistency measures

We evaluate the ENS forecasts of TC activity and genesis using the Brier score (Wilks 2020) which is a measure of the mean squared error of the forecast probability:

$$\mathrm{b} = \frac{1}{N}\sum_{i=1}^{N}(p_n - y_n)^2 \qquad (5\text{-}1)$$

where p_n is the forecast probability (proportion of ENS members that predict the event), y_n is 1 if the event occurs and 0 otherwise and N is the total number of cases.

In the assessment of overall performance using the gridded data (section 4d), we use the observed sample climate probability of genesis $\bar{y}$ as a reference forecast:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5-2}$$

This sample climate includes all dates in our evaluation data and is computed separately for each grid point. By construction, the sample climate has the lowest Brier score of any fixed reference forecast and so is harder to beat than a long-term climate; using this as a reference for the Brier skill score hence provides a conservative indication of forecast skill.

The Brier score of the climate forecast is given as

$$\mathrm{b_c} = \frac{1}{N}\sum_{i=1}^{N} (\bar{y} - y_i)^2 \tag{5-3}$$

and the Brier skill score is then given as

$$\mathrm{B} = \frac{b_c - b}{b_c} \tag{5-4}$$

Positive values of B indicate positive skill relative to the sample climate. Maximum skill $B = 1$ is achieved for perfect deterministic forecasts.

We evaluate the hits and false alarms associated with different forecast probability thresholds using the ROC (Mason 1982; Ben Bouallègue and Richardson 2022) and performance diagram (Roebber 2009). The ROC is a plot of the hit rate (proportion of observed events correctly forecast) against false alarm rate (proportion of observed non-events where genesis was forecast). The performance diagram plots the hit rate against the success ratio (proportion of genesis forecasts that were correct); the performance diagram also shows the frequency bias (number of forecast events divided by number of observed events) and the threat score (number of hits divided by the sum of hits, misses and false alarms).

To measure the jumpiness or consistency over a sequence of forecasts we measure the difference (divergence) d in probability between consecutive forecasts.

Here, we consider the forecasts initialized at 12 h intervals between 24 h and 216 h before a given verification time t_v. The probability of the given event (TC activity or TS genesis) in the ENS forecast initialized at $t_v - h$ is written as $p(t_v, t_v - h)$ and the difference between consecutive forecasts is

$$D(t_v, h) = d\big(f(t_v, h), f(t_v, h-12)\big) = |p(t_v, h) - p(t_v, h-12)| \tag{5-5}$$

The mean divergence over the full sequence of $\mathrm{L} = 17$ initial times is

$$\overline{D(t_v)} = \frac{1}{L-1}\left(\sum_{l=1}^{L-1} D(t_v, 24 + 12l)\right) \tag{5-6}$$

The minimum value of $\overline{D}$ is zero, indicating that the forecast probability does not change over the set of forecasts, while larger values indicate greater differences in probability between successive forecasts in the sequence.

For each observed genesis event, we expect that the forecast probability will be low at the longest forecast ranges (close to the climatological probability) and will increase, ideally reaching close to 100% at the shortest forecast ranges. To account for the expected increase in probability over the sequence of forecasts, we use the difference between the probabilities from the first and last forecasts of the sequence to represent this overall trend. We then subtract this difference from $\overline{D}$ to give the Divergence Index, DI (Richardson et al. 2020b, 2024):

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1}\,|p(t_v, 24 + 12(L-1)) - p(t_v, 24)| \qquad (5\text{-}7)$$

DI summarizes the jumpiness about the overall trend over the sequence of forecasts, with larger values of DI indicating more jumpy forecasts (bigger difference in probabilities).

5.4 Results

Firstly, we evaluate how far in advance the ENS can predict the observed genesis events with low, medium, and high probability. Next, we assess how consistent these probabilities are in the sequence of consecutive forecasts leading up to each observed genesis event. We then consider potential factors that may affect the jumpiness and skill of these forecasts. Finally, we assess the overall skill of the ENS probability forecasts for TC genesis and activity.

5.4.1 *How far in advance can we predict the observed Atlantic TS genesis events?*

Figure 5.1 shows the percentage of the 97 observed genesis events that were forecast with at least 5%, 35% and 65% probabilities at or before each forecast lead time from 216 to 24 hours in advance. The probability thresholds were chosen to be consistent with the categories used to indicate low, medium and high probability respectively in the NHC Tropical Weather Outlook. NHC genesis probabilities are given in 10% intervals and their low, medium and high probability categories are 10-30%, 40-60% and 70-100% respectively.

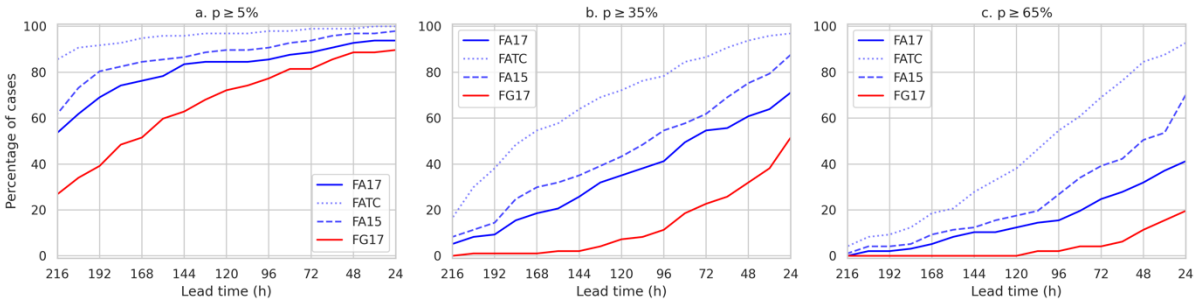


Figure 5.1. Lead time of ENS forecasts of TS genesis. The percentage of cases predicted with probability of at least (a) 5% (low), (b) 35% (medium) and (c) 65% (high) at lead times from 216 to 24 h before the observed TS genesis time.

The red curve shows the results for the FG17 probabilities where the forecast is required to match the observed genesis in both timing and intensity (within the specified 500 km and 24 h tolerances). Few cases are predicted with high probability (Figure 5.1c) and only 20% of cases can be predicted with medium probability more than 72 h ahead (Figure 5.1b). The low probability threshold is reached in over 50% of cases at 168 h lead time (Figure 5.1a), indicating that the ENS is capable of generating tropical storms a week in advance although the predictability is low.

The three blue curves in each panel of Figure 5.1 help to identify some of the reasons for this poor performance in the direct forecasting of the observed genesis. The solid blue curve shows the results for the FA17 probabilities. As well as the hits included in FG17, these allow for early genesis in the forecasts and indicate the proportion of ENS members that have TS activity at the observed genesis time and location. Many more cases are predicted for all three probability categories for FA17 than for FG17: more than 20% of observed events are predicted with high probability at least 72 h ahead, with the proportion increasing to over 50% for the medium probability threshold and over 80% for low probability. 25% of cases are predicted with medium probability at least 6 days (144 h) ahead. The higher probabilities for FA17 compared to FG17 show that the timing of TS genesis is one significant difference between ENS and observed genesis, with a substantial number of forecast tracks reaching TS strength before the observed genesis time. Comparing FA17 and FA15 (solid and dashed blue lines) shows that the choice of wind threshold for the forecast tracks also affects the performance. The relatively minor change of wind threshold from 17 to 15 m s^{-1} increases the proportion of correctly forecast cases by around 10 percentage points. Larger improvements are achieved when considering all forecast tracks without specifying a minimum wind speed (FATC, dotted blue line): around 60% of cases are predicted with medium probability at least 6 days (144 h) ahead and with high probability at least 4 days (96 h ahead). The sensitivity to wind thresholds agrees with results from other studies (Yamaguchi et al. 2015; Zhang et al. 2022).

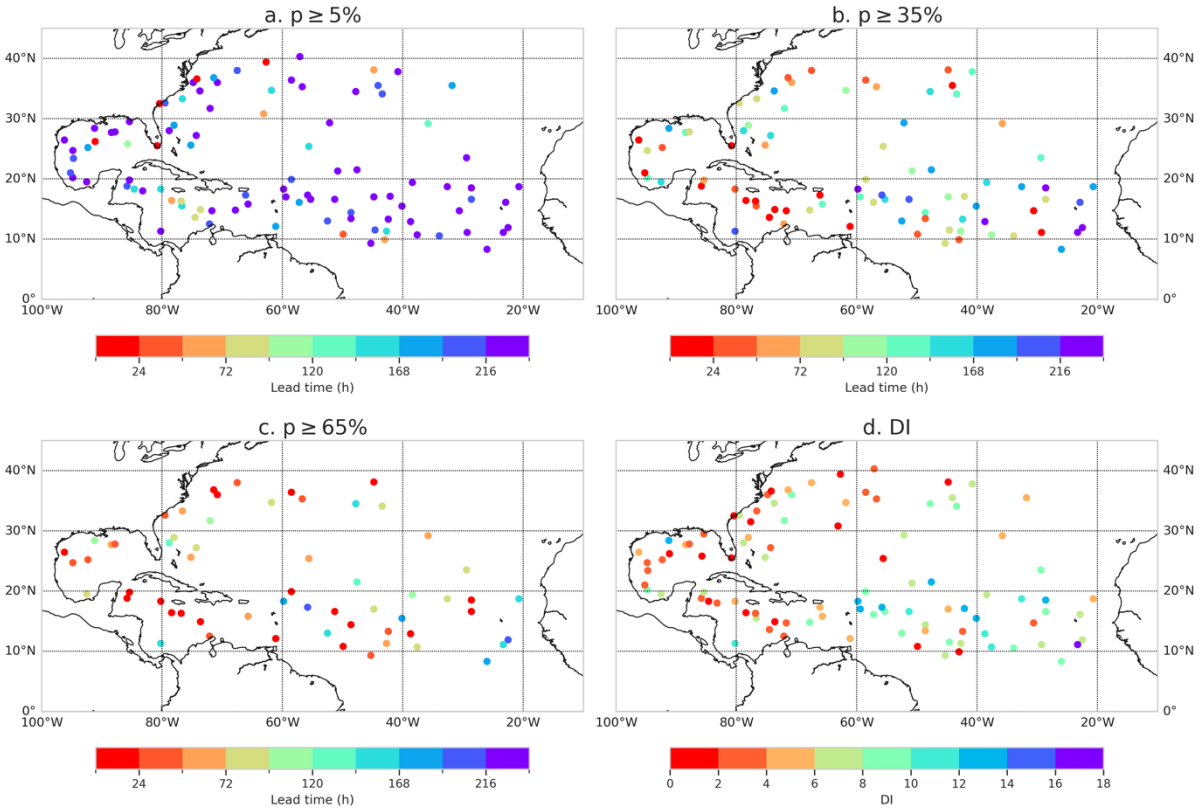


Figure 5.2. Lead time of ENS forecasts of observed TS genesis events. Longest lead time (hours) at which the probability of TS activity (FA17) at the observed TS genesis location was predicted with probability of at least (a) 5%, (b) 35% and (c) 65%. (d) shows the jumpiness in forecast probability for these cases, as measured by DI.

The geographical distribution of the FA17 results is shown in Figure 5.2 for each of the low/medium/high probability thresholds. The TS in the eastern Atlantic tend to be predicted earlier than those in the Caribbean and the Gulf of Mexico. In the central and eastern Atlantic (east of 60°W and south of 30°N) the median lead time for the first indications of TS activity (low 5% probability threshold) is 228 h (the longest lead we have considered here). For medium and high probability thresholds the corresponding median lead times are 132 h and 72 h, respectively. In contrast, the equivalent median lead times for the western Atlantic, Caribbean and Gulf of Mexico (west of 60°W, south of 30°N) are 204, 48 and 36 h. respectively. In other words, the observed genesis events in the eastern Atlantic are predicted 2-3 days earlier than those further west. The predictability for the genesis >30°N is generally similar to that for the western Atlantic. The consistency or jumpiness of these forecasts as measured by DI is shown in Figure 5.2d. Again, there are strong geographical variations, with the highest DI (jumpiest cases) in the central and east Atlantic. The median DI for this region is 8.75, more than twice the median DI value of the western and northern regions (3.5 and 4.0 respectively).

The regional differences may be associated with different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008). The more predictable (and also more jumpy) cases tend to occur in regions

dominated by non-baroclinic developments, although some of the most predictable and jumpiest genesis events occur in the Cape Verde region associated with the low-level baroclinic pathway (baroclinic development under the African easterly jet). The less predictable cases further west and north are in regions where other baroclinic pathways (tropical transition TT (Davis and Bosart 2003, 2004); trough interaction) are more common developments. This is consistent with results from (Wang et al. 2018) who found lower predictability in the TT pathways in an evaluation of reforecasts from the NCEP GEFS ensemble. It is however notable that there are very few predictable cases in the Caribbean and Gulf of Mexico despite the non-baroclinic pathway also being a significant development category in this region. These non-baroclinic pathways often originate from barotropic breakdown of vorticity along stalled fronts, which are smaller and could be less predictable, especially for a lower-resolution model. Environmental factors influencing TC genesis in the western Atlantic have been discussed by Klotzbach et al. (2022) and (in the wider context of cyclonic circulations over central America) by Papin et al. (2017). Additional factors, such as land interactions, may also affect the model ability to correctly predict genesis and would have a more significant impact on genesis forecasts in the western Atlantic and Caribbean rather than eastern Atlantic; this is an area for future research.

5.4.2 *Consistency – the jumpiest forecasts of observed TS genesis events*

The run-to-run consistency of the ENS forecast probabilities is shown in Figure 5.3 for the 12 cases with highest DI for the FA17 forecasts. For each case, the forecast probabilities from the forecasts initialized every 12 hours from 24 to 216 h before the observed genesis are shown for each forecast set FG17, FA17, FA15, FATC.

Most of these jumpy cases occur in September (August for Laura) and there are cases for each of the five years in our sample. As seen from Figure 5.2, the jumpy cases are typically in the central to east Atlantic and between 10°N and 20°N. The two exceptions to both time and location are Bonnie and Claudette which were both early season TCs in the west of the basin. Claudette was the only one of these cases that did not originate from an African easterly wave.

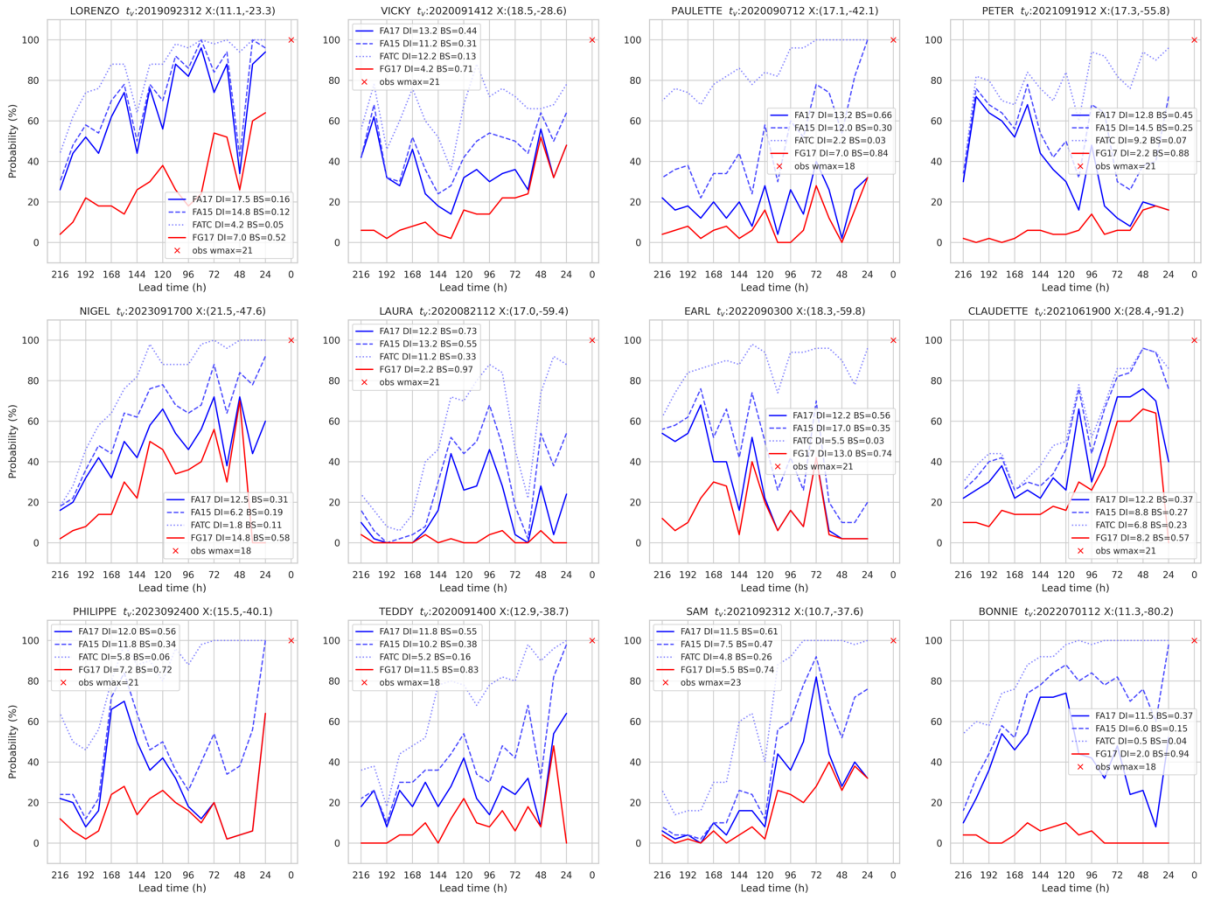


Figure 5.3. Forecast probability of TS activity for the jumpiest FA17 cases. Curves show the forecast probability of TC activity at the observed genesis time genesis time (t_v) and location X (latitude, longitude) for forecasts initialized at 12h intervals from 216 to 24 h before the observed genesis time. The probability for genesis (FA17) is shown by the red line, while the three blue curves show the probability of TC activity with different wind intensity thresholds FA17 (solid dark blue), FA15 (dashed blue), and FATC (dotted light blue). The legend shows the jumpiness (DI) and error (Brier score, BS) for each.

In most cases the jumpiness is related to the forecast intensity: the FATC probabilities are much more consistent from run to run than the FA17 probabilities, and the corresponding DI is consequently much lower. The two notable exceptions to this are Laura and Vicky, which both have substantial jumpiness for the lower wind thresholds. Interactions between African easterly waves or between these waves and other low-pressure systems have also been noted to affect the forecast probabilities of genesis for cases including Laura and Paulette (Magnusson et al. 2021). In the case of Vicky, we note that Teddy and Vicky originated from successive easterly waves that developed off the coast of Africa on 10 and 11 September 2020. The earlier ENS forecasts tended to favour a development associated with Vicky with tracks moving north-westwards away from the coast of Africa, while later forecasts produced more westward tracks associated with Teddy. This uncertainty about which would be the stronger development, together with potential interactions between the two, may account for the jumpiness seen in the predictions for both Vicky and Teddy.

A notable feature of several cases is the high probability for TS activity (FA17) at longer range that is not maintained in the following forecasts made closer to the observed genesis time. Peter, Earl,

Philippe, and Bonnie all have high probability (>65%) at some time 5 or more days ahead, but then have much lower probabilities for later forecasts. However, in all these cases the probability for TC activity (FATC) remains consistently high (well above 65%).

The jumpiest case in this sample is hurricane Lorenzo. There is a clear flip-flopping in the FA17 probabilities between the forecasts started at 0000 UTC and those started at 1200 UTC: the forecasts from 1200 UTC tend to have lower probability for TS activity than the forecasts from 0000 UTC made 12 h earlier and later. This suggests some systematic difference between the analyses for 0000 and 1200 UTC that affects the forecast intensity. Similar flip-flops, though not as large or long-lasting can be seen in some other cases (e.g. Nigel, Paulette).

These cases illustrate a number of different behaviours in the run-to-run consistency of the forecasts. In the next section we consider some of the factors that may contribute to these distinctive characteristics.

5.4.3 *Factors affecting forecast jumpiness and skill*

In this section we consider three factors that may affect the forecast jumpiness results discussed in the previous section. We look at the effect of ensemble size, the issue of flip-flops between 0000 and 1200 UTC analysis times and finally consider the early genesis noted in all results and how this model bias may affect the results for both jumpiness and skill. Although a detailed analysis of causes is beyond the scope of the present study, the aim of this initial assessment is to identify avenues for further research.

5.4.3.1 The effect of ensemble size

We compute the forecast probabilities as the proportion of ensemble members that predict TC activity at a given time and location. How much does the finite ensemble size affect the jumpiness in these probabilities? In this section we use a simple idealized framework to illustrate sampling effects and show the levels of jumpiness that might be expected in an ensemble of 50 members.

Figure 5.4a shows four idealized examples of how the probability of a TC increases over a set of 17 consecutive forecasts (such as the sequences of forecasts initialized every 12 hours from 216 to 24 h before a given observed genesis time, as used in this study). For each set of probabilities, we generate an idealized M-member ensemble by drawing a random sample with the given probability p at each step (Bernoulli process such that each member is either 1, representing forecast of genesis or 0, indicating genesis not forecast) and then compute the DI for this sequence of 17 ensemble forecasts. We repeat this to generate 10000 cases and summarize the distribution of DI over these 10000 cases in Figure 5.4b.

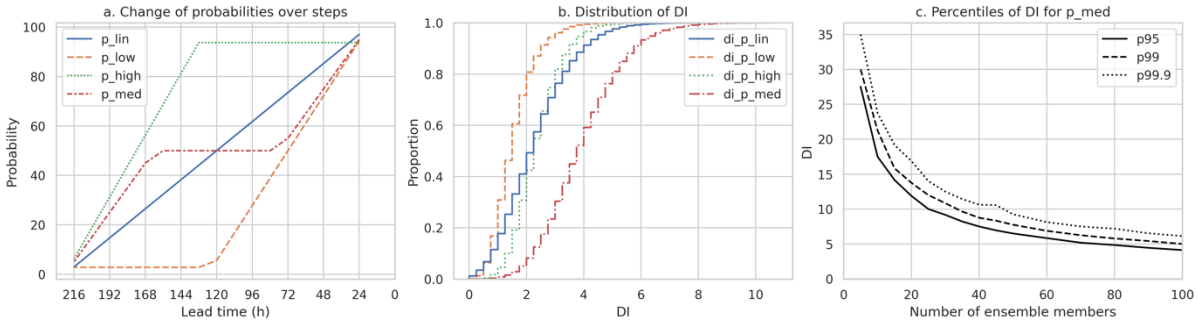


Figure 5.4. Effect of ensemble size on forecast jumpiness. (a) 4 idealized examples of how the probability of TC genesis might evolve over a sequence of 17 50-member ensemble forecasts initialized e.g., every 12 hours from 216 to 24 h before a given verification time. (b) the empirical cumulative distribution of DI for each of the probability sets shown in (a) based on 10000 cases. (c) the effect of ensemble size (number of members) on the extreme percentiles (95% solid, 99% dashed, 99.9% dotted) of the DI distribution for the probability set leading to the jumpiest cases (p_med).

The four examples represent different predictability: linear increase in probability with forecast lead time (p_lin); a high predictability situation (p_high) in which the genesis event is forecast with high probability from five days ahead; a low predictability situation (p_low) where there is no signal at longer range and medium probability (35%) is reached only around 3-4 days ahead; and finally an intermediate situation (p_med) where the signal for genesis is captured with medium probability more than 7 days ahead, and this level of predictability is maintained until the probability increases again closer to the event.

The expected jumpiness for a 50-member ensemble varies depending on the underlying predictability (Figure 5.4b). The low predictability situation is also the least jumpy of the four examples – when the probability of the event is low, there is little variability in the ensemble probability due to sampling (i.e. the finite ensemble size) and the jumpiness (DI) is also low. The intermediate predictability (p_med) situation is the jumpiest, with expected DI substantially higher than for the other examples. In general the sampling effects due to limited ensemble size are largest for probabilities close to 50%.

We have seen that the jumpiness of the ENS genesis forecasts is higher in the central and eastern Atlantic where the predictability is also higher than in other parts of the basin. This is consistent with the above results – the low predictability (p_low) situation is more typical in the west of the basin, while the intermediate (p_med) is more representative of the central and eastern Atlantic. Users should be aware that more predictable situations are likely to be more jumpy because of sampling effects from the finite size of the ensemble.

For all four idealized distributions, the maximum DI is less than 10. In section 5.4.1 we noted that the median DI for the observed genesis events in the eastern Atlantic was 8.75. This is much higher than would be expected from any of the idealized cases considered here. While still high compared to these idealized results, the median DI in the other parts of the Atlantic basin (3.5-4) is closer to the values suggested by these idealized cases.

Figure 5.4c shows how the ensemble size affects the results for the probability distribution that gives the jumpiest results overall (p_med, Figure 5.4b). As noted above, for a 50-member ensemble the probability of DI>10 is extremely small. However, for a 20-member ensemble the chance of having DI>10 is not negligible: we should expect that more than 5% of cases will have DI>10. In general sampling uncertainties will be larger for smaller ensembles (the proportion of members predicting genesis will be a less reliable estimate of the true underlying probability) and therefore the jumpiness from run to run will increase and more cases should be expected with large DI. Conversely, there is a steady decrease in the chances of high jumpiness as the ensemble size increases from 20 to 100 members: for a 100-member ensemble, the maximum DI is not likely to be above 5.

Overall, these idealized results suggest that for the ENS and the set of observed cases considered here, values of DI greater than 10 are unlikely to be due purely to ensemble size. The high median value of DI (8.75) for the cases in the eastern Atlantic suggests there is a substantial number of cases where factors other than pure sampling contribute to the jumpiness.

However, it should be noted that if the ensemble is under-dispersive, the effective ensemble size could be lower than the nominal 50 members and this could significantly affect the DI. These idealized results also show that increasing ensemble size would be expected to reduce overall jumpiness and improve the overall consistency of the ENS predictions. This may be important for some decision-making applications (Jewson et al. 2022) such as deciding when to plan and initiate evacuation from areas at potential risk (Regnier and Harr 2006) or rerouting of transportation to avoid adverse weather (McLay 2008).

5.4.3.2 Analysis impacts – flip-flop between 0000 and 1200 UTC initial conditions

The case of Lorenzo demonstrated a marked jumpiness between the forecasts initialized at 0000 and at 1200 UTC. Figure 5.5 shows the forecast tracks for Lorenzo initialized from 36 to 168 h before the observed genesis time. The circle indicates locations within 500 km of the observed genesis location. The potential for TS activity is predicted at all lead times, and the earliest forecast with high probability was initialized 7 days before the observed genesis time (Figure 5.3). Most of the forecast TCs intensify to TS strength very soon after the track leaves land and moves over the sea off the African coast. This is generally earlier than the observed genesis, consistent with the low probabilities shown in the FG17 curve in Figure 5.3. A notable feature of the forecast probabilities (both FA17 and FA15) is the long sequence of flip-flops in the probabilities between successive forecasts: the forecasts started from 0000 UTC have higher probability than those started 12 h earlier and 12 h later at 1200 UTC.

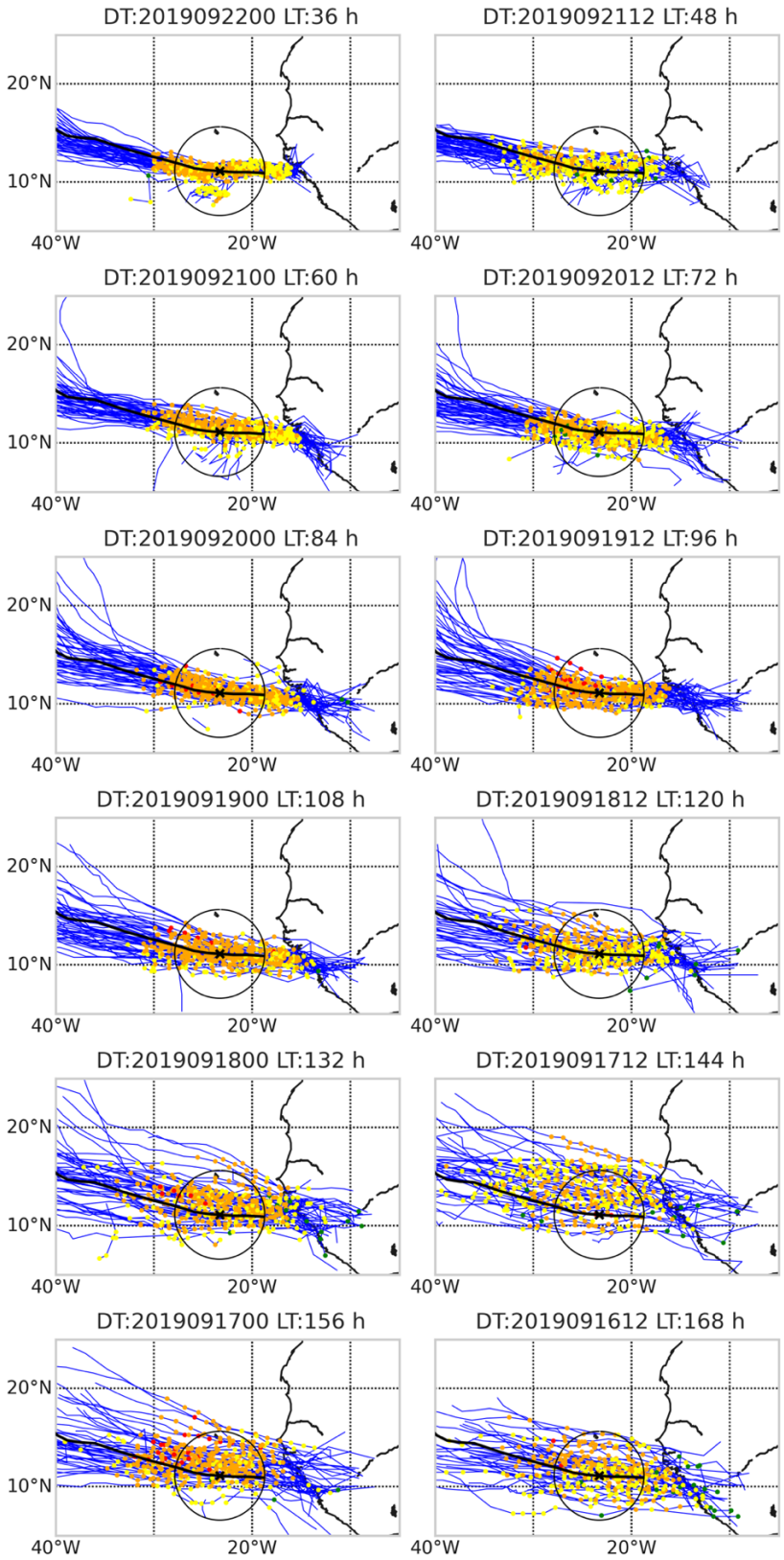


Figure 5.5. ENS forecasts for the genesis of Lorenzo, 1200 UTC 23 September 2019. ECMWF ensemble forecast tracks (blue) and observed track (black). Forecast start dates (DT) from 1200 UTC on 16 September to 0000 UTC on 22 September 2019 (LT: forecast lead time in hours to observed genesis time). Coloured symbols show forecast intensity (maximum wind speed) at all times within 24 h of the observed genesis time (1200 UTC 21 September to 1200 UTC 23 September); Colours represent the maximum wind speed: yellow (<17 m s^{-1}), orange (17-32 m s^{-1}), red (>32 m s^{-1}). Observed genesis location at 1200 UTC 23 September marked (x) and circle indicates locations within 500 km radius of this location.

We extracted the maximum wind for each forecast TC position within 500 km and 24 h of the observed genesis position and time of Lorenzo for all ENS forecasts started from 0000 UTC and compared the distribution of these winds with those from the forecasts started at 1200 UTC. There is a statistically significant shift towards stronger winds in the forecasts from 0000 UTC analysis times (Figure 5.6). This suggests that there is some systematic difference in the assimilation at 0000 and 1200 UTC that affects the intensification of the forecasts in this case. One possibility is the analysis over West Africa where a systematic difference in analysis increments has been identified in the ECMWF assimilation system (Bormann et al. 2023). The reasons for this are not yet understood and are the subject of further investigation.

While some other cases in the same region also have some flip-flops between 0000 and 1200 UTC initial conditions, this is not a common occurrence. Therefore, while assimilation differences may be one factor, it is likely that a combination of factors may be involved to make the large and significant impact found in this Lorenzo case. Further evaluation of this case is beyond the scope of this paper, but the results suggest that additional investigation into the differences between 0000 and 1200 UTC analyses may be relevant.

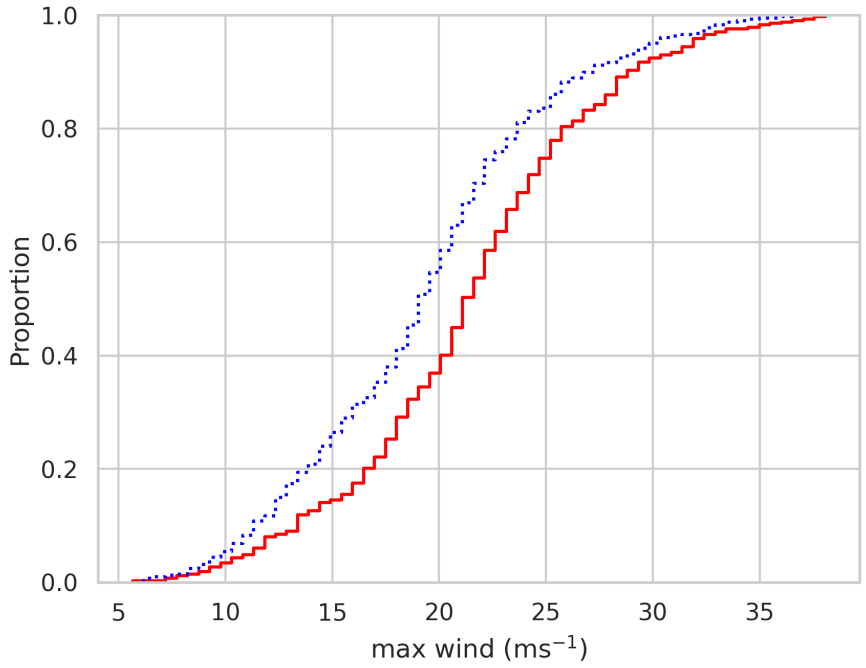


Figure 5.6. Sensitivity of TC intensity to analysis time in ENS forecasts for the genesis of Lorenzo, 1200 UTC 23 September 2019. Empirical cumulative distribution functions of maximum wind speed for ENS TC forecasts initialized at 0000 UTC (solid red line) and at 1200 UTC (dotted blue line) that are within 500 km and 24 h of the observed genesis event of Lorenzo (at 11.1°N, 23.3°W). All forecast start dates between 0000 UTC 14 and 1200 UTC 22 September.

5.4.3.3 Model bias (systematic error)

In many cases that develop from tropical waves over Africa, the forecast tracks intensified to TS strength before the observed TS genesis time. The example of Lorenzo above shows that the forecast tracks often intensified to TS strength immediately after leaving the African continent and moving over sea.

To investigate how typical this early intensification is, we consider all forecast tracks in the 5-year sample. Figure 5.7a shows the location of the first time each forecast track reaches TS strength, accumulated on a 1°x1° grid. Figure 5.7b shows the observed locations for the equivalent first time that the observed TC is reported as TS. There is a substantial peak in the number of forecast TCs that intensify to TS strength immediately after leaving the African coast. In contrast, none of the observed cases are reported to reach TS intensity east of 20°W. There are fewer forecast TS genesis events in the central and western areas (60-80°W, 10-20°N). Overall, there is a shift eastwards of the genesis locations in the forecasts. A similar bias in overforecasting TC genesis was found in the NCEP GEFS reforecasts, associated with overactivity of African easterly waves in that system (Li et al. 2016; Wang et al. 2018).

Overdevelopment of initial wave activity over Africa and the quick intensification to TS soon after the waves move over the open sea may also account for some of the high DI cases shown in Figure 5.3. Peter and Philippe were two cases predicted with high probability at longer lead times, but for both the probability for TS intensity dropped at shorter leads. In each case the higher probabilities occurred for forecasts initialized when the wave activity was still over the African continent, and TS genesis occurred soon after the system left the coast. In the later forecasts where the forecast TC developed further to the west, the probabilities for more intense developments (both FA17 and FA15) were lower.

In summary, there is a tendency in the ENS for TC development to occur too quickly in TCs that develop from African easterly waves and for the intensification to TS to occur soon after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause of the jumpy behaviour seen in some cases.

We hypothesize that this bias is associated with overdevelopment of African easterly wave activity in the ENS and identify this as an important area for future research.

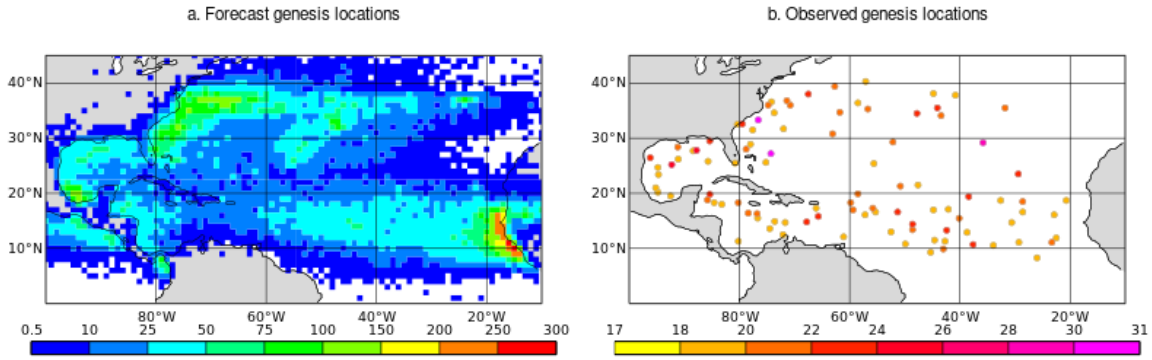


Figure 5.7. Locations of TS genesis in forecasts and observations. (a) forecast genesis: location of the first point on each forecast track with maximum wind speed >17 m s^{-1}; map shows total number of forecast genesis events in each 1°x1° grid box over the full set of forecasts May-December 2019-2023. (b) observed TS genesis locations for all 97 observed cases; colour indicates the reported maximum wind at genesis time in the IBTrACS data (m s^{-1})

5.4.4 *Overall skill of TS genesis forecasts*

So far, we have focused on the results for observed TS genesis events. Although these results show the performance for hits and misses of observed events, they do not take account of false alarms in the forecasts.

To assess the overall performance of the ENS genesis probability forecasts, we now include all forecast tracks, including those false alarm cases where a TS did not actually occur. For each case, and at each grid point, the forecast is the probability that a TS genesis event will occur within 500km and between 24h and 216h ahead

Figure 5.8 shows the Brier skill score (B, Eq. (5.4)) of these ENS forecasts of TS genesis. This shows that there is skill in some areas. The highest skill is in the eastern Atlantic, consistent with the regions where genesis was found to be more predictable at longer lead for the observed cases (Figure 5.2). Although BSS is lower in more western areas, there are still some regions with positive skill. The low overall skill is consistent with the findings in the earlier sections that FG17 skill is limited because of the tendency in the ENS to predict TS genesis earlier than observed.

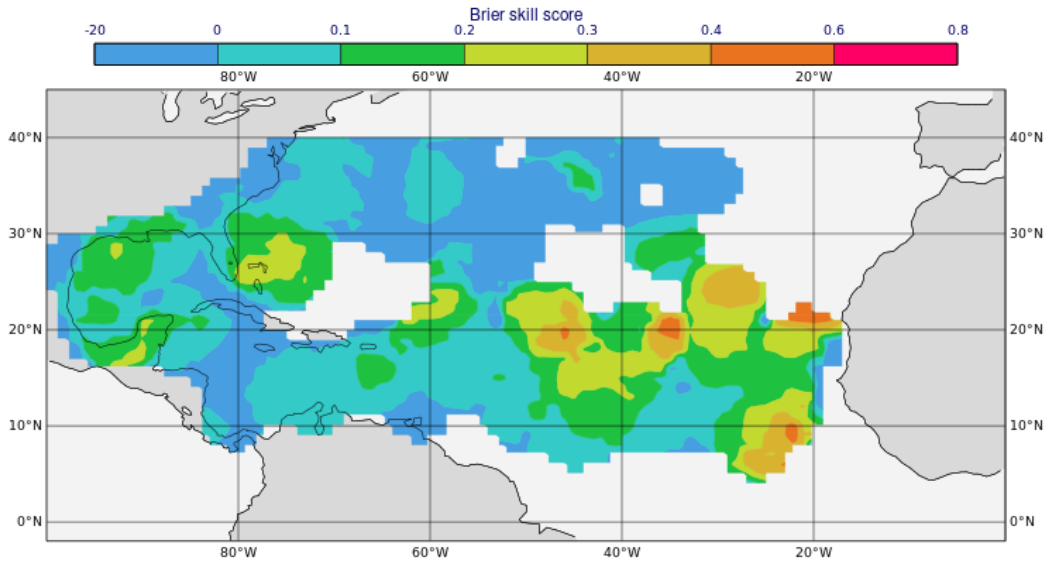


Figure 5.8. Skill of ENS forecasts of TS genesis. Brier skill score for the forecast probability that TS genesis will occur within 500 km of each grid point during the forecast, between 24 h and 216 h lead time; score computed over all forecasts in 5 years sample 2019-2023.

Figure 5.9a shows the reliability diagram for the TS genesis forecasts; the ENS probabilities are grouped into 10% probability intervals and accumulated over all grid points and over the full 5-year sample. The curve is below the diagonal, indicating that the genesis forecasts are overconfident and lack reliability. While this can be a result of lack of spread in the ensemble, it is also consistent with our results that the ENS tends to predict TS genesis earlier than observed. A similar overconfidence is also found in the operational ECMWF verification of TC activity (Haiden et al. 2023) and in corresponding TC activity forecasts from other ensemble systems (Magnusson et al. 2021).

The positive slope of the reliability curve shows that, while lacking reliability, the forecasts do have some resolution: the ability to distinguish between more and less likely genesis events. This discrimination ability is confirmed in Figure 5.9b which shows the ROC diagram for the genesis forecasts. In the ROC computation, all possible forecast probabilities are considered (Ben Bouallègue and Richardson 2022). In Figure 5.9b, the ROC for all grid points is compared with the corresponding ROC curves for three sub-regions: the skill is greater in the eastern Atlantic (east of 60°W and south of 30°N) and lower in the western (west of 60°W and south of 30°N) and northern (north of 30°N) areas. This confirms the regional differences in skill noted in the evaluation of the observed cases (Figure 5.2). Although the reliability diagrams for the sub-areas are more noisy due to the smaller sample size in each sub-area, they also indicate better performance for the eastern region and lowest reliability in the northern region.

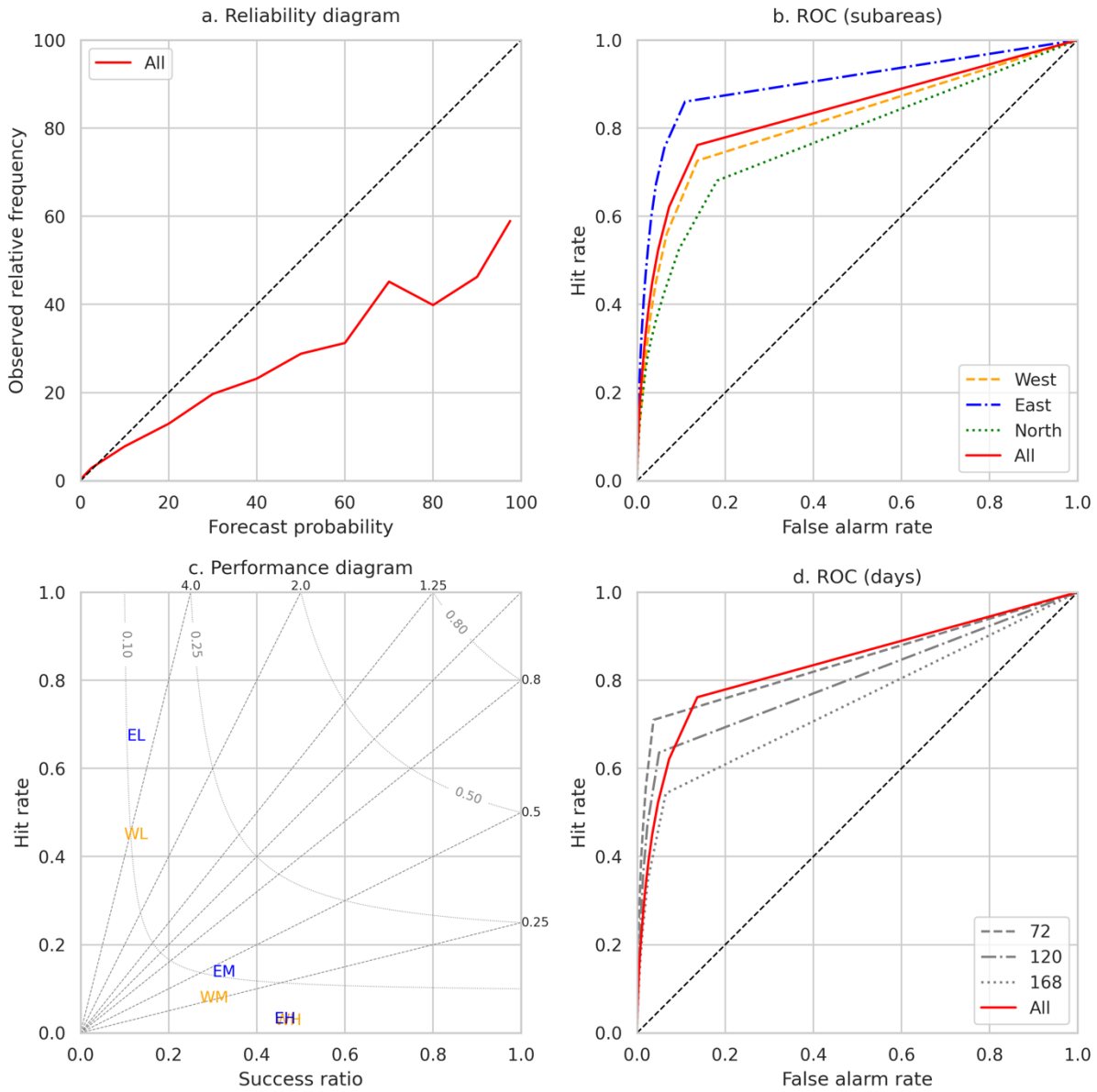


Figure 5.9. Evaluation of ENS forecasts of TS genesis to occur between 24 h and 216 h lead time; scores computed over all forecasts in 5 years sample 2019-2023. a) reliability diagram, results accumulated over all grid points; b) ROC diagram for all grid points (solid red) and for western (orange dashed), eastern (blue dash-dotted) and northern (dotted green) sub-regions (see text for details); c) performance diagram for eastern (E) and western (W) regions and for the low (L), medium (M) and high (H) probability thresholds (first letter indicates region and second letter indicates the probability threshold), grey diagonal lines show bias and grey curved lines show threat score; d) ROC diagram comparing overall results (all, solid red, same as in panel b) with FG17 forecasts of TS genesis at lead times of 72, 120 and 168 h.

To highlight the false alarms as a proportion of the genesis forecasts, the skill of the genesis forecasts for the low, medium and high probability thresholds in the eastern and western regions is shown on a performance diagram in Figure 5.9c. As for the reliability diagram and ROC, Figure 5.9c shows a substantial difference in performance between eastern and western areas, especially for the low and medium probabilities, with substantially better hit rate for a similar false alarm ratio. As for the other performance measures, the northern region has the poorest performance (not shown).

Figure 5.9d shows the ROC curves for the FG17 forecasts for days 3, 5 and 7 (72, 120, 168 h in grey) together with the overall ROC (same as in Figure 5.9b). The discrimination skill decreases at longer lead, although there is still substantial discrimination ability at 168 h. The overall ROC (for genesis between 24 and 216 h) lies between the curves for 120 h and 168 h, suggesting the overall results are reasonably indicative of the medium-range performance.

The results in this section have been based on the comparison of the forecast and observed genesis of tropical storms, defined as the first point on forecast or observed track with wind speed of 17 m s^{-1}. To investigate the sensitivity of the results to the forecast wind speed threshold, we recomputed the ROC results using alternative forecast wind speed thresholds of 8 m s^{-1}, 15 m s^{-1} and 19 m s^{-1}, all verified against the operational genesis of TS (17 m s^{-1}). We found that the results are relatively insensitive to small changes (+/-2 m s^{-1}) in the forecast wind speed threshold, but a large reduction in the forecast threshold (to 8 m s^{-1}) substantially reduces the forecast skill. This section has focused on whether TS genesis will occur at some point during the forecast, and this may be why these results are not too sensitive to the wind threshold – a given threshold will likely be exceeded as the tropical cyclone intensifies during the forecast. A more detailed investigation of the definition of genesis in the forecast and the effect on forecast skill will be a topic for future research.

5.5 Conclusions

We have investigated the ability of the ECMWF ensemble forecasts ENS to predict the genesis of tropical cyclones in the Atlantic basin up to 10 days ahead. We compared the ENS operational TC track forecasts to observed tracks from the IBTrACS archive for all named tropical storms for the 5 years 2019-2023.

We focused on the probabilistic performance of the ENS rather than the evaluation of deterministic forecasts that has been more typically the subject of previous studies.

Defining a genesis event as the first time the TC reached tropical storm strength (winds at least 17 m s^{-1}), the ENS probability forecasts (FG17, Table 5-1) of the observed genesis events had relatively low skill with only 20% of the observed cases predicted with medium or high probability (probability 35% or more) more than 72 h ahead. In many cases the forecast track reached TS strength more than 24 h before the observed TS genesis time. Allowing for this early genesis in the forecasts increased the forecast probabilities (FA17, Table 5-1) for the observed event.

In part, this may reflect differences between the IBTrACS reports and the ECMWF TC tracker - the ECMWF tracker tends to pick up the TC at an earlier stage than the official designation as a TC. Differences in feature identification between different TC trackers can have a significant impact on

the number of TCs identified by a forecast model (Conroy et al. 2023) and there is currently no generally agreed best practice for the definition and evaluation of TC genesis (Dunion et al. 2023).

We also found substantial geographical variation in the performance of the ENS probabilities: observed genesis events were predicted 2-3 days earlier in the central and eastern Atlantic than in other parts of the basin. The regional differences may be associated with intrinsic differences in predictability in different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008; Wang et al. 2018). Investigation of the ENS skill and jumpiness in the different pathways is an area of future research.

We assessed the run-to-run consistency of the ENS probabilities of genesis using the divergence index DI (Richardson et al 2020, 2024). The DI also varied between different regions, with the jumpiest cases being in the central and eastern Atlantic. The median DI here was more than twice that found in the western and northern parts of the basin. The most jumpy cases occurred in different years but almost always in late August or September. In most of these cases the jumpiness depended on the forecast intensity: the forecasts were consistent in predicting the existence of the TC, but the probability for the TC to be at tropical storm strength varied from run to run.

Understanding the causes of jumpiness is important to inform both users and model developers. Forecast jumpiness is a measure of the internal consistency of the forecasting system. Although we used the observed genesis events as reference, the computation of DI does not depend on the observations. Hence, the results for jumpiness are not directly affected by the differences between the model and observed definitions of genesis discussed above. Examining the issues affecting jumpiness can therefore help to identify potential weakness in the modelling system. Based on consideration of the most jumpy cases in our sample, we considered a number of factors that could affect the ENS jumpiness in predicting TC genesis.

One possible cause of large jumpiness is the sampling uncertainty associated with the limited ensemble size. We found that the DI for the most jumpy cases is significantly higher than should be expected for a well-constructed 50-member ensemble. However, jumpiness is sensitive to ensemble size and the highest values of DI found in our results may occur for ensembles with around 20 members. While ENS track forecasts are well-calibrated, the forecast intensity is overall underdispersive (Haiden et al. 2023) and in some situations this may reduce the effective ensemble size, contributing to increased jumpiness. In certain situations with intrinsically low predictability, there may be particular sensitivity to ensemble size and substantially more than 50 members may be needed to properly represent the underlying distribution (Leutbecher 2019; Craig et al. 2022; Kondo

and Miyoshi 2019). This may be important in some genesis situations involving complex interactions between waves where the ENS showed large jumpiness.

In some cases, there was a notable sequence of flip-flops between the forecasts started from 0000 and 1200 UTC analyses. Lorenzo was a particularly strong example, and for this case we found a significant difference between the forecast maximum winds associated with the TCs initialized at the two analysis times, with higher winds from the 0000 UTC analysis. We hypothesize that this may be associated with a known systematic difference in analysis increments at 0000 and 1200 UTC over West Africa in the ECMWF assimilation system (Bormann et al, 2023). However, this flip-flop behaviour was not a common feature across cases, suggesting that a combination of factors in addition to the analysis differences may be involved to make the large and significant impact found in this case. This is an area requiring further investigation.

A significant difference between the observed and forecast TS genesis is that the ENS TC tracks tend to intensify to TS strength earlier than the observed TS genesis event. ENS tracks that develop from African easterly waves often reach TS soon after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause of the jumpy behaviour seen in some cases (for example Peter and Philippe) where earlier forecasts had high probability for TS development, while later forecasts that were initialized after the disturbance moved over the ocean had lower probability. The association with jumpy behaviour lends weight to this being a systematic error in the forecasting system and not just an artifact of the differences between forecast and observed genesis identification methods. We hypothesize that this bias is associated with overdevelopment of African easterly wave activity in the ENS and identify this as an important area for future research.

Finally, we provided a baseline evaluation of the skill of the ENS TS genesis forecasts including all forecasts from the 5-year sample to take account of both hits and false alarms. Overall, forecasts were overconfident but showed good discrimination ability, with higher skill in the east of the basin (particularly for low to medium probabilities) consistent with the results for the observed genesis cases. The ECMWF forecasting system is typically upgraded annually and some of these changes affect the tropical cyclone performance, for example the increase in ensemble resolution in 2023 (Haiden et al. 2023). Given that TS genesis is a relatively rare event, skill evaluation generally needs to be carried out over a sample of several seasons, inevitably covering a number of different model versions (Leonardo and Colle 2021). We found cases of large jumpiness in each year of our sample, and this suggests that the underlying causes still need to be addressed. The overall results can be seen as a general assessment of recent model performance and provide a benchmark against which to evaluate future model developments.

Chapter 6 Additional research towards the aim of this thesis

In addition to the three main papers of this thesis (Chapters 3-5), I have contributed to several other publications that are relevant to the overall aim of the thesis: **to carry out research to improve the use and understanding of ensemble forecasts through the evaluation of run-to-run consistency together with existing verification methods.**

These are summarised below. The first paper, Magnusson et al. (2021), made a significant contribution to the focus of the second and third main objectives of the thesis addressed in Chapters 4 and 5. The remaining papers are grouped under six main headings that correspond to the topics discussed under recommendations and next steps in the next chapter (section 7.4). Some papers contribute to several topics; they are included under the most relevant heading, and Table 6-1 at the end of this chapter indicates which topics are relevant for each paper.

6.1 Contribution to main objectives of the thesis

6.1.1 *Magnusson et al. (2021)*

Magnusson, L., Majumdar, S., Emerton, R., Richardson, D., et. al. (2021) 'Tropical cyclone activities at ECMWF', *ECMWF Technical Memorandum 888*. ECMWF. Available at: https://doi.org/10.21957/zzxzzygwv.

This comprehensive review of TC activities at ECMWF was prepared and presented as a Special Topic paper to the ECMWF Scientific and Technical Advisory Committees in October 2021.

The study identified occasional jumpiness of TC tracks as a significant issue that have caused challenges for forecasters, in particular for the landfall of Hurricane Laura in 2020. This provided an important opportunity to demonstrate the benefit of using the DI consistency measure developed as the first objective of this PhD (Chapter 3) to a significant high-impact weather hazard. The result is shown in Figure 6.1 – applying the DI to the ECMWF ensemble track positions of all 2020 Atlantic TCs shows that the forecasts for hurricane Laura stand out as the most inconsistent cases. This confirmed the subjective feedback from forecasters who were trying to assess the areas most at risk along the US Gulf coast. The DI results showed that inconsistency was especially large around the time of landfall.

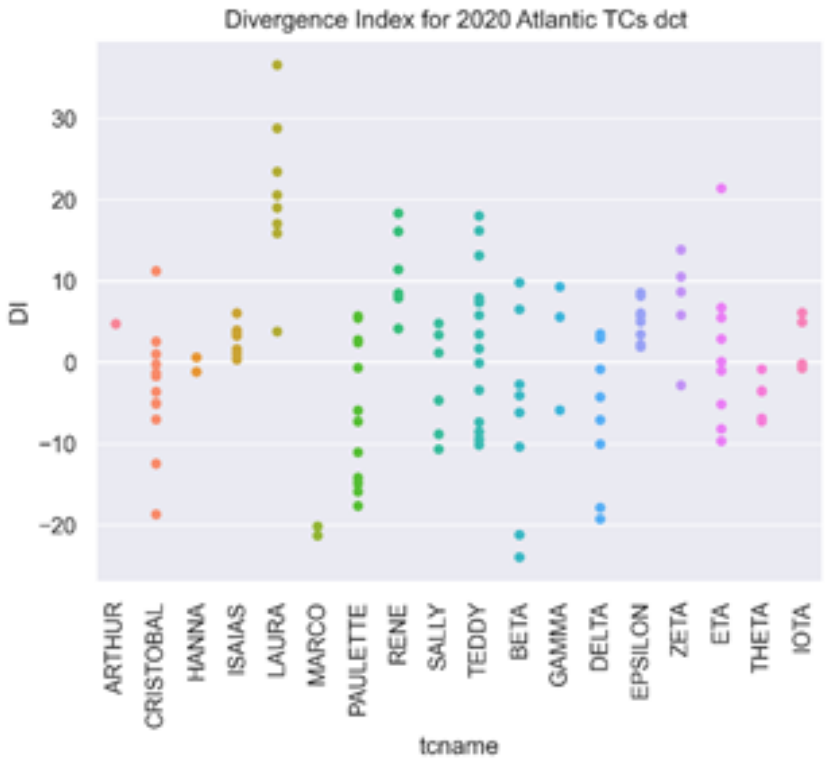


Figure 6.1. Run-to-run jumpiness of ECMWF ensemble TC track forecasts for the Atlantic basin in 2020; all cases with at least 20 members for lead times 12, 24, 36, …, 60 h and longer if available). Divergence index (DI) – large positive numbers indicate inconsistent cases (large negative numbers indicate high consistency). Each dot represents the DI for a sequence of forecasts valid for one valid time of the observed TC. Reproduced from Fig 18 in Appendix A4 (Magnusson et al, 2021).

This preliminary work on TC track jumpiness laid the foundation for objective 2 of the thesis and the research to address this objective was presented in Chapter 4.

The review also highlighted key issues and research gaps in the prediction of TC genesis, including occasional run-to-run jumpiness and the lack of routine verification. This was instrumental in setting the third objective for the PhD which was addressed in the research contained in Chapter 5 of the thesis.

My contribution: I was a member of the editorial team that coordinated the aims, objectives and contents of the report and reviewed all the writing. I contributed to the section on forecast challenges, conducting the analysis and writing the text for the forecast jumpiness (figure 18 and accompanying text), biases in cross-track and along-track errors (figure 19 and accompanying text), and the preliminary evaluation of genesis forecasts (figures 23, 24 and accompanying text).

This report is included in Appendix A4.

6.2 Guidance for forecast users

6.2.1 *Ben Bouallègue et al. (2019)*

Ben Bouallègue, Z., Magnusson, L., Haiden, T. and Richardson, D.S. (2019) 'Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events', *Quarterly Journal of the Royal Meteorological Society*, 145(721), pp. 1741–1755. Available at: https://doi.org/10.1002/qj.3523.

This paper discussed the challenges involved in monitoring trends in ensemble forecast performance, especially for high-impact weather events. It investigated the relative benefits of different choices of methodology, including definition of events, impact of representativeness error and selection of reference benchmark. In Section 7.4.1. it is recommended that trends in ensemble forecast consistency be monitored at operational centres and the results from this paper will provide useful guidance in establishing the appropriate method to do that.

My contribution: I contributed to the discussion and interpretation of results and editing of the paper.

6.3 Causes of jumpiness

6.3.1 *Ben Bouallégue et al. (2020)*

Ben Bouallégue, Z., Ferro, C.A.T., Leutbecher, M. and Richardson, D.S. (2020) 'Predictive verification for the design of partially exchangeable multi-model ensembles', *Tellus A: Dynamic Meteorology and Oceanography*, 72(1), pp. 1–12. Available at: https://doi.org/10.1080/16000870.2019.1697165

This paper develops a methodology to account for different ensemble sizes in verification of multi-model ensemble configurations. It shows that the performance of different ensemble combinations can be robustly estimated based on a small subset of members from each model.

Ensemble size is one factor influencing run-to-run jumpiness. The research in this thesis has shown the benefit of comparing ensembles from different centres (with different numbers of members) and it is also recommended that the use multi-model combinations to mitigate jumpiness be assessed. It will be useful to see if the results from this paper can also be applied to the measures of jumpiness to account for the different ensemble sizes.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of the paper.

6.3.2 *Day et al. (2020)*

Day, J.J., Arduini, G., Sandu, I., Magnusson, L., Beljaars, A., Balsamo, G., Rodwell, M. and Richardson, D. (2020) 'Measuring the Impact of a New Snow Model Using Surface Energy Budget Process

Relationships', *Journal of Advances in Modeling Earth Systems*, 12(12). Available at: https://doi.org/10.1029/2020MS002144.

Diagnosing the causes of model errors for a single variable such as surface temperature can be difficult because of the range of processes involved. This paper developed a set of diagnostic tools that are useful for evaluating the energy exchange at the Earth's surface in an Earth System Model, from a process-based perspective, using in situ observations. These tools were used to show that the improvements to surface temperature following the introduction of a new multi-layer snow scheme in the ECMWF model were a result of a better representation of the surface energy balance, showing that the model bias is improved for the right reasons.

When addressing causes of ensemble jumpiness, it will be important to understand the underlying model weaknesses and confirm that any improvements from model developments are being made for the right reasons. Processed-based diagnostics, such as the one developed in this study, have an important role to play in this process.

My contribution: I contributed to the discussion and interpretation of results and editing of the paper.

6.4 Mitigation of ensemble forecast jumpiness

An outcome of this thesis is a recommendation to investigate the use of multi-model combinations and statistical post-processing to mitigate the impact on users of ensemble jumpiness. The following papers relate to these two approaches.

6.4.1 *Gascón et al. (2019)*

Gascón, E., Lavers, D., Hamill, T.M., Richardson, D.S., Bouallègue, Z.B., Leutbecher, M. and Pappenberger, F. (2019) 'Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe', *Quarterly Journal of the Royal Meteorological Society*, 145(724), pp. 3218–3235. Available at: https://doi.org/10.1002/qj.3615.

This paper assessed combinations of raw and post-processed output from two different ensembles and developed a new method for the statistical calibration that takes account of differences in ensemble size between training data (reforecasts) and the real-time forecasts.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of the paper.

6.4.2 *Feldmann et al. (2019)*

Feldmann, K., Richardson, D.S. and Gneiting, T. (2019) 'Grid- Versus Station-Based Postprocessing of Ensemble Temperature Forecasts', *Geophysical Research Letters*, 46(13), pp. 7744–7751. Available at: https://doi.org/10.1029/2019GL083189.

This paper compared the benefits of postprocessing ensemble temperature forecasts with gridded analyses against postprocessing at observation sites. It showed that the statistical postprocessing improves on the raw model output and that the relative improvement achieved by postprocessing is greater when trained and verified against station observations.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of the paper.

6.4.3 *Korhonen et al. (2020)*

Korhonen, N., Hyvärinen, O., Kämäraïnen, M., Richardson, D.S., Järvinen, H. and Gregow, H. (2020) 'Adding value to extended-range forecasts in northern Europe by statistical post-processing using stratospheric observations', *Atmospheric Chemistry and Physics*, 20(14). Available at: https://doi.org/10.5194/acp-20-8441-2020.

This paper demonstrated the potential to improve extended-range forecasts for weeks 3-4 and 5-6 by better accounting for the influence of the stratospheric polar vortex. Post-processing the ensemble forecasts of near-surface temperature using information about the stratospheric winds at the start of the forecast was shown to improve the skill of the temperature forecasts over northern Europe in winter.

My contribution: I contributed to the discussion and interpretation of results.

6.4.4 *WMO (2021)*

WMO (2021) 'Guidelines on Ensemble Prediction System Postprocessing', Geneva: WMO. Available at: https://library.wmo.int/records/item/57510-guidelines-on-ensemble-prediction-system-postprocessing#.YOW-x-gzYuV (Accessed: 9 July 2024).

This report was developed to provide WMO Members with practical guidelines about the range of postprocessing methods by which they can use information from available EPS forecasts to enhance and improve forecasts for their own specific regions and applications.

My contribution: I contributed to the conceptualization, writing and editing of the report.

6.5 Measures of jumpiness and skill

6.5.1 *Rodwell et al. (2020)*

Rodwell, M.J., Hammond, J., Thornton, S. and Richardson, D.S. (2020) 'User decisions, and how these could guide developments in probabilistic forecasting', *Quarterly Journal of the Royal Meteorological Society*, 146(732), pp. 3266–3284. Available at: https://doi.org/10.1002/qj.3845.

This paper investigated how users combine objective probabilities with their own subjective feelings when deciding how to act on weather forecast information. The audience at a Live Science event held by the Royal Meteorological Society was asked to make yes/no decisions on the basis of a range of forecast probabilities. The results were used to build a picture of the distribution of cost-loss ratios across the audience and to calculate a 'User Brier Score' (UBS) to measure the overall utility to society (represented by the audience as a whole), and which could be used to guide forecast system development. Differences between results for the UBS and Brier score demonstrate how forecast utility depends of the decision-making requirements of the user community.

This study provided valuable insight into users' decision-making based on probabilistic forecast information and on the choice of user-relevant scores to monitor and guide forecast system developments. It would be equally valuable to develop and carry out a corresponding study to investigate users behaviour when making choices over a sequence of forecasts and to assess the consequences for users actions of jumpiness in that forecast sequence.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of the paper.

This paper is included in Appendix A5

6.5.2 *Ben Bouallègue and Richardson (2022)*

Ben Bouallègue, Z. and Richardson, D.S. (2022) 'On the ROC Area of Ensemble Forecasts for Rare Events', *Weather and Forecasting*, 37(5), pp. 787–796. Available at: https://doi.org/10.1175/WAF-D-21-0195.1.

The ROC is a common verification tool used to assess the discrimination ability of ensemble forecasts. Interpretation of ROC results can be difficult for rare events. This paper investigated and provided recommendations to facilitate the use and interpretation of the ROC in these situations. It also introduced a new approach to use the concept of imprecise probabilities and to subdivide the lowest ensemble probability category to address the issue of capturing low-probability events with limited ensemble size.

The ROC has been used in some previous studies of ensemble performance for TC genesis (a rare event) and it would be a useful extension of the research carried out for Chapter 5 of the thesis to extend that verification (using the Brier score) to also include ROC scores. That should be done following the recommendation and applying the methodology in this study. The same approach should be used in integrated evaluation studies of skill and jumpiness of ensemble forecasts of other high-impact weather hazards (section 7.4.5)

My contribution: I contributed to the conceptualization, writing, discussion and interpretation of results of the paper.

This paper is included in Appendix A6

6.6 Applications to other hazards

6.6.1 *Vitart et al. (2019a)*

Vitart, F., Alonso-Balmaseda, M., Ferranti, L., Benedetti, A., Balan-Sarojini, B., Tietsche, S., Yao, J., Janousek, M., Balsamo, G., Leutbecher, M., Bechtold, P., Polichtchouk, I., Richardson, D., Stockdale, T. and Roberts, C. (2019a) 'Extended-range prediction', *ECMWF Technical Memorandum 854*. ECMWF. Available at: https://doi.org/10.21957/pdivp3t9m.

This comprehensive review of extended-range forecasting at ECMWF was prepared and presented as a Special Topic paper to the ECMWF Scientific and Technical Advisory Committees in October 2019. The paper reviewed the progress of ECMWF forecasting for the extended range (1-2 months ahead) over the last five years. It provided a summary of lessons learned from participating in the WMO Subseasonal to Seasonal (S2S) project. It also discussed the considerations and choices made in preparation for upgrading the extended-range forecasting system, including the balance between ensemble size, resolution and production frequency to make best use of the additional computation resources expected from the forthcoming new HPC at ECMWF.

It includes an example showing that run-to-run consistency can be an issue in the extended-range forecasts, and further research on this would be useful: this is discussed further in Chapter 7 (section 7.4.5) of this thesis.

My contribution: I contributed to the strategy for the proposed new configuration, the discussion and interpretation of results and editing of the paper.

6.7 Data availability and challenges

6.7.1 *Ben Bouallegue et al. (2020)*

Ben Bouallegue, Z., Haiden, T., Weber, N.J., Hamill, T.M. and Richardson, D.S. (2020) 'Accounting for Representativeness in the Verification of Ensemble Precipitation Forecasts', *Monthly Weather Review*, 148(5), pp. 2049–2062. Available at: https://doi.org/10.1175/MWR-D-19-0323.1.

This paper addresses the representativeness issue that occurs when comparing point measurements of precipitation to model gridded data representative of the mean precipitation over a larger area. A new methodology is applied to account for this mismatch in spatial scale and the representativeness is shown to have a large impact on verification results.

Although evaluation of ensemble consistency does not rely on observations, they are essential to the overall evaluation of forecast quality. When incorporating consistency as part of an integrated approach to evaluation of ensemble performance it will be important to take account of representativeness when evaluating the ensemble against observations.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of the paper.

6.7.2 *Lavers et al. (2019, 2020)*

Lavers, D.A., Harrigan, S., Andersson, E., Richardson, D.S., Prudhomme, C. and Pappenberger, F. (2019) 'A vision for improving global flood forecasting', *Environmental Research Letters*, 14(12), p. 121002. Available at: https://doi.org/10.1088/1748-9326/AB52B2.

Lavers, D.A., Ramos, M.-H., Magnusson, L., Pechlivanidis, I., Klein, B., Prudhomme, C., Arnal, L., Crochemore, L., Van Den Hurk, B., Weerts, A.H., Harrigan, S., Cloke, H.L., Richardson, D.S. and Pappenberger, F. (2020) 'A Vision for Hydrological Prediction', *Atmosphere*, 11(3), p. 237. Available at: https://doi.org/10.3390/atmos11030237.

Both papers highlight how a lack of hydrological observations in many areas of the world present a severe limitation on the evaluation of hydrological forecasts and can severely restrict the useability of the forecasts and diagnosis and understanding of model weaknesses.

Evaluation of the run-to-run jumpiness of the hydrological forecasts is one approach that may help to identify model issues and is an area for future research. Nevertheless observations will still be essential to evaluate the skill and usefulness of the forecasts.

My contribution: I contributed to the conceptualization, discussion and interpretation of results and editing of Lavers et al. (2019) and to the reviewing and editing of Lavers et al. (2020).

6.8 Summary table of co-authored publications

paper	Guidance for forecast users	Causes of jumpiness	Mitigation of ensemble forecast jumpiness	Measures of jumpiness and skill	Applications to other hazards	Data availability and challenges
Magnusson et al. (2021)						
Ben Bouallègue et al. (2019)						
Ben Bouallégue et al. (2020)						
Day et al. (2020)						
Gascón et al. (2019)						
Feldmann et al. (2019)						
Korhonen et al. (2020)						
WMO (2021)						
Rodwell et al. (2010)						
Ben Bouallègue and Richardson (2022)						
Vitart et al. (2019a)						
Ben Bouallegue et al. (2020)						
Lavers et al. (2019)						
Lavers et al. (2020)						

Table 6-1. Co-authored papers and their connection to this thesis. Column heading correspond to the section titles above and in the discussion of recommendations for next steps in section 7.4

Chapter 7 Discussion and recommendations

This thesis has addressed three objectives relating to the run-to-run jumpiness of ensemble forecasts. The overall aim was to respond to user concerns and to demonstrate how evaluation of jumpiness can complement existing verification methods in understanding ensemble forecast performance and identifying aspects where future research may improve ensemble prediction systems.

The three main objectives were:

1. Develop a suitable index to measure the run-to-run consistency in a sequence of ensemble forecasts and demonstrate how this can identify important cases of high ensemble forecast jumpiness.
2. Evaluate and compare the jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks from three operational centres, identify any common factors and provide guidance to users.
3. Evaluate the skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis, provide guidance to users and identify factors affecting forecast performance.

Each objective was addressed in a research paper, included in the thesis as Chapters 3, 4, and 5. Additional work carried out over the course of the PhD was summarised in Chapter 6. This Chapter discusses the outcomes and limitations of the research of each of the three main papers (Chapters 3, 4 and 5) and presents several recommendations for next steps.

7.1 Identifying run-to-run consistency in ensemble forecasts

This study was designed to answer the first key research question: How can we identify run-to-run consistency in a sequence of ensemble forecasts?

It addressed the first objective: Develop a suitable index to measure the run-to-run consistency in a sequence of ensemble forecasts and demonstrate how this can identify important cases of high ensemble forecast jumpiness.

The study, entitled “Evaluation of the consistency of ECMWF ensemble forecasts”, forms Chapter 3 of this thesis and was published in Geophysical Research Letters in 2020.

7.1.1 *Key findings*

The first objective was to develop and demonstrate an appropriate score to evaluate the run-to-run jumpiness of ensemble forecasts. This needed to be able to quantify both individual jumps and the overall consistency in a sequence of consecutive forecasts valid for a given time. The aim was to account for the full ensemble distribution rather than just considering the ensemble mean (EM) as in

previous studies. The research in this study developed the Divergence Index (DI) to fulfil these criteria and provided the first systematic, objective evaluation of ECMWF ensemble (ENS) jumpiness.

The study identified important general characteristics of run-to-run consistency and compared the ensemble jumpiness with that of the deterministic control (CTRL) and EM forecasts. The DI was substantially lower for the ENS than for the control forecast and the EM, demonstrating how by representing the range of possible scenarios, the ensemble distribution as a whole mitigates the jumpiness seen in the deterministic forecasts. On average, the error and spread of ensemble forecasts increase with forecast lead time, reaching an asymptotic limit dependent on the climatological atmospheric variability. The jumpiness of a deterministic forecast such as the ensemble control also increases throughout the forecast. However the ensemble as a whole behaves differently – at long forecast lead times when predictability is lost, ensemble forecasts will represent samples from the climate distribution and jumpiness will only be due to the sampling effects of finite ensemble size. The largest run-to-run jumps between ensemble forecasts will occur for some intermediate forecast lead time. In Chapter 3, this was found to be at 7-9 days ahead for forecasts of the large-scale flow, but this will be different for different forecasting applications.

The study focused on two key characteristics of the large-scale flow over the European-Atlantic region: the North Atlantic Oscillation (NAO) and Scandinavian Blocking (BLO). Predicting transitions between such weather regimes is a significant scientific challenge at the frontier of NWP and identifying and understanding EPS behaviour in these situations is important to guide research to improve the predictions. The DI identified occasional cases where successive ensemble forecasts give contradictory indications about the probability for a change in weather type. To understand the reasons for this jumpiness, more detailed investigation is needed to identify what aspects of the ensemble forecast configuration lead to such behaviour. This was illustrated for one high-DI case: error tracking showed that the jumpiness was related to the initial mishandling of developing trough-ridge patterns over eastern North America. This provides useful information for developers investigating the causes of poor forecasts. It was also shown that care is needed in the interpretation of jumpiness using a single index - an apparent clear flip-flop in a single index may hide a more complex predictability issue.

7.1.2 *Limitations*

The key aim of this study was to develop and demonstrate a methodology to measure the run-to-run consistency in a sequence of ensemble forecasts. The main limitations of the study are that only a single evaluation measure (chosen to correspond to the standard CRPS verification score) was used to assess jumpiness, and the methodology was only applied to two indices representing the large-scale flow over Europe. The second of these limitations was addressed in the following studies, where the

DI was applied to TCs. Application to different weather hazards and forecasting timescales will be important to evaluate how the methodology and conclusions apply in other circumstances. Forecast verification employs a range of different scores that have different characteristics, which can be used to address different aspects of forecast performance. Similarly, alternative scores to evaluate run-to-run consistency may be useful. This is discussed further in section 7.4.4, including the extension to the multi-variate case identified as a specific limitation of the study.

An additional limitation is that this study only addressed run-to-run consistency and did not consider the potential relationships between jumpiness, error and skill. It is important to demonstrate how evaluation of consistency complements existing verification procedures, and these aspects were addressed in the following studies (Chapters 4 and 5).

Finally, although the study did investigate the causes of two specific cases of high jumpiness, it did not consider the range of factors that may cause ensemble jumpiness or discuss methods to identify such factors. Again, these limitations were addressed in the subsequent research presented in Chapters 4 and 5.

7.2 Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks

This study was designed to answer the second key research question: How does run-to-run consistency vary between ensemble forecasts from different centres, and do these differences shed light on the causes of jumpiness?

It addressed the second objective: Evaluate and compare the jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks from three operational centres, identify any common factors and provide guidance to users.

The study, entitled “Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks”, forms Chapter 4 of this thesis and was published in Weather and Forecasting in 2024.

7.2.1 *Key findings*

This study applied the DI developed in Chapter 3 to investigate the run-to-run consistency of ensemble TC track forecasts from three operational NWP centres (ECMWF, the Met Office and NCEP). It was found that the jumpiness varied substantially between cases and that the jumpiest cases were different for each centre. This suggests that the ensemble jumpiness is not strongly linked to the atmospheric situation or to the availability of observations, which would be expected to affect all centres. Instead, the results suggest that sampling uncertainties (due to limited ensemble size) or individual model deficiencies are more likely causes for the jumpiness. Although earlier versions of the Met Office and NCEP ensembles were shown to be overall more jumpy than the ECMWF ensemble,

recent upgrades significantly reduced jumpiness and the most recent operational ensembles of each centre had similar overall levels of jumpiness. The different centres use different methodologies to generate initial perturbations and account for model uncertainties and the results from this study suggest that the differences in methodology are not a major factor in determining the run-to-run consistency of the ensemble systems.

The study provided quantitative information to users on the expected change in cross-track position from one forecast to the next. This information has practical applications in supporting users' decision making, for example in deciding whether to act now or wait for the next forecast. Finally, the study investigated the link between ensemble forecast jumpiness and probabilistic skill.

There was no clear association between jumpiness and skill, indicating that users should not rely on the consistency between successive forecasts as a measure of confidence; there is no indication that users should expect less jumpy cases to be more skilful. This provides important guidance for forecast users.

7.2.2 *Limitations*

By comparing the jumpiness of three different ensemble systems, this study showed that the causes of jumpiness are most likely due either to sampling uncertainty (finite ensemble size) or to specific issues in the data assimilation, model or ensemble configuration. These issues were not investigated further in this study and this limitation is one area where further research is needed. An initial analysis of the expected impact of ensemble size on ensemble forecast jumpiness was subsequently carried out in Chapter 5 and this was shown to provide valuable context, demonstrating that sampling effects due to ensemble size were not a major factor in that study.

A second limitation of the study is that it only evaluated three North Atlantic TC seasons (2019-2021). Since NWP models are regularly upgraded, selecting an appropriate set of cases is always a balance between generating a large enough sample and ensuring that the results are representative of the current operational forecasts. Although all three centres upgraded their ensemble systems during the three year period of the study, the ECMWF upgrades were neutral in terms of their impact on TC track performance, allowing the ECMWF ensemble to be used as a reference against which to assess the impacts of the upgrades at the other centres.

One approach to investigate further the causes jumpiness in each centre would be to carry out a more detailed examination of individual cases, as was done using error tracking in Chapter 3. However, investigation of the jumpiness for hurricane Laura in the ECMWF ensemble did not identify any specific cause (Magnusson et al. 2021) and further work is needed to investigate this.

An alternative approach is to look for commonalities across the most jumpy cases. However, this would require a substantially larger sample from a given ensemble model version. This could be done by including TCs from other basins, although the model performance is known to vary between basins and the processes involved are different so this also may not produce a homogeneous sample.

Operational ensemble reforecasts have been developed to address the sampling issues discussed here. ECMWF produces ensemble reforecasts for the past 20 years using each operational version of the ensemble system and these are used in statistical post-processing of the operational forecasts and in verification of forecast performance. Because of the limitations of computational resources the reforecasts are run with reduced ensemble size (11 instead of 50 members) and are only produced twice a week from initial conditions on Monday and Thursday. This configuration precludes their use in assessing run-to-run consistency in the operational (twice-daily) forecasts.

Approaches to mitigate jumpiness, such as multi-model or lagged ensembles combinations were not assessed in this study and are topics for future research (see section 7.4.3).

A final limitation to note for this study is that it only addressed the ensemble TC track jumpiness of three operational centres in the North Atlantic. While it provides valuable guidance to users on the jumpiness of the three global ensemble systems and identifies likely factors influencing the ensemble jumpiness, the conclusions may not be applicable to other forecast centres, weather hazards or forecast time-scales. It is therefore important to undertake equivalent studies to provide appropriate guidance in other applications.

7.3 Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

This study was designed to answer the third key research question: Can an integrated approach using both skill and consistency measures be beneficial in evaluation of ensemble forecast performance for weather hazards with significant forecasting challenges and significant observational representativeness or uncertainty issues?

It addressed the third objective: Evaluate the skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis, provide guidance to users and identify factors affecting forecast performance.

The study, entitled “Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis”, forms Chapter 5 of this thesis and was submitted to Weather and Forecasting in 2024.

7.3.1 *Key findings*

This study addressed two key knowledge gaps that limit the effective use of ECMWF ensemble forecasts of TC genesis (Magnusson et al. 2021):

- lack of routine evaluation of the operational ENS forecasts of TC genesis
- lack of guidance to users and poor understanding of occasional run-to-run jumpiness

The study evaluated the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting TC genesis in the Atlantic basin. The study provided a first quantitative evaluation of the probabilistic performance of the ENS, finding that first indications of genesis are picked up at least 7 days ahead in 50% of the observed cases, although strong signals may not appear until 3 days before genesis. There are significant regional differences, with observed genesis events predicted 2-3 days earlier in the eastern Atlantic than in other parts of the basin. In some cases, genesis probabilities are jumpy from run to run; the jumpiest cases occur in the more skilful regions (central and eastern Atlantic) and for situations where the initial signal for genesis appears at longer lead time. In the eastern Atlantic, there is a tendency for the ENS tracks to reach tropical storm strength earlier and further east than observed; this model bias can affect both skill and jumpiness of the genesis forecasts.

The results are sensitive to the choice of wind threshold in the forecasts and this may be related to differences between the observed (IBTrACS) and ECMWF model definitions of TCs. The lack of an agreed best practice for the definition of TC genesis is a significant challenge that makes it difficult to carry out effective evaluation of TC genesis forecasts (Dunion et al. 2023). However, forecast jumpiness is a measure of the internal consistency of the forecasting system, and the results for jumpiness are not directly affected by these differences between model and observed definitions of genesis. Hence evaluation of jumpiness can provide important complementary information, especially in situations with substantial model-observation representativeness issues. Examination of factors that affect ensemble jumpiness may identify potential weakness in the modelling system.

Addressing one of the limitations of the previous study (Chapter 4), this study used a simple idealized framework to demonstrate the expected impact of ensemble size on jumpiness. It was shown that the DI for the most jumpy cases in the study was significantly higher than should be expected for a well-constructed 50-member ensemble, indicating that sampling effects were not the main factor affecting the jumpiness of the genesis forecasts.

Examination of the most jumpy cases in the sample of genesis events identified several alternative factors that could affect both the ENS jumpiness and skill in predicting TC genesis. These include the link between early intensification in the eastern Atlantic and African easterly wave activity, the impact of systematic differences between 0000 and 1200 UTC analyses on forecast intensity, the impact of interactions between tropical waves, and the relationship between skill and the different TC development pathways.

7.3.2 *Limitations*

This study used data from five Atlantic TC seasons (2019-2023) to provide a sample of around 100 observed TC genesis events. This was a compromise between having a large enough sample of events and using as up-to-date versions of the operational ensemble system as possible. Because of the small sample of cases available for each model version, the impact of system upgrades on the results was not assessed, although it was shown that the jumpiest cases were spread across the 5 years of the sample. While including data for other TC basins would increase the overall sample size, the TC development processes involved are different and there are also inconsistencies in the observations between different basins (the IBTRaCS observations are subjectively determined and the criteria used vary between basins); therefore the analysis was restricted to the Atlantic basin. It would, though, be valuable to repeat the study for other basins to provide corresponding guidance for users and to identify factors affecting ensemble performance in those regions.

Although this study included a first assessment of the effect of ensemble size on ensemble jumpiness, further work is needed to provide a more comprehensive evaluation. The study identified a number of factors that may affect both the skill and jumpiness of the ECMWF ensemble forecasts, but further research is needed to understand and address the model issues involved and to improve the forecasting system. This will need more in depth diagnostic analysis of the model behaviour.

The study provided a first assessment of the operational ENS probabilistic skill in predicting TC genesis using the BSS, reliability diagram, performance diagram and ROC. It would be useful to extend the evaluation to include other verification measures. For example, decomposition of the BSS will provide a more understanding of the resolution and reliability aspects of the forecast. It would also be valuable to conduct more conditional verification to better evaluate temporal and spatial variations in skill. However, a larger sample than was used in this study would be needed to give robust results for these additional evaluations.

This study only investigated the performance of the ECMWF ensemble. Although the analysis in Chapter 4 did not identify any common causes of TC jumpiness across the different centres, it would be interesting to carry out a similar comparison for TC genesis. Equivalent studies to Chapter 5 for other centres are also important to provide corresponding guidance to forecasters on their ensemble genesis capabilities. Multi-model and lagged approaches as well as statistical post-processing could also be investigated as ways to mitigate the effects of the biases and jumpiness identified in the ECMWF ensemble.

A key limitation and important issue to address is the lack of a common definition of TC genesis in the different models and in the observed datasets. This complicates the process of comparing the

different models with each other and with the observed genesis events. Although evaluation of run-to-run consistency of each model can mitigate this, intercomparison of results for different centres will still be affected. Work to address the issues of best practice for the definition and evaluation of TC genesis has been proposed as part of the WMO TC-PFP project (Dunion et al 2023).

7.4 Recommendations

7.4.1 *Guidance for forecast users*

A key motivation for this thesis was the concerns raised by users about run-to -run jumpiness in ensemble forecasts and a lack of available guidance to address these concerns.

In Section 2.2, I presented a hypothetical example of a jumpy sequence of ensemble TC forecasts which raised a number of challenges and questions for the forecaster. The research carried out in this thesis has provided general guidance on ensemble jumpiness as well as more specific guidance on the jumpiness of ensemble TC track forecasts, neither of which was previously available to forecast users.

Faced with a similar set of jumpy ensemble forecasts, a user will now have relevant information to help them make optimal use of the ensemble. Firstly, with reference to the results from Chapter 4, the user will know the typical size of jumps between successive forecasts (e.g. 80-90 km for the difference between 120-h and 108-h ensemble mean forecast TC positions for a fixed verification time) and will be able to put the current situation in context (is this an unusually jumpy situation?). This quantitative information will be applicable in the type of decision-making situations (act now or wait for the next forecast) discussed by Regnier and Harr (2006) and Jewson et al. (2022, 2021).

Users will also know from Chapter 4 that if the ensemble forecast from one global centre is particularly jumpy, it may be worth looking at the forecasts from other centres since the jumpy cases tend to occur at different times for the different ensemble systems. Depending on the users decision making, it may be appropriate to consider multi-model ensemble combinations to mitigate jumpiness (see Section 7.4.3).

Users will know that the current ensemble TC track forecasts from the three leading global centres have similar overall run-to-run consistency. If the consistency assessment had been applied as part of the routine evaluation of new model cycles, the global centres would have been able inform users about significant improvements to ensemble consistency in recent model upgrades at both the Met Office and NCEP. Users will also be aware that there is not a strong link between jumpiness and forecast skill and that they should not rely on consistency between consecutive forecasts as a measure of confidence in the forecasts.

Specific guidance will be different for different forecasting applications. The results from Chapter 5 provide additional guidance for users of ECMWF TC genesis forecasts, including that jumpy cases tend to occur more in the eastern Atlantic than the rest of the basin (partly associated with systematic model errors in that region) and that in jumpy cases it can be useful to consider lower wind thresholds in the forecasts (which tend to be more consistent).

This thesis has shown that occasional cases of large run-to-run inconsistency occur in current operational EPS for both the large-scale flow (Chapter 3) and for TC tracks (Chapter 4) and genesis (Chapter 5). All of these have consequences for user decisions. The outcomes of this research provide guidance for forecasters on the expected jumpiness and advice on how to manage this in the forecasting process. It is recommended that operational centres provide guidance on ensemble jumpiness alongside existing guidance on the skill and interpretation of ensemble forecasts. As with standard skill scores, a routine monitoring of forecast jumpiness will allow monitoring of trends and inform users of any changes over time due to model cycle upgrades (Ben Bouallègue et al. 2019).

7.4.2 *Causes of jumpiness*

A second key motivation for the thesis was to identify factors affecting ensemble jumpiness. In Chapter 3, the jumpiest cases were found to be associated with transitions between large-scale regimes. For TC tracks (Chapter 4), jumpiness in some ensemble systems was associated with a lack of spread (indicating insufficient representation of model or analysis uncertainty), while in Chapter 5 model bias was found to be a factor affecting jumpiness in TC genesis. Ensemble size was also shown to be an important factor affecting jumpiness – the ensemble needs to be large enough to give a reliable estimate of the forecast uncertainty. External factors that may affect run-to-run jumpiness include missing or erroneous observations at certain analysis times or a change in atmospheric predictability.

The research in this thesis has focused on ensemble jumpiness in the Euro-Atlantic region, where the prevalence of large-scale weather regimes is well-documented (Ferranti et al. 2015, 2018; Woollings et al. 2010; Straus et al. 2017; Hannachi et al. 2017), and in Chapter 3 the jumpiest forecasts were found to occur in cases of regime transition. Is jumpiness more likely to occur in regions with substantial regime-like behaviour? Ensemble size is likely to be a factor in such cases – if the predicted distribution is multi-modal as would occur for a situation where regime transition is possible, then a larger ensemble will be needed to give robust estimates of the transition probabilities than would be needed in a situation with a more unimodal predictive distribution. Also, conditional bias in forecasting changes of regime (Ferranti et al. 2018) may contribute to situation-dependent jumpiness. However, how much the presence of distinct weather regimes or forecast scenarios (where ensemble

members cluster into distinct weather types) affects the occurrence or magnitude of run-to-run jumpiness is an open question and would merit further research. This would help to identify whether jumpiness is more likely to occur in some regions than others.

Identifying the circumstances in which jumpiness occurs is an important step towards addressing the underlying cause. It is therefore recommended that run-to-run consistency and the associated causes be investigated as part of the routine evaluation of ensemble strengths and weaknesses. The approaches used in the thesis, including comparison between different centres, assessment of common factors in outliers and more in-depth evaluation of individual cases, can be used in such evaluations to analyse the different factors involved.

Further research is recommended on the effects of limited ensemble size on jumpiness. As well as understanding how important this is for a given operational system (and providing guidance on optimal ensemble configuration) the effect of different ensemble sizes needs to be taken into account when comparing across different centres or evaluating research experiments with reduced ensemble size. Work on adapting the DI to take account of ensemble size would be valuable, following equivalent work on verification scores (Richardson 2001; Ferro et al. 2008; Ferro 2014; Ben Bouallégue et al. 2020).

7.4.3 *Mitigation of ensemble forecast jumpiness*

There are several approaches that can be taken to mitigate the impact of ensemble jumpiness for forecast users.

The most straightforward option is to generate lagged ensembles by combining consecutive forecasts (DelSole et al. 2017; Mittermaier 2007). By construction this will reduce run-to-run inconsistencies. Lagged ensembles are already used in some operational settings, typically as a way to increase ensemble size as an efficient use of computational resources (Shanker et al. 2022; Roberts et al. 2023).

While the lagged approach is bound to reduce jumpiness, the impact on forecast skill is a balance between the benefit of increasing the number of ensemble members and the detrimental impact of including older forecasts in the combined ensemble:

- On average the most recent ensemble forecast will have higher skill; adding older forecasts may reduce the forecast skill, especially in cases where the most recent observations lead to a significant increase in predictability.
- Adding earlier forecasts increases the ensemble size and so may better account for sampling uncertainties. This is likely to be particularly useful where either the ensemble size for a given

initial time is small, or in situations with high uncertainty and potential extreme event at the tail of the forecast distribution (low probability extreme event)

Lagged ensembles have been shown to improve skill when combining small ensembles (Ben Bouallègue et al. 2013; Shanker et al. 2022) but do not always provide improvements for larger ensembles (Buizza 2008a). The impact is also likely to depend on the forecast lead time and the difference in initial time between the components of the lagged system (Vitart and Takaya 2021). It is therefore important to assess the impact of a lagged approach on forecast skill and if necessary to balance a potential decrease in skill against a guaranteed reduction in jumpiness.

A second approach is to combine forecasts from different centres to create a multi-model ensemble (Johnson and Swinbank 2009; Swinbank et al. 2016; Hagedorn et al. 2012; Gascón et al. 2019). As for lagged ensembles, a multi-model combination naturally increases ensemble size and so addresses sampling issues. However if the different contributing ensemble systems are not equally skilful, the combined ensemble may have reduced skill compared to the best single-model system. An additional factor is different error characteristics of the component ensemble systems. The multi-model combination may cancel out biases in the different models and compensate for specific flow-dependent errors in individual systems.

In Chapter 4 it was shown that the jumpiest cases for TC track forecasts were different in the different centres' ensembles, suggesting that a multi-model combination would be beneficial, while the skill of multi-model ensemble track forecasts has already been demonstrated (Titley et al. 2020). Many studies have investigated the skill of multi-model ensembles and shown that benefits vary depending on the forecast variable, forecast range, relative skill and size of contributing models (e.g. see the review of TIGGE-based multi-model studies in Swinbank et al. (2016)).

An additional approach, that can be applied independently or in combination with either the lagged or multi-model combination, is statistical post-processing of the ensemble distribution. This can include weighting of the individual models that contribute to the combined multi-model ensemble. There are many methods for post-processing ensemble forecasts: see (WMO 2021; Vannitsem et al. 2018, 2021) for recent comprehensive reviews. The main focus of ensemble post-processing has been on improving forecast skill. It would be beneficial to also investigate the impact of different post-processing methodologies on forecast jumpiness.

Further research is recommended to explore the benefits of lagged and multi-model ensembles as well as statistical ensemble post-processing to mitigate ensemble jumpiness.

7.4.4 *Measures of jumpiness and skill*

In this thesis the run-to-run consistency of ensemble forecasts was evaluated using DI and $\overline{D}$. The difference between two consecutive ensemble forecasts valid for the same time was measured by the quadratic divergence corresponding to the CRPS error measure (Chapters 3, 4) to take account of the ensemble distribution, and by directly using the absolute difference in probability for the discrete event of TC genesis in Chapter 5. The quadratic divergence has the added advantage of being applicable to deterministic forecasts as well as to EPS, and the ENS, EM and CTRL were compared in Chapter 3. The CRPS and BS used in this thesis are two common measures of ensemble probabilistic skill. CRPS has not been used in TC track verification before and it is recommended to add it to the routine verification of ensemble TC track forecasts.

CRPS and BS can both be interpreted in terms of potential economic value of forecasts to users in an idealised cost-loss decision-making model. However, this assumes a uniform distribution of users across all possible cost-loss ratios. Studies show that users tend to be more concentrated towards lower cost-loss ratios, and appropriate skill scores to focus on a specific set of users can be derived (Richardson 2001; Rodwell et al. 2020). It is recommended to carry out research to investigate the impact of run-to-run jumpiness on users with different decision-making criteria; this would require using a more complex decision model than the simple cost-loss model, to take account of a user's choice to act now or wait for the next forecast.

A range of verification measures are used to focus on different aspects of ensemble performance. The relative operating characteristic (ROC) is another common verification measure, useful in comparing the performance of different forecasting systems, including comparing deterministic and probabilistic systems. Although care is needed in the application and interpretation when evaluating rare events (Ben Bouallègue and Richardson 2022), it will be interesting to explore its use an integrated investigation of skill and jumpiness.

The logarithmic or ignorance score (Good 1952; Roulston and Smith 2002) is an alternative error measure that can be related to a user's expected profit from betting on forecast outcomes (Hagedorn and Smith 2009). The logarithmic score can be more sensitive than the CRPS to some flow-dependent variations in ensemble spread (Leutbecher 2019). The divergence function associated with the logarithmic score is the Kullback-Leibler divergence (Gneiting and Raftery 2007) and it would be interesting to see whether this would be pick out different aspects of forecast jumpiness. However, the logarithmic score and Kullback-Leibler divergence both require non-zero probabilities and therefore are not directly applicable to the raw ensemble members. Hence some further research is needed to derive the best way to assess the ensemble jumpiness using these measures.

In Chapter 3 it was shown that jumpiness is different for the NAO and BLO indices. A multivariate measure of jumpiness would allow assessment of both indices together and may lead to more insights. The multivariate equivalent of the CRPS is known as the energy score (Gneiting and Raftery 2007) and a corresponding measure of the difference between two ensembles could be made in the same way.

7.4.5 *Application to other hazards*

This thesis has demonstrated the benefit of assessing the run-to-run jumpiness of ensemble forecasts in the large-scale flow over Europe and for TC tracks and genesis. It would be valuable to extend the evaluation to other weather hazards.

In 2022, the United Nations launched the Early Warnings for All (EW4All) initiative (WMO 2022) with the aim to ensure that within five years early warning systems are in place to protect all people from hazardous weather, water, and climate events. The WMO has identified EW4All priority hazards, including tropical cyclones, extra-tropical storms, heatwaves and cold-waves, drought, thunderstorms, floods, coastal inundation, storm surges, and glacial lake outflows.

Many of these priority hazards present both forecasting and evaluation challenges as was the case for TC genesis. Observations are often either lacking (data-sparse regions), differ in scale from the model (heavy localised precipitation) or are not directly comparable to model output variables (e.g. thunderstorms).

Heatwaves and cold spells are often related to changes in large-scale weather patterns. The jumpiness in medium-range forecasts for two such patterns (NAO and blocking) related to the occurrence of wintertime cold spells in Europe was considered in Chapter 3, and the methodology could be applied to summer heatwaves using appropriate summer weather regimes. Prediction of heatwaves and cold spells beyond the medium range is a major challenge (Brunet et al. 2023; ECMWF 2015) while low skill and jumpiness in the ECMWF extended-range forecasts for the 3-4 week range have been raised as areas of concern for the ECMWF forecast users (Hewson 2021, 2020). The configuration of the ECMWF extended-range forecasts was recently changed from 50-member ensembles run twice a week to 100 members run daily (Vitart et al. 2022, 2019a). This increase in ensemble size and forecast frequency improves the skill and should also help to improve jumpiness. However, skill is still limited at weeks 3-4 and more work is needed to identify causes for the lack of predictability. An evaluation of run-to-run consistency in the new configuration will provide useful guidance to users on the current jumpiness characteristics and may also help to identify model issues.

Vitart et al. (2019a) present an example where unusually mild temperatures over Europe were well forecast 2-3 weeks ahead, while longer-range forecasts predicted unusually cold temperatures for the same period. A comparison with the ensemble forecasts from other centres showed a similar

behaviour, suggesting that in this case there may be a common driver of the jumpy behaviour. Although this is a single case, it demonstrates that the factors affecting forecast jumpiness may be different if different variables and timescales are considered.

It is recommended that the integrated approach used in Chapter 5 be applied to evaluate the capability of ensemble prediction systems to provide appropriate products to support early-warnings for the WMO priority hazards. Including evaluation of run-to-run consistency in the evaluation will help to mitigate identified problems with observations, contribute to understand factors affecting model predictability and provide guidance to users to allow then to make best use of the available ensemble forecasts.

7.4.6 *Data availability and challenges*

In Chapter 4 the run-to-run consistency of ensemble forecasts from three global operational centres was compared using data from the TIGGE archive. The TIGGE archive for medium-range ensemble forecasts (Swinbank et al. 2016) and the corresponding S2S archive for sub-seaonal to seasonal forecasts (Vitart et al. 2017) have been invaluable resources for research in evaluation of forecast performance, multi-model ensemble combinations and identification of model systematic errors. It is strongly recommended to use these datasets to help assess differences in jumpiness between ensemble systems, to identify factors affecting jumpiness (as done in Chapter 4) as well as to demonstrate mitigation approaches through use of multi-model ensembles. It is important that both TIGGE and S2S are continued to facilitate this research.

In both Chapters 4 and 5, having a sufficiently large sample was an issue and necessitated compromising between the number of cases in the sample and ensuring that the results are relevant to the performance of the current operational forecasts. Reforecasts are designed to address this data availability issue.

The ECMWF operational reforecasts comprise 11-member ensembles run twice weekly for the past 20 years to provide a homogeneous dataset of model integrations using the same model version as used for the real-time forecasts (Vitart et al. 2019b). They are used to calibrate the real-time forecasts and to evaluate the skill of the operational system over a much larger number of years than is possible with the real-time forecasts (the ECMWF forecasting system is typically upgraded each year). Because they are currently produced only twice a week (because of constraints computational resources), they cannot be used to evaluate the run-to-run jumpiness between the medium-range forecasts initialized every 12 hours as needed for the research in this thesis. Although the configuration will remain the same for the medium-range reforecasts, the reforecast configuration for the extended-range system will be changed to run every 2 days in the next model cycle (49r1) due to be implemented in late 2024.

Although this is still less frequent than the operational extended-range forecasts that now run once a day, it may be useful to investigate the jumpiness in the extended-range reforecasts, especially to assess the jumpiness at the longer forecast ranges 3-5 weeks ahead.

The ECMWF reforecast ensembles have substantially fewer members than the operational forecasts (11 compared to 51 members) and this difference needs to be accounted for in both verification (Ben Bouallégue et al. 2020) and post-processing (Gascón et al. 2019). Evaluation of run-to-run consistency will also need to take account of this difference if the reforecasts are to be used to provide guidance on the expected jumpiness of the operational forecasts. It is recommended that research is carried out to address the feasibility to extend predictive verification approach (Ben Bouallégue et al. 2020) to the evaluation of ensemble jumpiness.

The NOAA reforecast dataset (Hamill et al. 2013) is run daily and although again not at the same frequency as the operational medium-range forecasts may be a useful dataset to assess some aspects of the jumpiness in the GEFS. Again, it has significantly fewer members than the real-time forecasts and additional research will be needed to interpret how results would apply to the real-time system.

It is recommended to use reforecast datasets where feasible to investigate run-to-run consistency and to carry out necessary research to account for differences in ensemble size.

The availability, quality and representativeness of observational datasets is another challenge for forecast evaluation. Regional differences in TC reporting and differences between the definition of TC genesis in the model and observation datasets were a significant limitation in assessing the skill of the TC genesis forecasts in Chapter 5. A key benefit of assessing consistency is that it does not rely on observations. Nevertheless, observations are essential to quantify and understand the skill and value of ensemble forecasts. It is recommended that a strong case be made for the observational requirements for forecast evaluation as well as for forecast initialization (Lavers et al. 2020, 2019), and that observational representativeness be accounted for in verification (Ben Bouallegue et al. 2020).

7.4.7 *Data-driven models*

In the last two years, there has been rapid development of weather forecast models based on Machine Learning (ML). Following the pioneering work of Keisler (2022), several groups have developed data-driven ML weather forecast models (Pathak et al. 2022; Bi et al. 2023; Lam et al. 2023a,b; Chen et al. 2023), typically trained on the ECMWF ERA5 reanalysis (Hersbach et al. 2020). Although requiring significant computing resources to train the model, these data-driven models can produce real-time deterministic medium-range forecasts much more quickly and for a fraction of the cost of running an NWP model. The ML models do though require the initial conditions from the NWP-based data assimilation to initialise their forecasts. The skill of these models is comparable to that of the current

operational global NWP models for both the large-scale flow patterns and for extreme events (Ben Bouallègue et al. 2024).

The physical realism of these ML models is an area of significant current research. While they have been shown to produce physically realistic dynamical behaviour in some studies (Hakim and Masanam 2023), other authors have noted substantially different behaviour, for example in initial error growth (Selz and Craig 2023) and some physical inconsistency between variables (Bonavita 2024).

Several data-driven models, including the AIFS developed at ECMWF (Lang et al. 2024) are displayed in real time on the ECMWF web site and are regularly monitored by ECMWF forecast analysts. Although examples of jumpy behaviour have been noticed, there is an impression that overall, the ML models tend to be less jumpy from run to run than the deterministic ECMWF NWP model (Magnusson, 2023).

The main focus to date has been on deterministic ML forecasts, but experimental ensemble systems are also now being explored (Hu et al. 2023) and ECMWF recently introduced the first experimental version of its data-driven ensemble forecast model, AIFS-ENS.

Exploring the run-to-run consistency of these data-driven systems, in particular the ensemble forecasts, may give additional insights into the behaviour of these ML models. It is recommended to investigate this once sufficient data is available from the ML ensembles. Depending on the results of this proposed work it may be useful to update the Weatherbench benchmark procedures for evaluation of data-driven models (Rasp et al. 2020, 2024).

7.5 Summary

This Chapter has discussed the key outcomes and limitations of the research conducted for this thesis and presented recommendations on next steps to apply these findings and on directions for further research.

The scientific contribution of this thesis is a new objective diagnostic approach to quantify the run-to-run consistency of a sequence of ensemble forecasts. This includes the development of a new divergence index (DI) that was used to evaluate the ensemble consistency for the first time in a way that accounted for all aspects of the ensemble distribution. A second new score, $\overline{D}$, was developed as a complement to DI, and it was shown that DI and $\overline{D}$ can be used together to distinguish inconsistency due to trends in the forecasts from inconsistency due to flip-flopping between different solutions. The new scores were used to provide new insights into the relationship between jumpiness, skill and spread, and to show how the asymptotic behaviour of ensemble consistency differs from that of deterministic forecasts.

Another important scientific contribution of this thesis has been to provide practical guidance to address user concerns over ensemble jumpiness. In particular, Chapters 4 and 5 have provided specific guidance for users that will enable them to make better use of the available operation ensemble tropical cyclone track and genesis forecasts. The new diagnostic approach developed in this thesis can and should be used to provide equivalent guidance for other ensemble forecast applications. It was shown that evaluation of forecast consistency is complementary to the current focus on skill and ensemble spread, and that an integrated approach using both skill and consistency measures can be beneficial in evaluation of ensemble forecast performance.

The work in this thesis and the recommended next steps will contribute to improving the utilisation of ensemble forecasts to provide early warnings of significant weather hazards, especially important in the context of the UN Early Warnings for All initiative.

Chapter 8 Conclusions

There is a growing recognition that ensemble forecasts have a key role to play in improving the skill and use of early warnings to enable actionable decisions to mitigate the impact of hazardous weather events. This thesis has presented a range of research to improve the use and understanding of ensemble forecasts through the evaluation of run-to-run consistency together with existing verification methods. This chapter summarises the main scientific contributions of the thesis.

A new diagnostic approach was developed to quantify the run-to-run jumpiness (inconsistency) in a sequence of ensemble forecasts (Chapter 3). The divergence index (DI) enabled the consistency of the ECMWF ensemble forecasts (ENS) to be assessed for the first time in a way that accounted for all aspects of the ensemble distribution. It was shown that the DI was much lower for the ENS than for the EM and the CTRL, demonstrating how the ensemble as a whole mitigated the jumpiness of the deterministic forecasts by representing the range of possible forecast scenarios. The jumpiness of the CRTL increased throughout the forecast, while the ENS jumpiness peaked around forecast days 7-9. This difference is a consequence of the different asymptotic behaviour of the single CTRL and the ENS distribution as a whole, and is important for both users and developers to understand.

The benefit of the DI was demonstrated for two indices representing different large-scale weather regimes over the north Atlantic and Europe. The study found that peaks of high and low consistency occur at different times for NAO and BLO; there was no strong correlation between the forecast jumpiness for the two regimes. A more detailed investigation of the jumpiest cases found that the inconsistency in these cases was related to uncertainty in the transitions between the two regimes, originating in mishandling of developing trough-ridge patterns over eastern North America.

Investigation of the jumpiness in ensemble TC track forecasts from 3 different global centres found that each centre had occasional cases of high jumpiness, but that the jumpiest cases were different for each centre (Chapter 4). This implies that the cause of the ensemble jumpiness is not strongly related to either the prevailing atmospheric conditions or the available observations, since these would be expected to affect the forecasts from all three centres. It is more likely that the jumpiness is a result of specific issues in the data assimilation, modelling or ensemble configurations of each centre. It was shown that recent model upgrades to both MOGREPS-G and GEFS did significantly reduce their jumpiness (probably as a result of improvements to the ensemble spread) and that the overall level of jumpiness is similar for all three centres in their current operational configurations. The study introduced a second score, $\overline{D}$, to help identify different aspects of forecast consistency. Used together, DI and $\overline{D}$ can distinguish between run-to-run consistency due to trends in the forecast and cases of flip-flopping between different solutions.

The association between jumpiness and skill was also investigated. No clear link was found, indicating that users should not rely on the consistency between successive forecasts as a measure of confidence. However, the ensemble spread does provide useful situation-dependent information on the forecast uncertainty. The study also provided quantitative guidance to forecasters on the expected jumpiness between successive forecasts, addressing specific user requirements for decision makers who need to decide between acting now and waiting for the next forecast.

The skill and consistency of the operational ENS forecasts of TC genesis were evaluated for the first time, revealing significant regional differences in performance across the Atlantic basin (Chapter 5). Observed genesis events were predicted 2-3 days earlier in the eastern Atlantic than in other regions. However, the forecast genesis probabilities were not always consistent from run to run, with the jumpiest cases occurring in the more skilful regions and in situations where the initial signal for genesis appeared at longer lead time. A notable bias for TCs to develop to tropical storm strength earlier and further east in the model than in the observations was identified and shown to affect both skill and jumpiness.

A new aspect of this study was the assessment of the expected impact of ensemble size on the forecast jumpiness. This showed that the DI for the jumpiest cases in the study was significantly higher than expected for a reliable (well-tuned) 50-member ensemble and is therefore likely to indicate a deficiency in the ensemble system. Examination of the jumpiest cases identified several potential contributing factors, including the link between early intensification in the eastern Atlantic and African easterly wave activity, and the impact of systematic analysis differences between 0000 UTC and 1200 UTC on forecast intensity. The relationship between skill and the TC development pathways was highlighted as another area for future work.

A key aim of the study was to provide users with guidance on the ENS probabilistic performance in predicting Atlantic TC genesis, to support wider use of the operational ENS forecasts. As well as providing the first quantitative skill assessment of the ENS genesis forecasts, this guidance included recommendations on the differences between forecast and observed genesis (timing and location), situations where jumpiness is more likely to occur, and practical steps to mitigate the impact of the jumpiness, including the consideration of different wind thresholds in the forecasts.

The research carried out in this thesis has been directed towards improving the use and understanding of ensemble forecasts through the evaluation of ensemble forecast consistency together with existing verification methods. Recommended next steps were discussed in Section 7.4. Implementation of these recommendations at NWP centres will ensure that users have the necessary information and guidance to mitigate the impact of occasional run-to-run jumpiness and will provide additional

feedback to model developers on model weaknesses, complementing the use of existing evaluation tools. The recommendations for future research will help to develop the range of evaluation tools to address ensemble consistency, improve understanding of EPS behaviour and extend the applicability of the approach.

It is hoped that the research carried out in this thesis together with the recommended next steps will be beneficial in future research and operational activities to improve the utilisation of ensemble forecasts to provide early warnings of significant weather hazards, contributing to the UN Early Warnings for All initiative.

References

Avila, L. A., R. J. Pasch, and J.-G. Jiing, 2000: Atlantic Tropical Systems of 1996 and 1997: Years of Contrasts. *Mon Weather Rev*, **128**, 3695–3706, https://doi.org/10.1175/1520-0493(2000)128<3695:ATSOAY>2.0.CO;2.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC Area of Ensemble Forecasts for Rare Events. *Weather Forecast*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-0195.1.

——, S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, **22**, 49–59, https://doi.org/10.1127/0941-2948/2013/0374.

——, L. Magnusson, T. Haiden, and D. S. Richardson, 2019: Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, **145**, 1741–1755, https://doi.org/10.1002/qj.3523.

Ben Bouallégue, Z., C. A. T. Ferro, M. Leutbecher, and D. S. Richardson, 2020: Predictive verification for the design of partially exchangeable multi-model ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, **72**, 1–12, https://doi.org/10.1080/16000870.2019.1697165.

Ben Bouallegue, Z., T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for Representativeness in the Verification of Ensemble Precipitation Forecasts. *Mon Weather Rev*, **148**, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1.

Ben Bouallègue, Z., and Coauthors, 2024: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bull Am Meteorol Soc*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Berrisford, P., and Coauthors, 2011: The ERA-Interim archive Version 2.0. *ERA Report Series*, 1. Available at: https://www.ecmwf.int/en/elibrary/73682-era-interim-archive-version-20 (Accessed: 20 August 2024).

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature 2023 619:7970*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bidlot, J.-R., F. Prates, R. Ribas, A. Mueller-Quintino, M. Crepulja, and F. Vitart, 2020: Enhancing tropical cyclone wind forecasts. *ECMWF Newsletter*, **164**, 33–37, https://doi.org/10.21957/k0w4fp581h.

Bonavita, M., 2024: On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophys Res Lett*, **51**, e2023GL107377, https://doi.org/10.1029/2023GL107377.

Bormann, N., L. Magnusson, D. Duncan, and M. Dahoui, 2023: Characterisation and correction of orbital biases in AMSU-A and ATMS observations in the ECMWF system. *ECMWF Technical Memorandum*, **912**, https://doi.org/10.21957/d281dc221a.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull Am Meteorol Soc*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Bowler, N. E., 2006: Explicitly Accounting for Observation Error in Categorical Verification of Forecasts. *Mon Weather Rev*, **134**, 1600–1606, https://doi.org/10.1175/MWR3138.1.

——, 2008: Accounting for the effect of observation errors on verification of MOGREPS. *Meteorological Applications*, **15**, 199–205, https://doi.org/10.1002/met.64.

Brannan, A. L., and J. M. Chagnon, 2020: A Climatology of the Extratropical Flow Response to Recurving Atlantic Tropical Cyclones. *Mon Weather Rev*, **148**, 541–558, https://doi.org/10.1175/MWR-D-19-0216.1.

Brier, G. W., and R. A. Allen, 1951: Verification of Weather Forecasts. *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, T.F. Malone, Ed., American Meteorological Society, 841–848.

Broad, K., A. Leiserowitz, J. Weinkle, and M. Steketee, 2007: Misinterpretations of the "Cone of Uncertainty" in Florida during the 2004 Hurricane Season. *Bull Am Meteorol Soc*, **88**, 651–668, https://doi.org/10.1175/BAMS-88-5-651.

Brunet, G., and Coauthors, 2023: Advancing Weather and Climate Forecasting for Our Changing World. *Bull Am Meteorol Soc*, **104**, E909–E927, https://doi.org/10.1175/BAMS-D-21-0262.1.

Buizza, R., 2008a: Comparison of a 51-Member Low-Resolution (TL399L62) Ensemble with a 6-Member High-Resolution (TL799L91) Lagged-Forecast Ensemble. *Mon Weather Rev*, **136**, 3343–3362, https://doi.org/10.1175/2008MWR2430.1.

——, 2008b: The value of probabilistic prediction. *Atmospheric Science Letters*, **9**, 36–42, https://doi.org/10.1002/asl.170.

——, and D. Richardson, 2017: 25 years of ensemble forecasting at ECMWF. *ECMWF Newsletter*, **153**, 20–31, https://doi.org/10.21957/bv418o.

——, M. Leutbecher, and L. Isaksen, 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, **134**, 2051–2066, https://doi.org/10.1002/qj.346.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **134**, 959–971, https://doi.org/10.1002/QJ.268.

Cangialosi, J. P., 2022: National Hurricane Center forecast verification report: 2021 hurricane season. *NOAA/National Hurricane Center Report*, NOAA National Hurricane Center. Available at: https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf (Accessed: 20 August 2024).

——, E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent Progress in Tropical Cyclone Intensity Forecasting at the National Hurricane Center. *Weather Forecast*, **35**, 1913–1922, https://doi.org/10.1175/WAF-D-20-0059.1.

Casati, B., M. Dorninger, C. A. S. Coelho, E. E. Ebert, C. Marsigli, M. P. Mittermaier, and E. Gilleland, 2022: The 2020 International Verification Methods Workshop Online: Major Outcomes and Way Forward. *Bull Am Meteorol Soc*, **103**, E899–E910, https://doi.org/10.1175/BAMS-D-21-0126.1.

Cassou, C., 2008: Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, **455**, 523–527, https://doi.org/10.1038/nature07286.

Charlton-Perez, A. J., R. W. Aldridge, C. M. Grams, and R. Lee, 2019: Winter pressures on the UK health system dominated by the Greenland Blocking weather regime. *Weather Clim Extrem*, **25**, 100218, https://doi.org/10.1016/j.wace.2019.100218.

Chen, L., X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li, 2023: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *NPJ Clim Atmos Sci*, **6**, 190, https://doi.org/10.1038/s41612-023-00512-1.

Conroy, A., and Coauthors, 2023: Track forecast: Operational capability and new techniques - Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10). *Tropical Cyclone Research and Review*, **12**, 64–80, https://doi.org/10.1016/J.TCRR.2023.05.002.

Craig, G. C., M. Puh, C. Keil, K. Tempest, T. Necker, J. Ruiz, M. Weissmann, and T. Miyoshi, 2022: Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, **148**, 2325–2343, https://doi.org/10.1002/QJ.4305.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus A*, **39 A**, 474–491, https://doi.org/10.1111/j.1600-0870.1987.tb00322.x.

Davis, C. A., and L. F. Bosart, 2003: Baroclinically Induced Tropical Cyclogenesis. *Mon Weather Rev*, **131**, 2730–2747, https://doi.org/10.1175/1520-0493(2003)131<2730:BITC>2.0.CO;2.

——, and ——, 2004: The TT Problem: Forecasting the Tropical Transition of Cyclones. *Bull Am Meteorol Soc*, **85**, 1657–1662, https://doi.org/10.1175/BAMS-85-11-1657.

Day, J. J., G. Arduini, I. Sandu, L. Magnusson, A. Beljaars, G. Balsamo, M. Rodwell, and D. Richardson, 2020: Measuring the Impact of a New Snow Model Using Surface Energy Budget Process Relationships. *J Adv Model Earth Syst*, **12**, https://doi.org/10.1029/2020MS002144.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597, https://doi.org/10.1002/qj.828.

DelSole, T., L. Trenary, and M. K. Tippett, 2017: The Weighted-Average Lagged Ensemble. *J Adv Model Earth Syst*, **9**, 2739–2752, https://doi.org/10.1002/2017MS001128.

Dorninger, M., P. Friederichs, S. Wahl, M. P. Mittermaier, C. Marsigli, and B. G. Brown, 2018: Editorial: Forecast verification methods across time and space scales – Part I. *Meteorologische Zeitschrift*, **27**, 433–434, https://doi.org/10.1127/METZ/2018/0955.

Dunion, J. P., and Coauthors, 2023: Recommendations for improved tropical cyclone formation and position probabilistic Forecast products. *Tropical Cyclone Research and Review*, **12**, 241–258, https://doi.org/10.1016/J.TCRR.2023.11.003.

Dvorak, V. F., 1984: Tropical cyclone intensity analysis using satellite data. *NOAA technical report*, **11**, 45. Available at: https://repository.library.noaa.gov/view/noaa/19322 (Accessed: 16 May 2024).

Ebert, E., and Coauthors, 2013: Progress and challenges in forecast verification. *Meteorological Applications*, **20**, 130–139, https://doi.org/10.1002/met.1392.

Ebert, E., and Coauthors, 2018: The WMO Challenge to Develop and Demonstrate the Best New User-Oriented Forecast Verification Metric. *Meteorologische Zeitschrift*, **27**, 435–440, https://doi.org/10.1127/METZ/2018/0892.

ECMWF, 2015: The strength of a common goal. Strategy 2016-2025. Available at: https://www.ecmwf.int/sites/default/files/ECMWF_Strategy_2016-2025.pdf (Accessed: 9 February 2020).

Ehret, U., 2010: Convergence Index: a new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift*, **19**, 441–451, https://doi.org/10.1127/0941-2948/2010/0480.

Elsberry, R. L., and P. H. Dobos, 1990: Time Consistency of Track Prediction Aids for Western North Pacific Tropical Cyclones. *Mon Weather Rev*, **118**, 746–754, https://doi.org/10.1175/1520-0493(1990)118<0746:TCOTPA>2.0.CO;2.

Emanuel, K., 2022: Tropical Cyclone Seeds, Transition Probabilities, and Genesis. *J Clim*, **35**, 3557–3566, https://doi.org/10.1175/JCLI-D-21-0922.1.

Feldmann, K., D. S. Richardson, and T. Gneiting, 2019: Grid- Versus Station-Based Postprocessing of Ensemble Temperature Forecasts. *Geophys Res Lett*, **46**, 7744–7751, https://doi.org/10.1029/2019GL083189.

Feng, X., G. Y. Yang, K. I. Hodges, and J. Methven, 2023: Equatorial waves as useful precursors to tropical cyclone occurrence and intensification. *Nat Commun*, **14**, 1–11, https://doi.org/10.1038/s41467-023-36055-5.

Ferranti, L., S. Corti, and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, **141**, 916–924, https://doi.org/10.1002/QJ.2411.

——, L. Magnusson, F. Vitart, and D. S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, **144**, 1788–1802, https://doi.org/10.1002/qj.3341.

Ferro, C. A. T., 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1917–1923, https://doi.org/10.1002/qj.2270.

Ferro, C. A. T., 2017: Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2665–2676, https://doi.org/10.1002/qj.3115.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, **15**, 19–24, https://doi.org/10.1002/met.45.

Fowler, T. L., B. G. Brown, J. H. Gotway, and P. Kucera, 2015: Spare Change : Evaluating revised forecasts. *MAUSAM*, **66**, 635–644, https://doi.org/https://doi.org/10.54302/mausam.v66i3.572.

Frank, W. M., and G. S. Young, 2007: The Interannual Variability of Tropical Cyclones. *Mon Weather Rev*, **135**, 3587–3598, https://doi.org/10.1175/MWR3435.1.

Froude, L. S. R., L. Bengtsson, and K. I. Hodges, 2013: Atmospheric predictability revisited. *Tellus A: Dynamic Meteorology and Oceanography*, **65**, 19022, https://doi.org/10.3402/tellusa.v65i0.19022.

Gascón, E., D. Lavers, T. M. Hamill, D. S. Richardson, Z. B. Bouallègue, M. Leutbecher, and F. Pappenberger, 2019: Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe. *Quarterly Journal of the Royal Meteorological Society*, **145**, 3218–3235, https://doi.org/10.1002/qj.3615.

Gneiting, T., and A. E. Raftery, 2007: Strictly Proper Scoring Rules, Prediction, and Estimation. *J Am Stat Assoc*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Goerss, J. S., 2000: Tropical Cyclone Track Forecasts Using an Ensemble of Dynamical Models. *Mon Weather Rev*, **128**, 1187–1193, https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

Good, I. J., 1952: Rational Decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, **14**, 107–114, https://doi.org/10.1111/J.2517-6161.1952.TB00104.X.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017: Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nat Clim Chang*, **7**, 557–562, https://doi.org/10.1038/nclimate3338.

——, L. Magnusson, and E. Madonna, 2018: An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study. *Quarterly Journal of the Royal Meteorological Society*, **144**, 2577–2591, https://doi.org/10.1002/qj.3353.

Griffiths, D., M. Foley, I. Ioannou, and T. Leeuwenburg, 2019: Flip-Flop Index: Quantifying revision stability for fixed-event forecasts. *Meteorological Applications*, **26**, 30–35, https://doi.org/10.1002/met.1732.

Hagedorn, R., and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, **16**, 143–155, https://doi.org/10.1002/MET.92.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1814–1827, https://doi.org/10.1002/QJ.1895.

Haiden, T., D. Richardson, M. Janousek, Z. Ben Bouallegue, L. Ferranti, and F. Vitart, 2018: ECMWF introduces two additional headline scores. *ECMWF Newsletter*, **154**, 8. Available at: https://www.ecmwf.int/en/newsletter/154/news/ecmwf-introduces-two-additional-headline-scores (Accessed: 18 June 2024).

——, M. Janousek, F. Vitart, L. Ferranti, and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. *ECMWF Technical Memorandum*, **853**, https://doi.org/10.21957/mlvapkke.

——, ——, ——, Z. Ben-Bouallegue, L. Ferranti, and F. Prates, 2021: Evaluation of ECMWF forecasts, including the 2021 upgrade. *ECMWF Technical Memorandum*, **884**, https://doi.org/10.21957/90pgicjk4.

——, ——, ——, ——, ——, ——, and D. Richardson, 2022: Evaluation of ECMWF forecasts, including the 2021 upgrade. *ECMWF Technical Memorandum*, **902**, https://doi.org/10.21957/xqnu5o3p.

——, ——, ——, ——, and F. Prates, 2023: Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memorandum*, **911**, https://doi.org/10.21957/d47ba5263c.

Hakim, G. J., and S. Masanam, 2023: Dynamical Tests of a Deep-Learning Weather Prediction Model. *Artificial Intelligence for the Earth Systems*, https://doi.org/10.1175/aies-d-23-0090.1.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of Tropical cyclone genesis forecasts from global numerical models. *Weather Forecast*, **28**, 1423–1445, https://doi.org/10.1175/WAF-D-13-00008.1.

——, ——, ——, and ——, 2016: Verification of tropical cyclone genesis forecasts from global numerical models: Comparisons between the North Atlantic and Eastern North Pacific Basins. *Weather Forecast*, **31**, 947–955, https://doi.org/10.1175/WAF-D-15-0157.1.

——, R. E. Hart, H. E. Fuelberg, and J. H. Cossuth, 2017: The development and evaluation of a statistical-dynamical tropical cyclone genesis guidance tool. *Weather Forecast*, **32**, 27–46, https://doi.org/10.1175/WAF-D-16-0072.1.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bull Am Meteorol Soc*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Hannachi, A., D. M. Straus, C. L. E. Franzke, S. Corti, and T. Woollings, 2017: Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, **55**, 199–234, https://doi.org/10.1002/2015RG000509.

Heming, J. T., 2017: Tropical cyclone tracking and verification techniques for Met Office numerical weather prediction models. *Meteorological Applications*, **24**, 1–8, https://doi.org/10.1002/MET.1599.

Heming, J. T., and Coauthors, 2019: Review of Recent Progress in Tropical Cyclone Track Forecasting and Expression of Uncertainties. *Tropical Cyclone Research and Review*, **8**, 181–218, https://doi.org/10.1016/j.tcrr.2020.01.001.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049, https://doi.org/10.1002/QJ.3803.

Hewson, T., 2020: Use and Verification of ECMWF products in Member and Co-operating States (2019). *ECMWF Technical Memorandum*, **860**, https://doi.org/10.21957/80s471ib1.

——, 2021: Use and Verification of ECMWF products in Member and Co-operating States (2021). *ECMWF Technical Memorandum*, **885**, https://doi.org/10.21957/vp4z0x4yo.

Hewson, T. D., and H. A. Titley, 2010: Objective identification, typing and tracking of the complete life-cycles of cyclonic features at high spatial resolution. *Meteorological Applications*, **17**, 355–381, https://doi.org/10.1002/MET.204.

Hon, K. K., and Coauthors, 2023: Recent advances in operational tropical cyclone genesis forecast. *Tropical Cyclone Research and Review*, **12**, 323–340, https://doi.org/10.1016/J.TCRR.2023.12.001.

Hu, Y., L. Chen, Z. Wang, and H. Li, 2023: SwinVRNN: A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation. *J Adv Model Earth Syst*, **15**, e2022MS003211, https://doi.org/10.1029/2022MS003211.

Inverarity, G. W., and Coauthors, 2023: Met Office MOGREPS-G initialisation using an ensemble of hybrid four-dimensional ensemble variational (En-4DEnVar) data assimilations. *Quarterly Journal of the Royal Meteorological Society*, **149**, 1138–1164, https://doi.org/10.1002/QJ.4431.

Janjić, T., and Coauthors, 2018: On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **144**, 1257–1278, https://doi.org/10.1002/QJ.3130.

Jewson, S., S. Scher, and G. Messori, 2021: Decide Now or Wait for the Next Forecast? Testing a Decision Framework Using Real Forecasts and Observations. *Mon Weather Rev*, **149**, 1637–1650, https://doi.org/10.1175/MWR-D-20-0392.1.

——, ——, and ——, 2022: Communicating Properties of Changes in Lagged Weather Forecasts. *Weather Forecast*, **37**, 125–142, https://doi.org/10.1175/WAF-D-21-0086.1.

Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, **135**, 777–794, https://doi.org/10.1002/QJ.383.

Jolliffe, I. T., and D. B. Stephenson, 2011: Introduction. *Forecast Verification*, 2nd edition, I.T. Jolliffe and D.B. Stephenson, Eds., John Wiley & Sons, Ltd, 1–9. Available at: https://doi.org/https://doi.org/10.1002/9781119960003.ch1.

Jones, S. C., and Coauthors, 2003: The Extratropical Transition of Tropical Cyclones: Forecast Challenges, Current Understanding, and Future Directions. *Weather Forecast*, **18**, 1052–1092, https://doi.org/10.1175/1520-0434(2003)018<1052:TETOTC>2.0.CO;2.

Jung, T., 2011: Diagnosing remote origins of forecast error: relaxation versus 4D-Var data-assimilation experiments. *Quarterly Journal of the Royal Meteorological Society*, **137**, 598–606, https://doi.org/10.1002/QJ.781.

Jung, T., M. J. Miller, and T. N. Palmer, 2010: Diagnosing the Origin of Extended-Range Forecast Errors. *Mon Weather Rev*, **138**, 2434–2446, https://doi.org/10.1175/2010MWR3255.1.

Kawabata, Y., and M. Yamaguchi, 2020: Probability Ellipse for Tropical Cyclone Track Forecasts with Multiple Ensembles. *Journal of the Meteorological Society of Japan. Ser. II*, **98**, 821–833, https://doi.org/10.2151/jmsj.2020-042.

Keisler, R., 2022: Forecasting Global Weather with Graph Neural Networks. *arXiv preprint*, **arXiv:2202.07575**, https://doi.org/https://doi.org/10.48550/arXiv.2202.07575.

Keller, J. H., and Coauthors, 2019: The extratropical transition of tropical cyclones. Part II: Interaction with the midlatitude flow, downstream impacts, and implications for predictability. *Mon Weather Rev*, **147**, 1077–1106, https://doi.org/10.1175/MWR-D-17-0329.1.

Klotzbach, P. J., and Coauthors, 2022: A Hyperactive End to the Atlantic Hurricane Season October–November 2020. *Bull Am Meteorol Soc*, **103**, E110–E128, https://doi.org/10.1175/BAMS-D-20-0312.1.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data. *Bull Am Meteorol Soc*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

Knapp, K. R., H. J. Diamond, M. C. Kossin, and C. J. Schreck, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4. https://doi.org/10.25921/82ty-9e16.

Komaromi, W. A., and S. J. Majumdar, 2014: Ensemble-Based Error and Predictability Metrics Associated with Tropical Cyclogenesis. Part I: Basinwide Perspective. *Mon Weather Rev*, **142**, 2879–2898, https://doi.org/10.1175/MWR-D-13-00370.1.

——, and ——, 2015: Ensemble-Based Error and Predictability Metrics Associated with Tropical Cyclogenesis. Part II: Wave-Relative Framework. *Mon Weather Rev*, **143**, 1665–1686, https://doi.org/10.1175/MWR-D-14-00286.1.

Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: a study with a 10 240-member ensemble Kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Process Geophys*, **26**, 211–225, https://doi.org/10.5194/npg-26-211-2019.

Korhonen, N., O. Hyvärinen, M. Kämäraïnen, D. S. Richardson, H. Järvinen, and H. Gregow, 2020: Adding value to extended-range forecasts in northern Europe by statistical post-processing using stratospheric observations. *Atmos Chem Phys*, **20**, https://doi.org/10.5194/acp-20-8441-2020.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society*, **129**, 3037–3057, https://doi.org/10.1256/qj.02.152.

Lam, R., and Coauthors, 2023a: GraphCast: Learning skillful medium-range global weather forecasting. *arXiv preprint*, **arXiv:2212.12794v2**, https://doi.org/https://doi.org/10.48550/arXiv.2212.12794.

——, and Coauthors, 2023b: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1422, https://doi.org/10.1126/SCIENCE.ADI2336/SUPPL_FILE/SCIENCE.ADI2336_SM.PDF.

Landsea, C. W., 1993: A Climatology of Intense (or Major) Atlantic Hurricanes. *Mon Weather Rev*, **121**, 1703–1713, https://doi.org/10.1175/1520-0493(1993)121<1703:ACOIMA>2.0.CO;2.

——, and J. L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. *Mon Weather Rev*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

——, and J. P. Cangialosi, 2018: Have We Reached the Limits of Predictability for Tropical Cyclone Track Forecasting? *Bull Am Meteorol Soc*, **99**, 2237–2243, https://doi.org/10.1175/BAMS-D-17-0136.1.

Lang, S., and Coauthors, 2024: AIFS -- ECMWF's data-driven forecasting system. *arXiv preprint*, **arXiv:2406.01465**, https://doi.org/https://doi.org/10.48550/arXiv.2406.01465.

Lavers, D. A., S. Harrigan, E. Andersson, D. S. Richardson, C. Prudhomme, and F. Pappenberger, 2019: A vision for improving global flood forecasting. *Environmental Research Letters*, **14**, 121002, https://doi.org/10.1088/1748-9326/AB52B2.

——, and Coauthors, 2020: A Vision for Hydrological Prediction. *Atmosphere*, **11**, 237, https://doi.org/10.3390/atmos11030237.

Lawton, Q. A., S. J. Majumdar, K. Dotterer, C. Thorncroft, and C. J. Schreck, 2022: The Influence of Convectively Coupled Kelvin Waves on African Easterly Waves in a Wave-Following Framework. *Mon Weather Rev*, **150**, 2055–2072, https://doi.org/10.1175/MWR-D-21-0321.1.

Leonardo, N. M., and B. A. Colle, 2017: Verification of multimodel ensemble forecasts of North Atlantic tropical cyclones. *Weather Forecast*, **32**, 2083–2101, https://doi.org/10.1175/WAF-D-17-0058.1.

——, and ——, 2020: An Investigation of Large Along-Track Errors in Extratropical Transitioning North Atlantic Tropical Cyclones in the ECMWF Ensemble. *Mon Weather Rev*, **148**, 457–476, https://doi.org/10.1175/MWR-D-19-0044.1.

——, and ——, 2021: An Investigation of Large Cross-Track Errors in North Atlantic Tropical Cyclones in the GEFS and ECMWF Ensembles. *Mon Weather Rev*, **149**, 395–417, https://doi.org/10.1175/MWR-D-20-0035.1.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, **145**, 107–128, https://doi.org/10.1002/QJ.3387.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J Comput Phys*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Leutbecher, M., and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2315–2339, https://doi.org/10.1002/qj.3094.

Li, W., Z. Wang, and M. S. Peng, 2016: Evaluating Tropical Cyclone Forecasts from the NCEP Global Ensemble Forecasting System (GEFS) Reforecast Version 2. *Weather Forecast*, **31**, 895–916, https://doi.org/10.1175/WAF-D-15-0176.1.

Liang, M., J. C. L. Chan, J. Xu, and M. Yamaguchi, 2021: Numerical prediction of tropical cyclogenesis part I: Evaluation of model performance. *Quarterly Journal of the Royal Meteorological Society*, **147**, 1626–1641, https://doi.org/10.1002/qj.3987.

Lillo, S. P., and D. B. Parsons, 2017: Investigating the dynamics of error growth in ECMWF medium-range forecast busts. *Quarterly Journal of the Royal Meteorological Society*, **143**, 1211–1226, https://doi.org/10.1002/qj.2938.

Lorenz, E. N., 1963: Deterministic Nonperiodic Flow. *J Atmos Sci*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2129–2142, https://doi.org/10.1002/qj.3072.

Magnusson, L., 2023: Exploring machine-learning forecasts of extreme weather. *ECMWF Newsletter*, **176**, 8–9. Available at: https://www.ecmwf.int/en/newsletter/176/news/exploring-machine-learning-forecasts-extreme-weather (Accessed: 17 July 2024).

——, and E. Källén, 2013: Factors influencing skill improvements in the ECMWF forecasting system. *Mon Weather Rev*, **141**, 3142–3153, https://doi.org/10.1175/MWR-D-12-00318.1.

Magnusson, L., and Coauthors, 2019: ECMWF Activities for Improved Hurricane Forecasts. *Bull Am Meteorol Soc*, **100**, 445–458, https://doi.org/10.1175/BAMS-D-18-0044.1.

Magnusson, L., and Coauthors, 2021: Tropical cyclone activities at ECMWF. *ECMWF Technical Memorandum*, **888**, https://doi.org/10.21957/zzxzzygwv.

Majumdar, S. J., and P. M. Finocchio, 2010: On the Ability of Global Ensemble Prediction Systems to Predict Tropical Cyclone Track Probabilities. *Weather Forecast*, **25**, 659–680, https://doi.org/10.1175/2009WAF2222327.1.

——, and R. D. Torn, 2014: Probabilistic Verification of Global and Mesoscale Ensemble Forecasts of Tropical Cyclogenesis. *Weather Forecast*, **29**, 1181–1198, https://doi.org/10.1175/WAF-D-14-00028.1.

Marchok, T., 2021: Important Factors in the Tracking of Tropical Cyclones in Operational Models. *J Appl Meteorol Climatol*, **60**, 1265–1284, https://doi.org/10.1175/JAMC-D-20-0175.1.

Marsigli, C., and Coauthors, 2021: Review article: Observations for high-impact weather and their use in verification. *Natural Hazards and Earth System Sciences*, **21**, 1297–1312, https://doi.org/10.5194/NHESS-21-1297-2021.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291–303.

McLay, J. G., 2008: Markov Chain Modeling of Sequences of Lagged NWP Ensemble Probability Forecasts: An Exploration of Model Properties and Decision Support Applications. *Mon Weather Rev*, **136**, 3655–3670, https://doi.org/10.1175/2008MWR2376.1.

——, 2011: Diagnosing the Relative Impact of “Sneaks,” “Phantoms,” and Volatility in Sequences of Lagged Ensemble Probability Forecasts with a Simple Dynamic Decision Model. *Mon Weather Rev*, **139**, 387–402, https://doi.org/10.1175/2010MWR3449.1.

McTaggart-Cowan, R., G. D. Deane, L. F. Bosart, C. A. Davis, and T. J. Galarneau, 2008: Climatology of Tropical Cyclogenesis in the North Atlantic (1948–2004). *Mon Weather Rev*, **136**, 1284–1304, https://doi.org/10.1175/2007MWR2245.1.

——, T. J. Galarneau, L. F. Bosart, R. W. Moore, and O. Martius, 2013: A Global Climatology of Baroclinically Influenced Tropical Cyclogenesis. *Mon Weather Rev*, **141**, 1963–1989, https://doi.org/10.1175/MWR-D-12-00186.1.

Met Office, 2019: Parallel Suite 43 release notes - Met Office. https://www.metoffice.gov.uk/services/data/met-office-data-for-reuse/ps43_ftp (Accessed August 20, 2024).

Mittermaier, M., 2012: A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1794–1807, https://doi.org/10.1002/QJ.1918.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society*, **133**, 1487–1500, https://doi.org/10.1002/QJ.135.

Mittermaier, M. P., 2008: Introducing uncertainty of radar-rainfall estimates to the verification of mesoscale model precipitation forecasts. *Natural Hazards and Earth System Sciences*, **8**, 445–460, https://doi.org/10.5194/nhess-8-445-2008.

Mittermaier, M. P., 2014: A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites. *Weather Forecast*, **29**, 185–204, https://doi.org/10.1175/WAF-D-12-00075.1.

——, and D. B. Stephenson, 2015: Inherent Bounds on Forecast Accuracy due to Observation Uncertainty Caused by Temporal Sampling. *Mon Weather Rev*, **143**, 4236–4243, https://doi.org/10.1175/MWR-D-15-0173.1.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, **122**, 73–119, https://doi.org/10.1002/QJ.49712252905.

Neal, R., D. Fereday, R. Crocker, and R. E. Comer, 2016: A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, **23**, 389–400, https://doi.org/10.1002/MET.1563.

——, J. Robbins, R. Crocker, D. Cox, K. Fenwick, J. Millard, and J. Kelly, 2024: A seamless blended multi-model ensemble approach to probabilistic medium-range weather pattern forecasts over the UK. *Meteorological Applications*, **31**, e2179, https://doi.org/10.1002/MET.2179.

Núñez Ocasio, K. M., J. L. Evans, and G. S. Young, 2020: A Wave-Relative Framework Analysis of AEW–MCS Interactions Leading to Tropical Cyclogenesis. *Mon Weather Rev*, **148**, 4657–4671, https://doi.org/10.1175/MWR-D-20-0152.1.

——, A. Brammer, J. L. Evans, G. S. Young, and Z. L. Moon, 2021: Favorable Monsoon Environment over Eastern Africa for Subsequent Tropical Cyclogenesis of African Easterly Waves. *J Atmos Sci*, **78**, 2911–2925, https://doi.org/10.1175/JAS-D-20-0339.1.

Olander, T. L., and C. S. Velden, 2007: The Advanced Dvorak Technique: Continued Development of an Objective Scheme to Estimate Tropical Cyclone Intensity Using Geostationary Infrared Satellite Imagery. *Weather Forecast*, **22**, 287–298, https://doi.org/10.1175/WAF975.1.

Papin, P. P., L. F. Bosart, and R. D. Torn, 2017: A Climatology of Central American Gyres. *Mon Weather Rev*, **145**, 1983–2000, https://doi.org/10.1175/MWR-D-16-0411.1.

Pappenberger, F., K. Bogner, F. Wetterhall, Y. He, H. L. Cloke, and J. Thielen, 2011a: Forecast convergence score: a forecaster's approach to analysing hydro-meteorological forecast systems. *Advances in Geosciences*, **29**, 27–32, https://doi.org/10.5194/adgeo-29-27-2011.

——, H. L. Cloke, A. Persson, and D. Demeritt, 2011b: HESS Opinions "On forecast (in)consistency in a hydro-meteorological chain: curse or blessing?" *Hydrol Earth Syst Sci*, **15**, 2391–2400, https://doi.org/10.5194/hess-15-2391-2011.

Pathak, J., and Coauthors, 2022: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *arXiv preprint*, **arXiv:2202.11214v1**, https://doi.org/10.48550/arxiv.2202.11214.

Pinson, P., and R. Hagedorn, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications*, **19**, 484–500, https://doi.org/10.1002/MET.283.

Rajasree, V. P. M., and Coauthors, 2023: Tropical cyclogenesis: Controlling factors and physical mechanisms. *Tropical Cyclone Research and Review*, **12**, 165–181, https://doi.org/10.1016/J.TCRR.2023.09.004.

Rappaport, E. N., and Coauthors, 2009: Advances and Challenges at the National Hurricane Center. *Weather Forecast*, **24**, 395–419, https://doi.org/10.1175/2008WAF2222128.1.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *J Adv Model Earth Syst*, **12**, e2020MS002203, https://doi.org/10.1029/2020MS002203.

——, and Coauthors, 2024: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *J Adv Model Earth Syst*, **16**, e2023MS004019, https://doi.org/10.1029/2023MS004019.

Regnier, E., and P. A. Harr, 2006: A Dynamic Decision Model Applied to Hurricane Landfall. *Weather Forecast*, **21**, 764–780, https://doi.org/10.1175/WAF958.1.

Richardson, D., R. Neal, R. Dankers, K. Mylne, R. Cowling, H. Clements, and J. Millard, 2020a: Linking weather patterns to regional extreme precipitation for highlighting potential flood events in medium- to long-range forecasts. *Meteorological Applications*, **27**, e1931, https://doi.org/10.1002/met.1931.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, **127**, 2473–2489, https://doi.org/10.1002/QJ.49712757715.

Richardson, D. S., 2011: Economic Value and Skill. *Forecast Verification*, 2nd edition, I.T. Jolliffe and D.B. Stephenson, Eds., John Wiley & Sons, Ltd, 167–184.

Richardson, D. S., H. L. Cloke, and F. Pappenberger, 2020b: Evaluation of the Consistency of ECMWF Ensemble Forecasts. *Geophys Res Lett*, **47**, e2020GL087934, https://doi.org/10.1029/2020GL087934.

——, ——, J. A. Methven, and F. Pappenberger, 2024: Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks. *Weather Forecast*, **39**, 203–215, https://doi.org/10.1175/WAF-D-23-0113.1.

Roberts, N., and Coauthors, 2023: IMPROVER: The New Probabilistic Postprocessing System at the Met Office. *Bull Am Meteorol Soc*, **104**, E680–E697, https://doi.org/10.1175/BAMS-D-21-0273.1.

Rodwell, M., and Coauthors, 2021: IFS upgrade provides more skilful ensemble forecasts. *ECMWF Newsletter*, **168**, 18–23, https://doi.org/10.21957/m830hnz27r.

Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **136**, 1344–1363, https://doi.org/10.1002/qj.656.

Rodwell, M. J., and Coauthors, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bull Am Meteorol Soc*, **94**, 1393–1405, https://doi.org/10.1175/BAMS-D-12-00099.1.

Rodwell, M. J., S. T. K. Lang, N. B. Ingleby, N. Bormann, E. Hólm, F. Rabier, D. S. Richardson, and M. Yamaguchi, 2016: Reliability in ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **142**, 443–454, https://doi.org/10.1002/qj.2663.

Rodwell, M. J., D. S. Richardson, D. B. Parsons, H. Wernli, M. J. Rodwell, D. S. Richardson, D. B. Parsons, and H. Wernli, 2018: Flow-Dependent Reliability: A Path to More Skillful Ensemble Forecasts. *Bull Am Meteorol Soc*, **99**, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1.

——, J. Hammond, S. Thornton, and D. S. Richardson, 2020: User decisions, and how these could guide developments in probabilistic forecasting. *Quarterly Journal of the Royal Meteorological Society*, **146**, 3266–3284, https://doi.org/10.1002/qj.3845.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Weather Forecast*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Roulston, M. S., and L. a. Smith, 2002: Evaluating Probabilistic Forecasts Using Information Theory. *Mon Weather Rev*, **130**, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Russell, J. O., A. Aiyyer, J. D. White, and W. Hannah, 2017: Revisiting the connection between African Easterly Waves and Atlantic tropical cyclogenesis. *Geophys Res Lett*, **44**, 587–595, https://doi.org/10.1002/2016GL071236.

Ruth, D. P., B. Glahn, V. Dagostaro, K. Gilbert, D. P. Ruth, B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The Performance of MOS in the Digital Age. *Weather Forecast*, **24**, 504–519, https://doi.org/10.1175/2008WAF2222158.1.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon Weather Rev*, **132**, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Sánchez, C., J. Methven, S. Gray, and M. Cullen, 2020: Linking rapid forecast error growth to diabatic processes. *Quarterly Journal of the Royal Meteorological Society*, **146**, 3548–3569, https://doi.org/10.1002/QJ.3861.

Schreck, C. J., K. R. Knapp, and J. P. Kossin, 2014: The Impact of Best Track Discrepancies on Global Tropical Cyclone Climatologies using IBTrACS. *Mon Weather Rev*, **142**, 3881–3899, https://doi.org/10.1175/MWR-D-14-00021.1.

Selz, T., and G. C. Craig, 2023: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophys Res Lett*, **50**, e2023GL105747, https://doi.org/10.1029/2023GL105747.

Shanker, G., A. Sarkar, A. Mamgain, S. K. Prasad, R. Bhatla, and A. K. Mitra, 2022: Contribution of lagged members to the performance of a global ensemble prediction system. *Atmos Res*, **280**, 106451, https://doi.org/10.1016/J.ATMOSRES.2022.106451.

Sherman-Morris, K., H. Lussenden, A. Kent, and C. Macdonald, 2018: Perceptions about Social Science among NWS Warning Coordination Meteorologists. *Weather, Climate, and Society*, **10**, 597–612, https://doi.org/10.1175/WCAS-D-17-0079.1.

Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **128**, 647–677, https://doi.org/10.1256/003590002321042135.

Straus, D. M., F. Molteni, and S. Corti, 2017: Atmospheric Regimes: The Link between Weather and the Large-Scale Circulation. *Nonlinear and Stochastic Climate Dynamics*, C.L.E. Franzke and T.J.E. O'Kane, Eds., Cambridge University Press, 105–135.

Swinbank, R., and Coauthors, 2016: The TIGGE Project and Its Achievements. *Bull Am Meteorol Soc*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Tang, B. H., and Coauthors, 2020: Recent advances in research on tropical cyclogenesis. *Tropical Cyclone Research and Review*, **9**, 87–105, https://doi.org/10.1016/J.TCRR.2020.04.004.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using Proper Divergence Functions to Evaluate Climate Models. *SIAM/ASA Journal on Uncertainty Quantification*, **1**, 522–534, https://doi.org/10.1137/130907550.

Titley, H. A., M. Yamaguchi, and L. Magnusson, 2019: Current and potential use of ensemble forecasts in operational TC forecasting: results from a global forecaster survey. *Tropical Cyclone Research and Review*, **8**, 166–180, https://doi.org/10.1016/J.TCRR.2019.10.005.

Titley, H. A., R. L. Bowyer, and H. L. Cloke, 2020: A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Quarterly Journal of the Royal Meteorological Society*, **146**, 531–545, https://doi.org/10.1002/qj.3712.

Torn, R. D., and G. J. Hakim, 2009: Initial condition sensitivity of Western Pacific extratropical transitions determined using ensemble-based sensitivity analysis. *Mon Weather Rev*, **137**, 3388–3406, https://doi.org/10.1175/2009MWR2879.1.

——, and C. Snyder, 2012: Uncertainty of Tropical Cyclone Best-Track Information. *Weather Forecast*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

Toth, Z., and E. Kalnay, 1993: Ensemble Forecasting at NMC: The Generation of Perturbations. *Bull Am Meteorol Soc*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Tsonevsky, I., C. A. Doswell III, and H. E. Brooks, 2018: Early Warnings of Severe Convection Using the ECMWF Extreme Forecast Index. *Weather Forecast*, **33**, 857–871, https://doi.org/10.1175/WAF-D-18-0030.1.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research: Atmospheres*, **106**, 11775–11784, https://doi.org/10.1029/2001JD900066.

Vannitsem, S., D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. S. Vannitsem, D.S. Wilks, and J.W. Messner, Eds. Elsevier, https://doi.org/10.1016/C2016-0-03244-8.

——, and Coauthors, 2021: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bull Am Meteorol Soc*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Velden, C., and Coauthors, 2006: The Dvorak Tropical Cyclone Intensity Estimation Technique: A Satellite-Based Method that Has Endured for over 30 Years. *Bull Am Meteorol Soc*, **87**, 1195–1210, https://doi.org/10.1175/BAMS-87-9-1195.

Vitart, F., and Y. Takaya, 2021: Lagged ensembles in sub-seasonal predictions. *Quarterly Journal of the Royal Meteorological Society*, **147**, 3227–3242, https://doi.org/10.1002/QJ.4125.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull Am Meteorol Soc*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Vitart, F., and Coauthors, 2019a: Extended-range prediction. *ECMWF Technical Memorandum*, **854**, https://doi.org/10.21957/pdivp3t9m.

——, G. Balsamo, J.-R. Bidlot, S. Lang, I. Tsonevsky, D. Richardson, and M. Alonso-Balmaseda, 2019b: Use of ERA5 to Initialize Ensemble Re-forecasts. *ECMWF Technical Memorandum*, **841**, https://doi.org/10.21957/w8i57wuz6.

——, M. Alonso-Balmaseda, L. Ferranti, and M. Fuentes, 2022: The next extended-range configuration for IFS Cycle 48r1. *ECMWF Newsletter*, **173**, 21–26, https://doi.org/10.21957/fv6k37c49h.

Walters, D., and Coauthors, 2019: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *Geosci Model Dev*, **12**, 1909–1963, https://doi.org/10.5194/GMD-12-1909-2019.

Wang, Z., W. Li, M. S. Peng, X. Jiang, R. McTaggart-Cowan, and C. A. Davis, 2018: Predictive Skill and Predictability of North Atlantic Tropical Cyclogenesis in Different Synoptic Flow Regimes. *J Atmos Sci*, **75**, 361–378, https://doi.org/10.1175/JAS-D-17-0094.1.

Wilks, D. S., 2020: *Statistical Methods in the Atmospheric Sciences, Fourth Edition*. Elsevier, 1–818 pp. https://doi.org/10.1016/C2017-0-03921-6.

WMO, 2013: Verification Methods for Tropical Cyclone Forecasts. WWRP 2013-7. Available at: https://filecloud.wmo.int/share/s/Fotf-7H1RLyK1lkSC_onBA (Accessed: 17 May 2024)

——, 2017: Global Guide to Tropical Cyclone Forecasting. **WMO-No.1194**. Available at: https://library.wmo.int/records/item/53583-global-guide-to-tropical-cyclone-forecasting (Accessed: 17 May 2024)

——, 2021: Guidelines on Ensemble Prediction System Postprocessing. **WMO-No. 1254**. Available at: https://library.wmo.int/records/item/57510-guidelines-on-ensemble-prediction-system-postprocessing#.YOW-x-gzYuV (Accessed: 9 July 2024).

——, 2022: *Early Warnings for All Executive Action Plan 2023-27*. 56 pp. https://library.wmo.int/idurl/4/58209 (Accessed July 8, 2024).

——, 2023: Manual on the WMO Integrated Processing and Prediction System (WMO-No. 485): Annex IV to the WMO Technical Regulations. *WMO*, 166 pp. Available at: https://library.wmo.int/idurl/4/35703 (Accessed: 17 June 2024).

Woollings, T., A. Hannachi, and B. Hoskins, 2010: Variability of the North Atlantic eddy-driven jet stream. *Quarterly Journal of the Royal Meteorological Society*, **136**, 856–868, https://doi.org/10.1002/QJ.625.

Yamaguchi, M., and N. Koide, 2017: Tropical Cyclone Genesis Guidance Using the Early Stage Dvorak Analysis and Global Ensembles. *Weather Forecast*, **32**, 2133–2141, https://doi.org/10.1175/WAF-D-17-0056.1.

——, R. Sakai, M. Kyoda, T. Komori, and T. Kadowaki, 2009: Typhoon ensemble prediction system developed at the Japan meteorological agency. *Mon Weather Rev*, **137**, 2592–2604, https://doi.org/10.1175/2009MWR2697.1.

——, T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quarterly Journal of the Royal Meteorological Society*, **138**, 2019–2029, https://doi.org/10.1002/QJ.1937.

——, F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T. Nakazawa, 2015: Global Distribution of the Skill of Tropical Cyclone Activity Forecasts on Short- to Medium-Range Time Scales. *Weather Forecast*, **30**, 1695–1709, https://doi.org/10.1175/WAF-D-14-00136.1.

——, S. T. K. Lang, M. Leutbecher, M. J. Rodwell, G. Radnoti, and N. Bormann, 2016: Observation-based evaluation of ensemble reliability. *Quarterly Journal of the Royal Meteorological Society*, **142**, 506–514, https://doi.org/10.1002/qj.2675.

——, J. Ishida, H. Sato, and M. Nakagawa, 2017: WGNE intercomparison of tropical cyclone forecasts by operational NWP models: A quarter century and beyond. *Bull Am Meteorol Soc*, **98**, 2337–2349, https://doi.org/10.1175/BAMS-D-16-0133.1.

Yiou, P., and M. Nogaj, 2004: Extreme climatic events and weather regimes over the North Atlantic: When and where? *Geophys Res Lett*, **31**, L07202, https://doi.org/10.1029/2003GL019119.

Zhang, F., and Coauthors, 2019: What Is the Predictability Limit of Midlatitude Weather? *J Atmos Sci*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhang, X., J. Fang, and Z. Yu, 2022: The Forecast Skill of Tropical Cyclone Genesis in Two Global Ensembles. *Weather Forecast*, **38**, 83–97, https://doi.org/10.1175/WAF-D-22-0145.1.

Zhou, X., and Coauthors, 2022: The Development of the NCEP Global Ensemble Forecast System Version 12. *Weather Forecast*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zsoter, E., R. Buizza, and D. Richardson, 2009: “Jumpiness” of the ECMWF and Met Office EPS Control and Ensemble-Mean Forecasts. *Mon Weather Rev*, **137**, 3823–3836, https://doi.org/10.1175/2009MWR2960.1.

——, F. Pappenberger, and D. Richardson, 2015: Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index. *Meteorological Applications*, **22**, 236–247, https://doi.org/10.1002/met.1447.

Appendices

These appendices contain the typeset versions of each of the published chapters presented in this thesis, together with additional important publications during this PhD. All author contribution statements (provided in the respective chapters) have been approved by Professor Hannah Cloke, supervisor.

Hannah L. Cloke

A1. Published Article: Evaluation of the consistency of ECMWF ensemble forecasts

This appendix contains the published version of Chapter 3 of this thesis, with the following reference:

Richardson, D.S., Cloke, H.L. and Pappenberger, F. (2020) 'Evaluation of the Consistency of ECMWF Ensemble Forecasts', *Geophysical Research Letters*, 47(11), p. e2020GL087934. Available at: https://doi.org/10.1029/2020GL087934.

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL087934

Evaluation of the Consistency of ECMWF Ensemble Forecasts

David S. Richardson[1,2], Hannah L. Cloke[2,3,4], and Florian Pappenberger[1]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK, [2]Department of Geography and Environmental Science, University of Reading, Reading, UK, [3]Department of Meteorology, University of Reading, Reading, UK, [4]Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Key Points:
- A new divergence index is introduced to measure inconsistency (jumpiness) in a sequence of ensemble forecasts
- The ECMWF ensemble has occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7–9 days lead
- To understand the causes of jumpiness it is important to consider the time evolution of each ensemble (e.g., using phase-space trajectories)

Supporting Information:
- Supporting Information S1

Correspondence to:
D. S. Richardson,
david.richardson@ecmwf.int

Citation:
Richardson, D. S., Cloke, H. L., & Pappenberger, F. (2020). Evaluation of the consistency of ECMWF ensemble forecasts. *Geophysical Research Letters*, *46*, e2020GL087934. https://doi.org/10.1029/2020GL087934

Received 12 MAR 2020
Accepted 29 APR 2020
Accepted article online 3 MAY 2020

Abstract An expected benefit of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts. Inconsistent (jumpy) forecasts can cause users to lose confidence in the forecasting system. We present a first systematic, objective evaluation of the consistency of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble using a measure of forecast divergence that takes account of the full ensemble distribution. Focusing on forecasts of the North Atlantic Oscillation and European Blocking regimes up to 2 weeks ahead, we identify occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7–9 days lead. However, care is needed in the interpretation of ensemble jumpiness. An apparent clear flip-flop in a single index may hide a more complex predictability issue which may be better understood by examining the ensemble evolution in phase space.

Plain Language Summary Ensemble forecasts show the range of weather scenarios that can occur, allowing users to make appropriate risk-based decisions. An ensemble forecast made 2 weeks in advance will show a range of possible outcomes. New observations included in subsequent forecasts will eliminate some of these scenarios, and the forecast will become more certain. Occasionally, a new forecast seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the earlier forecast. Such inconsistencies can cause users to lose confidence in the forecasting system. We present a new method to assess the consistency of ensemble forecasts of large-scale weather patterns over Europe made by the European Centre for Medium-Range Weather Forecasts. We show that a careful analysis of each forecast is needed to understand how and why these jumps occur. Understanding and reducing the occurrence of inconsistent ensemble forecasts will increase user confidence and improve decision making.

1. Introduction

The chaotic nature of the atmosphere means that numerical weather prediction (NWP) forecasts are sensitive to small changes in their initial conditions. Operational NWP centers address this by running a number of forecasts from similar starting conditions. The resulting ensemble of forecasts shows the range of future atmospheric states consistent with the known uncertainties in the initial conditions (Leutbecher & Palmer, 2008; Swinbank et al., 2016). One of the expected benefits of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts (Buizza, 2008; Zsoter et al., 2009). Inconsistent (or jumpy) forecasts are difficult to handle and can cause users to lose confidence in the forecasting system (Hewson, 2020; Pappenberger et al., 2011). However, this aspect of ensemble forecasts has received little attention in the literature.

The inconsistency between successive ensemble-mean (EM) forecasts valid for the same time was investigated by Zsoter et al. (2009). They define an inconsistency index as the difference between two fields over a given area, divided by their average standard deviation over the area. They consider cases of large jumps (inconsistency greater than a chosen threshold) and focus on sequences of jumps of opposite sign (flip-flops). Using this methodology, they showed that EM forecasts are more consistent than the corresponding ensemble control forecasts. Zsoter et al. (2009) conclude by noting that to further investigate the benefit of ensemble forecasts compared to single forecast, an index for probabilistic forecasts will need to be developed. Forecast consistency has also been considered in the context of model output statistics

Geophysical Research Letters 10.1029/2020GL087934

(Ruth et al., 2009), comparing automated with manual forecasts (Griffiths et al., 2019), comparing deterministic rainfall forecasts from different models (Ehret, 2010) and in forecasts of river flow (Pappenberger et al., 2011).

None of the above methods are directly applicable to assess the consistency of a sequence of ensemble forecasts taking account of the full ensemble distribution. In this work, for the first time, we investigate the consistency of the European Centre for Medium-Range Forecasts (ECMWF) ensemble (ENS) using a measure of forecast divergence that accounts for all aspects of the ensemble empirical distribution.

We focus on two key characteristics of the large-scale flow over the European-Atlantic region: the North Atlantic Oscillation (NAO) and Scandinavian Blocking (BLO). Predicting transitions between such large-scale weather regimes 2 weeks or more ahead is a significant scientific challenge and at the frontier of NWP (ECMWF, 2015). These transitions are associated with large-scale changes in temperature and winds over Europe (Ferranti et al., 2018; Yiou & Nogaj, 2004) and hence have significant societal impacts, for example, on health (Charlton-Perez et al., 2019) and on energy production (Grams et al., 2017). We consider the full 15-day forecast range of the operational ENS.

The data and indices used are introduced in section 2. Methods, including the definition of the forecast divergence, are described in section 3. We then evaluate the inconsistency of the ENS forecasts for NAO and BLO and compare the jumpiness of the ENS with that of the EM and control forecasts in section 4. We present concluding remarks and avenues for future work in section 5.

2. Data

We study the time evolution of the NAO and BLO patterns that are associated with high-impact temperature anomalies over Europe (Ferranti et al., 2018). Following the approach of Ferranti et al. (2018), we use a two-dimensional phase space based on the two leading Empirical Orthogonal Functions (EOFs) of mid-tropospheric flow computed over the Euro-Atlantic region. The EOFs are computed using daily geopotential height at 500 hPa computed for the Euro-Atlantic region (30°N to 88.5°N, 80°W to 40°E) from 29 years of extended winter periods (October to March) of ECMWF ERA-Interim data (Berrisford et al., 2011; Dee et al., 2011). For the EOF computation, a 5-day running mean was used, and the mean seasonal cycle was removed. The first EOF represents the positive phase of the NAO (NAO+): a negative anomaly over Iceland and positive anomaly to the south (Cassou, 2008). The second EOF has a positive anomaly (high pressure) over Scandinavia, and a low to the east over the Atlantic, representing the flow pattern associated with blocking events over northern Europe (Ferranti et al., 2015). We refer to Ferranti et al. (2018) for further details.

We study the consistency of the operational ECMWF ensemble forecasts (ENS; Ben Bouallègue et al., 2019; Buizza & Richardson, 2017) of the large-scale flow over the North Atlantic Europe region for DJF 2016–2019, that is, 1 December 2015 to 28 February 2019, a total of 361 cases. All forecasts verifying at 00 UTC between 1 December and 28/29 February are included in the evaluation. The ENS comprises 50 perturbed members and one control member. The forecasts are valid for lead times of 1 to 15 days (at 24-hr intervals). The 500 hPa fields of each ENS forecast are extracted on a 1×1 degree grid and projected onto the two EOFs. The projections describe the magnitude of the NAO and BLO in each forecast, calculated relative to the climatological standard deviation. Following Ferranti et al. (2018), cases with projections greater than one standard deviation are considered large amplitude events.

3. Methods

We consider a sequence of ensemble forecasts valid for the same time t_v and started from initial conditions between 1 and L days before, $f(t_v, i), i = 1, \ldots L$. Each ensemble consists of M members, $f_m(t_v, i), m = 1, \ldots M$. We consider NAO and BLO separately, so f_m are univariate and real-valued.

To measure the difference between two ensembles f and g with M and N members, respectively, we use the divergence function given by

Geophysical Research Letters 10.1029/2020GL087934

$$d(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|f_i - g_j\right| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|f_i - f_j\right| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_i - g_j\right|.$$

d is the divergence function associated with the Continuous Ranked Probability Score (CRPS), which is widely used to measure of the quality of ensemble forecasts (Gneiting & Raftery, 2007). If either M or N is equal to one, then d reduces to the CRPS, while if both are one, d is simply the absolute distance $|f - g|$. This means that d can also be used to measure the difference between two EM or control forecasts. d shares the important property of propriety with CRPS (Gneiting & Raftery, 2007), and as shown by Thorarinsdottir et al. (2013), these properties make d a particularly suitable choice.

The difference between two ensemble forecasts initialized on consecutive days and valid for the same time is

$$D(t_v, i) = d(f(t_v, i), f(t_v, i-1)),\ i = 1, \ldots, L,$$

where $f(t_v, 0)$ is the set of initial perturbed ensemble members at time t_v.

To measure the overall divergence (or inconsistency) between the sequence of forecasts valid for a given time, we sum the divergence between successive pairs of forecasts. To focus on the jumpiness within the sequence rather than a general trend across lead times (or a single large jump representing a one-time change in predictability), we subtract the difference between the first and last forecast of the sequence and define the divergence index (DI) for a given case as

$$\mathrm{DI}(t_v) = \frac{1}{L-1}\left(\left(\sum_{l=1}^{L} D(t_v, l)\right) - d(f(t_v, L), f(t_v, 0))\right).$$

The DI is calculated for the ENS and also for the ensemble control forecast (CTRL) and the EM. We refer to DI (ENS), DI (CTRL), and DI (EM), respectively. In this study, all ensemble forecasts have $M = 50$ members (control not included), and we consider forecasts up to lead time of $L = 15$ days.

As noted above, for a single forecast such as CTRL and EM, the divergence is equal to the absolute difference. For these forecasts, DI is similar (though not identical) to the flip-flop index of Griffiths et al. (2019).

4. Results

Figure 1 (upper panel) shows the DI (ENS) for NAO (solid) and BLO (dashed) for each day of the last four winters (December–February 2015–2016 to 2018–2019; the vertical dotted lines indicate the start of each season). Positive values indicate higher inconsistency. There is similar variability in DI for both regimes (standard deviation of 0.027 for NAO and 0.028 for BLO). However, the peaks of high/low consistency occur at different times. Winter 2018–2019 was more inconsistent than usual for blocking, while forecasts for NAO were not unusually inconsistent in this season. Overall, there is no strong correlation between inconsistency in forecasts of blocking and NAO (correlation = −0.1 over the full set of cases).

To illustrate the different levels of consistency associated with the high and low DI, three example cases are shown in the lower panel (labeled A, B, and C on top panel). B (center) shows an example of a case with very good consistency in forecasting the BLO regime. The plot shows the amplitude of blocking for 14 December 2017 predicted by forecasts initialized between 30 November and 13 December. The 15-day ENS forecast has a broad distribution (large spread), similar to the climate distribution. Subsequent forecasts show smaller spread and a consistent shift of the ENS towards negative BLO.

C (right) shows a contrasting case with poor consistency in forecasting blocking. The plot shows the amplitude of blocking for 14 December 2018 predicted by forecasts initialized on 30 November to 13 December. The longest range forecasts are similar to the climate distribution; there is a trend over the following days showing an increasing probability for blocking. However, there is then an abrupt change in the forecast to a strong signal for neutral conditions, followed by an equally abrupt change back to blocking. This is the most inconsistent BLO case of this whole period.

A (left) is a case of large inconsistency for the NAO. This occurs at the end of an extended period of strong NAO− (and associated cold weather over NW Europe). The forecasting challenge in this case is to identify

Geophysical Research Letters 10.1029/2020GL087934

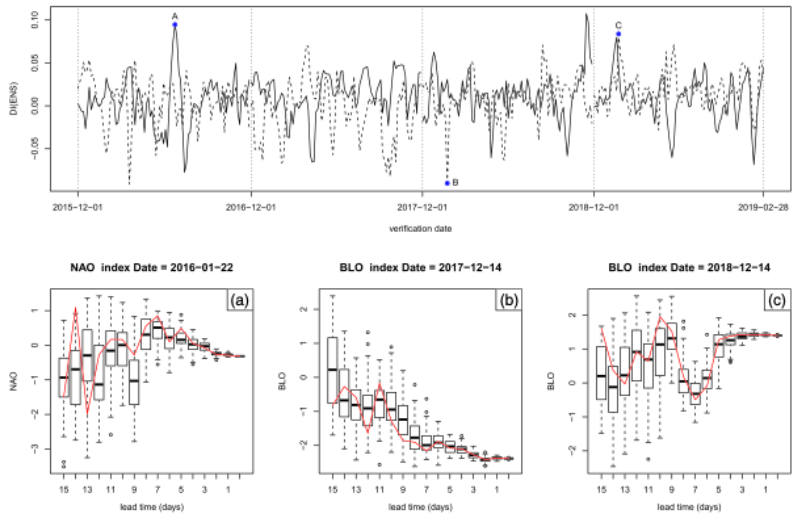


Figure 1. Consistency in ENS forecasts of the NAO and BLO regimes. Upper panel: time series of overall consistency DI (ENS) of 1–15 day forecasts verifying during winters (December–February) 2015–2019. Positive values indicate lower consistency. Lower panels show examples of both consistent and inconsistent cases for each regime. Each example shows the distribution of ENS forecasts verifying for a given date with lead time of 1 to 15 days; box and whiskers show min, max, and 25, 50, and 75 percentiles of the ENS distribution (50 perturbed members); red line shows the ENS control.

when this cold event will end. The longest range forecasts show large uncertainty but with probability of around 50% for a return to near-normal conditions (NAO magnitude <1). The forecasts from 11 January onwards show much higher probability for the end of the NAO− event, with the exception of the forecast from 13 January which again gives a higher probability for the cold spell to continue beyond 21 January.

These cases of large inconsistency illustrate the challenge for users—in both, there is an apparent increase in certainty for a change in weather type (regime). But this is thrown into doubt by a large change in a subsequent forecast. The following jump back is also difficult for the user to manage—can it be trusted, or will the following forecast jump again? While such cases are uncommon in the ENS (Figure 1, top), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

The consistency of ENS is compared with that of the control forecast and of the EM in Figure 2 for NAO (results for BLO are similar). Overall, DI is much larger for the EM (mean DI 0.14) and especially CTRL (0.42) than for ENS (0.01), reflecting how the full ENS distribution does mitigate the jumpiness seen in the deterministic forecasts. The cases with large DI (ENS) also tend to have large DI (EM), and vice versa. The examples of inconsistent ENS forecasts in Figure 1 are typical—there is a substantial shift of the whole ENS distribution, which is reflected in both DI (EM) and DI (ENS). For more consistent cases, the correlation is less strong. When the whole ENS distribution is very consistent, the EM must also be consistent. However, when the EM is consistent, there may still be variation in the ENS distribution as a whole (for example, changes in spread) that can lead to larger DI (ENS).

There is much less correlation between DI (ENS) and DI (CTRL). The most inconsistent cases for ENS tend to be associated with a substantial shift in the whole ENS distribution, and the control also shows large inconsistency as expected. However, there are also cases with large DI (CTRL) but small DI (ENS)—large jumps in CTRL are not reflected in the ENS as a whole, as seen in the examples. This is an important result that demonstrates that jumpiness in the ENS is not simply a consequence of a corresponding jumpiness in the CTRL.

Figure 3 shows the distribution of magnitude of the individual jumps ($|D(t_v, i)|$, absolute value of difference between forecasts started 1 day apart) at each lead time for both ENS and CTRL. The two inconsistent cases

Geophysical Research Letters 10.1029/2020GL087934

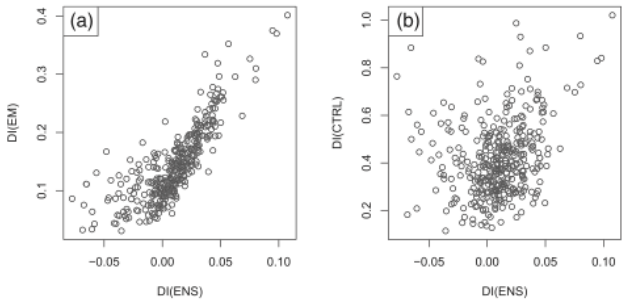


Figure 2. Comparison of consistency of ENS, CTRL, and EM for NAO. Each panel shows a scatter plot of DI (ENS) on the *x*-axis against (a) DI (EM) and (b) DI (CTRL); 361 cases verifying during winters (December–February) 2015–2019.

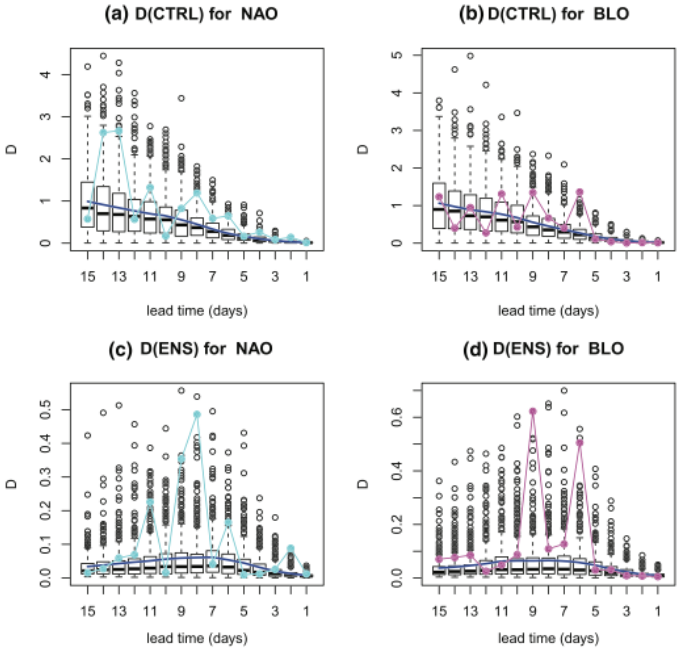


Figure 3. Distribution of jumps ($|D(t_0, i)|$) at each forecast lead time (i days) for CTRL (top) and ENS (bottom) for the NAO (left) and BLO (right) regimes. Box and whiskers show 25, 50, and 75 percentiles of the ENS distribution, with outliers shown by open circles; thick blue lines show the mean value. The values for the sequence of forecasts verifying on 22 January 2016 for NAO (cyan) and 14 December 2018 (magenta) correspond to the two examples of inconsistent forecasts shown in Figure 1.

A and C from Figure 1 are highlighted. As well as having large overall DI (ENS), both cases have some of the largest individual ENS jumps between consecutive forecasts at any lead time. As for DI, the magnitude of the individual jumps is much larger for CTRL than for ENS.

Geophysical Research Letters 10.1029/2020GL087934

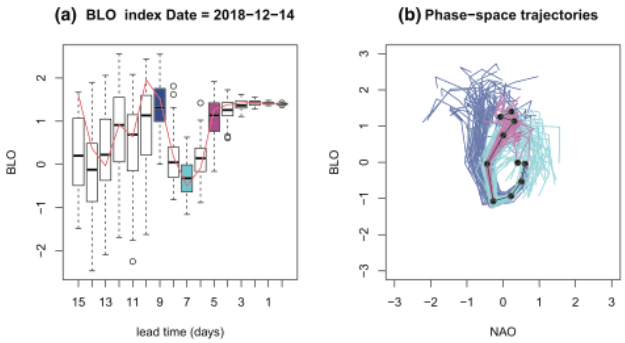


Figure 4. Phase-space trajectories of ENS forecasts initialized on 5, 7, and 9 December 2018. (a) Amplitude of blocking for 14 December 2018 predicted by forecasts from different initial times up to 15 days ahead with forecasts from 5, 7, and 9 December highlighted; box and whiskers show min, max, and 25, 50, and 75 percentiles of the ENS distribution, with outliers shown by open circles; red line shows the ENS control. (b) Phase-space trajectories of ENS forecasts initialized on 5, 7, and 9 December 2018 (blue, cyan, and magenta, respectively) and verifying analysis trajectory (black).

Figure 3 highlights another important difference between the jumpiness of the ENS and CTRL. For CTRL, $|D(t_v, i)|$ increases with lead time, with the mean jump approaching 1 by day 15. However, for ENS, the largest mean value and most extreme jumps tend to occur at around 7–9 days lead. At longer lead times, as memory of the initial conditions is lost, the limit of predictability is reached and each forecast behaves like a random draw from the climate distribution. This means that at long lead, the difference between two control forecasts will be on average the same as the difference between two randomly selected states from the climate (see Text S1 in the supporting information for details). In contrast, at this range, two ENS forecasts will represent two statistically indistinguishable samples from the same climate distribution. Any difference between them will only be due to sampling, and for a sufficiently large ensemble, $D(t_v, i)$ will be small.

We have seen that DI can identify cases of high inconsistency in the ENS. A more detailed investigation of such cases is merited to understand what aspects of the ensemble forecast configuration lead to such behavior. The high-DI cases, A and C (Figure 1) both occur in situations of transitions between large-scale regimes. A compact way to visualize these transitions is in a phase-space plot which can be used to examine how the magnitude of both BLO and NAO evolve through the forecast for each ensemble member (Ferranti et al., 2018). Following this approach for high-DI cases also brings some new insight into the jumpiness itself.

To illustrate this, we consider the BLO case of 14 December 2018 (C in Figure 1) and examine the phase-space trajectories of the relevant forecasts. We compare the forecasts started on 5 and 9 December (which both predict a positive BLO pattern) with the contrasting forecast from 7 December which has largest probability for a negative BLO to occur (Figure 4a). Figure 4b (and Figure S1) shows the phase-space evolution of the forecasts from 5, 7, and 9 December 2018. The forecast from 9 December follows the observed trajectory with only a few members moving too quickly away from the block. The forecast from 7 December also follows the observed trajectory for the first 4–5 days of the forecast, but then most members fail to maintain the blocking and evolve too quickly towards the more mobile NAO+ pattern, leading to the poor 7-day forecast for BLO (Figure 4a, cyan). The forecast from 5 December does not follow the observed trajectory so well from 9 December onwards: most ENS members move too quickly into a strong blocking and NAO−. Although this forecast gives a strong indication of blocking for 14 December (day 9 forecast, Figure 4a, blue), the evolution leading to this is clearly inconsistent with the observed development. While Figure 4a suggests that the forecast from 7 December has lost the signal that was present in earlier forecasts, the analysis of the phase-space trajectories shows that the situation was more complex. In fact, the forecast from 7 December better captured

Geophysical Research Letters 10.1029/2020GL087934

the observed evolution up to 11 December, with significantly smaller ENS spread. Neither the 5 December nor the 7 December forecast captured the observed trajectory after this time. It was only the later forecasts, from 9 December onwards that correctly predicted the observed evolution.

This shows us that care is needed in the interpretation of the ensemble jumpiness. An apparent clear flip-flop in a single index may hide a more complex predictability issue. When investigating the cause of a case of high DI, it is important to frame the analysis in the right context, as shown by Figure 4. From a diagnostic point of view, Figure 4a raises the question: why do the forecasts from 7 December lose the signal that was present in the earlier forecast from 5 December? In contrast, looking at the wider context of Figure 4b raises the question: what mechanism caused the two successive changes in predictability, first to avoid the too strong NAO−/BLO (5 December forecast) and second to maintain the block and not move too quickly to NAO+ (7 December forecast). Error tracking (Grams et al., 2018; Magnusson, 2017) shows that both these errors can be traced back to the initial mishandling of developing trough-ridge patterns over eastern North America (Figures S2 and S3).

5. Conclusions

Predicting transitions between large-scale weather regimes 2 weeks ahead is a significant forecasting challenge. Occasionally, successive ensemble forecasts can give contradictory indications about the probability for a change in weather type. Such jumpiness or "flip-flopping" is difficult for users to manage since the forecast does not give a consistent message for decision making. While such cases are uncommon (Figure 1), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

For the first time, we have carried out a systematic, objective evaluation of the consistency of ECMWF ensemble forecasts that takes account of the full ensemble distribution. This extends the earlier work of Zsoter et al., 2009 who focused specifically on flip-flops of the EM.

We investigated the ENS consistency for two key flow patterns for Europe, NAO and blocking. We used a measure of the divergence between two ensembles started at different times but valid for the same time. This allowed us to quantify both individual jumps and the overall consistency of a sequence of ENS forecasts valid for a given time. Our main conclusions are the following:

- In general, the peaks of high and low consistency occur at different times for NAO and BLO; there is no strong correlation between inconsistency for NAO and BLO (Figure 1).
- DI for the ENS is on average much lower than for EM and especially for CTRL (Figure 2) demonstrating benefit of the ensemble in mitigating the jumpiness of the deterministic forecasts by representing the range of possible scenarios.
- The largest individual jumps for ENS tend to be days 7–9, while for the CTRL the magnitude of individual jumps continues to increase throughout the forecast (Figure 3). This is associated with the different asymptotic behavior of the (deterministic) CTRL forecast and the ENS at long forecast lead.
- Care is needed in the interpretation of the ensemble jumpiness. What looks at first sight to be a clear case of flip-flopping in a single index (BLO or NAO) may be a more complex predictability issue. This may be better understood by examining the phase-space evolution of both components together (Figure 4).

In this work, we assessed the consistency of the univariate forecast of NAO and BLO separately. However, we also showed how it is important to consider the ensemble trajectories in the two-dimensional phase to properly understand the reason for apparent jumpiness. It will therefore be valuable to extend the divergence and DI methodology to the multivariate situation so that the consistency of NAO and BLO can be evaluated together. This will also enable investigation of the consistency of other aspects of ensemble performance such as for tropical cyclone tracks.

The DI allows us to identify important cases of high ensemble forecast inconsistency and to routinely monitor the occurrence of such cases. Careful diagnosis of these cases will help to identify the causes of the inconsistency and hence to address the relevant aspects of ensemble configuration and modeling. Reducing the occurrence of inconsistent (or jumpy) ensemble forecasts will increase user confidence and improve decision making.

Acknowledgments
David Richardson is supported by a Wilkie Calvert PhD Studentship at the University of Reading. The data used for the EOF computations is available from the ECMWF ERA-Interim archive (Berrisford et al., 2011): https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim. The ECMWF ensemble forecast data used in this research is available on the TIGGE archive (Swinbank et al., 2016): https://confluence.ecmwf.int/display/TIGGE.

References

Ben Bouallègue, Z., Magnusson, L., Haiden, T., & Richardson, D. S. (2019). Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, *145*(721), 1741–1755. https://doi.org/10.1002/qj.3523

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., et al. (2011). The ERA-Interim archive Version 2.0. ERA Report Series, (1), 23. Retrieved from https://www.ecmwf.int/node/8174.

Buizza, R. (2008). The value of probabilistic prediction. *Atmospheric Science Letters*, *9*(2), 36–42. https://doi.org/10.1002/asl.170

Buizza, R, & Richardson, D. (2017). 25 years of ensemble forecasting at ECMWF|ECMWF (No. 153). ECMWF Newsletter. https://doi.org/10.21957/bv418o

Cassou, C. (2008). Intraseasonal interaction between the Madden-Julian oscillation and the North Atlantic oscillation. *Nature*, *455*(7212), 523–527. https://doi.org/10.1038/nature07286

Charlton-Perez, A. J., Aldridge, R. W., Grams, C. M., & Lee, R. (2019). Winter pressures on the UK health system dominated by the Greenland Blocking weather regime. *Weather and Climate Extremes*, *25*, 100218. https://doi.org/10.1016/j.wace.2019.100218

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

ECMWF. (2015). The strength of a common goal. Strategy 2016-2025. Retrieved from https://www.ecmwf.int/sites/default/files/ECMWF_Strategy_2016-2025.pdf.

Ehret, U. (2010). Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift*, *19*(5), 441–451. https://doi.org/10.1127/0941-2948/2010/0480

Ferranti, L., Corti, S., & Janousek, M. (2015). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 916–924.

Ferranti, L., Magnusson, L., Vitart, F., & Richardson, D. S. (2018). How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, *144*(715), 1788–1802. https://doi.org/10.1002/qj.3341

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., & Wernli, H. (2017). Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, *7*(8), 557–562. https://doi.org/10.1038/nclimate3338

Grams, C. M., Magnusson, L., & Madonna, E. (2018). An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2577–2591. https://doi.org/10.1002/qj.3353

Griffiths, D., Foley, M., Ioannou, I., & Leeuwenburg, T. (2019). Flip-flop index: Quantifying revision stability for fixed-event forecasts. *Meteorological Applications*, *26*(1), 30–35. https://doi.org/10.1002/met.1732

Hewson, T. (2020). Use and verification of ECMWF products in member and co-operating states (2019). *ECMWF Technical Memorandum*, (860). https://doi.org/10.21957/80s471ib1

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Magnusson, L. (2017). Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2129–2142. https://doi.org/10.1002/qj.3072

Pappenberger, F., Bogner, K., Wetterhall, F., He, Y., Cloke, H. L., & Thielen, J. (2011). Forecast convergence score: A forecaster's approach to analysing hydro-meteorological forecast systems. *Advances in Geosciences*, *29*, 27–32. https://doi.org/10.5194/adgeo-29-27-2011

Pappenberger, F., Cloke, H. L., Persson, A., & Demeritt, D. (2011). HESS opinions "on forecast (in)consistency in a hydro-meteorological chain: Curse or blessing?". *Hydrology and Earth System Sciences*, *15*(7), 2391–2400. https://doi.org/10.5194/hess-15-2391-2011

Ruth, D. P., Glahn, B., Dagostaro, V., Gilbert, K., Ruth, D. P., & Glahn, B. (2009). The performance of MOS in the digital age. *Weather and Forecasting*, *24*(2), 504–519. https://doi.org/10.1175/2008WAF2222158.1

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, *97*(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534. https://doi.org/10.1137/130907550

Yiou, P., & Nogaj, M. (2004). Extreme climatic events and weather regimes over the North Atlantic: When and where? *Geophysical Research Letters*, *31*, L07202. https://doi.org/10.1029/2003GL019119

Zsoter, E., Buizza, R., & Richardson, D. (2009). "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Monthly Weather Review*, *137*(11), 3823–3836. https://doi.org/10.1175/2009MWR2960.1

Geophysical Research Letters

Supporting Information for

Evaluation of the consistency of ECMWF ensemble forecasts

David S. Richardson[1,2], Hannah L. Cloke[2,3,4], and Florian Pappenberger[1]

[1]European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, UK.

[2]Department of Geography and Environmental Science, University of Reading, Reading, UK.

[3]Department of Meteorology, University of Reading, Reading, UK.

[4]Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Corresponding author: David Richardson (david.richardson@ecmwf.int)

Contents of this file

Text S1 to S2
Figures S1 to S4

Introduction

This supporting information provides details on the asymptotic limit for jumps of the CRTL forecast and four figures with explanatory text detailing the evolution of the errors in the forecasts from 5, 7, 9 December 2018.

Text S1 explains the theoretical limit for the magnitude of individual jumps in the CTRL forecast.

Text S2 describes the evolution of the errors in the ensemble mean forecasts from 5, 7, 9 December 2018 which are shown in Figures S2-S4. Figure S1 shows the magnitude of the BLO and NAO projections for all ensemble members of these forecasts at 24-hour intervals. This is the same information as shown in Figure 4b, but with each lead time shown separately for 9-14 December for extra clarity.

Text S1.

Figure 3 shows the distribution of magnitude of the individual jumps $|D(t_v, i)|$ at each lead time for both ENS and CTRL. For CTRL, For CTRL, $|D(t_v, i)|$ increases with lead time, with the mean value approaching 1 by day 15. Here we consider the asymptotic limit for this mean value.

At long lead times, each forecast behaves like a random draw from the climate distribution, ie two control forecasts $f(t_v, i)$ and $f(t_v, i-1)$ will be uncorrelated for sufficiently large i. The average distance (divergence) between two such random states, f and g, is

$$\overline{d_r} = \overline{|f-g|}$$

where the overbar denotes the average of all cases, the subscript r indicates this is for random selection of states, and recall that for the deterministic forecast the divergence d is the absolute distance $|f-g|$.

If the climatology is normally distributed then we can compute $\overline{d_r}$ analytically

$$\overline{d_r} = \frac{2}{\sqrt{\pi}}\sigma \approx 1.13\sigma$$

where σ is the climate standard deviation. In our study the NAO and BLO projections are already normalized by the climate standard deviation, so that $\sigma = 1$. Hence we could expect the mean curves for CTRL shown in Figure 3 to tend to 1.13 in the long-range as predictability is lost. The fact that this limit has not quite been reached by day 15 is suggests there is some predictability still at this range.

The above relies on the assumption that the climate distribution is normal. If we make no assumption about the climatological distribution of the projections, then we cannot make an analytic value for d_r but we can derive an upper limit.

The mean absolute distance $\overline{d_r}$ is related to the mean squared distance $\overline{d_r^2}$:

$$\overline{(d_r - \overline{d_r})^2} = \overline{d_r^2} + \overline{d_r}^2 - 2\overline{d_r}^2 = \overline{d_r^2} - \overline{d_r}^2$$

The mean-squared difference between two random states can be written as

$$\overline{d_r^2} = \overline{(f-g)^2} = \overline{f^2} + \overline{g^2} - 2\overline{fg} = 2\sigma^2$$

since by definition the random states f and g are uncorrelated, $\overline{fg} = 0$, and $\overline{f^2} = \overline{g^2} = \sigma^2$. This is the standard result that asymptotically the control forecast error will be equal to twice the climatological variance.

Hence, from the above two equations we see that

$$\overline{d_r}^2 = \overline{d_r}^2 - \overline{\left(d_r - \overline{d_r}\right)^2} = 2\sigma^2 - \overline{\left(d_r - \overline{d_r}\right)^2}$$

This shows that the mean absolute distance $\overline{d_r}$ is never greater than the root mean squared distance $\sqrt{\overline{d_r}^2}$. As noted above, in this study $\sigma = 1$, and so the upper limit for the asymptotic (long-lead) limit for the mean curves for CTRL in Figure 3 is $\sqrt{2}$.

Text S2.

The forecast from 5 December develops too strong BLO and NAO- for 11-12 December (Figure S1). This can be seen clearly in the ensemble mean forecast (Figure S2) – there is a large error in the ridge over the E Atlantic, the axis of the ridge is too far west and there is a too strong north-westwards extension over Iceland and towards Greenland (consistent with the NAO- signal). The troughs either side of this main ridge are too deep, giving overall a much too strong omega block (enhanced omega pattern). The error pattern established by 11 December remains through the rest of the forecast. This error pattern can be traced back through the forecast to an over amplification of the trough-ridge structure over eastern North America the western Atlantic in the 72-hour forecast: enhanced ridge over the Hudson Bay, slightly extended trough off the eastern seaboard and small overdevelopment of the ridge in the mid-Atlantic.

The forecast from 7 December captures much better the initial trough ridge structure (on 9 December) and does not overextend the meridional pattern, resulting in much lower errors over N Atlantic/Europe on 11-12 December (Figure S3). However, upstream errors in a following trough-ridge pattern (also originating with positive error over the Hudson Bay and negative errors over the east coast) amplify downstream. In this case though the interaction with the pre-existing ridge appears to speed up the anticyclonic wave breaking and the high pressure moves further downstream. This results in especially large error over western Europe.

It is worth noting that the forecast from 5 December also has very similar error structure that develops in this second trough-ridge pattern (compare the centers of the error positive and negative over the west Atlantic on 12 December in Figures S2 and S3). However, the much extended and higher amplitude pre-existing block in the central Atlantic appears to limit the impact of this second error pattern.

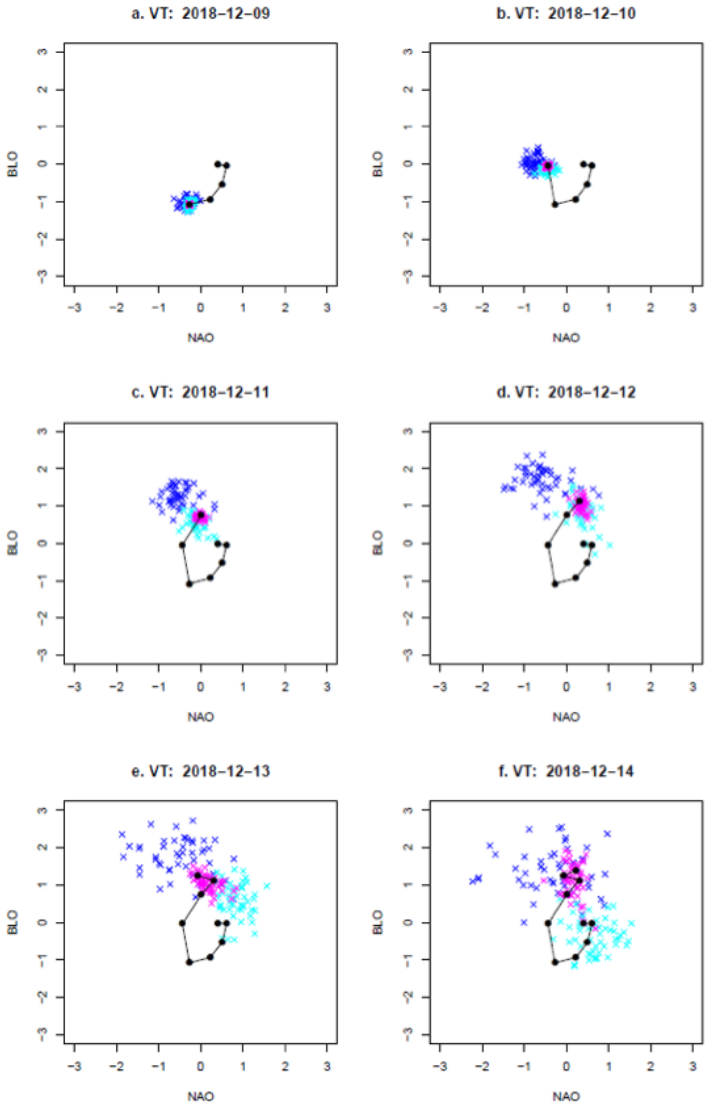


Figure S1. Phase space plots of ENS forecasts initialized on 5, 7, 9 December 2018 (blue, cyan, magenta respectively) verifying on 9-14 December (panels a to f). In each panel the

4

verifying analysis trajectory is shown at 24-hour intervals from 5 December to the verifying date (black line).

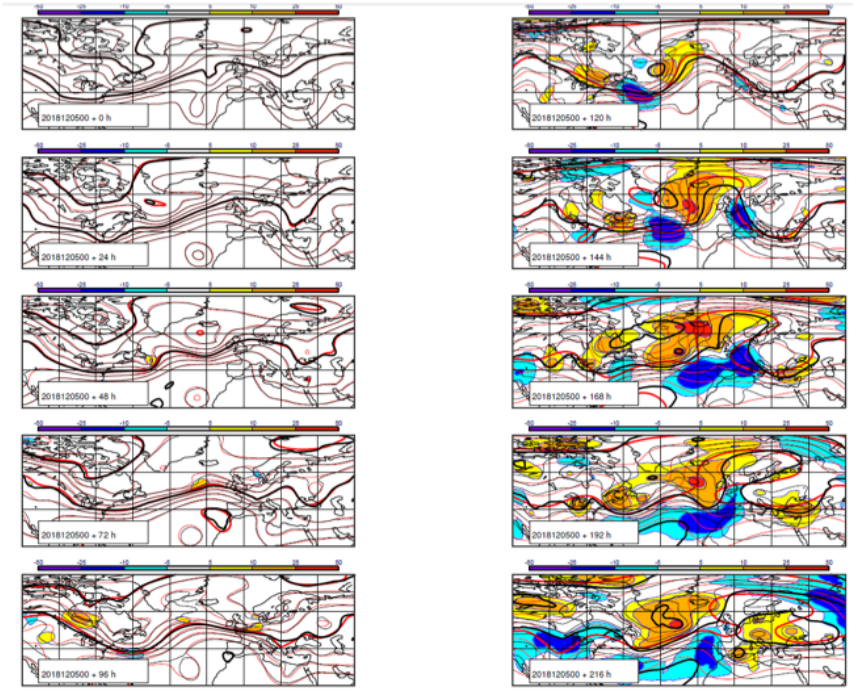


Figure S2. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 9 December 2018 together with the verifying analysis (black), at 24-hour intervals.

5

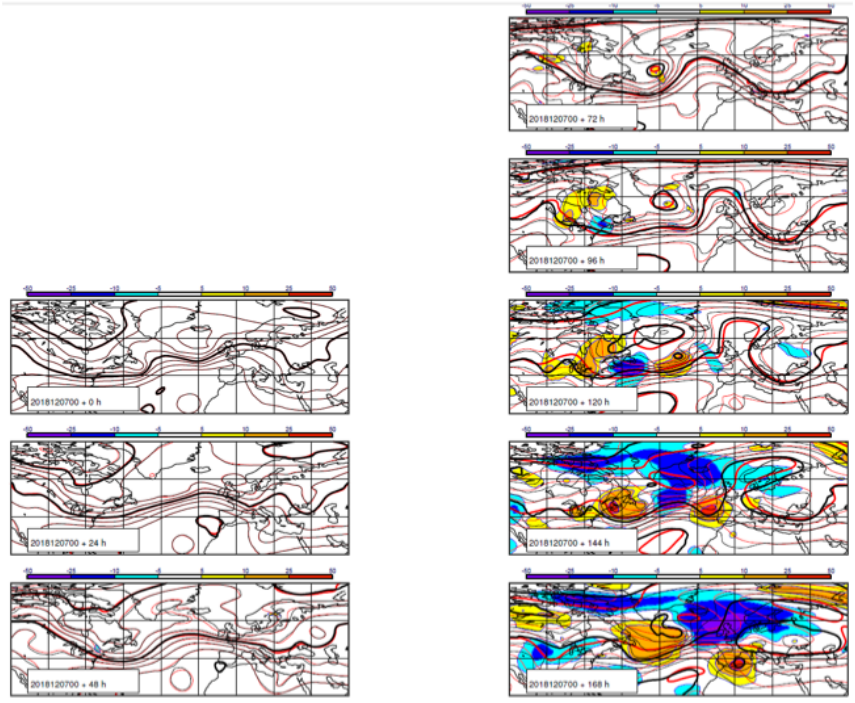


Figure S3. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 7 December 2018 together with the verifying analysis (black), at 24-hour intervals.

6

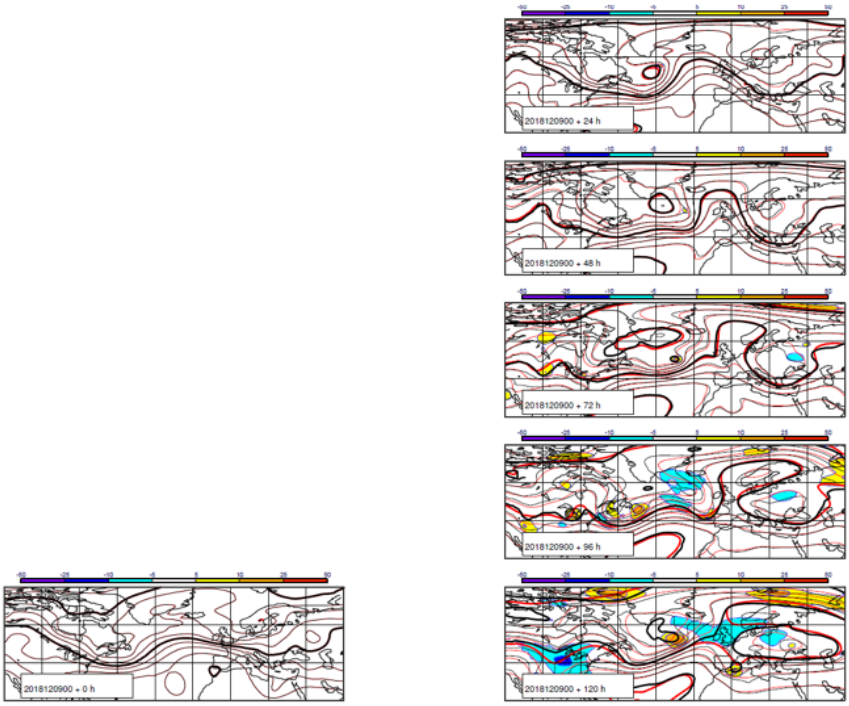


Figure S4. 500 hPa geopotential height error (shaded) for the ensemble mean forecast (red) from 0 UTC on 9 December 2018 together with the verifying analysis (black), at 24-hour intervals.

7

A2. Published Article: Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks

This appendix contains the published version of chapter 4 of this thesis, with the following reference:

Richardson, D.S., Cloke, H.L., Methven, J.A. and Pappenberger, F. (2024) 'Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks', *Weather and Forecasting*, 39(1), pp. 203–215. Available at: https://doi.org/10.1175/WAF-D-23-0113.1.

Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks

DAVID S. RICHARDSON,[a,b] HANNAH L. CLOKE,[a,c,d] JOHN A. METHVEN,[c] AND FLORIAN PAPPENBERGER[b]

[a] *Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom*
[b] *ECMWF, Reading, United Kingdom*
[c] *Department of Meteorology, University of Reading, Reading, United Kingdom*
[d] *Department of Earth Sciences, Uppsala University, Uppsala, Sweden*

(Manuscript received 4 July 2023, in final form 25 October 2023, accepted 25 November 2023)

ABSTRACT: We investigate the run-to-run consistency (jumpiness) of ensemble forecasts of tropical cyclone tracks from three global centers: ECMWF, the Met Office, and NCEP. We use a divergence function to quantify the change in cross-track position between consecutive ensemble forecasts initialized at 12-h intervals. Results for the 2019–21 North Atlantic hurricane season show that the jumpiness varied substantially between cases and centers, with no common cause across the different ensemble systems. Recent upgrades to the Met Office and NCEP ensembles reduced their overall jumpiness to match that of the ECMWF ensemble. The average divergence over the set of cases provides an objective measure of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80–90 km in the cross-track direction between a forecast for 120 h ahead and the updated forecast made 12 h later for the same valid time. This quantitative information can support users' decision-making, for example, in deciding whether to act now or wait for the next forecast. We did not find any link between jumpiness and skill, indicating that users should not rely on the consistency between successive forecasts as a measure of confidence. Instead, we suggest that users should use ensemble spread and probabilistic information to assess forecast uncertainty, and consider multimodel combinations to reduce the effects of jumpiness.

SIGNIFICANCE STATEMENT: Forecasting the tracks of tropical cyclones is essential to mitigate their impacts on society. Numerical weather prediction models provide valuable guidance, but occasionally there is a large jump in the predicted track from one run to the next. This jumpiness complicates the creation and communication of consistent forecast advisories and early warnings. In this work we aim to better understand forecast jumpiness and we provide practical information to forecasters to help them better use the model guidance. We show that the jumpiest cases are different for different modeling centers, that recent model upgrades have reduced forecast jumpiness, and that there is not a strong link between jumpiness and forecast skill.

KEYWORDS: Tropical cyclones; Ensembles; Forecast verification/skill

1. Introduction

Official forecasts of tropical cyclone (TC) tracks are typically based on guidance from numerical weather prediction (NWP) models (Conroy et al. 2023). NWP ensemble forecasts are increasingly being used. Although their use in official forecasts is often limited to the ensemble mean (EM) track, there is increasing evidence of the benefits of using more of the ensemble probabilistic information (Titley et al. 2019, 2020; Kawabata and Yamaguchi 2020; Leonardo and Colle 2017). One benefit of using ensembles is the increased consistency between consecutive forecasts (Buizza 2008; Zsoter et al. 2009). There are nevertheless occasions where an ensemble is unexpectedly jumpy with the predicted TC locations flip-flopping over several consecutive forecasts (Magnusson et al. 2021). Such cases can be difficult to interpret, complicating the creation of consistent forecast advisories and early warning communications. Understanding the frequency and reasons for these cases as well as information about the overall levels of consistency in operational ensemble forecasts can help forecasters to better use the available ensemble track data.

As new forecast information arrives (usually every 6–12 h for global NWP models), forecasters need to decide how to revise their forecasts to take account of the new forecast information. National Hurricane Center (NHC) Tropical Cyclone Advisories often discuss the change in forecast track due to updated guidance, making adjustments to the path depending on the new information. There is a balance to be struck between closely following the changed model guidance and taking a more conservative approach of making a smaller change to minimize the potential need to make a change in the opposite direction later, that is to avoid a so-called windshield-wiper effect (Broad et al. 2007). Contradictory messages from such jumpiness can cause difficulties for decision-makers and reduce users' confidence in the forecasts (Hewson 2020; Pappenberger et al. 2011b; McLay 2011;

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-23-0113.s1.

Corresponding author: David S. Richardson, d.s.richardson@pgr.reading.ac.uk

DOI: 10.1175/WAF-D-23-0113.1

Elsberry and Dobos 1990). Information quantifying the consistency between successive probabilistic forecasts can be important to inform optimal decision-making, such as whether to act now or wait for the next forecast (Regnier and Harr 2006; Jewson et al. 2021, 2022). Both noted that such information is not readily available to users.

Evaluation of operational ensemble TC track forecasts includes EM track errors, ensemble spread, and strike probability (e.g., Cangialosi 2022; Haiden et al. 2022; Titley et al. 2020; Heming et al. 2019; Leonardo and Colle 2017). However, few authors have addressed the jumpiness of TC track forecasts. Elsberry and Dobos (1990) investigate consistency of TC guidance for the western North Pacific by using the difference in cross-track errors between successive forecasts. Fowler et al. (2015) assess consistency of Atlantic TC track forecasts by counting forecast crossovers—how often in a sequence of forecasts the predicted position changes from one side to the other of a fixed reference track, for example the observed track. However, they caution that biased forecasts may appear to be consistent since successive forecasts may jump considerably without crossing the observed track. Both Elsberry and Dobos (1990) and Fowler et al. (2015) recommend the regular evaluation of forecast consistency in addition to the standard assessments of forecast accuracy.

More generally, there has been limited investigation of forecast jumpiness, especially for ensemble forecasts. Zsoter et al. (2009) considered flip-flops in sequences of forecasts all valid for a given time and showed that EM forecasts are more consistent than the corresponding ensemble control forecasts. Griffiths et al. (2019) introduced a flip-flop index to compare the consistency of automated and manual forecasts, while Ruth et al. (2009) assessed how model output statistics improved forecast consistency. Forecast consistency has been considered for rainfall (Ehret 2010) and river flow (Pappenberger et al. 2011a).

These previous studies were mainly focused on deterministic forecasts (either single runs or EM) and the methods are not directly applicable to assess the jumpiness in sequence of ensemble forecasts taking account of the full ensemble distribution. Recently, Richardson et al. (2020) introduced a measure of forecast jumpiness based on forecast divergence that accounts for all aspects of the ensemble empirical distribution. They used this to investigate jumpiness of ensemble forecasts for the large-scale flow over the Euro-Atlantic region.

In the present study we apply the forecast jumpiness measure introduced by Richardson et al. (2020) to ensemble forecasts of Atlantic TCs, focusing on the run-to-run consistency in the cross-track direction which is most important in determining the location of TC landfall. The aim is to provide forecasters and model developers with information about the jumpiness of ensemble TC forecasts. This will help forecasters and decision-makers better understand the expected changes between successive forecasts. We address the following questions:

- How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centers?
- Is there a common cause of "jumpy" cases—are the ensembles from different centers particularly jumpy for the same TC cases and if so what is the reason?
- Have recent ensemble model upgrades had a noticeable effect on the forecast consistency?
- What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness – what information is practically useful? Is there any useful link between jumpiness and skill?

We investigate these questions using ensemble forecast data from three global NWP centers. The data used in this study and the methods to assess forecast jumpiness are introduced in sections 2 and 3. Results are presented in section 4. We start with a case study to illustrate the issues of ensemble TC track jumpiness. Then we look at the overall jumpiness over the 2019, 2020 and 2021 Atlantic hurricane seasons. Finally, we consider the relationship between jumpiness, error and spread. We conclude with a summary, recommendations for forecasters and avenues for future work in section 5.

2. Data

In this study we investigate the run-to-run consistency of ensemble tropical cyclone track forecasts from three global centers: the European Centre for Medium-Range Weather Forecasts (ECMWF), the U.S. National Centers for Environmental Prediction (NCEP) and the Met Office. Each center runs its own tropical cyclone tracker (Conroy et al. 2023) and the resulting track forecasts are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the TIGGE forecast tracks for all available dates from the Atlantic basin for 2019, 2020 and 2021 for forecasts initialized at 0000 and 1200 UTC from the ECMWF ensemble (ENS, 51 members integrated on ~18-km grid), NCEP ensemble (GEFS, 21 members, ~34-km grid until 22 September 2020; 31 members, ~25-km grid from 23 September 2020 onward), and Met Office ensemble (MOGREPS-G, 36 members, ~20-km grid). A given TC is not always tracked in every ensemble member (e.g., because the system dissipates in that member or the forecast intensity is below the threshold used in the tracking algorithm) and we exclude cases where a center has fewer than 10 members that track the TC at each forecast step.

We use the observed TC positions from International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010, 2018). We concentrate our analysis on named Atlantic tropical cyclones and for each cyclone include all 0000 and 1200 UTC verification times when the observed system is at least tropical storm strength (winds at least 34 kt; 1 kt $\approx$ 0.51 m s^{-1}) and the system is reported as tropical in IBTrACS (Titley et al. 2020; Goerss 2000). For each of these verification times we consider all available TIGGE forecasts. These include forecasts initialized when the TC is still a tropical depression. However, TIGGE forecast tracks are only generated for existing TCs, so longer lead-time forecasts are not always available for verification times close to when the TC is first analyzed as a tropical storm. This means that overall there are fewer forecasts for longer lead times than for shorter lead times in our sample.

We make a homogeneous sample by only including a case if the ensemble data are available from each of the three centers. This ensures that we are comparing the different centers

over the same set of cases. The total number of cases decreases with forecast lead time from 356 for 12-h forecasts to 91 for 120-h forecasts. To maintain a reasonable sample we restrict the study to forecasts of 120 h or less.

Our focus is on the changes between successive forecasts for a given verification time. We therefore need to set a minimum number of consecutive initial times over which we can assess these changes. For a given verification time t_v, we require a minimum of six consecutive forecasts, initialized at (t_v − 12 h), (t_v − 24 h), up to (t_v − 72 h), all valid for t_v. To ensure homogeneity, the same cases must be available from all three centers. With these conditions, the total number of available cases to assess the run-to-run jumpiness is 139 over the 3-yr period.

Each NWP center has made upgrades to their operational ensemble system during the 2019–21 period used in this study. A major upgrade to the GEFS was implemented on 23 September 2020, including the introduction of a new forecast model and an increase in the number of ensemble members from 20 to 30 (Zhou et al. 2022). This upgrade brought significant improvements to the ensemble performance, including for tropical cyclone forecasts. The MOGREPS-G ensemble was upgraded on 4 December 2019, including a major change to the generation of the ensemble perturbations (Inverarity et al. 2023) and revised model physics (Walters et al. 2019). This upgrade improved TC track errors (Met Office 2019).

Upgrades to the ECMWF ENS in June 2019 (Haiden et al. 2019), June 2020 (Haiden et al. 2021), and May 2021 (Rodwell et al. 2021) were neutral in terms of TC track performance, although the latter two brought improvements to intensity forecasts (Bidlot et al. 2020; Rodwell et al. 2021). A later upgrade in October 2021 did also improve TC track forecasts (Haiden et al. 2022); however, there was only one Atlantic TC in 2021 after this date. Overall, the ECMWF ensemble track forecast performance can be considered relatively stable over the period of this study. We therefore use the ENS as a reference against which to evaluate the impact of the upgrades of the other centers on ensemble jumpiness.

3. Methods

For each tropical cyclone, the observed track provides a convenient frame of reference. We consider jumpiness in a sequence of forecasts in terms of changes in the predicted cross-track location (Elsberry and Domos 1990). A positive cross-track position indicates that the forecast is to the right of observed track (facing the observed direction of travel). We also consider the links between jumpiness, ensemble error and spread. All scores—error, spread, and jumpiness—are computed in terms of the cross-track distance and are defined below.

We measure the cross-track error of the ensemble forecasts using the continuous ranked probability score (CRPS). The CRPS is widely used for evaluation of ensemble forecasts. It is a so-called proper score: if the "true" forecast probability distribution is F, a proper score ensures that the best expected score will be achieved using the forecast F rather than any other forecast distribution $G \neq F$. Hence forecasters are rewarded for honest forecasts reflecting their true beliefs. As a proper score, CRPS discourages hedging (Gneiting and Raftery 2007) and rewards both reliability and resolution (Hersbach 2000).

For an ensemble of M members f_i, $i = 1, \ldots M$ the CRPS is given in its kernel representation by

$$\mathrm{CRPS}(f) = \frac{1}{M}\sum_{i=1}^{M}|f_i - y| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j|, \quad (1)$$

where y is the verifying observation (Gneiting and Raftery 2007). The first term is the mean of the absolute error of the individual ensemble members and the second term is the mean of the distances between the different ensemble members, which accounts for the ensemble spread.

The ensemble mean forecast is given by

$$\bar{f} = \frac{1}{M}\sum_{i=1}^{M} f_i. \quad (2)$$

For a single deterministic forecast, the CRPS is equal to the mean absolute error, so the error of the ensemble mean is

$$\mathrm{CRPS}(\bar{f}) = |\bar{f} - y|. \quad (3)$$

To allow us to compare the mean spread and error over the sample of cases, we use a measure of ensemble spread that is also based on the mean absolute difference. The spread measure which corresponds to the mean absolute error of the ensemble mean is the mean absolute deviation of ensemble members from the ensemble mean:

$$s = \frac{1}{M}\sum_{i=1}^{M}|f_i - \bar{f}|. \quad (4)$$

On average over a large sample of cases the ensemble mean error [Eq. (3)] and spread [Eq. (4)] should be equal for a well-tuned ensemble system.

To measure the "jump" from one forecast to the next we follow Richardson et al. (2020) and use the divergence function d associated with the CRPS. For two ensembles f and g with M and N members, respectively, d is given by

$$d(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|f_i - g_j| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}|f_i - f_j| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|g_i - g_j|. \quad (5)$$

The first term measures the distance between the two ensembles f and g, while the second and third terms reflect the variability (spread) in each ensemble, f and g, respectively. Comparing Eq. (5) to Eq. (1) shows that the divergence reduces to the CRPS if either M or N is equal to one. If both M and N are one, then d is the absolute distance $|f - g|$. The divergence d takes account of both location and spread differences between f and g and, like the CRPS, d is a proper score (Gneiting and Raftery 2007; Thorarinsdottir et al. 2013) which discourages hedging.

Consider a given verification time t_v: an ensemble forecast f valid for this time and initialized h hours before is written

$f(t_v, h)$ and individual ensemble members are $f_i(t_v, h)$. In this study $f_i(t_v, h)$ represents the distance (in km) in the cross-track direction from the observed TC location at verification time t_v. The difference between two consecutive ensemble forecasts initialized at time $(t_v - h)$ and $[t_v - (h - 12)]$ and valid for the same time t_v is

$$D(t_v, h) = d[f(t_v, h), f(t_v, h - 12)], \tag{6}$$

where d is the divergence function [Eq. (5)].

To measure the overall divergence between the sequence of L forecasts valid for a given time we use the mean divergence between successive pairs of forecasts:

$$\overline{D(t_v)} = \frac{1}{L-1}\left[\sum_{l=2}^{L} D(t_v, 12l)\right]. \tag{7}$$

Larger values of $\overline{D}$ indicate greater change (in position, spread or both) between successive forecasts in the sequence. However, it does not necessarily indicate jumpiness in the sense of flip-flopping back and forth between different solutions. For example, if in the initial ensemble forecast all members are far to the right of the observed position and subsequent forecasts become progressively closer to the observed location, this will result in large $\overline{D}$. To distinguish between "trend" cases and "flip-flop" cases, we use the difference between the first and last forecasts of the sequence to represent this overall change (trend). Subtracting this difference from $\overline{D}$ gives the divergence index (DI) introduced by Richardson et al. (2020), which highlights jumpiness (flip-flops) in the sequence:

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1} d[f(t_v, 12L), f(t_v, 12)]. \tag{8}$$

In this way, DI will be less sensitive than $\overline{D}$ to trends caused by bias or to cases with single large jumps (resulting for example from a sudden increase in predictability). This means that the larger values of DI will be more closely related to flip-flops in the sequence of forecasts.

Our focus is on the performance of the ensemble forecast distribution and both D and DI are computed using all available ensemble members. However, because the ensemble mean (EM) track is also often used in operational forecasting we also compute the same measures for the ensemble mean. Note that for tropical cyclone tracks, the ensemble mean refers to the Euclidean mean position of the tracks from the individual ensemble members and not to a track calculated from the ensemble mean spatial fields.

The statistical significance of differences between the different centers' distributions of $\overline{D}$ and DI are assessed using the Kolmogorov–Smirnov (KS) and Mann–Whitney U (MWU) tests (Wilks 2019). Both tests are nonparametric statistical methods to compare the empirical cumulative distributions of two samples. The MWU test is mainly sensitive to differences in location (e.g., differences in the median), while the KS test is sensitive to differences in both location and shape of the distributions.

4. Results

We start with an example to illustrate the issues of jumpiness and sampling. Then we look at the overall jumpiness over 2019, 2020 and 2021 seasons. Finally, we consider the relationship between jumpiness, error and spread.

a. Example: Hurricane Laura, August 2020

Hurricane Laura formed initially as a tropical storm in the western tropical Atlantic on 20 August 2020 and affected several Caribbean countries. After traveling across the Caribbean, it reached hurricane strength on 25 August as it entered the Gulf of Mexico. It made landfall in Louisiana at 0600 UTC 27 August. Here we focus on the ECMWF ensemble (ENS) forecasts for 0000 UTC 27 August, just before the Louisiana landfall. Figure 1 shows the ENS tracks for Laura from forecasts initialized every 12 h between 21 and 25 August. The earliest forecasts, from 1200 UTC 20 August (not shown) to 0000 UTC 21 August were almost all to the northeast (right-hand side) of the observed track throughout the forecast, and predicted landfall most likely along the central and eastern Gulf coast. From 1200 UTC 21 August, the forecasts showed a higher probability for landfall further west, although with a large uncertainty as shown by the distribution of the tracks from the individual ensemble members. Between 0000 UTC 22 August and 0000 UTC 24 August, successive forecasts exhibited a "flip-flop" behavior, alternating between the western or more central Gulf coast as the most likely landfall location. Finally, from 1200 UTC 24 August onward, the forecasts more consistently indicated the western solution as most likely and it turned out that the observed track was at the eastern (right-hand) end of the range of predicted locations.

We can summarize the variations in successive forecasts for a fixed valid time in a box-and-whisker meteogram (Fig. 2). This shows the distribution of the position in the cross-track direction for all ensemble members valid for 0000 UTC 27 August, from forecasts initialized every 12 h between 1200 UTC 20 August (the first available forecast) and 1200 UTC 26 August. Each ENS forecast has one control forecast and 50 perturbed members. However, the number of members that successfully track Laura until 27 August is substantially below this, especially for the earlier forecasts. Figure 2 clearly shows the jumpiness of the ENS forecasts. The earlier forecasts are mainly to the right of the observed track (too far east), while the shorter-range forecasts are too far west (left of observed track). Intermediate forecasts flip-flop between left and right of the observed position. For each lead time (except the 48-h forecast from 0000 UTC 25 August), the observed track does lie within the ensemble distribution. However, the jumpiness (lack of consistency) between successive forecasts poses a challenge for forecasters trying to assess the most likely location of landfall.

This was a particularly jumpy case for the ENS (Magnusson et al. 2021) which merits further investigation. Comparing with other ensemble forecasts may help to identify possible causes. For example, if all centers display the same flip-flop behavior it might suggest a common cause, such as changes in available observational data between the different analysis times.

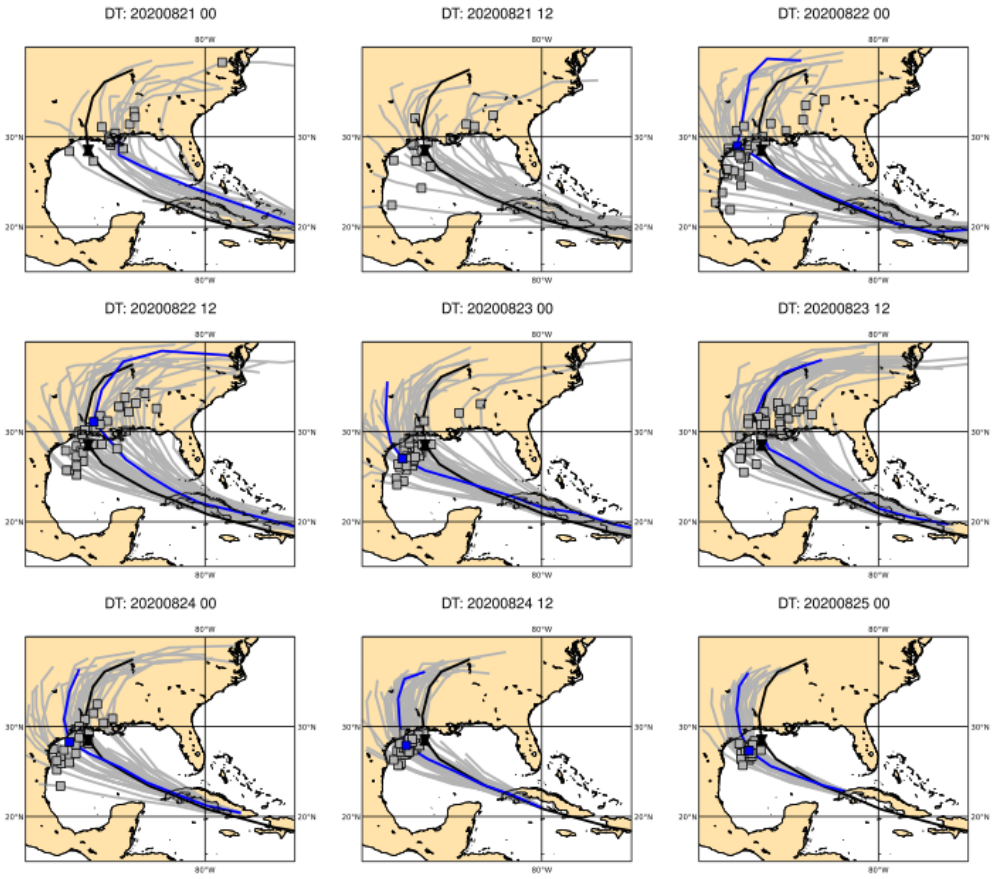


FIG. 1. Hurricane Laura: ECMWF ensemble forecast tracks (blue: control; gray: perturbed members) and observed track (black). Forecast start dates (DT) from 0000 UTC 21 Aug to 0000 UTC 25 Aug 2020. Colored symbols show forecast and observed (hourglass) position at 0000 UTC 27 Aug.

Figures 2b and 2c show the corresponding cross-track position forecasts for the MOGREPS-G and GEFS ensembles. Note that the MOGREPS-G ensemble data are missing from the TIGGE archive for forecast start times 1200 UTC 21 August and 0000 UTC 22 August. There are some similarities between all three centers: a general right bias for earlier forecasts (initialized at 0000 UTC 21 August and earlier), with a substantial proportion of members not able to track Laura as far as the verification time of 0000 UTC 27 August. Short-range forecasts for all centers are slightly left of the observed position. However, neither MOGREPS-G nor GEFS shows the same degree of flip-flop behavior as ENS.

The MOGREPS-G forecasts are the most consistent from 1200 UTC 22 August onward, with relatively small changes between successive forecasts. The GEFS forecasts maintain the initial right-hand bias for several successive forecasts, with a notable jump between 0000 and 1200 UTC 21 August. There is a second noticeable jump between 1200 UTC 23 August and 0000 UTC 24 August, after which the GEFS forecasts are generally close to the observed position, although with a small left bias. It is also worth noting that both MOGREPS-G and GEFS track Laura in all members for forecasts initialized from 1200 UTC 23 August onward, while the ECMWF ensemble does not, even for the shorter ranges. The three centers use different tracking algorithms, and this suggests differences in the sensitivity and robustness of the different trackers (Conroy et al. 2023).

This example was chosen to illustrate jumpiness in the ECMWF ENS, and in particular the flip-flops between successive forecasts. Comparison with the other centers shows that

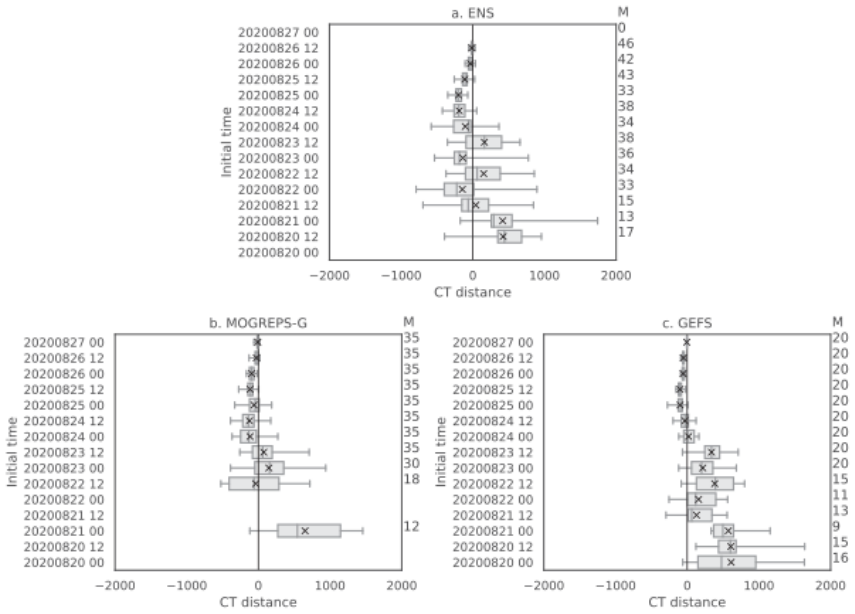


FIG. 2. Jumpiness of ensemble forecasts for hurricane Laura, valid at 0000 UTC 27 Aug 2020. Each boxplot summarizes the distribution of the cross-track (CT) errors (error at right angles to the observed direction of travel; negative values indicate left-of-track error) for one ensemble forecast (distance measured in km). Forecasts started every 12 h from 1200 UTC 20 Aug; the *y* axis shows the forecast initial time. The box-and-whisker plot shows the min, max and 25th, 50th, and 75th percentiles of the ensemble distribution (number of members shown to right of plot). The ensemble mean is shown as *X*. (a) ECMWF ENS, (b) Met Office MOGREPS-G, and (c) NCEP GEFS.

this was not a feature common to all centers. The ENS jumpiness may be related to possible issues with the data assimilation or initial perturbations, but further work is needed to investigate this (Magnusson et al. 2021). Alternatively, this could be just a chance occurrence due to the limited number of ensemble members. For each of the initial times before 25 August, 20%–30% of the ENS members did not track Laura as far as the verification time of 0000 UTC 27 August. In some cases, especially for initial times on 24 and 25 August, the ECMWF tracker misassigned some of the later forecast steps to hurricane Marco. However, this does not account for the majority of the missing tracks. These may be related to difficulties in initializing the cyclone due to the land interactions as Laura passed Puerto Rico, Hispaniola, and Cuba, while at earlier initial times, Laura was a relatively weak tropical storm and there was relatively large uncertainty in the initial analyzed position (Magnusson et al. 2021). We have recomputed the results including the corrected misassigned tracks and confirmed that this does not affect any of our conclusions.

How typical is this Laura case? To investigate how often such jumpy cases occur and whether jumpiness tends to occur for the same or different cases in different ensemble systems, the following sections consider the run-to-run consistency over all Atlantic tropical cyclones from 2019 to 2021.

b. *Ensemble jumpiness 2019–21*

To summarize the run-to-run inconsistency for a single case, we use the mean divergence $\overline{D}$ and DI, both computed over all forecasts verifying at a given time for a given tropical cyclone. The mean divergence $\overline{D}$ measures the overall change in each sequence of forecasts, while DI accounts for the trend over the sequence and highlights any flip-flop behavior.

Figure 3 shows the distribution of $\overline{D}$ and DI over all available cases for Atlantic tropical cyclones from 2019 to 2021 for the ENS, MOGREPS-G and GEFS ensembles. For $\overline{D}$, ENS has the lowest median value and smallest interquartile range, while the distribution for GEFS is noticeably broader than for the other centers. The difference between the distributions of GEFS and the other centers are statistically significant at the 1% level for both the KS and MWU tests. Although much closer to each other, the difference between ENS and MOGREPS-G distributions is significant at the 5% level for MWU test (but not significant for KS). For DI, GEFS also has the broadest distribution and ENS has the narrowest distribution. The difference between MOGREPS-G and GEFS is not statistically significant. ENS is significantly different from both MOGREPS-G and GEFS at the 5% level.

In general, a larger ensemble should give a more robust representation of the predicted distribution while a smaller

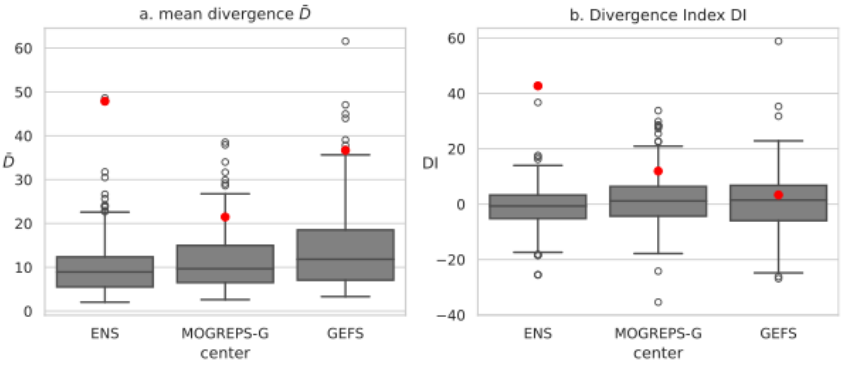


FIG. 3. Run-to-run inconsistency (jumpiness) of ensemble forecasts for Atlantic tropical cyclone tracks (2019–21). Boxplots show the distribution over all cases for the two divergence-based measures: (a) mean divergence ($\overline{D}$) and (b) divergence index (DI). Boxplots show the interquartile range and the median; the whiskers indicate the minimum and maximum values that are within 1.5 times the interquartile range; any more extreme points are shown with open circles as outliers. For both $\overline{D}$ and DI, larger positive values indicate the most inconsistent cases. The points for the example case of Hurricane Laura shown in Figs. 1 and 2 (verification time at 0000 UTC 27 Aug 2020) are marked as red filled circles.

ensemble will be more susceptible to sampling uncertainties and therefore may be expected to jump more from run to run. The above results are therefore consistent with the GEFS ensemble having fewer members than the other centers, especially before the upgrade to 31 members in September 2020. However, other factors can also influence the run-to-run consistency of the ensemble. For example, a lack of spread due to underrepresentation of either initial condition or model uncertainties would also tend to make the ensemble more jumpy. The impact of the upgrade is considered in the next subsection.

High positive values indicate the most inconsistent cases for both $\overline{D}$ and DI. For each center, points that are more than 1.5 times the interquartile range above the upper quartile are classed as outliers (marked with open circles in Fig. 3). The example case for Hurricane Laura discussed in the previous section is highlighted—this is an extreme outlier for ENS for both measures, highlighting the unusually large jumpiness for this case.

For MOGREPS-G and GEFS, this case was not an outlier for DI, consistent with the absence of flip-flops that characterized the ENS forecasts. Although not the most extreme case, this case was an outlier for GEFS using the $\overline{D}$ measure. This was due to the large right bias in the earlier GEFS forecasts. This example illustrates the difference between $\overline{D}$ and DI: ENS had several flip-flops between successive forecasts, while changes between GEFS forecasts were more associated with a trend away from the initial right bias. Both centers had large mean divergence $\overline{D}$, but the underlying cause was different. MOGREPS-G was more consistent than the other centers.

We have seen that while Laura was an example of extreme jumpiness for ENS, this was not such an extreme case for the other centers, especially for DI. Scatterplots of $\overline{D}$ and DI for pairs of centers (Fig. 4) show that this is a typical example. For each pair of centers, the number of cases that are outliers (high positive values, the most inconsistent cases) for either one center or both centers are indicated in the figure. The dashed lines in the figures indicate the threshold used for the outliers (1.5 times the interquartile range above the upper quartile). The jumpiest cases (high positive DI) for one center are in general not extremes for the other centers. For DI, none of the other ENS outliers are also outliers for either of the other centers. The results are similar for the outliers from MOGREPS-G and GEFS. There is only one case which is an outlier for more than one center, MOGREPS-G and GEFS, but that case is not an outlier for ENS. For $\overline{D}$, the highlighted Laura case is unusual in that it has high $\overline{D}$ for both ENS and GEFS, although the cause is different for each center as discussed above. However, more typically the cases of high $\overline{D}$ for one center are not exceptional for the other centers. In the scatterplots, the outliers with high $\overline{D}$ tend to lie away from the diagonal so that there are substantially more cases in the upper-left and lower-right quadrants than in the upper right.

These results suggest that the ensemble jumpiness is not strongly linked to the atmospheric situation or to the availability of observations. Rather, they suggest that individual model deficiencies or sampling uncertainties are more likely causes for the jumpiness. Sampling uncertainties will lead to run-to-run jumpiness if the ensemble is not large enough to fully represent the distribution of possible outcomes; a larger ensemble would better sample this underlying distribution and improve consistency from run to run. Alternatively, an ensemble may fail to properly represent the range of possible outcomes because the perturbations to initial conditions are not adequate or because the uncertainties in the model formulation are not sufficiently represented. Either of these will result in the ensemble spread being too small and may lead to jumpy behavior.

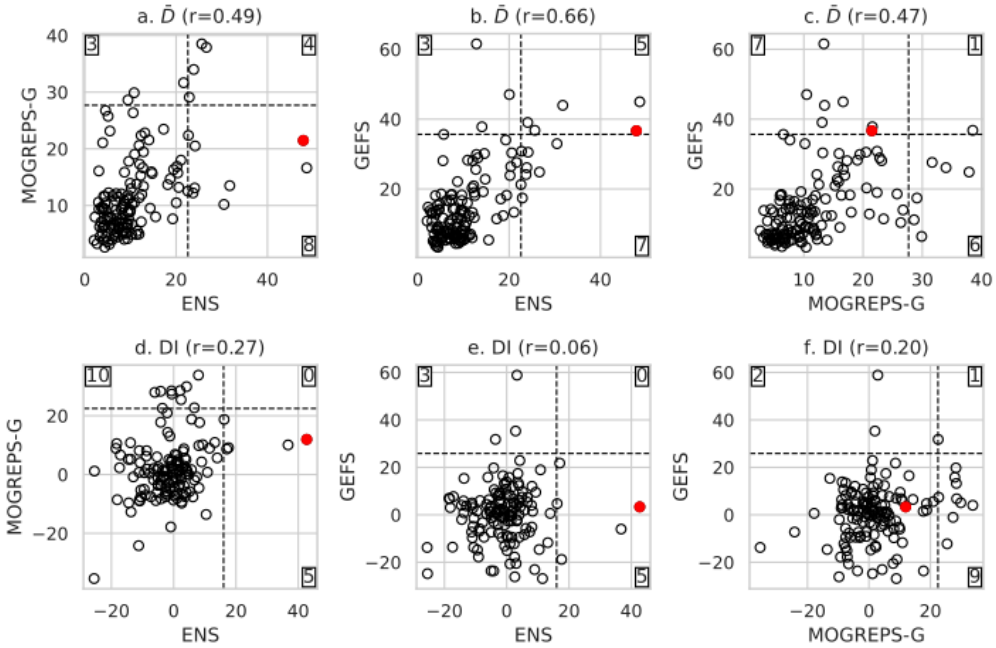


FIG. 4. Comparison of jumpiness between different centers' ensemble forecasts for Atlantic tropical cyclone tracks (2019–21). Scatterplots show the distribution of the two divergence-based measures: (top) mean divergence ($\overline{D}$) and (bottom) divergence index (DI) over all cases for pairs of centers. For both $\overline{D}$ and DI, larger positive values indicate the most inconsistent cases. Dashed lines mark the threshold for the most inconsistent outliers (1.5 times the interquartile range above the upper quartile). In each panel, the number of cases that are outliers for both centers or just one of the centers is indicated in the corresponding quadrant. The points for the example case of Hurricane Laura shown in Figs. 1 and 2 (verification time at 0000 UTC 27 Aug 2020) are marked as red filled circles.

c. *The effect of recent NWP system upgrades on ensemble jumpiness*

The results of the previous section showed that overall GEFS was more jumpy than the other centers. The GEFS upgrade in September 2020 was the most substantial upgrade of any of the centers during the study period, including a new forecast model, changes to the ensemble perturbations and an increase in the number of ensemble members. It brought a substantial improvement in the spread of tropical cyclone track forecasts (Zhou et al. 2022). Here we consider the impact of the upgrade on the jumpiness of ensemble track forecasts.

We separate our sample into two subsets initialized before (64 cases) and after (75 cases) the GEFS upgrade. In Fig. 5 we compare the empirical cumulative distribution of the mean divergence $\overline{D}$ for the three centers before (Fig. 5a) and after (Fig. 5b) the upgrade. Overall, $\overline{D}$ is significantly lower after the upgrade (comparing Figs. 5a,b). However, this applies also to the results from the other centers, suggesting that the difference is at least partly due to the differences between the observed samples. To mitigate this sampling effect, we focus on the difference between the GEFS ensemble and the other centers for the two subsets of cases.

Before the upgrade, the GEFS had substantially more cases with high values of $\overline{D}$ compared to ENS and MOGREPS (Fig. 5a). The difference in distribution compared to the other centers is highly significant at well below the 1% level for both KS and MWU tests. Differences in the distributions for ENS and MOGREPS-G are not statistically significant. After the upgrade, the GEFS distribution was much closer to those of the other centers (Fig. 5b) and there were no statistically significant differences between the distributions of any of the centers. These results show that the upgrade to the GEFS did make a significant difference to the consistency in terms of mean divergence $\overline{D}$. As for the full sample, differences in the distributions of DI are smaller (not shown); the only statistically significant difference between GEFS and either of the other centers is with ENS before the GEFS upgrade.

The GEFS upgrade brought a substantial improvement in the spread of tropical cyclone track forecasts. This was considerably underdispersive in the previous version and the upgrade resulted in a much better spread–error relationship, due to the upgrade to the stochastic model perturbations (Zhou et al. 2022). The change in $\overline{D}$ is consistent with this increase in spread for the GEFS system. In general, a larger

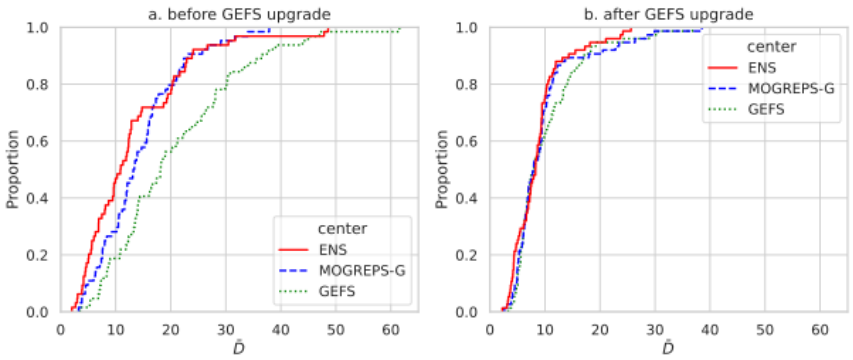


FIG. 5. Effect of GEFS v12 cycle upgrade, 23 Sep 2020. Empirical cumulative distribution function of $\overline{D}$ for subsamples of cases (a) before and (b) after the upgrade.

spread will give a broader distribution of tropical cyclone positions and the change between the set of positions for successive forecasts would tend to be less than for a less dispersive ensemble. For the same reason, the improved spread might also be expected to affect DI. Although there was some indication of this in our results (the ENS and GEFS distributions were closer and not significantly different after the upgrade), it was not such a clear change as for $\overline{D}$.

It is possible that additional factors as well as the increased spread also helped to improve $\overline{D}$. For example, a reduction in cross-track bias in the longer-lead forecasts would help to reduce $\overline{D}$, but would not tend to affect DI. Leonardo and Colle (2021) showed that the GEFS had larger cross-track errors than ENS in a large sample of Atlantic tropical cyclones for 2008–16. We were not able to identify any significant changes in the GEFS bias after the upgrade in our sample of cases. While the change in ensemble spread was large enough to identify in our sample, it may be that other differences require larger samples. Leonardo and Colle (2021) also noted that large year-to-year variability made it difficult to identify any changes due to model upgrades.

The MOGREPS-G upgrade in December 2019 also improved TC track errors and spread (Met Office 2019; Titley et al. 2020). Taking the same approach as above we found that for the subset of cases before the MOGREPS-G upgrade there was a significant difference between the ENS and MOGREPS-G distributions for both $\overline{D}$ and DI (with the MOGREPS-G having overall higher jumpiness). After the upgrade there was no significant difference between the two centers. See Fig. S1 in the online supplemental material.

We conclude that the recent upgrades to the MOGREPS-G and GEFS systems both improved the run-to-run consistency of the ensemble track forecasts, and that since these upgrades the overall jumpiness is similar for the three ensemble systems.

d. Comparison of error, spread, and divergence

We now compare the mean scores over all cases for the three different aspects of ensemble performance: error, spread and divergence. The upper panel of Fig. 6 shows the ensemble error (CRPS, left), divergence (D, center) and spread (s, right) at lead times out to 5 days ahead for the three centers. The vertical bars indicate the bootstrapped 95% confidence intervals for each center's scores. Overall, the three centers have similar performance and most differences between scores are not statistically significant.

The larger divergences in the short range for ENS and GEFS (Fig. 6b) are consistent with the lower spread (Fig. 6c) at these time steps for these centers. MOGREPS-G has larger initial spread (maybe partly due to the time-lagging of the initial conditions of the MOGREPS-G system), and this will tend to reduce the difference (divergence) between consecutive forecasts as seen in Fig. 6b.

For each center, the mean ensemble divergence (Fig. 6b) is approximately equal to the mean difference in CRPS between consecutive forecasts (difference between successive points on the curves in Fig. 6a). The agreement is particularly strong at short range for all centers, and for ENS at all forecast ranges. In other words, on average the divergence gives an indication of the expected change in error for the next forecast. However, this does not apply in individual cases.

Table 1 shows the Pearson correlation between divergence and CRPS across all available cases for each forecast lead time. For comparison, the correlation between ensemble spread and CRPS is also shown. Corresponding scatterplots are shown in Figs. S2–S5 in the online supplemental material. The association between divergence and error is in general substantially weaker than the link between spread and error. These results are consistent with previous studies that show the benefit of using spread as a measure of forecast uncertainty (Majumdar and Finocchio 2010; Yamaguchi et al. 2009; Kawabata and Yamaguchi 2020; Titley et al. 2019). However, the low correlation for divergence suggests that it does not provide useful case-to-case guidance: there is no indication that users should expect less jumpy cases to be more skillful.

Table 2 shows the Pearson correlation over all cases between the two overall measures, $\overline{D}$ and DI, and the corresponding mean error over all forecast lead times $\overline{\text{CRPS}}$. Although for $\overline{D}$ the correlation is somewhat higher than for the individual forecast steps (Table 1), the corresponding scatterplots show large

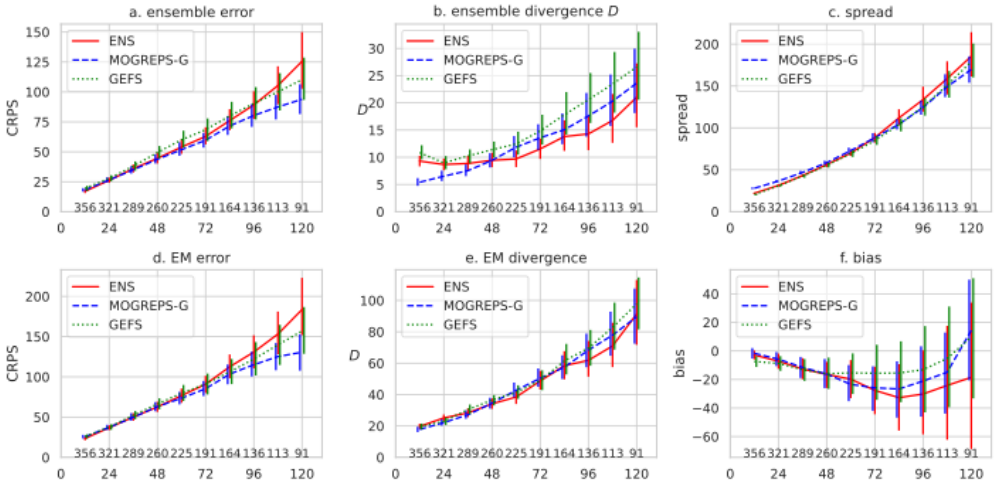


FIG. 6. Error, spread, and divergence for forecast lead time from 12 to 120 h. Scores for the (top) full ensemble and (bottom) corresponding error and divergence for the ensemble means. (a),(d) CRPS error; (b),(e) divergence; (c) ensemble spread; and (f) bias. Vertical bars indicate 95% confidence intervals. Mean scores over all available cases for each forecast lead time: number of cases indicated above the x axis.

variations in error for cases of both low and high $\overline{D}$. This again suggests that users should be cautious in individual cases—a consistent case with relatively low jumpiness may still have large overall error.

We can do the same analysis for the ensemble-mean forecasts, which are often used in operational TC forecasting (Figs. 6d,e; lower panel). Again, the divergence gives useful additional information for forecast users. For example, for ENS the ensemble mean cross-track error is around 175 km for 120-h forecasts (Fig. 6d), and the ensemble spread is similar (showing that the ensemble system is overall well-tuned; Fig. 6c). The mean expected change in cross-track EM position between $T + 120$ and $T + 108$ is ~80 km (Fig. 6e). This is similar for all three centers.

The forecast systematic error (bias) is shown in Fig. 6f. Overall, each center has a negative bias, that is the forecast positions tend to be to the left of the observed position. However, there is large uncertainty as indicated by the large confidence intervals shown on the plot. Magnusson et al. (2021) show that the ENS tends to have a left-of-track bias for northward-moving TCs, but a right-of track bias for westward moving systems and this situation-dependent variation in bias may partly explain the large confidence intervals at longer lead times. As for the other scores, the confidence intervals indicate that there is no significant difference between the biases of the different centers. Comparing Figs. 6d and 6f shows that for all centers the bias is relatively small compared to the total error.

TABLE 1. Correlation between divergence and error. Each row shows the correlation between the CRPS error at a given forecast lead time h and the divergence D between h- and $(h + 12)$-h forecasts. For comparison the correlation between the CRPS and the ensemble spread for the h-h forecasts is shown in parentheses.

Step (h)	ENS	MOGREPS-G	GEFS
72	0.18 (0.45)	0.22 (0.38)	0.07 (0.29)
84	0.25 (0.56)	0.32 (0.47)	0.05 (0.27)
96	0.19 (0.58)	0.36 (0.47)	−0.01 (0.32)
108	0.29 (0.67)	0.42 (0.41)	0.19 (0.44)

5. Conclusions

We have carried out an investigation of the jumpiness or run-to-run consistency of ensemble forecasts of tropical cyclone tracks. We used ensemble forecasts from the TIGGE tropical cyclone track archive for three global centers: ECMWF (ENS), Met Office (MOGREPS-G), and NCEP (GEFS). The forecasts were compared to the observed tracks for all named tropical cyclones from the IBTrACS archive for the Atlantic basin for 2019, 2020, and 2021.

We looked at the change in the distribution of cross-track position (relative to the observed track) for tropical cyclones in consecutive ensemble forecasts initialized at 12-h intervals.

TABLE 2. Correlation between overall jumpiness and error ($\overline{\text{CRPS}}$).

Center	$\overline{D}$ vs $\overline{\text{CRPS}}$	DI vs $\overline{\text{CRPS}}$
ENS	0.54	−0.30
MOGREPS-G	0.56	−0.01
GEFS	0.67	−0.30

This was quantified using the divergence function D associated with the CRPS error score following Richardson et al. (2020). The overall jumpiness of a sequence of forecasts all verifying at the same time was summarized using the mean divergence $\overline{D}$ and the divergence index (DI).

We present our conclusions in the framework of the questions posed in the introduction.

a. How does run-to-run jumpiness vary from case to case and between the ensemble systems of different NWP centers?

The distribution of DI was similar for each center, showing substantial variation between centers with a few significant outliers. There was no strong agreement between the centers on which cases were most jumpy. The case shown for Hurricane Laura was a typical example: this was the most extreme case of jumpiness (largest DI) for the ECMWF ENS, showing a clear flip-flopping of the ensemble between being left and right of the observed track in successive forecasts. This behavior was not apparent in either the MOGREPS or GEFS ensembles. This case also illustrated the difference between the two summary measures $\overline{D}$ and DI. Earlier GEFS forecasts were substantially to the right of the observed track and this right-of-track bias decreased in later forecasts. The large trend over successive forecasts is indicated in the relatively high mean divergence. However, the absence of the flip-flop behavior seen in the ECMWF ENS results in the DI being close to the overall median value. Using the combination of both $\overline{D}$ and DI can help to distinguish these different behaviors in a sequence of forecasts.

b. Is there a common cause of "jumpy" cases—Are the ensembles from different centers particularly jumpy for the same cases and if so, what is the reason?

The jumpiest cases were different for each center for both $\overline{D}$ and DI, indicating that there is not a common cause of jumpiness across the different ensemble systems. This suggests that the ensemble jumpiness is not strongly related to the prevailing atmospheric conditions or to the available observations.

Outliers for the different centers may be due more to specific issues in the data assimilation, models or ensemble configurations. Recent studies highlight both continuing progress and ongoing challenges in each of these areas (e.g., Magnusson et al. 2019, 2021). However, a deeper analysis of outliers would require a substantially larger sample than we have used and is beyond the scope of the present work. Leonardo and Colle (2021) used 9 years (2008–16) of Atlantic TC data to investigate the causes of large cross-track errors in the GEFS and ENS. However, we have also seen that recent upgrades to ensemble systems have led to a significant reduction in the ensemble jumpiness and therefore including a longer sample of earlier years may not be representative of the current ensemble capabilities.

Another possible reason for the occasional cases of large jumpiness is sampling uncertainty due to finite ensemble size. This would be consistent with outliers occurring at different times for the different centers. Richardson (2001) showed how even a well-tuned ensemble will appear unreliable if it has insufficient members and that the required number of ensemble members depends on both the underlying distribution and the needs of the users. Leutbecher (2019) and Craig et al. (2022) have demonstrated substantial sensitivity to ensemble size in studies using large ensembles of 200 members and 1000 members, respectively. Kondo and Miyoshi (2019) suggest that up to 1000 ensemble members are necessary to represent important aspects of some forecast distributions. The impact of ensemble size on forecast jumpiness has not been investigated and is a topic for future work.

c. Have recent ensemble model upgrades had a noticeable effect on the forecast jumpiness?

In this study we used a 3-yr period to provide a sufficient number of cases to assess. During this period upgrades to both the MOGREPS-G and GEFS ensembles resulted in substantial improvements to their predictions of TC tracks. Using the ECMWF ENS as a reference, we found that both these upgrades significantly reduced the jumpiness of the ensembles. Before the upgrades the ENS was significantly less jumpy than the other centers. However, after the upgrades there was no significant different between the centers. Both upgrades increased the spread of the ensembles, and the improved jumpiness is consistent with this change. These results suggest that it is the overall level of ensemble spread that is important and that differences in initialization and perturbation methodology between the current systems are not a major factor in determining the overall level of ensemble jumpiness.

The more recent upgrade to the ENS at the end of 2021 improved TC track errors by 10% but had little impact on the overall spread (Haiden et al. 2022). This improved the statistical reliability of the TC track. The impact on jumpiness of this upgrade has not been assessed but can be done once a sufficient sample of cases is available.

d. What guidance should be provided to forecasters and decision-makers on the ensemble jumpiness—What information is practically useful? Is there any useful link between jumpiness and skill?

The divergence D gives an indication of the expected change in cross-track position from one forecast to the next. For example, a user should expect on average that the ensemble mean position will change by around 80–90 km in the cross-track direction between a forecast for 120 h ahead and the 108-h forecast for the same time made 12 h later. The expected change between a 72- and 60-h forecast is around 50 km. These expected changes were similar for all three centers. Corresponding values for the expected divergence for the full ensemble distributions are 20–25 and 10–15 km, respectively. These results address the user requirements identified for example by Regnier and Harr (2006) and Jewson et al. (2022) to provide objective measures of the expected change from run to run so that users can take account of this in their decision-making.

We did not find any strong link between either $\overline{D}$ or DI and error (CRPS). This indicates that users should not rely on the jumpiness or consistency between successive forecasts as measure of confidence in the forecasts. This is consistent with the

work of Zsoter et al. (2009) who found only a weak link between jumpiness and error in ensemble forecasts for Europe. In contrast, ensemble spread and the ensemble probabilistic information (e.g., strike probabilities) have been shown to provide useful situation-dependent guidance on forecast uncertainty (Majumdar and Finocchio 2010; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020).

Although we note that the effect of more recent system upgrades has not yet been evaluated, users should expect generally similar levels of jumpiness in the three ensemble systems considered in this study. The jumpiest cases will tend to be different for the different centers, likely to be a result of sampling uncertainties or specific deficiencies in the individual ensemble configurations.

One practical approach for users to adopt to address both these potential sources of jumpiness would be to combine the ensemble forecasts from the different centers into multimodel ensembles. Such multimodel combinations have already been shown to improve probabilistic TC track prediction (Yamaguchi et al. 2012; Leonardo and Colle 2017; Titley et al. 2020; Kawabata and Yamaguchi 2020). Another option would be to use lagged ensembles, combining consecutive forecasts from one center. By construction this will reduce jumpiness and this is already used in the MOGREPS-G system to increase ensemble size. Although our aim in this study was to evaluate and compare the jumpiness in the individual systems, the effect of multimodel combinations on ensemble jumpiness is an area for future work.

Acknowledgments. This work is based on TIGGE data. The International Grand Global Ensemble (TIGGE) is an initiative of the World Weather Research Programme (WWRP). David Richardson is supported by a Wilkie Calvert Ph.D. studentship at the University of Reading. We thank Linus Magnusson, Sharanya Majumdar, and two anonymous reviewers for their valuable comments.

Data availability statement. The forecast data used in this study are available from The International Grand Global Ensemble (TIGGE) Model Tropical Cyclone Track Data, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory at https://doi.org/10.5065/D6GH9GSZ (Bougeault et al. 2010; Swinbank et al. 2016). The observed tropical cyclone tracks are available from NOAA's International Best Track Archive for Climate Stewardship (IBTrACS) archive at https://doi.org/10.25921/82ty-9e16 (Knapp et al 2010, 2018).

REFERENCES

Bidlot, J.-R., F. Prates, R. Ribas, A. Mueller-Quintino, M. Crepulja, and F. Vitart, 2020: Enhancing tropical cyclone wind forecasts. *ECMWF Newsletter*, No. 164, ECMWF, Reading, United Kingdom, 33–37, https://www.ecmwf.int/en/elibrary/81182-enhancing-tropical-cyclone-wind-forecasts.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Broad, K., A. Leiserowitz, J. Weinkle, and M. Steketee, 2007: Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bull. Amer. Meteor. Soc.*, **88**, 651–668, https://doi.org/10.1175/BAMS-88-5-651.

Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, https://doi.org/10.1002/asl.170.

Cangialosi, J. P., 2022: National Hurricane Center forecast verification report: 2021 hurricane season. NOAA/National Hurricane Center Rep., 76 pp., https://www.nhc.noaa.gov/verification/pdfs/Verification_2021.pdf.

Conroy, A., and Coauthors, 2023: Track forecast: Operational capability and new techniques—Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10). *Trop. Cyclone Res. Rev.*, **12**, 64–80, https://doi.org/10.1016/j.tcrr.2023.05.002.

Craig, G. C., M. Puh, C. Keil, K. Tempest, T. Necker, J. Ruiz, M. Weissmann, and T. Miyoshi, 2022: Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quart. J. Roy. Meteor. Soc.*, **148**, 2325–2343, https://doi.org/10.1002/qj.4305.

Ehret, U., 2010: Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts. *Meteor. Z.*, **19**, 441–451, https://doi.org/10.1127/0941-2948/2010/0480.

Elsberry, R. L., and P. H. Dobos, 1990: Time consistency of track prediction aids for western North Pacific tropical cyclones. *Mon. Wea. Rev.*, **118**, 746–754, https://doi.org/10.1175/1520-0493(1990)118<0746:TCOTPA>2.0.CO;2.

Fowler, T. L., B. G. Brown, J. H. Gotway, and P. Kucera, 2015: Spare change: Evaluating revised forecasts. *MAUSAM*, **66**, 635–644, https://doi.org/10.54302/mausam.v66i3.572.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

Griffiths, D., M. Foley, I. Ioannou, and T. Leeuwenburg, 2019: Flip-flop index: Quantifying revision stability for fixed-event forecasts. *Meteor. Appl.*, **26**, 30–35, https://doi.org/10.1002/met.1732.

Haiden, T., M. Janousek, F. Vitart, L. Ferranti, and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. 853, 56 pp., https://doi.org/10.21957/mlvapkke.

——, ——, ——, Z. Ben-Bouallegue, L. Ferranti, C. Prates, and D. Richardson, 2021: Evaluation of ECMWF forecasts, including the 2020 upgrade. ECMWF Tech. Memo. 880, 56 pp., https://doi.org/10.21957/6njp8byz4.

——, ——, ——, ——, ——, F. Prates, and D. Richardson, 2022: Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Tech. Memo. 902, 56 pp., https://doi.org/10.21957/xqnu5o3p.

Heming, J. T., and Coauthors, 2019: Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. *Trop. Cyclone Res. Rev.*, **8**, 181–218, https://doi.org/10.1016/j.tcrr.2020.01.001.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hewson, T., 2020: Use and verification of ECMWF products in member and co-operating states (2019). ECMWF Tech. Memo. 860, 42 pp., https://doi.org/10.21957/80s471ib1.

Inverarity, G. W., and Coauthors, 2023: Met Office MOGREPS-G initialisation using an Ensemble of Hybrid Four-Dimensional Ensemble Variational (En-4DEnVar) data assimilations. *Quart. J. Roy. Meteor. Soc.*, **149**, 1138–1164, https://doi.org/10.1002/qj.4431.

Jewson, S., S. Scher, and G. Messori, 2021: Decide now or wait for the next forecast? Testing a decision framework using real forecasts and observations. *Mon. Wea. Rev.*, **149**, 1637–1650, https://doi.org/10.1175/MWR-D-20-0392.1.

——, ——, and ——, 2022: Communicating properties of changes in lagged weather forecasts. *Wea. Forecasting*, **37**, 125–142, https://doi.org/10.1175/WAF-D-21-0086.1.

Kawabata, Y., and M. Yamaguchi, 2020: Probability ellipse for tropical cyclone track forecasts with multiple ensembles. *J. Meteor. Soc. Japan*, **98**, 821–833, https://doi.org/10.2151/jmsj.2020-042.

Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Amer. Meteor. Soc.*, **91**, 363–376, https://doi.org/10.1175/2009BAMS2755.1.

——, H. J. Diamond, J. P. Kossin, M. C. Kruk, and C. J. Schreck III, 2018: International Best Track Archive for Climate Stewardship (IBTrACS) Project, version 4. NOAA/National Centers for Environmental Information, accessed 26 May 2022, https://doi.org/10.25921/82ty-9e16.

Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: A study with a 10240-member ensemble Kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Processes Geophys.*, **26**, 211–225, https://doi.org/10.5194/npg-26-211-2019.

Leonardo, N. M., and B. A. Colle, 2017: Verification of multimodel ensemble forecasts of North Atlantic tropical cyclones. *Wea. Forecasting*, **32**, 2083–2101, https://doi.org/10.1175/WAF-D-17-0058.1.

——, and ——, 2021: An investigation of large cross-track errors in North Atlantic tropical cyclones in the GEFS and ECMWF ensembles. *Mon. Wea. Rev.*, **149**, 395–417, https://doi.org/10.1175/MWR-D-20-0035.1.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145** (Suppl. 1), 107–128, https://doi.org/10.1002/qj.3387.

Magnusson, L., and Coauthors, 2019: ECMWF activities for improved hurricane forecasts. *Bull. Amer. Meteor. Soc.*, **100**, 445–458, https://doi.org/10.1175/BAMS-D-18-0044.1.

——, and Coauthors, 2021: Tropical cyclone activities at ECMWF. ECMWF Tech. Memo. 888, 140 pp., https://www.ecmwf.int/sites/default/files/elibrary/2021/20228-tropical-cyclone-activities-ecmwf.pdf.

Majumdar, S. J., and P. M. Finocchio, 2010: On the ability of global ensemble prediction systems to predict tropical cyclone track probabilities. *Wea. Forecasting*, **25**, 659–680, https://doi.org/10.1175/2009WAF2222327.1.

McLay, J. G., 2011: Diagnosing the relative impact of "sneaks," "phantoms," and volatility in sequences of lagged ensemble probability forecasts with a simple dynamic decision model. *Mon. Wea. Rev.*, **139**, 387–402, https://doi.org/10.1175/2010MWR3449.1.

Met Office, 2019: Parallel Suite 43 release notes. Met Office, accessed 18 May 2023, https://www.metoffice.gov.uk/services/data/met-office-data-for-reuse/ps43_ftp.

Pappenberger, F., K. Bogner, F. Wetterhall, Y. He, H. L. Cloke, and J. Thielen, 2011a: Forecast convergence score: A forecaster's approach to analysing hydro-meteorological forecast systems. *Adv. Geosci.*, **29**, 27–32, https://doi.org/10.5194/adgeo-29-27-2011.

——, H. L. Cloke, A. Persson, and D. Demeritt, 2011b: HESS opinions "On forecast (in)consistency in a hydro-meteorological chain: Curse or blessing?" *Hydrol. Earth Syst. Sci.*, **15**, 2391–2400, https://doi.org/10.5194/hess-15-2391-2011.

Regnier, E., and P. A. Harr, 2006: A dynamic decision model applied to hurricane landfall. *Wea. Forecasting*, **21**, 764–780, https://doi.org/10.1175/WAF958.1.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, https://doi.org/10.1002/qj.49712757715.

——, H. L. Cloke, and F. Pappenberger, 2020: Evaluation of the consistency of ECMWF ensemble forecasts. *Geophys. Res. Lett.*, **47**, e2020GL087934, https://doi.org/10.1029/2020GL087934.

Rodwell, M. J., and Coauthors, 2021: IFS upgrade provides more skilful ensemble forecasts. *ECMWF Newsletter*, No. 168, ECMWF, Reading, United Kingdom, 18–23, https://www.ecmwf.int/sites/default/files/elibrary/2021/20115-ifs-upgrade-provides-more-skilful-ensemble-forecasts.pdf.

Ruth, D. P., B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The performance of MOS in the digital age. *Wea. Forecasting*, **24**, 504–519, https://doi.org/10.1175/2008WAF2222158.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertainty Quantif.*, **1**, 522–534, https://doi.org/10.1137/130907550.

Titley, H. A., M. Yamaguchi, and L. Magnusson, 2019: Current and potential use of ensemble forecasts in operational TC forecasting: Results from a global forecaster survey. *Trop. Cyclone Res. Rev.*, **8**, 166–180, https://doi.org/10.1016/j.tcrr.2019.10.005.

——, R. L. Bowyer, and H. L. Cloke, 2020: A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Quart. J. Roy. Meteor. Soc.*, **146**, 531–545, https://doi.org/10.1002/qj.3712.

Walters, D., and Coauthors, 2019: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations. *Geosci. Model Dev.*, **12**, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences.* 4th ed. Elsevier, 840 pp.

Yamaguchi, M., R. Sakai, M. Kyoda, T. Komori, and T. Kadowaki, 2009: Typhoon ensemble prediction system developed at the Japan Meteorological Agency. *Mon. Wea. Rev.*, **137**, 2592–2604, https://doi.org/10.1175/2009MWR2697.1.

——, T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quart. J. Roy. Meteor. Soc.*, **138**, 2019–2029, https://doi.org/10.1002/qj.1937.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zsoter, E., R. Buizza, and D. Richardson, 2009: "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Mon. Wea. Rev.*, **137**, 3823–3836, https://doi.org/10.1175/2009MWR2960.1.

Supplemental Material

Weather and Forecasting
Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks
https://doi.org/10.1175/WAF-D-23-0113.1

Jumpiness in ensemble forecasts of Atlantic tropical cyclone tracks

David S. Richardson,[a,b] Hannah L. Cloke,[a,c,d] John A. Methven,[c] Florian Pappenberger,[b]

[a] *Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom*

[b] *ECMWF, Reading, United Kingdom*

[c] *Department of Meteorology, University of Reading, Reading, United Kingdom*

[d] *Department of Earth Sciences, Uppsala University, Uppsala, Sweden*

Corresponding author: David S. Richardson, d.s.richardson@pgr.reading.ac.uk

Supplementary material

This supplementary material provides 5 additional figures to complement the results shown in the paper.

Figure S1 shows the impact of the MOGREPS-G upgrade in December 2019 on the mean divergence $\overline{D}$ and Divergence Index DI. Before the upgrade, the differences between the empirical cumulative distributions for ENS and MOGREPS-G are statistically significant at the 5% level ($p < 0.03$) for $\overline{D}$ and at the 1% level ($p < 0.005$) for DI using both the Kolmogorov-Smirnov and Mann-Whitney U tests. After the upgrade there was no significant difference between ENS and MOGREPS-G.

Figures S2-S5 show scatter plots of divergence against error and of spread against error for lead times of 72, 84, 96 and 108 hours. These figures complement the correlations shown in Table 1.

File generated with AMS Word template 2.0

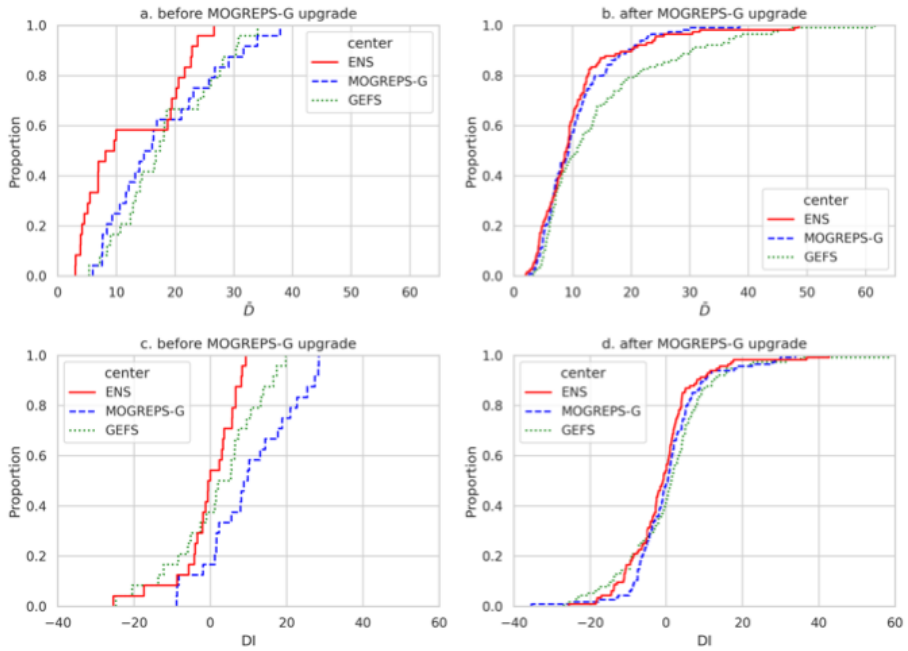


Fig. S1. Effect of MOGREPS-G cycle upgrade, 4 December 2019. Top row: empirical cumulative distribution function of $\bar{D}$ for subsamples of cases (a) before and (b) after the upgrade. Bottom row: empirical cumulative distribution function of DI for subsamples of cases (c) before and (d) after the upgrade.

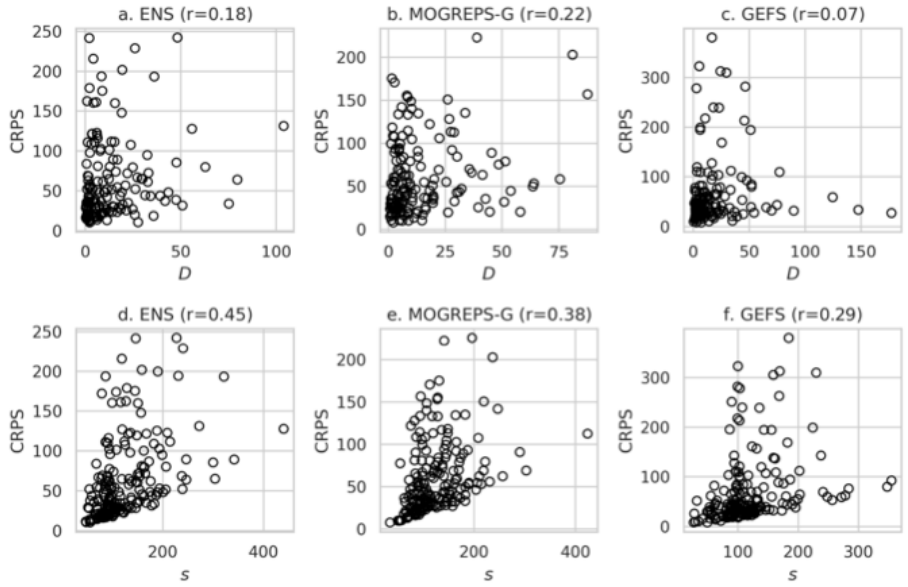


Fig. S2. Correlation between divergence and error (top row) and between spread and error (bottom row). On the top row, each panel shows a scatter plot over all cases of the CRPS error at a forecast lead time of 72 hours and the divergence D between 72-hour and 84-hour forecasts for a: ENS, b: MOGREPS-G, and c: GEFS . For comparison the correlation between the CRPS and the ensemble spread s for the 72-hour forecast is shown in the panel below. The Pearson correlation coefficient r for each sample is shown in the title of each panel.

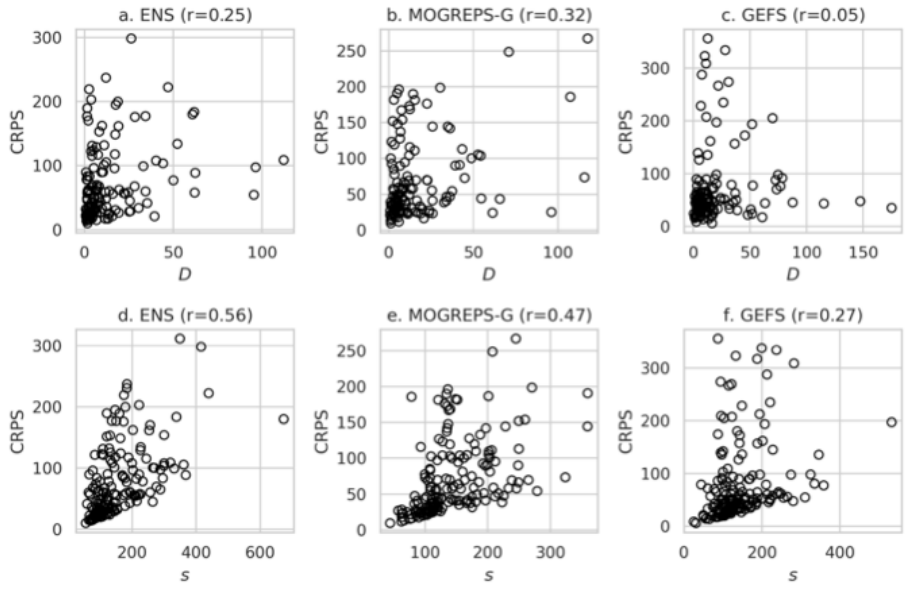


Fig. S3. As Fig. S2 for 84-hour forecast.

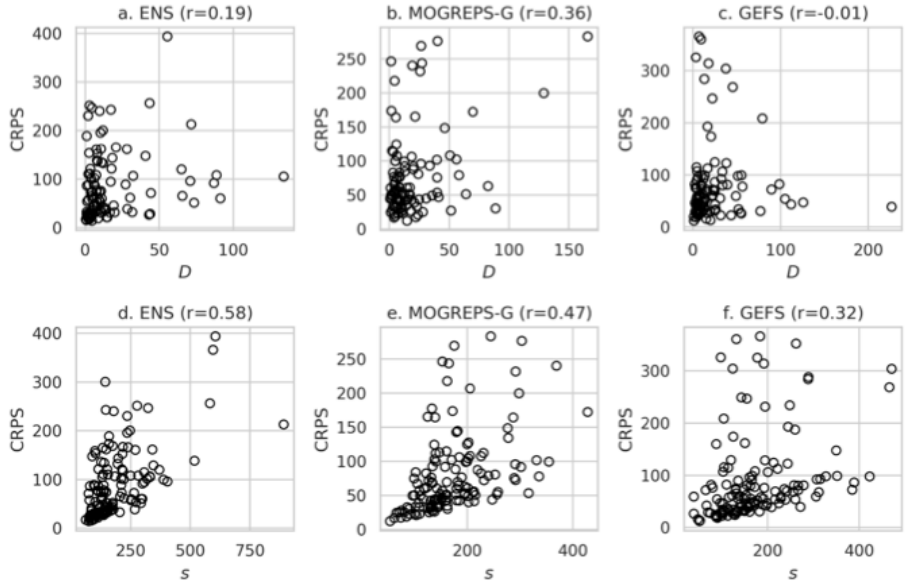


Fig. S4. As Fig. S2 for 96-hour forecast.

File generated with AMS Word template 2.0

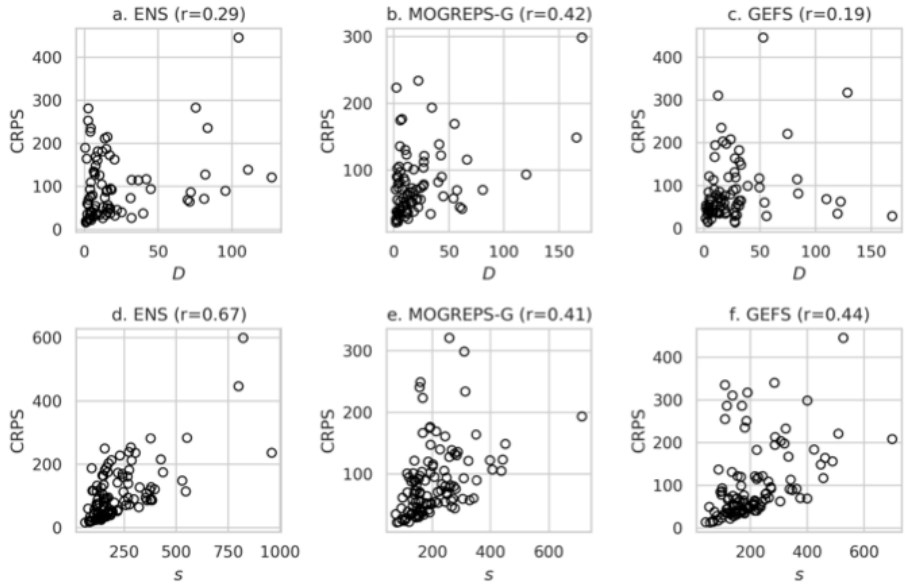


Fig. S5. As Fig. S2 for 108-hour forecast.

5

File generated with AMS Word template 2.0

A3. Submitted Article: Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

This appendix contains the formatted version of chapter 5 of this thesis that was submitted to Weather and Forecasting, with the following reference:

Richardson, D.S., Cloke, H.L., Magnusson, L., Majumdar., S. J., Methven, J.A. and Pappenberger, F. (2024) 'Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis', *Weather and Forecasting* (resubmitted November 2024 following revison)

Skill and consistency of ECMWF forecasts of Atlantic tropical cyclone genesis

David S. Richardson,[a,b] Hannah L. Cloke,[a,c] Linus Magnusson,[b] Sharanya J. Majumdar[d], John A. Methven,[c] Florian Pappenberger[b]

[a] *Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom*

[b] *ECMWF, Reading, United Kingdom*

[c] *Department of Meteorology, University of Reading, Reading, United Kingdom*

[d] *Department of Atmospheric Sciences, University of Miami, Miami, Florida*

Corresponding author: David S. Richardson, d.s.richardson@pgr.reading.ac.uk

ABSTRACT

We evaluate the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting tropical cyclone genesis in the Atlantic basin. Focusing on the probabilistic performance of the ENS, we assess how far in advance the ENS can predict genesis, quantify the consistency (jumpiness) from run to run and investigate what factors influence the skill and consistency. We find that first indications of genesis are picked up at least 7 days ahead in 50% of the observed cases, although strong signals often only appear less than 3 days before genesis. There are significant regional differences, with observed genesis events predicted 2-3 days earlier in the eastern Atlantic than in other areas. The genesis probabilities can be jumpy from run to run and the jumpiest cases are in the more skilful regions (central and eastern Atlantic) and for situations where the initial signal for genesis appears at longer lead time. In the eastern Atlantic, there is a tendency for the ENS tracks to reach tropical storm strength earlier and further east than observed; this model bias can affect both skill and jumpiness of the genesis forecasts. Our results provide guidance to forecasters on how to use and interpret the ENS predictions. Areas for future work include the link between early intensification in the eastern Atlantic and African easterly wave activity, the relationship between skill and the TC development pathways, and the impact of systematic analysis differences between 0000 UTC and 1200 UTC on forecast intensity.

SIGNIFICANCE STATEMENT

Forecasting where and when tropical cyclones will appear increases the lead time at which decision makers can begin to take preparatory mitigating action. Numerical weather prediction models can provide important guidance, but sometimes are not consistent from one run to the next. We evaluate the skill and consistency of a state-of-the-art global model in predicting the formation of tropical cyclones up to ten days ahead and provide guidance to forecasters on how to use and interpret the model predictions. We show that the formation of tropical cyclones can be predicted 2-3 days earlier in the eastern Atlantic than in the western Atlantic and identify some of the factors influencing both skill and consistency.

1. Introduction

Following significant progress in forecasting tropical cyclone (TC) tracks (Landsea and Cangialosi 2018) and intensity (Cangialosi et al. 2020), there is increasing focus on predicting TC genesis (Hon et al. 2023). For the Atlantic basin, the US National Hurricane Center (NHC) Tropical Weather Outlook provides forecasts of TC genesis for 2 and 7 days ahead (Hon et al. 2023). By providing information about the likely development of TCs before they have formed, skillful genesis forecasts can effectively increase the lead time at which decision makers can begin to take preparatory mitigating action.

Numerical weather prediction (NWP) forecasts including ensemble forecasts are used in operational genesis forecasts (Titley et al. 2019; Hon et al. 2023), often in combination with statistical methods (Halperin et al. 2017). Use and verification of NWP genesis forecasts has focused on deterministic aspects, assessing hits and false alarms using standard contingency-table measures such as hit rate or probability of detection, success ratio, and the threat score or critical success index (Wilks 2020). These have been applied to the high-resolution global forecasts from different centres (Halperin et al. 2016, 2013; Liang et al. 2021) to ensemble mean forecasts (Li et al. 2016; Wang et al. 2018) and to individual ensemble members (Zhang et al. 2022).

Recently there has been increasing development of probabilistic TC genesis forecast products for operational centres (Hon et al. 2023). For example, Halperin et al. (2017) developed a statistical–dynamical tool to generate TC genesis probabilities using logistic regression models applied to the outputs from several high-resolution global NWP models. A consensus probability is also provided when more than one model predicts a genesis event. Verification using Brier scores and reliability diagrams showed that these provide useful guidance (Halperin et al. 2017), and the products are regularly used in the NHC (Hon et al. 2023). The use of probabilistic information from the ensembles is more limited, although ensemble forecasts have been shown to have skill in predicting TC genesis (Komaromi and Majumdar 2014, 2015; Majumdar and Torn 2014; Yamaguchi and Koide 2017; Yamaguchi et al. 2015).

One of the key issues limiting the uptake of ensemble TC forecasts is the run-to-run jumpiness that can occur in some situations (Dunion et al. 2023; Magnusson et al. 2021). Large jumps in the predicted probability of TC genesis between successive ensemble forecasts present a significant challenge to forecast centres and lessen users' confidence in the prediction system (McLay 2008; Elsberry and Dobos 1990; Hewson 2020; Dunion et al.

2023; Pappenberger et al. 2011). Although approaches such as multi-model combinations or lagged ensembles can help mitigate such jumpiness, it is important to identify and understand the underlying causes of such jumpy behaviour. Quantifying the level of jumpiness in an ensemble system provides valuable information to the forecast user. This can be important for example in helping the user to decide between acting now or waiting for the next forecast (Regnier and Harr 2006; Jewson et al. 2022, 2021). Identifying the circumstances in which jumpiness occurs is an important step towards addressing the underlying cause – is it related to model or analysis uncertainty (lack of spread in the ensemble perturbations) or model bias, or is it an indication of insufficient ensemble size to give a reliable uncertainty estimate? Jumpiness of TC track forecasts has been investigated for the western North Pacific (Elsberry and Dobos 1990) and the Atlantic (Fowler et al. 2015; Richardson et al. 2024). However there has been no corresponding assessment of TC genesis forecasts. In this study we conduct a first assessment of the jumpiness of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (ENS) forecasts for TC genesis.

Another factor limiting the use of ensemble TC genesis forecasts is the lack of routine evaluation of the products provided by the global centers. Although ECMWF regularly publishes verification results for ensemble forecasts of the track and intensity of existing TCs (Haiden et al. 2023), it does not routinely evaluate genesis forecasts, so users do not have a clear picture of ENS performance (Magnusson et al. 2021).

These knowledge gaps are addressed in this study which evaluates the skill and jumpiness of the ECMWF medium-range ensemble (ENS) in predicting TC genesis in the Atlantic basin. We address the following questions:

- How far in advance can the ENS forecast TC genesis the Atlantic basin?
- How consistent from run to run are the forecasts of the observed genesis events?
- What are the factors that influence the skill and consistency of the ENS genesis forecasts and what future work will help to improve these forecasts?

In each case, we focus on the probabilistic performance of the ENS. The data we use in this study and the methods we apply to identify genesis events are described in section 2, with verification scores and consistency measures introduced in section 3. Results are presented in section 4, addressing each of the three key questions in turn. We conclude with a summary and discussion of directions for future work in section 5.

2. Data

We investigate the ability of the ECMWF ensemble (ENS) to predict the genesis of tropical cyclones over the Atlantic. ENS comprises 50 perturbed members integrated on ~18km grid until 27 June 2023 and thereafter on ~9km grid. The ECMWF tropical cyclone tracker (Magnusson et al. 2021) identifies and tracks both existing TCs and those that develop during the forecast. The tracker is applied to all ensemble members. These operational forecast tracks are archived on the TIGGE database (Bougeault et al. 2010; Swinbank et al. 2016). We retrieve the operational forecast tracks for ENS forecasts initialized at 0000 and 1200 UTC from May to December 2019-2023 and consider forecast lead times from one to ten days ahead.

We evaluate the forecasts against the observed TC data from the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2018, 2010). We extract the observed positions and maximum winds from all named Atlantic tropical storms (i.e. tropical cyclones that reach tropical storm strength during their life cycle). We focus our evaluation on the first time the observed system is reported as a tropical system of at least tropical storm strength (winds at least 34 kt; 1kt ~ 0.51 m s^{-1}), which we define as the genesis time for the tropical storm (TS) (Magnusson et al. 2021; Zhang et al. 2022). To ensure a consistent set of forecast lead times throughout the evaluation, we limit the verification times to also be 0000 and 1200 UTC and so the observed genesis time is the first 0000 or 1200 UTC time with wind >17 m s^{-1}. There were 98 observed tropical storms in the Atlantic basin during the 5-year study period. However, TS Imelda (2019) was a TS for less than 12 h and was not included in the verification, therefore we used 97 observed TS genesis in this work.

To investigate how well and how consistently the ENS can forecast the observed TS genesis events, we compute the probability of TS genesis or TS activity at the observed genesis time and location for each of the 97 observed TS.

For a given verification time t_v, we refer to an ensemble forecast f valid for this time and initialized h hours earlier as $f(t_v, h)$ and write the individual ensemble members as $f_m(t_v, h)$. Given the inherent limitations of predictability as well as uncertainties in both forecast and observations (Landsea and Franklin 2013; Torn and Snyder 2012), we do not expect the forecast to predict genesis at exactly the time and location of the reported observed genesis event. Therefore, we define tolerances in both space and time. Several different choices have been used in previous studies (Halperin et al. 2016, 2013; Zhang et al. 2022; Magnusson et al.

2021; Yamaguchi et al. 2015). For each observed TS genesis event, we use the following procedure where t_v represents the observed genesis time:

- For the ENS forecast $f(t_v, h)$ we count how many members m have TC tracks that pass within 500 km of the observed genesis location at any time between $t_v - 24$ h and $t_v + 24$ h. We define the proportion of members m/M as the forecast probability of TC activity at the observed genesis event. This gives the probability for TC but does not address the intensity or the location of genesis in the forecast. We refer to this set of forecast probabilities as FATC.
- To address the intensity, we select the subset of the forecast tracks that have maximum wind greater than a given threshold. We use 17 m s^{-1} for a direct comparison with the observed intensity, but also consider lower thresholds (e.g., 15 m s^{-1}) to account for potential differences in intensity in the forecasts. We refer to these forecast activity probabilities as FA17 and FA15, respectively.
- Finally, to address the timing of the genesis we again subset the forecast tracks to keep only those that have forecast genesis within 24 h and 500 km of the observed genesis event. We define the forecast genesis event as the first point on the track with wind greater than 17 m s^{-1} and refer to this set of forecast probabilities as FG17.

Table 1 summarizes the different sets of forecast probabilities that we consider in this study and the naming convention that we use.

Identifier	set	description
FG17	Forecast TS genesis 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and the first time that wind is >17 m s^{-1} along this track is within this time/location tolerance
FA17	Forecast TS activity 17 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind >17 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind >17 m s^{-1} may have occurred

		more than 24 h before t_v) and more than 500 km from the given location.
FA15	Forecast TC activity 15 m s^{-1}	Forecast TC track passes within 500 km and 24 h of given location and has wind >15 m s^{-1}. But forecast genesis may have occurred earlier (i.e., first step with wind >15 m s^{-1} may have occurred more than 24 h before t_v) and more than 500 km from the given location. Accounts for overall lower intensity in forecasts
FATC	Forecast TC activity	Forecast TC track passes within 500 km and 24 h of given location (forecast wind may not reach TS strength)

160 Table 1. Different forecast sets considered in this study. Identifier used to refer to each set
161 of forecast probabilities.

162

163 For a broader perspective, to consider the overall forecast probabilities of TC genesis and
164 to include assessment of false alarms, we also conduct some evaluation on a regular 1°x1°
165 latitude-longitude grid. At each grid point, the forecast TS genesis probability is defined as
166 the proportion of ENS members that predict a TS genesis event to occur within 500km of that
167 grid point (center of the 1°x1° box) and between 24h and 216h ahead. Similarly, we define
168 TS genesis to occur if there is an observed TS genesis event within 500 km of the grid point
169 and within the same 192 h (8-day) time window.

170

171 **3. Verification and consistency measures**

172 We evaluate the ENS forecasts of TC activity and genesis using the Brier score (Wilks
173 2020) which is a measure of the mean squared error of the forecast probability:

174 $$\mathrm{b} = \frac{1}{N}\sum_{i=1}^{N}(p_n - y_n)^2 \qquad (1)$$

175

where p_n is the forecast probability (proportion of ENS members that predict the event), y_n is 1 if the event occurs and 0 otherwise and N is the total number of cases.

In the assessment of overall performance using the gridded data (section 4d), we use the observed sample climate probability of genesis $\bar{y}$ as a reference forecast:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2}$$

This sample climate includes all dates in our evaluation data and is computed separately for each grid point. By construction, the sample climate has the lowest Brier score of any fixed reference forecast and so is harder to beat than a long-term climate; using this as a reference for the Brier skill score hence provides a conservative indication of forecast skill.

The Brier score of the climate forecast is given as

$$\mathrm{b_c} = \frac{1}{N}\sum_{i=1}^{N} (\bar{y} - y_i)^2 \tag{3}$$

and the Brier skill score is then given as

$$\mathrm{B} = \frac{b_c - b}{b_c} \tag{4}$$

Positive values of B indicate positive skill relative to the sample climate. Maximum skill $B = 1$ is achieved for perfect deterministic forecasts.

We evaluate the hits and false alarms associated with different forecast probability thresholds using the ROC (Mason 1982; Ben Bouallègue and Richardson 2022) and performance diagram (Roebber 2009). The ROC is a plot of the hit rate (proportion of observed events correctly forecast) against false alarm rate (proportion of observed non-events where genesis was forecast). The performance diagram plots the hit rate against the success ratio (proportion of genesis forecasts that were correct); the performance diagram also shows the frequency bias (number of forecast events divided by number of observed events) and the threat score (number of hits divided by the sum of hits, misses and false alarms).

To measure the jumpiness or consistency over a sequence of forecasts we measure the difference (divergence) d in probability between consecutive forecasts.

Here, we consider the forecasts initialized at 12 h intervals between 24 h and 216 h before a given verification time t_v. The probability of the given event (TC activity or TS genesis) in the ENS forecast initialized at $t_v - h$ is written as $p(t_v, t_v - h)$ and the difference between consecutive forecasts is

$$D(t_v, h) = d\big(f(t_v, h), f(t_v, h-12)\big) = |p(t_v, h) - p(t_v, h-12)| \quad (5)$$

The mean divergence over the full sequence of L = 17 initial times is

$$\overline{D(t_v)} = \frac{1}{L-1}\left(\sum_{l=1}^{L-1} D(t_v, 24+12l)\right) \quad (6)$$

The minimum value of $\overline{D}$ is zero, indicating that the forecast probability does not change over the set of forecasts, while larger values indicate greater differences in probability between successive forecasts in the sequence.

For each observed genesis event, we expect that the forecast probability will be low at the longest forecast ranges (close to the climatological probability) and will increase, ideally reaching close to 100% at the shortest forecast ranges. To account for the expected increase in probability over the sequence of forecasts, we use the difference between the probabilities from the first and last forecasts of the sequence to represent this overall trend. We then subtract this difference from $\overline{D}$ to give the Divergence Index (DI; (Richardson et al. 2020, 2024).

$$\mathrm{DI}(t_v) = \overline{D(t_v)} - \frac{1}{L-1}\,|p(t_v, 24+12(L-1)) - p(t_v, 24)| \quad (7)$$

DI summarizes the jumpiness about the overall trend over the sequence of forecasts, with larger values of DI indicating more jumpy forecasts (bigger difference in probabilities).

4. Results

Firstly, we evaluate how far in advance the ENS can predict the observed genesis events with low, medium, and high probability. Next, we assess how consistent these probabilities are in the sequence of consecutive forecasts leading up to each observed genesis event. We then consider potential factors that may affect the jumpiness and skill of these forecasts. Finally, we assess the overall skill of the ENS probability forecasts for TC genesis and activity.

a. How far in advance can we predict the observed Atlantic TS genesis events?

Figure 1 shows the percentage of the 97 observed genesis events that were forecast with at least 5%, 35% and 65% probabilities at or before each forecast lead time from 216 to 24 hours in advance. The probability thresholds were chosen to be consistent with the categories used to indicate low, medium and high probability respectively in the NHC Tropical Weather Outlook: NHC genesis probabilities are given in 10% intervals and their low, medium and high probability categories are 10-30%, 40-60% and 70-100% respectively.

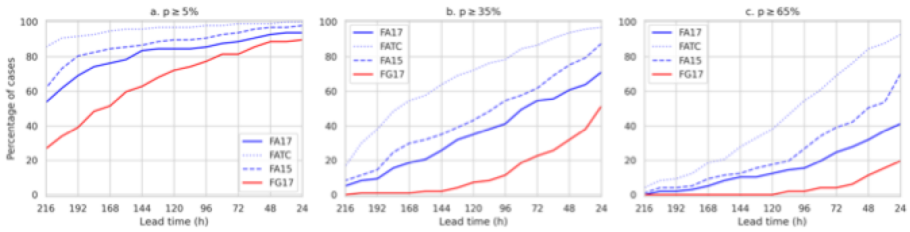


Fig. 1. Lead time of ENS forecasts of TS genesis. The percentage of cases predicted with probability of at least (a) 5% (low), (b) 35% (medium) and (c) 65% (high) at lead times from 216 to 24 h before the observed TS genesis time.

The red curve shows the results for the FG17 probabilities where the forecast is required to match the observed genesis in both timing and intensity (within the specified 500 km and 24 h tolerances). Few cases are predicted with high probability and only 20% of cases can be predicted with medium probability more than 72 h ahead. The low probability threshold is reached in over 50% of cases at 168 h lead time, indicating that the ENS is capable of generating tropical storms a week in advance although the predictability is low.

The three blue curves in Fig. 1 help to identify some of the reasons for this poor performance in the direct forecasting of the observed genesis. The solid blue curve shows the results for the FA17 probabilities. As well as the hits included in FG17, these allow for early genesis in the forecasts and indicate the proportion of ENS members that have TS activity at the observed genesis time and location. Many more cases are predicted for all three probability categories for FA17 than for FG17: more than 20% of observed events are predicted with high probability at least 72 h ahead, with the proportion increasing to over 50% for the medium probability threshold and over 80% for low probability. 25% of cases are predicted

with medium probability at least 6 days (144 h) ahead. The higher probabilities for FA17 compared to FG17 show that the timing of TS genesis is one significant difference between ENS and observed genesis, with a substantial number of forecast tracks reaching TS strength before the observed genesis time. Comparing FA17 and FA15 (solid and dashed blue lines) shows that the choice of wind threshold for the forecast tracks also affects the performance. The relatively minor change of wind threshold from 17 to 15 m s^{-1} increases the proportion of correctly forecast cases by around 10 percentage points. Larger improvements are achieved when considering all forecast tracks without specifying a minimum wind speed (FATC, dotted blue line): around 60% of cases are predicted with medium probability at least 6 days (144 h) ahead and with high probability at least 4 days (96 h ahead). The sensitivity to wind thresholds agrees with results from other studies (Yamaguchi et al. 2015; Zhang et al. 2022).

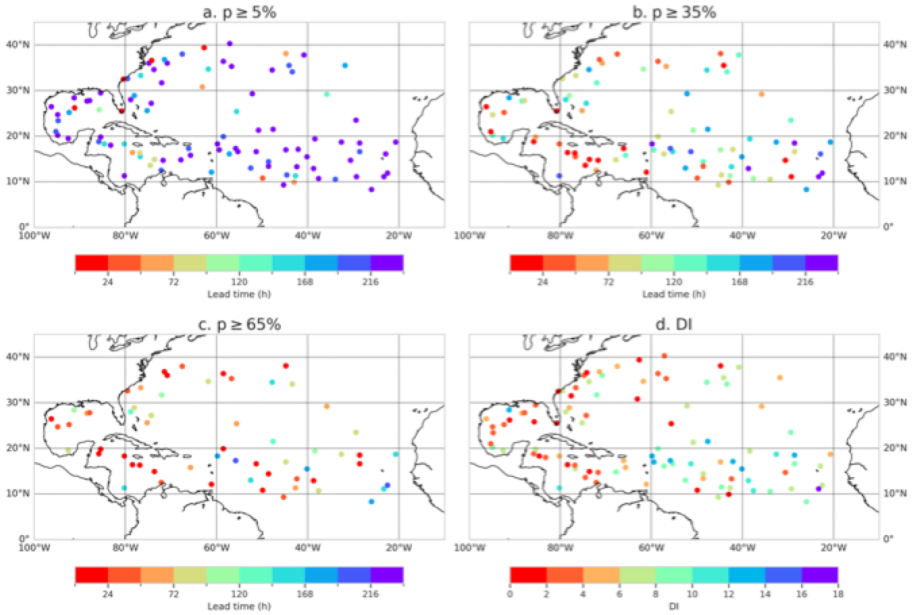


Fig. 2. Lead time of ENS forecasts of observed TS genesis events. Longest lead time (hours) at which the probability of TS activity (FA17) at the observed TS genesis location was predicted with probability of at least (a) 5%, (b) 35% and (c) 65%. (d) shows the jumpiness in forecast probability for these cases, as measured by DI.

The geographical distribution of the FA17 results is shown in Fig. 2 for each of the low/medium/high probability thresholds. The TS in the eastern Atlantic tend to be predicted earlier than those in the Caribbean and the Gulf of Mexico. In the central and eastern Atlantic (east of 60°W and south of 30°N) the median lead time for the first indications of TS activity (low 5% probability threshold) is 228 h (the longest lead we have considered here). For medium and high probability thresholds the corresponding median lead times are 132 h and 72 h, respectively. In contrast, the equivalent median lead times for the western Atlantic, Caribbean and Gulf of Mexico (west of 60°W, south of 30°N) are 204, 48 and 36 h respectively. In other words, the observed genesis events in the eastern Atlantic are predicted 2-3 days earlier than those further west. The predictability for the genesis >30°N is generally similar to that for the western Atlantic. The consistency or jumpiness of these forecasts as measured by DI is shown in Fig. 2d. Again, there are strong geographical variations, with the highest DI (jumpiest cases) in the central and east Atlantic. The median DI for this region is 8.75, more than twice the median DI value of the western and northern regions (3.5 and 4.0 respectively).

The regional differences may be associated with different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008). The more predictable (and also more jumpy) cases tend to occur in regions dominated by non-baroclinic developments, although some of the most predictable and jumpiest genesis events occur in the Cape Verde region associated with the low-level baroclinic pathway (baroclinic development under the African easterly jet). The less predictable cases further west and north are in regions where other baroclinic pathways (tropical transition TT (Davis and Bosart 2003, 2004); trough interaction) are more common developments. This is consistent with results from Wang et al. (2018) who found lower predictability in the TT pathways in an evaluation of reforecasts from the NCEP GEFS ensemble. It is however notable that there are very few predictable cases in the Caribbean and Gulf of Mexico despite the non-baroclinic pathway also being a significant development category in this region. These non-baroclinic pathways often originate from barotropic breakdown of vorticity along stalled fronts, which are smaller and could be less predictable, especially for a lower-resolution model. Environmental factors influencing TC genesis in the western Atlantic have been discussed by Klotzbach et al. (2022) and (in the wider context of cyclonic circulations over central America) by Papin et al. (2017). Additional factors, such as land interactions, may also affect the model ability to correctly predict genesis and would

have a more significant impact on genesis forecasts in the western Atlantic and Caribbean rather than eastern Atlantic; this is an area for future research.

b. Consistency – the jumpiest forecasts of observed TS genesis events

The run-to-run consistency of the ENS forecast probabilities is shown in Fig. 3 for the 12 cases with highest DI for the FA17 forecasts. For each case, the forecast probabilities from the forecasts initialized every 12 hours from 24 to 216 h before the observed genesis event are shown for each forecast set FG17, FA17, FA15, FATC.

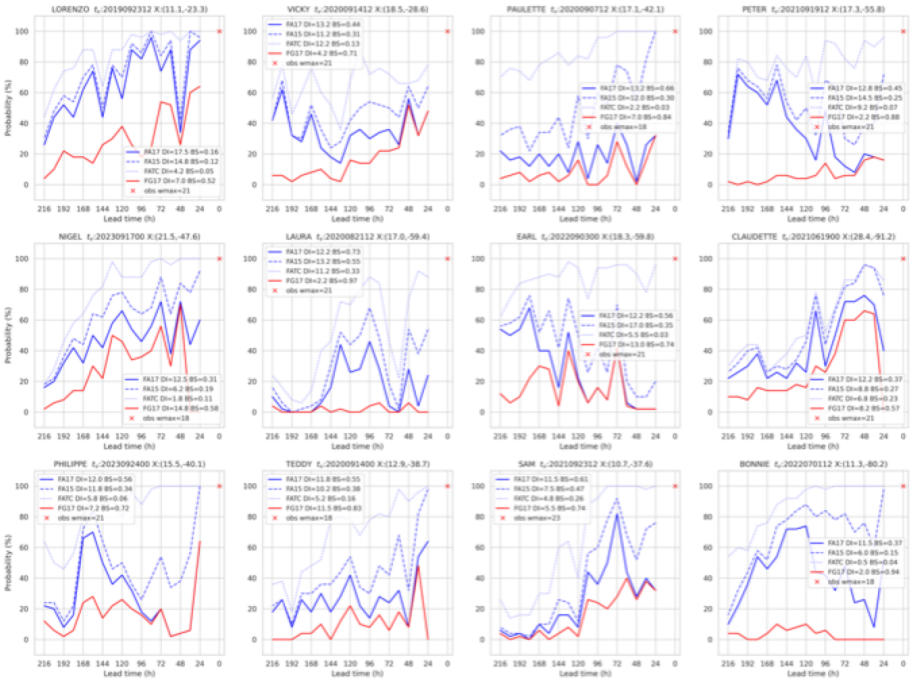


Fig. 3. Forecast probability of TS activity for the jumpiest FA17 cases. Curves show the forecast probability of TC activity at the observed genesis time (t_v) and location X (latitude, longitude) for forecasts initialized at 12h intervals from 216 to 24 h before the observed genesis time. The probability for genesis (FG17) is shown by the red line, while the three blue curves show the probability of TC activity with different wind intensity thresholds FA17 (solid dark blue), FA15 (dashed blue), and FATC (dotted light blue). The legend shows the jumpiness (DI) and error (Brier score, BS) for each.

Most of these jumpy cases occur in September (August for Laura) and there are cases for each of the five years in our sample. As seen from Fig. 2, the jumpy cases are typically in the central to east Atlantic and between 10°N and 20°N. The two exceptions to both time and location are Bonnie and Claudette which were both early season TCs in the west of the basin. Claudette was the only one of these cases that did not originate from an African easterly wave.

In most cases the jumpiness is related to the forecast intensity: the FATC probabilities are much more consistent from run to run than the FA17 probabilities, and the corresponding DI is consequently much lower. The two notable exceptions to this are Laura and Vicky, which both have substantial jumpiness for the lower wind thresholds. Interactions between African easterly waves or between these waves and other low-pressure systems have also been noted to affect the forecast probabilities of genesis for cases including Laura and Paulette (Magnusson et al. 2021). In the case of Vicky, we note that Teddy and Vicky originated from successive easterly waves that developed off the coast of Africa on 10 and 11 September 2020. The earlier ENS forecasts tended to favor a development associated with Vicky with tracks moving north-westwards away from the coast of Africa, while later forecasts produced more westward tracks associated with Teddy. This uncertainty about which would be the stronger development, together with potential interactions between the two, may account for the jumpiness seen in the predictions for both Vicky and Teddy.

A notable feature of several cases is the high probability for TS activity (FA17) at longer range that is not maintained in the following forecasts made closer to the observed genesis time. Peter, Earl, Philippe, and Bonnie all have high probability (>65%) at some time 5 or more days ahead, but then have much lower probabilities for later forecasts. However, in all these cases the probability for TC activity (FATC) remains consistently high (well above 65%).

The jumpiest case in this sample is hurricane Lorenzo. There is a clear flip-flopping in the FA17 probabilities between the forecasts started at 0000 UTC and those started at 1200 UTC: the forecasts from 1200 UTC tend to have lower probability for TS activity than the forecasts from 0000 UTC made 12 h earlier and later. This suggests some systematic difference between the analyses for 0000 and 1200 UTC that affects the forecast intensity. Similar flip-flops, though not as large or long-lasting can be seen in some other cases (e.g. Nigel, Paulette).

These cases illustrate a number of different behaviors in the run-to-run consistency of the forecasts. In the next section we consider some of the factors that may contribute to these distinctive characteristics.

c. Factors affecting forecast jumpiness and skill

In this section we consider three factors that may affect the forecast jumpiness results discussed in the previous section. We look at the effect of ensemble size, the issue of flip-flops between 0000 and 1200 UTC analysis times and finally consider the early genesis noted in all results and how this model bias may affect the results for both jumpiness and skill. Although a detailed analysis of causes is beyond the scope of the present study, the aim of this initial assessment is to identify avenues for further research.

1) THE EFFECT OF ENSEMBLE SIZE

We compute the forecast probabilities as the proportion of ensemble members that predict TC activity at a given time and location. How much does the finite ensemble size affect the jumpiness in these probabilities? In this section we use a simple idealized framework to illustrate sampling effects and show the levels of jumpiness that might be expected in an ensemble of 50 members.

Figure 4a shows four idealized examples of how the probability of a TC increases over a set of 17 consecutive forecasts (such as the sequences of forecasts initialized every 12 hours from 216 to 24 h before a given observed genesis time, as used in this study). For each set of probabilities, we generate an idealized M-member ensemble by drawing a random sample with the given probability p at each step (Bernoulli process such that each member is either 1, representing forecast of genesis or 0, indicating genesis not forecast) and then compute the DI for this sequence of 17 ensemble forecasts. We repeat this to generate 10000 cases and summarize the distribution of DI over these 10000 cases in Fig. 4b.

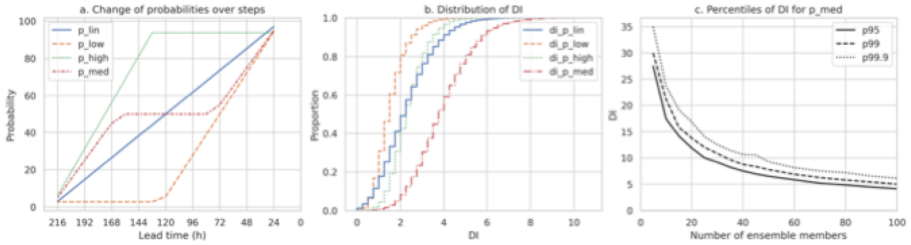


Fig. 4. Effect of ensemble size on forecast jumpiness. (a) 4 idealized examples of how the probability of TC genesis might evolve over a sequence of 17 50-member ensemble forecasts initialized e.g., every 12 hours from 216 to 24 h before a given verification time. (b) the empirical cumulative distribution of DI for each of the probability sets shown in (a) based on 10000 cases. (c) the effect of ensemble size (number of members) on the extreme percentiles (95% solid, 99% dashed, 99.9% dotted) of the DI distribution for the probability set leading to the jumpiest cases (p_med).

The four examples represent different predictability: linear increase in probability with forecast lead time (p_lin); a high predictability situation (p_high) in which the genesis event is forecast with high probability from five days ahead; a low predictability situation (p_low) where there is no signal at longer range and medium probability (35%) is reached only around 3-4 days ahead; and finally an intermediate situation (p_med) where the signal for genesis is captured with medium probability more than 7 days ahead, and this level of predictability is maintained until the probability increases again closer to the event.

The expected jumpiness for a 50-member ensemble varies depending on the underlying predictability (Fig. 4b). The low predictability situation is also the least jumpy of the four examples – when the probability of the event is low, there is little variability in the ensemble probability due to sampling (i.e. the finite ensemble size) and the jumpiness (DI) is also low. The intermediate predictability (p_med) situation is the jumpiest, with expected DI substantially higher than for the other examples. In general the sampling effects due to limited ensemble size are largest for probabilities close to 50%.

We have seen that the jumpiness of the ENS genesis forecasts is higher in the central and eastern Atlantic where the predictability is also higher than in other parts of the basin. This is consistent with the above results – the low predictability (p_low) situation is more typical in

404 the west of the basin, while the intermediate (p_med) is more representative of the central and
405 eastern Atlantic. Users should be aware that more predictable situations are likely to be more
406 jumpy because of sampling effects from the finite size of the ensemble.

407 For all four idealized distributions, the maximum DI is less than 10. In Section 4a we
408 noted that the median DI for the observed genesis events in the eastern Atlantic was 8.75.
409 This is much higher than would be expected from any of the idealized cases considered here.
410 While still high compared to these idealized results, the median DI in the other parts of the
411 Atlantic basin (3.5-4) is closer to the values suggested by these idealized cases.

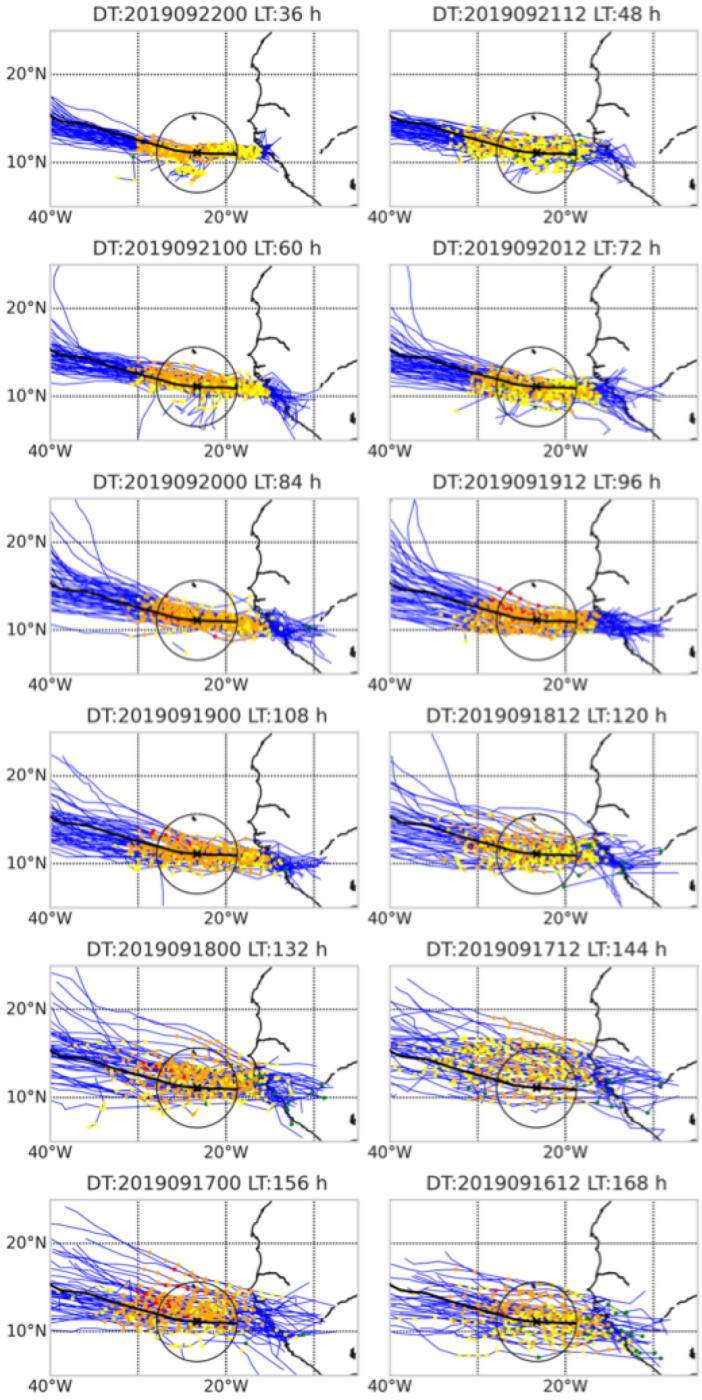
412 Figure 4c shows how the ensemble size affects the results for the probability distribution
413 that gives the jumpiest results overall (p_med, Fig. 4b). As noted above, for a 50-member
414 ensemble the probability of DI>10 is extremely small. However, for a 20-member ensemble
415 the chance of having DI>10 is not negligible: we should expect that more than 5% of cases
416 will have DI>10. In general sampling uncertainties will be larger for smaller ensembles (the
417 proportion of members predicting genesis will be a less reliable estimate of the true
418 underlying probability) and therefore the jumpiness from run to run will increase and more
419 cases should be expected with large DI. Conversely, there is a steady decrease in the chances
420 of high jumpiness as the ensemble size increases from 20 to 100 members: for a 100-member
421 ensemble, the maximum DI is not likely to be above 5.

422 Overall, these idealized results suggest that for the ENS and the set of observed cases
423 considered here, values of DI greater than 10 are unlikely to be due purely to ensemble size.
424 The high median value of DI (8.75) for the cases in the eastern Atlantic suggests there is a
425 substantial number of cases where factors other than pure sampling contribute to the
426 jumpiness.

427 However, it should be noted that if the ensemble is under-dispersive, the effective
428 ensemble size could be lower than the nominal 50 members and this could significantly affect
429 the DI. These idealized results also show that increasing ensemble size would be expected to
430 reduce overall jumpiness and improve the overall consistency of the ENS predictions. This
431 may be important for some decision-making applications (Jewson et al. 2022) such as
432 deciding when to plan and initiate evacuation from areas at potential risk (Regnier and Harr
433 2006) or rerouting of transportation to avoid adverse weather (McLay 2008).

434 2) ANALYSIS IMPACTS – FLIP-FLOP BETWEEN 00 AND 12 UTC INITIAL CONDITIONS

435 The case of Lorenzo demonstrated a marked jumpiness between the forecasts initialized
436 at 00 and at 12 UTC. Figure 5 shows the forecast tracks for Lorenzo initialized from 36 to
437 168 h before the observed genesis time. The circle indicates locations within 500 km of the
438 observed genesis location. The potential for TS activity is predicted at all lead times, and the
439 earliest forecast with high probability was initialized 7 days before the observed genesis time
440 (Fig. 3). Most of the forecast TCs intensify to TS strength very soon after the track leaves
441 land and moves over the sea off the African coast. This is generally earlier than the observed
442 genesis, consistent with the low probabilities shown in the FG17 curve in Fig. 3. A notable
443 feature of the forecast probabilities (both FA17 and FA15) is the long sequence of flip-flops
444 in the probabilities between successive forecasts: the forecasts started from 0000 UTC have
445 higher probability than those started 12 h earlier and 12 h later at 1200 UTC.
446



447 19

File generated with AMS Word template 2.0

448 Fig. 5. ENS forecasts for the genesis of Lorenzo, 1200 UTC 23 September 2019.
449 ECMWF ensemble forecast tracks (blue) and observed track (black). Forecast start dates
450 (DT) from 1200 UTC on 16 September to 0000 UTC on 22 September 2019 (LT: forecast
451 lead time in hours to observed genesis time). Colored symbols show forecast intensity
452 (maximum wind speed) at all times within 24 h of the observed genesis time (1200 UTC 21
453 September to 1200 UTC 23 September); Colors represent the maximum wind speed: yellow
454 (<17 m s^{-1}), orange (17-32 m s^{-1}), red (>32 m s^{-1}). Observed genesis location at 1200 UTC 23
455 September marked (x) and circle indicates locations within 500 km radius of this location.

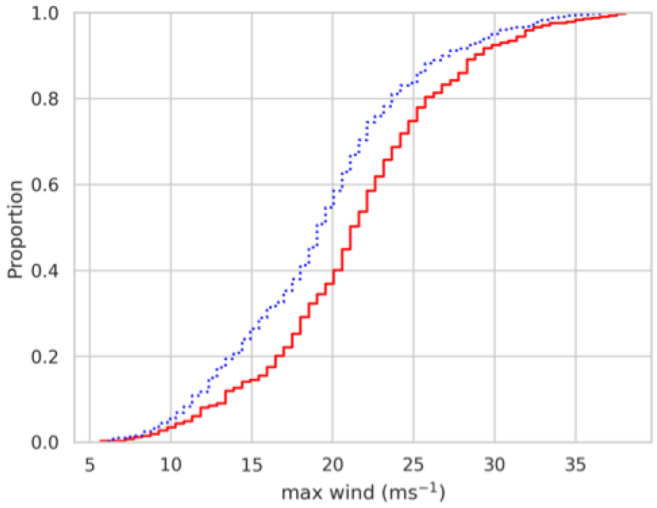
456

457 We extracted the maximum wind for each forecast TC position within 500 km and 24 h of
458 the observed genesis position and time of Lorenzo for all ENS forecasts started from 0000
459 UTC and compared the distribution of these winds with those from the forecasts started at
460 1200 UTC. There is a statistically significant shift towards stronger winds in the forecasts
461 from 0000 UTC analysis times (Fig. 6). This suggests that there is some systematic difference
462 in the assimilation at 0000 and 1200 UTC that affects the intensification of the forecasts in
463 this case. One possibility is the analysis over West Africa where a systematic difference in
464 analysis increments has been identified in the ECMWF assimilation system (Bormann et al.
465 2023). The reasons for this are not yet understood and are the subject of further investigation.

466 While some other cases in the same region also have some flip-flops between 0000 and
467 1200 UTC initial conditions, this is not a common occurrence. Therefore, while assimilation
468 differences may be one factor, it is likely that a combination of factors may be involved to
469 make the large and significant impact found in this Lorenzo case. Further evaluation of this
470 case is beyond the scope of this paper, but the results suggest that additional investigation
471 into the differences between 0000 and 1200 UTC analyses may be relevant.

472

473



474

475 Fig. 6. Sensitivity of TC intensity to analysis time in ENS forecasts for the genesis of
476 Lorenzo, 1200 UTC 23 September 2019. Empirical cumulative distribution functions of
477 maximum wind speed for ENS TC forecasts initialized at 0000 UTC (solid red line) and at
478 1200 UTC (dotted blue line) that are within 500 km and 24 h of the observed genesis event of
479 Lorenzo (at 11.1°N, 23.3°W). All forecast start dates between 0000 UTC 14 and 1200 UTC
480 22 September.

481

482 3) MODEL BIAS (SYSTEMATIC ERROR)

483 In many cases that develop from tropical waves over Africa, the forecast tracks
484 intensified to TS strength before the observed TS genesis time. The example of Lorenzo
485 above shows that the forecast tracks often intensified to TS strength immediately after
486 leaving the African continent and moving over sea.

487 To investigate how typical this early intensification is, we consider all forecast tracks in
488 the 5-year sample. Figure 7a shows the location of the first time each forecast track reaches
489 TS strength, accumulated on a 1°x1° grid. Figure 7b shows the observed locations for the
490 equivalent first time that the observed TC is reported as TS. There is a substantial peak in the
491 number of forecast TCs that intensify to TS strength immediately after leaving the African
492 coast. In contrast, none of the observed cases are reported to reach TS intensity east of 20°W.

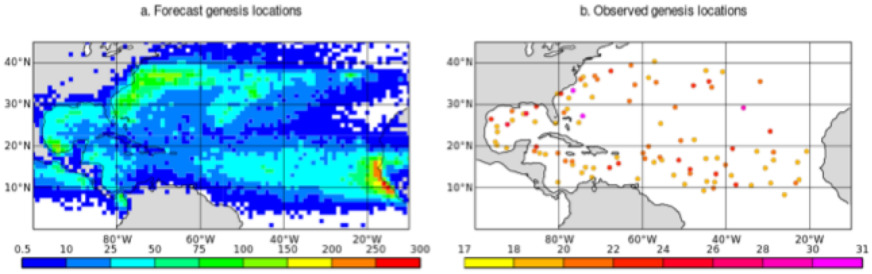
493 There are fewer forecast TS genesis events in the central and western areas (60-80°W, 10-
494 20°N). Overall, there is a shift eastwards of the genesis locations in the forecasts. A similar
495 bias in overforecasting TC genesis was found in the NCEP GEFS reforecasts, associated with
496 overactivity of African easterly waves in that system (Li et al. 2016; Wang et al. 2018).

497 Overdevelopment of initial wave activity over Africa and the quick intensification to TS
498 soon after the waves move over the open sea may also account for some of the high DI cases
499 shown in Fig. 3. Peter and Philippe were two cases predicted with high probability at longer
500 lead times, but for both the probability for TS intensity dropped at shorter leads. In each case
501 the higher probabilities occurred for forecasts initialized when the wave activity was still over
502 the African continent, and TS genesis occurred soon after the system left the coast. In the
503 later forecasts where the forecast TC developed further to the west, the probabilities for more
504 intense developments (both FA17 and FA15) were lower.

505 In summary, there is a tendency in the ENS for TC development to occur too quickly in
506 TCs that develop from African easterly waves and for the intensification to TS to occur soon
507 after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause
508 of the jumpy behavior seen in some cases.

509 We hypothesize that this bias is associated with overdevelopment of African easterly
510 wave activity in the ENS and identify this as an important area for future research.

511



512

513 Fig. 7. Locations of TS genesis in forecasts and observations. (a) forecast genesis:
514 location of the first point on each forecast track with maximum wind speed >17 m s^{-1}; map
515 shows total number of forecast genesis events in each 1°x1° grid box over the full set of
516 forecasts May-December 2019-2023. (b) observed TS genesis locations for all 97 observed

517 cases; color indicates the reported maximum wind at genesis time in the IBTrACS data (m s-
518 1).

519

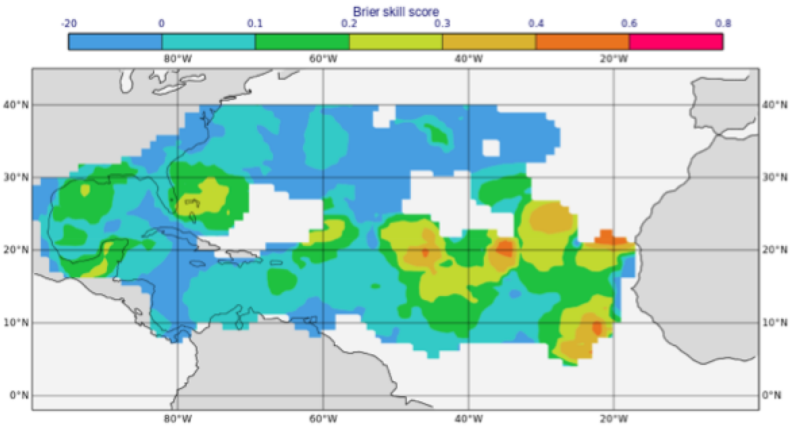
520 *d. Overall skill of TS genesis forecasts*

521 So far, we have focused on the results for observed TS genesis events. Although these
522 results show the performance for hits and misses of observed events, they do not take account
523 of false alarms in the forecasts.

524 To assess the overall performance of the ENS genesis probability forecasts, we now include
525 all forecast tracks, including those false alarm cases where a TS did not actually occur. For
526 each case, and at each grid point, the forecast is the probability that a TS genesis event will
527 occur within 500km and between 24h and 216h ahead.

528 Figure 8 shows the Brier skill score (B, Eq. (4)) of these ENS forecasts of TS genesis. This
529 shows that there is skill in some areas. The highest skill is in the eastern Atlantic, consistent
530 with the regions where genesis was found to be more predictable at longer lead for the observed
531 cases (Fig.2). Although BSS is lower in more western areas, there are still some regions with
532 positive skill. The low overall skill is consistent with the findings in the earlier sections that
533 FG17 skill is limited because of the tendency in the ENS to predict TS genesis earlier than
534 observed.

535



536

537 Fig. 8. Skill of ENS forecasts of TS genesis. Brier skill score for the forecast probability
538 that TS genesis will occur within 500 km of each grid point during the forecast, between 24 h
539 and 216 h lead time; score computed over all forecasts in 5 years sample 2019-2023.

540

541 Figure 9a shows the reliability diagram for the TS genesis forecasts; the ENS probabilities
542 are grouped into 10% probability intervals and accumulated over all grid points and over the
543 full 5-year sample. The curve is below the diagonal, indicating that the genesis forecasts are
544 overconfident and lack reliability. While this can be a result of lack of spread in the ensemble,
545 it is also consistent with our results that the ENS tends to predict TS genesis earlier than
546 observed. A similar overconfidence is also found in the operational ECMWF verification of
547 TC activity (Haiden et al. 2023) and in corresponding TC activity forecasts from other
548 ensemble systems (Magnusson et al. 2021).

549

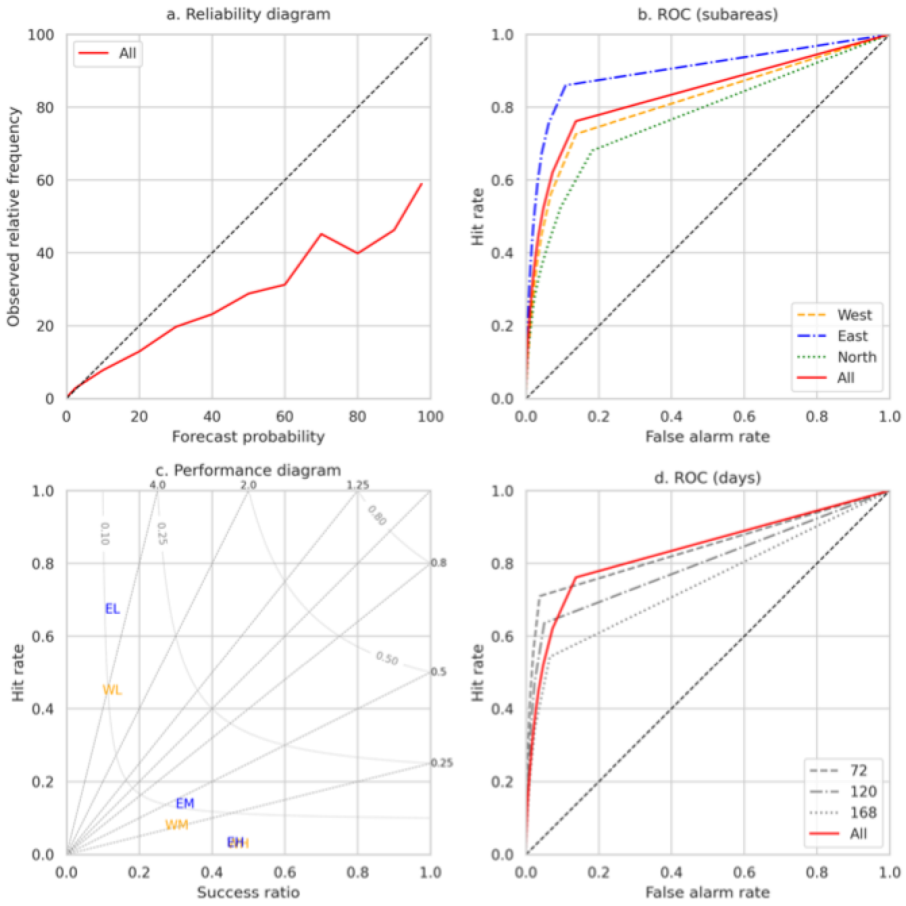


Fig. 9. Evaluation of ENS forecasts of TS genesis to occur between 24 h and 216 h lead time; scores computed over all forecasts in 5 years sample 2019-2023. a) reliability diagram, results accumulated over all grid points; b) ROC diagram for all grid points (solid red) and for western (orange dashed), eastern (blue dash-dotted) and northern (dotted green) sub-regions (see text for details); c) performance diagram for eastern (E) and western (W) regions and for the low (L), medium (M) and high (H) probability thresholds (first letter indicates region and second letter indicates the probability threshold), grey diagonal lines show bias and grey curved lines show threat score; d) ROC diagram comparing overall results (all, solid red, same as in panel b) with FG17 forecasts of TS genesis at lead times of 72, 120 and 168 h.

The positive slope of the reliability curve shows that, while lacking reliability, the forecasts do have some resolution: the ability to distinguish between more and less likely genesis events. This discrimination ability is confirmed in Fig. 9b which shows the ROC diagram for the

564 genesis forecasts. In the ROC computation, all possible forecast probabilities are considered
565 (Ben Bouallègue and Richardson 2022). In Fig. 9b, the ROC for all grid points is compared
566 with the corresponding ROC curves for three sub-regions: the skill is greater in the eastern
567 Atlantic (east of 60°W and south of 30°N) and lower in the western (west of 60°W and south
568 of 30°N) and northern (north of 30°N) areas. This confirms the regional differences in skill
569 noted in the evaluation of the observed cases (Fig.2). Although the reliability diagrams for the
570 sub-areas are more noisy due to the smaller sample size in each sub-area, they also indicate
571 better performance for the eastern region and lowest reliability in the northern region.

572 To highlight the false alarms as a proportion of the genesis forecasts, the skill of the genesis
573 forecasts for the low, medium and high probability thresholds in the eastern and western
574 regions is shown on a performance diagram in Fig. 9c. As for the reliability diagram and ROC,
575 Fig. 9c shows a substantial difference in performance between eastern and western areas,
576 especially for the low and medium probabilities, with substantially better hit rate for a similar
577 false alarm ratio. As for the other performance measures, the northern region has the poorest
578 performance (not shown).

579 Figure 9d shows the ROC curves for the FG17 forecasts for days 3, 5 and 7 (72, 120, 168
580 h in grey) together with the overall ROC (same as in Fig 9b). The discrimination skill decreases
581 at longer lead, although there is still substantial discrimination ability at 168 h. The overall
582 ROC (for genesis between 24 and 216 h) lies between the curves for 120 h and 168 h,
583 suggesting the overall results are reasonably indicative of the medium-range performance.

584 The results in this section have been based on the comparison of the forecast and observed
585 genesis of tropical storms, defined as the first point on forecast or observed track with wind
586 speed of 17 m s^{-1}. To investigate the sensitivity of the results to the forecast wind speed
587 threshold, we recomputed the ROC results using alternative forecast wind speed thresholds of
588 8 m s^{-1}, 15 m s^{-1} and 19 m s^{-1}, all verified against the operational genesis of TS (17 m s^{-1}). We
589 found that the results are relatively insensitive to small changes (+/-2 m s^{-1}) in the forecast wind
590 speed threshold, but a large reduction in the forecast threshold (to 8 m s^{-1}) substantially reduces
591 the forecast skill. This section has focused on whether TS genesis will occur at some point
592 during the forecast, and this may be why these results are not too sensitive to the wind threshold
593 – a given threshold will likely be exceeded as the tropical cyclone intensifies during the
594 forecast. A more detailed investigation of the definition of genesis in the forecast and the effect
595 on forecast skill will be a topic for future research.

5. Conclusions

We have investigated the ability of the ECMWF ensemble forecasts ENS to predict the genesis of tropical cyclones in the Atlantic basin up to 10 days ahead. We compared the ENS operational TC track forecasts to observed tracks from the IBTrACS archive for all named tropical storms for the 5 years 2019-2023. We focused on the probabilistic performance of the ENS rather than the evaluation of deterministic forecasts that has been more typically the subject of previous studies.

Defining a genesis event as the first time the TC reached tropical storm strength (winds at least 17 m s^{-1}), the ENS probability forecasts (FG17, Table 1) of the observed genesis events had relatively low skill with only 20% of the observed cases predicted with medium or high probability (probability 35% or more) more than 72 h ahead. In many cases the forecast track reached TS strength more than 24 h before the observed TS genesis time. Allowing for this early genesis in the forecasts increased the forecast probabilities (FA17, Table 1) for the observed event.

In part, this may reflect differences between the IBTrACS reports and the ECMWF TC tracker - the ECMWF tracker tends to pick up the TC at an earlier stage than the official designation as a TC. Differences in feature identification between different TC trackers can have a significant impact on the number of TCs identified by a forecast model (Conroy et al. 2023) and there is currently no generally agreed best practice for the definition and evaluation of TC genesis (Dunion et al. 2023).

We also found substantial geographical variation in the performance of the ENS probabilities: observed genesis events were predicted 2-3 days earlier in the central and eastern Atlantic than in other parts of the basin. The regional differences may be associated with intrinsic differences in predictability in different tropical cyclogenesis pathways (McTaggart-Cowan et al. 2013, 2008; Wang et al. 2018). Investigation of the ENS skill and jumpiness in the different pathways is an area of future research.

We assessed the run-to-run consistency of the ENS probabilities of genesis using the divergence index DI (Richardson et al 2020, 2024). The DI also varied between different regions, with the jumpiest cases being in the central and eastern Atlantic. The median DI here was more than twice that found in the western and northern parts of the basin. The most jumpy cases occurred in different years but almost always in late August or September. In

most of these cases the jumpiness depended on the forecast intensity: the forecasts were consistent in predicting the existence of the TC, but the probability for the TC to be at tropical storm strength varied from run to run.

Understanding the causes of jumpiness is important to inform both users and model developers. Forecast jumpiness is a measure of the internal consistency of the forecasting system. Although we used the observed genesis events as reference, the computation of DI does not depend on the observations. Hence, the results for jumpiness are not directly affected by the differences between the model and observed definitions of genesis discussed above. Examining the issues affecting jumpiness can therefore help to identify potential weakness in the modelling system. Based on consideration of the most jumpy cases in our sample, we considered a number of factors that could affect the ENS jumpiness in predicting TC genesis.

One possible cause of large jumpiness is the sampling uncertainty associated with the limited ensemble size. We found that the DI for the most jumpy cases is significantly higher than should be expected for a well-constructed 50-member ensemble. However, jumpiness is sensitive to ensemble size and the highest values of DI found in our results may occur for ensembles with around 20 members. While ENS track forecasts are well-calibrated, the forecast intensity is overall underdispersive (Haiden et al. 2023) and in some situations this may reduce the effective ensemble size, contributing to increased jumpiness. In certain situations with intrinsically low predictability, there may be particular sensitivity to ensemble size and substantially more than 50 members may be needed to properly represent the underlying distribution (Leutbecher 2019; Craig et al. 2022; Kondo and Miyoshi 2019). This may be important in some genesis situations involving complex interactions between waves where the ENS showed large jumpiness.

In some cases, there was a notable sequence of flip-flops between the forecasts started from 0000 and 1200 UTC analyses. Lorenzo was a particularly strong example, and for this case we found a significant difference between the forecast maximum winds associated with the TCs initialized at the two analysis times, with higher winds from the 0000 UTC analysis. We hypothesize that this may be associated with a known systematic difference in analysis increments at 0000 and 1200 UTC over West Africa in the ECMWF assimilation system (Bormann et al, 2023). However, this flip-flop behavior was not a common feature across cases, suggesting that a combination of factors in addition to the analysis differences may be

involved to make the large and significant impact found in this case. This is an area requiring further investigation.

A significant difference between the observed and forecast TS genesis is that the ENS TC tracks tend to intensify to TS strength earlier than the observed TS genesis event. ENS tracks that develop from African easterly waves often reach TS soon after the wave moves over the ocean, often before the TC reaches 20°W. This may be a cause of the jumpy behavior seen in some cases (for example Peter and Philippe) where earlier forecasts had high probability for TS development, while later forecasts that were initialized after the disturbance moved over the ocean had lower probability. The association with jumpy behavior lends weight to this being a systematic error in the forecasting system and not just an artifact of the differences between forecast and observed genesis identification methods. We hypothesize that this bias is associated with overdevelopment of African easterly wave activity in the ENS and identify this as an important area for future research.

Finally, we provided a baseline evaluation of the skill of the ENS TS genesis forecasts including all forecasts from the 5-year sample to take account of both hits and false alarms. Overall, forecasts were overconfident but showed good discrimination ability, with higher skill in the east of the basin (particularly for low to medium probabilities) consistent with the results for the observed genesis cases. The ECMWF forecasting system is typically upgraded annually and some of these changes affect the tropical cyclone performance, for example the increase in ensemble resolution in 2023 (Haiden et al. 2023). Given that TS genesis is a relatively rare event, skill evaluation generally needs to be carried out over a sample of several seasons, inevitably covering a number of different model versions (Leonardo and Colle 2021). We found cases of large jumpiness in each year of our sample, and this suggests that the underlying causes still need to be addressed. The overall results can be seen as a general assessment of recent model performance and provide a benchmark against which to evaluate future model developments.

Acknowledgments.

This work is based on TIGGE data. TIGGE (The International Grand Global Ensemble) is an initiative of the World Weather Research Programme (WWRP). David Richardson is supported by a Wilkie Calvert PhD Studentship at the University of Reading. Sharanya J. Majumdar gratefully acknowledges support from National Science Foundation Grant AGS-

691 1747781 and the University of Miami and ECMWF for jointly supporting a sabbatical year at
692 ECMWF. We thank three anonymous referees for their valuable comments.

693

694 *Data Availability Statement.*

695 The forecast data used in this study are available from The International Grand Global
696 Ensemble (TIGGE) Model Tropical Cyclone Track Data, Research Data Archive at the
697 National Center for Atmospheric Research, Computational and Information Systems
698 Laboratory at https://doi.org/10.5065/D6GH9GSZ (Bougeault et al. 2011; Swinbank et al.
699 2016).

700 The observed tropical cyclone tracks are available from NOAA's International Best Track
701 Archive for Climate Stewardship (IBTrACS) archive at https://doi.org/10.25921/82ty-9e16
702 (Knapp et al 2010, 2018).

703

704

705 REFERENCES

706 Bormann, N., L. Magnusson, D. Duncan, and M. Dahoui, 2023: Characterisation and
707 correction of orbital biases in AMSU-A and ATMS observations in the ECMWF
708 system. *ECMWF Technical Memorandum*, **912**, https://doi.org/10.21957/d281dc221a.

709 Ben Bouallègue, Z., and D. S. Richardson, 2022: On the ROC Area of Ensemble Forecasts
710 for Rare Events. *Weather Forecast*, **37**, 787–796, https://doi.org/10.1175/WAF-D-21-
711 0195.1.

712 Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble.
713 *Bull Am Meteorol Soc*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

714 Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V.
715 Tallapragada, 2020: Recent Progress in Tropical Cyclone Intensity Forecasting at the
716 National Hurricane Center. *Weather Forecast*, **35**, 1913–1922,
717 https://doi.org/10.1175/WAF-D-20-0059.1.

718 Conroy, A., and Coauthors, 2023: Track forecast: Operational capability and new techniques
719 - Summary from the Tenth International Workshop on Tropical Cyclones (IWTC-10).

Tropical Cyclone Research and Review, **12**, 64–80, https://doi.org/10.1016/J.TCRR.2023.05.002.

Craig, G. C., M. Puh, C. Keil, K. Tempest, T. Necker, J. Ruiz, M. Weissmann, and T. Miyoshi, 2022: Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, **148**, 2325–2343, https://doi.org/10.1002/QJ.4305.

Davis, C. A., and L. F. Bosart, 2003: Baroclinically Induced Tropical Cyclogenesis. *Mon Weather Rev*, **131**, 2730–2747, https://doi.org/10.1175/1520-0493(2003)131<2730:BITC>2.0.CO;2.

——, and ——, 2004: The TT Problem: Forecasting the Tropical Transition of Cyclones. *Bull Am Meteorol Soc*, **85**, 1657–1662, https://doi.org/10.1175/BAMS-85-11-1657.

Dunion, J. P., and Coauthors, 2023: Recommendations for improved tropical cyclone formation and position probabilistic Forecast products. *Tropical Cyclone Research and Review*, **12**, 241–258, https://doi.org/10.1016/J.TCRR.2023.11.003.

Elsberry, R. L., and P. H. Dobos, 1990: Time Consistency of Track Prediction Aids for Western North Pacific Tropical Cyclones. *Mon Weather Rev*, **118**, 746–754, https://doi.org/10.1175/1520-0493(1990)118<0746:TCOTPA>2.0.CO;2.

Fowler, T. L., B. G. Brown, J. H. Gotway, and P. Kucera, 2015: Spare Change : Evaluating revised forecasts. *MAUSAM*, **66**, 635–644, https://doi.org/https://doi.org/10.54302/mausam.v66i3.572.

Haiden, T., M. Janousek, F. Vitart, Z. Ben-Bouallegue, and F. Prates, 2023: Evaluation of ECMWF forecasts, including the 2023 upgrade. *ECMWF Technical Memorandum*, **911**, https://doi.org/10.21957/d47ba5263c.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of Tropical cyclone genesis forecasts from global numerical models. *Weather Forecast*, **28**, 1423–1445, https://doi.org/10.1175/WAF-D-13-00008.1.

——, ——, ——, and ——, 2016: Verification of tropical cyclone genesis forecasts from global numerical models: Comparisons between the North Atlantic and Eastern North Pacific Basins. *Weather Forecast*, **31**, 947–955, https://doi.org/10.1175/WAF-D-15-0157.1.

750 ——, R. E. Hart, H. E. Fuelberg, and J. H. Cossuth, 2017: The development and evaluation
751 of a statistical-dynamical tropical cyclone genesis guidance tool. *Weather Forecast*, **32**,
752 27–46, https://doi.org/10.1175/WAF-D-16-0072.1.

753 Hewson, T., 2020: Use and Verification of ECMWF products in Member and Co-operating
754 States (2019). *ECMWF Technical Memorandum*, **860**,
755 https://doi.org/10.21957/80s471ib1.

756 Hon, K. K., and Coauthors, 2023: Recent advances in operational tropical cyclone genesis
757 forecast. *Tropical Cyclone Research and Review*, **12**, 323–340,
758 https://doi.org/10.1016/J.TCRR.2023.12.001.

759 Jewson, S., S. Scher, and G. Messori, 2021: Decide Now or Wait for the Next Forecast?
760 Testing a Decision Framework Using Real Forecasts and Observations. *Mon Weather*
761 *Rev*, **149**, 1637–1650, https://doi.org/10.1175/MWR-D-20-0392.1.

762 ——, ——, and ——, 2022: Communicating Properties of Changes in Lagged Weather
763 Forecasts. *Weather Forecast*, **37**, 125–142, https://doi.org/10.1175/WAF-D-21-0086.1.

764 Klotzbach, P. J., and Coauthors, 2022: A Hyperactive End to the Atlantic Hurricane Season
765 October–November 2020. *Bull Am Meteorol Soc*, **103**, E110–E128,
766 https://doi.org/10.1175/BAMS-D-20-0312.1.

767 Knapp, K. R., M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann, 2010: The
768 International Best Track Archive for Climate Stewardship (IBTrACS): Unifying
769 Tropical Cyclone Data. *Bull Am Meteorol Soc*, **91**, 363–376,
770 https://doi.org/10.1175/2009BAMS2755.1.

771 Knapp, K. R., H. J. Diamond, M. C. Kossin, and C. J. Schreck, 2018: International Best
772 Track Archive for Climate Stewardship (IBTrACS) Project, Version 4.
773 https://doi.org/10.25921/82ty-9e16.

774 Komaromi, W. A., and S. J. Majumdar, 2014: Ensemble-Based Error and Predictability
775 Metrics Associated with Tropical Cyclogenesis. Part I: Basinwide Perspective. *Mon*
776 *Weather Rev*, **142**, 2879–2898, https://doi.org/10.1175/MWR-D-13-00370.1.

777 ——, and ——, 2015: Ensemble-Based Error and Predictability Metrics Associated with
778 Tropical Cyclogenesis. Part II: Wave-Relative Framework. *Mon Weather Rev*, **143**,
779 1665–1686, https://doi.org/10.1175/MWR-D-14-00286.1.

780 Kondo, K., and T. Miyoshi, 2019: Non-Gaussian statistics in global atmospheric dynamics: a
781 study with a 10 240-member ensemble Kalman filter using an intermediate atmospheric
782 general circulation model. *Nonlinear Process Geophys*, **26**, 211–225,
783 https://doi.org/10.5194/npg-26-211-2019.

784 Landsea, C. W., and J. L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and
785 Presentation of a New Database Format. *Mon Weather Rev*, **141**, 3576–3592,
786 https://doi.org/10.1175/MWR-D-12-00254.1.

787 ——, and J. P. Cangialosi, 2018: Have We Reached the Limits of Predictability for Tropical
788 Cyclone Track Forecasting? *Bull Am Meteorol Soc*, **99**, 2237–2243,
789 https://doi.org/10.1175/BAMS-D-17-0136.1.

790 Leonardo, N. M., and B. A. Colle, 2021: An Investigation of Large Cross-Track Errors in
791 North Atlantic Tropical Cyclones in the GEFS and ECMWF Ensembles. *Mon Weather*
792 *Rev*, **149**, 395–417, https://doi.org/10.1175/MWR-D-20-0035.1.

793 Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quarterly*
794 *Journal of the Royal Meteorological Society*, **145**, 107–128,
795 https://doi.org/10.1002/QJ.3387.

796 Li, W., Z. Wang, and M. S. Peng, 2016: Evaluating Tropical Cyclone Forecasts from the
797 NCEP Global Ensemble Forecasting System (GEFS) Reforecast Version 2. *Weather*
798 *Forecast*, **31**, 895–916, https://doi.org/10.1175/WAF-D-15-0176.1.

799 Liang, M., J. C. L. Chan, J. Xu, and M. Yamaguchi, 2021: Numerical prediction of tropical
800 cyclogenesis part I: Evaluation of model performance. *Quarterly Journal of the Royal*
801 *Meteorological Society*, **147**, 1626–1641, https://doi.org/10.1002/qj.3987.

802 Magnusson, L., and Coauthors, 2021: Tropical cyclone activities at ECMWF. *ECMWF*
803 *Technical Memorandum*, **888**, www.ecmwf.int/en/elibrary/81277-tropical-cyclone-
804 activities-ecmwf

805 Majumdar, S. J., and R. D. Torn, 2014: Probabilistic Verification of Global and Mesoscale
806 Ensemble Forecasts of Tropical Cyclogenesis. *Weather Forecast*, **29**, 1181–1198,
807 https://doi.org/10.1175/WAF-D-14-00028.1.

808 Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological*
809 *Magazine*, **30**, 291–303.

McLay, J. G., 2008: Markov Chain Modeling of Sequences of Lagged NWP Ensemble Probability Forecasts: An Exploration of Model Properties and Decision Support Applications. *Mon Weather Rev*, **136**, 3655–3670, https://doi.org/10.1175/2008MWR2376.1.

McTaggart-Cowan, R., G. D. Deane, L. F. Bosart, C. A. Davis, and T. J. Galarneau, 2008: Climatology of Tropical Cyclogenesis in the North Atlantic (1948–2004). *Mon Weather Rev*, **136**, 1284–1304, https://doi.org/10.1175/2007MWR2245.1.

——, T. J. Galarneau, L. F. Bosart, R. W. Moore, and O. Martius, 2013: A Global Climatology of Baroclinically Influenced Tropical Cyclogenesis. *Mon Weather Rev*, **141**, 1963–1989, https://doi.org/10.1175/MWR-D-12-00186.1.

Papin, P. P., L. F. Bosart, and R. D. Torn, 2017: A Climatology of Central American Gyres. *Mon Weather Rev*, **145**, 1983–2000, https://doi.org/10.1175/MWR-D-16-0411.1.

Pappenberger, F., H. L. Cloke, A. Persson, and D. Demeritt, 2011: HESS Opinions "On forecast (in)consistency in a hydro-meteorological chain: curse or blessing?" *Hydrol Earth Syst Sci*, **15**, 2391–2400, https://doi.org/10.5194/hess-15-2391-2011.

Regnier, E., and P. A. Harr, 2006: A Dynamic Decision Model Applied to Hurricane Landfall. *Weather Forecast*, **21**, 764–780, https://doi.org/10.1175/WAF958.1.

Richardson, D. S., H. L. Cloke, and F. Pappenberger, 2020: Evaluation of the consistency of ECMWF ensemble forecasts. *Geophys Res Lett*, https://doi.org/10.1029/2020GL087934.

——, ——, J. A. Methven, and F. Pappenberger, 2024: Jumpiness in Ensemble Forecasts of Atlantic Tropical Cyclone Tracks. *Weather Forecast*, **39**, 203–215, https://doi.org/10.1175/WAF-D-23-0113.1.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Weather Forecast*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Swinbank, R., and Coauthors, 2016: The TIGGE Project and Its Achievements. *Bull Am Meteorol Soc*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Titley, H. A., M. Yamaguchi, and L. Magnusson, 2019: Current and potential use of ensemble forecasts in operational TC forecasting: results from a global forecaster survey. *Tropical Cyclone Research and Review*, **8**, 166–180, https://doi.org/10.1016/J.TCRR.2019.10.005.

840 Torn, R. D., and C. Snyder, 2012: Uncertainty of Tropical Cyclone Best-Track Information.
841 *Weather Forecast*, **27**, 715–729, https://doi.org/10.1175/WAF-D-11-00085.1.

842 Wang, Z., W. Li, M. S. Peng, X. Jiang, R. McTaggart-Cowan, and C. A. Davis, 2018:
843 Predictive Skill and Predictability of North Atlantic Tropical Cyclogenesis in Different
844 Synoptic Flow Regimes. *J Atmos Sci*, **75**, 361–378, https://doi.org/10.1175/JAS-D-17-
845 0094.1.

846 Wilks, D. S., 2020: *Statistical Methods in the Atmospheric Sciences, Fourth Edition*.
847 Elsevier, 1–818 pp.

848 Yamaguchi, M., and N. Koide, 2017: Tropical Cyclone Genesis Guidance Using the Early
849 Stage Dvorak Analysis and Global Ensembles. *Weather Forecast*, **32**, 2133–2141,
850 https://doi.org/10.1175/WAF-D-17-0056.1.

851 ——, F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T.
852 Nakazawa, 2015: Global Distribution of the Skill of Tropical Cyclone Activity Forecasts
853 on Short- to Medium-Range Time Scales. *Weather Forecast*, **30**, 1695–1709,
854 https://doi.org/10.1175/WAF-D-14-00136.1.

855 Zhang, X., J. Fang, and Z. Yu, 2022: The Forecast Skill of Tropical Cyclone Genesis in Two
856 Global Ensembles. *Weather Forecast*, **38**, 83–97, https://doi.org/10.1175/WAF-D-22-
857 0145.1.

858

859

A4. Technical Memorandum: Tropical cyclone activities at ECMWF

This appendix contains the full version of the ECMWF Technical Memorandum summarised in Section 6.1.1:

Magnusson, L., Majumdar, S., Emerton, R., Richardson, D., et. al. (2021) 'Tropical cyclone activities at ECMWF', *ECMWF Technical Memorandum 888*. ECMWF. Available at: https://doi.org/10.21957/zzxzzygwv.

Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

888

Tropical cyclone activities at ECMWF

Linus Magnusson, Sharan Majumdar*, Rebecca Emerton, David Richardson, Magdalena Alonso-Balmaseda, Calum Baugh, Peter Bechtold, Jean Bidlot, Antonino Bonanni, Massimo Bonavita, Niels Bormann, Andy Brown, Phil Browne, Hilda Carr, Mohamed Dahoui, Giovanna De Chiara, Michail Diamantakis, David Duncan, Steve English, Richard Forbes, Alan Geer, Thomas Haiden, Sean Healy, Tim Hewson, Bruce Ingleby, Martin Janousek, Christian Kuehnlein, Simon Lang, Sarah-Jane Lock, Tony McNally, Kristian Mogensen, Florian Pappenberger, Inna Polichtchouk, Fernando Prates, Christel Prudhomme, Florence Rabier, Patricia de Rosnay, Tiago Quintino, Mike Rennie, Helen Titley**, Filip Vana, Frederic Vitart, Francis Warrick, Nils Wedi, Ervin Zsoter

* University of Miami
** Met Office and University of Reading

October 2021

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Contents

Abstract

ECMWF has a wide range of **users** of forecasts of tropical cyclones (TCs). Several member states have territories that are frequently hit by TCs, and Météo-France is the responsible Regional Specialized Meteorological Centre (RSMC) for the southern Indian Ocean. ECMWF forecasts of TCs are also routinely used by WMO member states. Additionally, ECMWF recently opened the operational charts to the public. Many commercial customers have activities related to TCs.

Traditionally, forecasters and users have relied on global models including ECMWF to predict the **position** (or "track") of the centre of the TC out to 5 days. Nowadays, users are increasingly demanding a wider range of products related to TCs, and NWP centres are working to meet their needs. These needs include **seasonal forecasts of TC activity**, sub-seasonal forecasts of the potential for **TC genesis** (formation), accurate medium-range forecasts of the **intensity and structure** of each TC, and **downstream influences** of TCs on extratropical weather (including over Europe). The interest in forecasting **impacts** of TCs has also increased. At ECMWF, the Copernicus Emergency Management Service (CEMS) run global **flood forecasts**. Given that an important fraction of flood events in the tropics are related to TCs, accurate forecasts of the magnitude, structure, and duration of rainfall from TCs are necessary. Several external institutes also use ECMWF forecasts to model **storm surges** induced by TCs; storm surges have historically been the largest cause of death from TCs.

Within this perspective of ever-increasing user needs, this report has been developed to document progress and challenges in the ECMWF forecasting system of special relevance to TCs. These included observations (Section 2), tracking of TCs (Section 3), verification (Section 4), predictions (Section 5), data assimilation (Section 6), modelling (Section 7), forecast products (Section 8), and impact forecasting applications (Section 9). Of note, especially in Sections 6 and 7, are new results from a series of special experiments conducted over a common 37-day period in August-September 2020 that explored potential future avenues in data assimilation and modelling. Section 10.1 summarizes key progress and challenges in each of the above areas, and Section 10.2 provides a larger-scale vision for future improvement. As this report was spawned by a 1-year visit by Prof. Sharan Majumdar from University of Miami, and due to the unusually high activity in the Atlantic basin during 2020, a special attention is paid to the performances in the Atlantic. The breadth of this report has resulted in a large amount of material. To improve its readability, the report is written in a way that each section can be treated as a stand-alone article on a specific topic, summarizing progress, challenges, and future directions.

1. Introduction

1.1. Background

Tropical cyclones (TCs) are deadly extreme events. They often bring extreme winds as they make landfall, leading to storm surges along the coastlines together with direct damage from the winds. They can also produce extreme rainfall, which can lead to coastal and inland flooding, especially if the TC moves slowly after landfall as in the cases of Harvey (2017; Blake and Zelinsky, 2018) and Idai (2019; Emerton et al., 2020). If a TC curves into the mid-latitudes, it may undergo extratropical

transition or be absorbed by an extratropical westerlies or a low-pressure system, which can lead to windstorms over western Europe or western North America, or downstream development of extratropical events with low predictability (Keller et al., 2019).

From an ECMWF perspective, accurate predictions of TCs are important for several reasons. Several of our Member States have overseas territories that are regularly affected by TCs and/or are responsible for the forecasting and warning of TC impacts. For example, during 2017, Hurricane Irma severely affected several of our Member States' territories on the Leeward Islands. In the southern Indian Ocean, Météo-France is the responsible Regional Specialized Meteorological Centre (RSMC). Furthermore, as mentioned above, TCs that recurve into the extratropics can affect Europe. Occasionally, former TCs can impact western Europe in their post-tropical stage (e.g., Leslie 2018, Lorenzo 2019). The forecasting and timely warning of events such as these is vital.

ECMWF forecasts are also widely used outside the European continent by WMO member states, commercial actors (providing forecasts to media) and private users via the recently gained access to the products on the ECMWF website or from other web pages that show ECMWF forecasts. Around the globe, it is also important to issue early warnings for events that will cause a demand for humanitarian aid. In recent years, the concept of forecast-based financing has been introduced. For example, the Red Cross aims to trigger humanitarian action ahead of an event, based on forecast information (IFRC, 2021; Coughlan de Perez et al., 2015).

ECMWF's global reputation is partly built on the performance for high-profile TCs such as Sandy (2012) and Irma (2017). TCs generate extensive media coverage before, during, and after the event. To respect the requirement not to comment on events as they happen so as not to distract affected populations from the local authoritative voice, ECMWF abstains as much as possible from commenting before and during the event. Even without commenting or issuing any statements, ECMWF is a part of conversations on social media and the forecasts are used on most TV channels in the United States. After the event has happened, and by capitalising on the existing momentum through accepting scientific interviews, TCs have proven to be an excellent way to improve awareness of ECMWF. Such interviews allow ECMWF to speak more broadly about the model and the science that sustains it, rather than focusing solely on TCs.

The primary metric upon which ECMWF's reputation in TC forecasting has been built is the **position** (or track) of the centre of the TC. To monitor progress from year to year, ECMWF has adopted the TC position error for the 3-day operational high-resolution (HRES) forecast as one of the supplementary headline scores. That score for 3-day and 5-day forecasts, together with verification of TC **central pressure ("Pmin")** mean absolute error and mean error (or bias) for 3-day forecasts is shown in Figure 1. The corresponding scores are also shown for the ensemble mean, derived from the track of each ensemble member. To account for the interannual variability of the sample size and overall forecast difficulty, the results for ERA5 forecasts are included as well. The position error at day 3 and 5 in HRES and ensemble mean has steadily improved between 2006-2015 (Figure 1(a)). Comparing with ERA5, which is based on a non-evolving forecasting system, the improvements are incremental with contributions from many upgrades of the operational forecasting system. In 2020, the position error was highest in the time-series for ERA5 for 5-day forecasts, indicating that it was the least predictable year, and this explains the relatively high position errors in HRES. Over the

years, the day 3 position error of the ensemble mean has approached the HRES position error. In contrast, the corresponding day 5 position error of the ensemble mean has been consistently lower than HRES, as a result of the filtering during unpredictable situations by the ensemble. For the Pmin mean absolute error (Figure 1(b)) and mean error (Figure 1(c)), discrete improvements in both quantities for Pmin are evident in 2010 and 2016, when the model resolution was increased. The ensemble mean has a lower mean absolute error than HRES, although the weak bias (Pmin too high) is larger due to its reduced resolution.

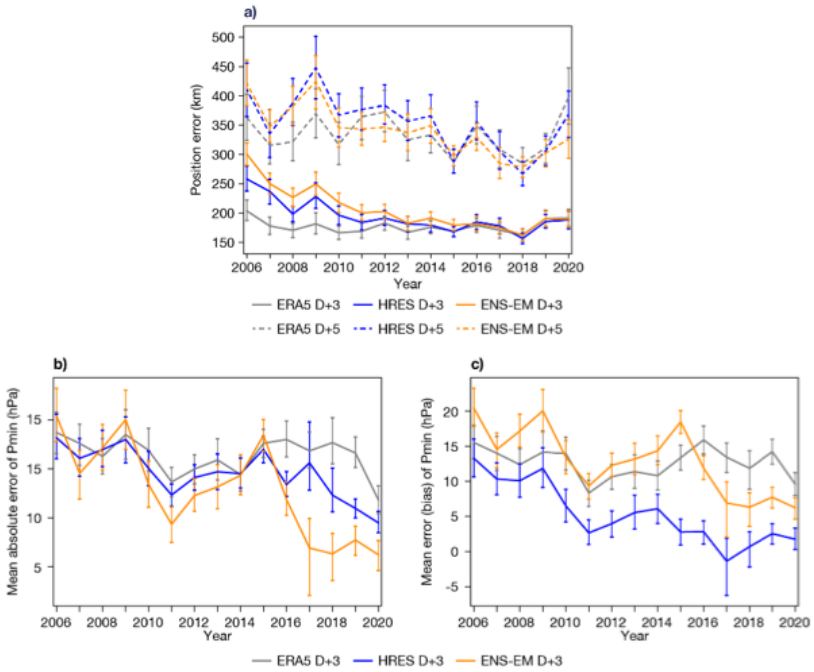


Figure 1: (a) 3-day (solid) and 5-day (dashed) forecast errors of TC position forecasts; (b) 3-day mean absolute error of TC central pressure (Pmin); (c) As for (b), for the mean error. Blue: HRES; Orange: ensemble mean; Grey: ERA5. The data is aggregated from 1 December the year before to 1 December of the labelled year.

It is a common practice for TC forecasts from different providers to be compared against each other. For example, Heming et al. (2019) summarised the verification of TC position in a report provided to the 2018 International Workshop on Tropical Cyclones. TC position forecasts from deterministic operational forecasts have been verified by the Japan Meteorological Agency (JMA) under an inter-comparison project of the Working Group on Numerical Experimentation (WGNE) since 1991 (Yamaguchi et al., 2017). The evaluation is based on meteorological fields provided to JMA who run the tracker, and an updated plot for the north-western Pacific basin is presented in Figure 2a. While there is a large year-to-year variability in the ordering of the models, ECMWF has always been

among the best if not the best. However, the spread among the group of models has decreased in recent years. For other basins, the sample is usually too small to draw conclusions for single years, but for the south-western Indian Ocean ECMWF has not been the superior model in recent years (not shown). The NOAA National Hurricane Center (NHC) in the United States routinely verifies the models that are used in their operational activities, as shown in Figure 2(b) for 3-day position forecasts in the Atlantic basin provided by ECMWF, Met Office, and NOAA (GFS and two regional models). For 2020, the position error of the three global models was very similar. Between 2010-2020, ECMWF had the lowest errors for 7 out of 11 years, UKMO 2 years and GFS 2 years. This indicates an interannual variability in the ranking even though ECMWF has the lowest errors averaged over the full period. However, there is also a large case-to-case variability for the ranking of the models, as illustrated in e.g Tang et al. (2021). Therefore, multi-model ensembles have been proven to be superior than a single (ensemble) forecast (Titley et al., 2020).

Since the 1990s, the use of **ensemble forecasts** has increased over a range of applications, including TCs. To better understand the use of these forecasts, Titley et al. (2019) conducted a survey about the use of ensemble forecasts in TC forecasting. The study showed that nearly all operational centres use ensemble forecasts, particularly for TC track and genesis forecasting. However, the information from the dynamical ensemble is to a lesser degree communicated in official graphical products and uncertainty information. While it was noted that the quality of the track forecasts is on a sufficient level, the quality of the intensity forecasts in operational ensembles is still poor.

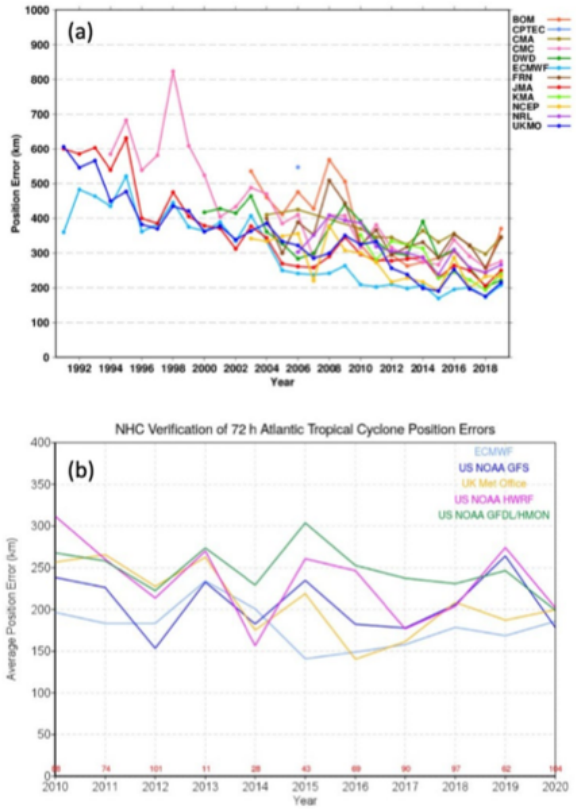


Figure 2: 3-day position error for different NWP centres for (a) the north-western Pacific basin provided by JMA via WGNE and (b) for the Atlantic basin provided by NHC. Panel (a) is taken with permission from the presentation on 35th WGNE meeting (http://wgne.meteoinfo.ru/wp-content/uploads/2020/10/WGNE35_Ujiie_tcverif.pdf).

1.2. Purpose and Structure of this Report

The goal of this report is to provide insights into the scientific and operational status of ECMWF's modelling and predictions of TCs, at the time during which the 47r1 cycle was operational. This includes understanding ECMWF's progress in TC simulation and prediction in comparison with other centres, to help provide directions on how ECMWF can further improve its TC predictions.

This report contains ten sections, comprising all parts of the forecast production chain relevant to TCs where ECMWF is involved. Each section can be treated as a stand-alone article on a specific topic, summarizing progress, challenges, and future directions. The report starts with background information, first outlining available **observations** relevant for assimilation and evaluation of TCs

(Section 2), then documenting how **TC tracking** is done at ECMWF (Section 3) to create operational products and perform **verification** (Section 4). The subsequent three sections describe new scientific results. Section 5 discusses various **predictability** aspects of the ECMWF forecasting system during the life cycle of the TC on different timescales. In Section 6, aspects of **assimilation of observations** are presented, including new results from a range of assimilation experiments that were conducted at ECMWF for a 37-day period in August-September 2020. In Section 7, **modelling activities** at ECMWF are described, assisted by modelling experiments conducted during the same 37-day period. The next two sections give examples of forecast and product usage and delivery. In Section 8, current and planned **forecast products** are presented, and progress and goals for applications in terms of **hazard and impact forecasts** are described in depth in Section 9. Finally, concluding remarks and priorities for future improvement are discussed in Section 10.

Throughout the report, most of the illustrated examples are provided from 2020. Due to the Atlantic basin having a record-breaking number of TCs (with the Pacific having relatively low activity), attention is paid more to Atlantic TCs. Reference materials in the Appendices include a list of abbreviations (Appendix A), the TCs during the experiment period in 2020 (Appendix B), and a summary of verification scores of wind and temperature in the tropics (Appendix C). Much of the work presented in this report was spawned by a one-year visit by Prof. Sharan Majumdar (University of Miami) between July 2020–July 2021.

2. Observations of Tropical Cyclones

2.1. Introduction

This section gives an overview of the different kinds of observations that provide information about the current conditions and can help to determine the future progression of TCs. The observations are important in our forecast process both for data assimilation and for post-event evaluation, although not all available observations are used for both purposes. For a more comprehensive report on the use of observations in verification, see WMO (2013). Table 1 lists the observation types that are used regularly at ECMWF for assimilation, and for diagnostics and evaluation of conditions in TCs and their vicinity. In Sections 4-5, the observation types for evaluation are presented, and in Section 6 the impact of assimilating selected observation types on subsequent TC analyses and forecasts is described.

Table 1: Use of TC-related observations at ECMWF.

		Assimilation	Diagnostics and Evaluation of TC
In-situ	Land/Ship/Buoy	X	X
	Radiosondes	X	
	Aircraft	X	

Satellite	Microwave (MW)	X	
	Infrared (IR)	X	X
	GPS Radio Occultation	X	
	Atmospheric Motion Vectors (AMV)	X	
	Aeolus	X	
	Scatterometer	X	X
	Synthetic Aperture Radar (SAR)		
	SMOS/SMAP		
Aircraft	Dropsondes	X	X
	In-flight data		
	Doppler Radar		
	Stepped Frequency Microwave Radiometer (SFMR)		
Cyclone estimate	Best Track		X
Sea-surface temperature	OSTIA	X	X

2.2. In-situ observations

TCs are very difficult to directly observe with conventional observations, due to their extreme conditions and since they often appear in remote locations. The availability of SYNOP observations of TCs from land stations, ships, and buoys are very sparse, due to the low probability of a TC passing over a permanent location and that ships avoid the vicinity of the TC. And if a TC passes close to a station, the risk is very high for the station or buoy to be damaged.

It is even more rare that a TC can be directly observed by a regular radiosonde or by commercial aircraft. However, the assimilation of data from both platforms is hypothesized to improve the predictions of the large-scale flow that determines the propagation of the TCs. It is therefore common practice in the United States (NOAA / National Weather Service) to launch special radiosondes at 06 UTC and 18 UTC ahead of a potentially landfalling TC.

2.3. Temperature and humidity from satellites

Infrared observations (including near infrared and visible) are provided by sensors onboard geostationary (GEO) and low-Earth orbit (LEO) satellites and are critically important for monitoring the current state of a TC and predicting its future evolution. High resolution imagery from LEO and GEO sensors provides the most spatially detailed view of the storm, primarily showing the cloud tops, with some GEO sensors (such as the Advanced Baseline Imager on the GOES satellites) capable of refreshing this view every five minutes. The imagery is used to first identify precursor disturbances that may develop into a TC, and then the TC itself by Regional Specialized Meteorological Centres (RSMCs) and Tropical Cyclone Warning Centres (TCWCs). It is also used to estimate the central pressure and maximum wind speed via the Dvorak technique (see Section 2.7). In clear air, the infrared radiances inform on the 3D atmospheric temperature field, mid and upper-level humidity and the ocean surface temperature.

Microwave observations are relatively insensitive to attenuation from clouds and can thus provide information on atmospheric temperature, humidity, and hydrometeors at levels throughout the troposphere. The large number of microwave sensors in LEO provides high spatiotemporal coverage. Microwave observations can thus constrain the steering flow of TCs while also providing unique information below cloud tops within the TC itself. In current ECMWF operational assimilation, microwave imager and humidity sounding channels are utilised in "all-sky" conditions (i.e., including regions with clouds and precipitation); these form a crucial constraint on the analysis of columnar water vapour and hydrometeors due to sensitivity to low- and mid-level moisture. This can also benefit wind forecasts through the tracer effect of humidity (Geer et al., 2018). Temperature sounding channels (e.g., AMSU-A) are assimilated in clear-sky conditions currently, meaning that temperature profile information is available in clear and lightly cloudy areas, but not within thicker clouds such as in deep convection. All-sky AMSU-A assimilation will be included in the 47r3 IFS cycle, so that temperature profile information will be assimilated even within TCs, and the impact of this change is tested in Section 6.

GNSS radio occultation (GNSS-RO) measurements provide information about temperature and humidity in the atmosphere. In 2020, there was a large increase in the number of GNSS-RO available for operational assimilation, with the active assimilation of COSMIC-2 data from 25 March 2020, and the use of Spire data from 13 May until 30 September 2020. GNSS-RO can be considered all-sky capable. The measurements have a limb geometry, and provide high vertical resolution information, but have poorer horizontal resolution (~100's km). They are assimilated with a two-dimensional observation operator, using NWP information in an "occultation plane", to mitigate errors caused by horizontal gradients in the atmosphere. The information content of GNSS-RO measurements is largest for upper-tropospheric and stratospheric temperatures. However, the inclusion of >4000 COSMIC-2 measurements per day, in the latitude band between ±40 degrees, had a clear positive impact on tropospheric temperature, humidity and wind forecasts in the tropics (Ruston and Healy, 2021). The positive impact on tropical tropospheric humidity was difficult to demonstrate prior to COSMIC-2, but we now see a clear signal in the departure statistics of almost all observations sensitive to tropospheric humidity. Therefore, some impact of GNSS-RO on the steering flow of TCs could be anticipated. In Section 6 we evaluate the impact of from COSMIC-2 on TCs.

2.4. Winds from satellites

Wind estimates can be derived from satellite observations in the free atmosphere and surface in different ways. For the free atmosphere, Atmospheric Motion Vectors (AMVs) are derived from cloud motion, or water vapour features, through a series of satellite images. In the tropics, AMVs account for most wind observations in the upper-level troposphere, particularly over the ocean. However, the impact of AMVs on TC forecasts at ECMWF is limited using 200x200 km thinning boxes (to mitigate spatially correlated errors), and uncertainties exist in attributing representative heights for AMVs. Restrictions on the assimilation of AMVs at high pressure levels further decrease the number of AMVs assimilated over TCs due to their very high cloud tops (not using water vapour winds with pressure below 200 hPa for latitudes above 25°N, or GOES winds anywhere with pressure < 150 hPa). In some cases, AMVs are not available at all as there is not enough contrast over the central dense overcast region of the TC for AMVs to be derived. These limitations affect the ability of assimilated AMVs to capture the curvature of the wind field and their overall impact. Nevertheless, previous research has shown some benefits in TC verification scores by assimilating AMVs.

Since 9 January 2020, horizontal line-of-sight winds from ESA's Earth Explorer satellite Aeolus have been operationally assimilated at ECMWF. Aeolus is the first Doppler wind lidar instrument in space. It produces two main types of wind information: Mie-cloudy measuring winds from the top of optically thick clouds, through semi-transparent cirrus and from aerosols with sufficiently large backscatter; and Rayleigh-clear winds which are based on backscatter from air molecules in clear air. Aeolus measures only a component of the vector wind along the line-of-sight of the lidar and we assimilate the horizontal line-of-sight wind component (see Figure 3 for example). The bulk of the coverage comes from Rayleigh-clear winds; however, Mie-cloudy winds are much less noisy than the Rayleigh-clear. There are some residual biases for the Mie-cloudy winds as a function of wind speed and an apparent temperature dependence to the Rayleigh-clear bias (more negative bias near surface) which could affect the impact.

Aeolus cannot measure below optically thick clouds and hence cannot see directly into a TC. It does however provide good quality Mie wind information in the outflow cirrus of TCs when a direct hit is made. Aeolus is probably more suited to capturing the winds around a TC, such as the broad circulation and vertical wind shear. One can expect these wind measurements to fill gaps in the observing system in the tropics, where vertical wind profile information is lacking, and potentially improve forecasts of tropical cyclogenesis, track, and the structure. Since Aeolus is a demonstration mission, it has not been available all the time in the latter half of 2020 and into 2021, due to various instrument test procedures. An ESA project started in July 2021 on the assessment of Aeolus winds on extreme events like TCs and extra-tropical storms.

For surface winds, there are several products available that derive winds from satellite instruments. Scatterometers measure the ocean surface roughness which is related to the local wind conditions. At ECMWF scatterometer winds have been used since 1995. Currently, the vector winds from ASCAT onboard MetOp-A/B/C and HY-2B (since June 2021) are operationally assimilated. However, the vectors have ambiguities, which are solved using information from the background winds, and this could lead to wrong selections if, for example, a TC is misplaced in the background fields (see Section 6.6). Another problem related to the use of scatterometer winds for TCs is the saturation of

the signal around 30 m/s (for the C-band instruments). Ku-band observations, like HY-2B, are less reliable at high winds and they are assimilated up to 25 m/s. Also, since Ku-band measurements are affected by the rain, the observations are filtered out around the centre of TCs.

There are also other satellite channels that provide wind information that are not currently used at ECMWF for wind assimilation. C-band Synthetic Aperture Radar (SAR) is the only instrument providing all-weather wind measurements at very high resolution. Thanks to its high sensitivity it can provide wind speeds up to 75 m/s in major hurricanes (Mouche et al., 2019). As for scatterometers, a wind direction ambiguity is removed by taking the closest direction to the ECMWF model winds. Due to other applications of the instrument, it is not possible to have a continuous sampling of the surface winds at such high resolution. Within the current CYMS (Cyclone Monitoring Service based on Sentinel-1) project, using TC track forecasts available from ECMWF, ESA plans the most suitable Sentinel-1 acquisition on the expected location of the hurricane's eye over the next five days. Currently the TC products from Sentinel 1 A/B and Radarsat-2 are available in near-real-time with a 3-km horizontal resolution. Thanks to the high resolution, SAR images demonstrate great potential for capturing a well-defined TC eye. A first assessment study of high-resolution Sentinel-1 data in an NWP assimilation, performed by Météo-France at the Regional Specialized Meteorological Center La Réunion (Duong et al., 2021) showed that the wind observations are biased at low (20–25°) or high (40–45°) incidence angles with sometimes large differences in maximum winds between SAR and Best Track. Their impact experiments, run with the LAM AROME OI, on a couple of cases showed that the assimilation of high-resolution Sentinel-1 wind data leads to a better TC positioning in the analysis and an improved representation of the outer vortex structure. Analysis increments were not confined to surface wind fields but are also visible in the upper layers up to 400 hPa in temperature, humidity and wind.

L-band passive sensors such as SMOS and SMAP provide all-weather wind information (speed only) near the ocean surface. At this frequency, there is no sensitivity saturating at moderately high wind speed, as occurs for scatterometer and higher frequency passive observations (De Chiara and English, 2016). SMOS wind retrievals have reduced sensitivity at low–moderate wind speeds. Cotton at al. (2018) showed that above 15 m s^{-1}, SMOS winds tend to be stronger than the Met Office model winds; it also reported the quality of the retrieved winds to be reduced in the presence of sea ice, radio-frequency interference (RFI) contamination and strong river plumes. The L-band passive sensors are limited by their coarse spatial resolution (40-50 km) but, in general, they do provide meaningful information on the wind radii of gale (34 kt), storm (50 kt), and hurricane (64 kt) force winds (Reul et al., 2017). These products could be considered in the future for the validation of the ECMWF tracker wind radii outputs (see Section 4). Assimilation experiments run at the Met Office showed that when the storm radius is small, SMOS is unable to resolve the eye structure present in the model. When the storm radius is much larger, SMOS can resolve the eye structure, and analysed and short-range forecast central pressures are closer to the Best Track when SMOS observations are assimilated (Cotton et al., 2018).

There are other ongoing satellite missions targeting surface wind near TCs. The Cyclone Global Navigation Satellite System (CYGNSS, Ruf et al., 2016), a constellation of eight micro-satellites launched in 2016 in LEO, is one such example. The radar receiver on each microsatellite measures GPS signals reflected from the ocean surface, and the wind speed is derived from these signals. Work

is still ongoing to determine if CYGNSS has significant value for data assimilation around TCs. Given that CYGNSS currently has a latency of around three days, the potential value of CYGNSS to ECMWF may lie in the evaluation of analyses and forecasts of surface wind fields in TCs.

2.5. Sea surface temperature, ocean waves and sub-surface ocean

ECMWF currently deploys separate data assimilations for the atmosphere, for the ocean waves and for the subsurface oceans. In this section we outline the observation information that is used in these different components and that also can be applied for evaluation.

The main source of information of sea-surface temperature (SST) comes from the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, which generates global, daily, gap-filled foundation SST fields from satellite data and in situ observations (Good et al., 2020). Although the system operates in near-real time, OSTIA is a daily product which results in ~1-day delay between the satellite observations used in the SST and the data assimilation. However, during cloudy conditions the effective delay could be longer, As discussed in Section 6, this could have a significant impact on the SST analysis in the wake of TCs.

Significant wave height observations from nadir looking space-borne altimeters are used by the ocean wave data assimilation. Most altimeters have a small footprint, therefore limiting the likelihood of a direct observation of TC's waves, yet their all-weather capabilities do offer insight into waves in a TC when there is a direct overpass (Magnusson and Bidlot, 2020). Nevertheless, waves propagate away from storm centres as swell and subsequent observations are very useful in correcting swell error. Novel and future altimeters have wider swath capabilities, enhancing their potential impact. Not assimilated yet at ECMWF, ocean wave data from SAR images potentially provide some more information on the wave spectrum, but they are also limited in their coverage. This relates to the discussion about surface winds from SAR discussed above, as the satellite cannot measure both the waves and the winds at the same time.

Moored buoys and platforms, usually deployed in near coastal areas, provide wave information. The wave data are currently used only for model validation and development due to their geographical distribution which has been shown to limit the impact such data can have at a global scale. On occasions, TC's conditions are observed at those locations and severe cases usually lead to the loss of the assets.

Drifting buoys and floats can now be equipped with ocean wave sensors. There is an increasing number of such devices in all ocean basins, some already on the GTS, others in private company's hands. The full potential of these data for both model diagnostics and assimilation has not yet been explored. Some deployments ahead of TCs have shown their potential for wave forecast evaluation, such as for TC Teddy in 2020 (see Section 7).

Subsurface information about ocean temperature and salinity is provided by four main sources: profiling floats, moored buoys, ship based XBTS/CTDS, and marine mammals. The latter two provide little information in the vicinity of TCs given that ships will avoid such situations and mammal borne observations tend to be from seals in high latitude regions. Moored buoys provide high frequency observations of temperature and salinity with low latency given their constant surface presence. Their coverage is higher in the tropical Pacific and Indian oceans compared with the Atlantic basin. The

backbone of ocean *in situ* observations are provided by Argo profiling floats. Argo profilers typically provide one profile in a 10-day period (to save battery life and provide timeseries for climate monitoring purposes). There are more experimental observation platforms available targeting TCs such as available surface drifter observations of ocean temperature, ocean gliders, Uncrewed Surface Vehicles, and Alamo floats. These are currently retained for verification purposes. Alamo floats are Argo-like floats that can be dropped from aircraft ahead of TCs (Sanabia and Jayne, 2020). In an ongoing collaboration with the US Navy and Woods Hole Oceanographic Institution, Alamo floats are used to evaluate the response in the ocean model behind TCs.

2.6. Aircraft reconnaissance and surveillance

Manned reconnaissance and surveillance flights into and around TCs are regularly performed in the Atlantic (and occasionally the Eastern North Pacific) by NOAA (Gulfstream G-IV; P-3) and the United States Air Force (USAF; C-130). The P-3 and C-130 flights provide a variety of observations to fix the location and intensity of the TC ("reconnaissance"), investigate the inner-core structure (using instrumentation such as airborne Doppler radars), and provide input to some NWP systems. The missions are undertaken with different constraints. The G-IV samples the synoptic environment of the TC from high altitude ("synoptic surveillance"), but cannot cross over the TC due to the risk of severe turbulence. On the other hand, the P-3 and C-130 are able to fly in the core of the TC but are usually flying at a much lower altitude (700 hPa) to avoid icing.

Dropsondes are usually launched during the flights. The data are provided on the GTS and are assimilated in the ECMWF forecasts. The assimilation has occasionally been problematic as discussed in Bonavita et al. (2017). One source of the problem was identified to be the horizontal drift of the sonde during the descent. Since that report, some of the dropsondes have started to include position information for each observation, and ECMWF has started to assimilate this information (see Section 6).

Recent improvements in the dropsonde technology also allow a higher temporal frequency of the dropsondes. For example, this was tested during the NASA/NOAA SHOUT campaign in 2015-16 with the unmanned Global Hawk aircraft. Another advantage with this observation platform was a very long operational range making it possible to sample systems in an early stage over the tropical Atlantic and the ability to cross the centre of the TC as it operated at very high altitude (~20 km). However, financial and logistical considerations have shelved the Global Hawk in recent years. On the other hand, small, low-altitude Unmanned Aircraft Systems (Cione et al., 2021) are expected to be significantly more cost-effective and could be used in the future for assimilation and evaluation.

Other products from the manned NOAA and USAF aircraft include measurements of precipitation and winds from airborne Doppler radar in the core of the TC, and surface wind speeds inferred from the Stepped Frequency Microwave Radiometer (SFMR). The aircraft also records winds and temperature at the flight level, in a similar way as commercial flights. The flight-level data are operationally assimilated by NOAA (Christophersen et al., 2021). As assimilation procedure exists for commercial flight data, first exploration of the observations from reconnaissance flights are under way at ECMWF.

ECMWF ensemble data are being used in a project that aims to develop products to identify sensitive regions for TC track forecasts, which could be targeted for aircraft surveillance and additional radiosonde observations. This method was first demonstrated, after a period of testing, by NOAA NHC forecasters using ECMWF forecasts during 2017 and 2018. Led by Ryan Torn at the University of Albany, this technique can be applied to metrics such as the TC latitude or longitude, or going beyond the TC track, could potentially be applied to intensity, wind field or precipitation. It allows ECMWF forecast data to inform on optimum aircraft surveillance flight paths and/or rawinsonde drop locations, to provide additional observation data for data assimilation.

2.7. "Best Track" estimates of position and intensity

At least four times a day, each of the RSMC/TCWCs produce estimates of the position and intensity for all present TCs in their basin. These observations are often referred to as Best Track, but the user needs to be aware of the ambiguity of this term (see below). In this report, we will use the term in a wide sense of the estimates from RSMC/TCWCs. Most of the information in this sub-section is based on WMO (2013).

The Best Track is a subjective human assessment of the TC centre location, intensity, and structure, using all observations available at the time of the analysis. As aircraft missions are generally only present in the Atlantic, the estimates are often only based on different satellite products. A common tool is the Dvorak technique (Dvorak, 1984) where the analyst identifies patterns in cloud features in satellite visible and enhanced IR imagery, and associates them with an intensity (T) number (Velden et al. 2006). From this, look-up tables are available to determine the minimum central pressure (Pmin) and maximum wind (Vmax). As this technique involves a human judgement, uncertainties naturally arise. To minimise this effect, the Advanced Dvorak Technique (Olander and Velden, 2007) is an automated technique to classify cloud patterns and apply the Dvorak rules.

A Best Track record typically consists of centre longitude and latitude, maximum surface wind speeds, central pressures, and may also include the radius of maximum wind, and the maximum radius to 34-, 50-, and 64-kt winds in each quadrant (often referred to as wind radii). The uncertainties in these estimates have been evaluated by Torn and Snyder (2012) and Landsea and Franklin (2013). Torn and Snyder (2012) found a strong case dependency with the position uncertainty being lower for strong TCs while the other way around for intensity (similar to the analysis errors discussed in Section 5.3). They also found a dependence on the presence of reconnaissance flights. The average uncertainties were found to be around 50 km for position, 8 kt for Vmax and 4hPa for Pmin. Landsea and Franklin (2013) finds similar results but it is worth pointing out the difference in position error for tropical storms without reconnaissance flights (63 km) and major hurricanes with flights (15 km).

Once the Best Track estimates are ready, the information is distributed on the GTS network. This product is referred to as working/operational Best Track. The data are also collected in real-time in "TCVitals" files. Note that in some basins, several institutes can issue estimates for the same TCs. After each season, the TCs are re-evaluated, and the estimates can be modified before the final Best Track is completed. The International Best Track Archive for Climate Stewardship (IBTrACS) combines track and intensity estimates from several RSMCs and other agencies to provide a central repository of both working and final Best Track (Knapp et al., 2010).

At ECMWF, the product that is internally referred to as Best Track is delivered from the Met Office and is based on real-time bulletins from RSMC/TCWCs. The current version includes information about position, intensity in terms of central pressure and category. This is used for triggering singular-vector calculations (see Section 6), product generation and for verification purposes. Due to the missing wind information, a different source for the operational verification may be considered for use in the future (see Section 4).

2.8. Summary and discussion

In this section, we have described various types of observations relevant to TCs, for assimilation and evaluation.

Given that TCs spend most or all their lifetimes over the ocean, there is a strong reliance on satellite data (radiances; retrieved profiles of humidity and temperature; wind retrievals). Data from new satellite platforms and instruments continue to be investigated. One example is SAR imagery, from which either surface wind or wave height data can be extracted and tested for both evaluation and assimilation. One current problem with the product is the limited availability, as the instrument needs to be dedicated to one parameter at the time.

Reconnaissance and surveillance aircraft for TCs, which normally only fly in the Atlantic basin, provide a range of observation types. Currently, ECMWF is mainly using dropsondes for assimilation and occasionally for evaluation. Recently, NOAA has implemented a data assimilation scheme in its operational limited-area Hurricane Weather Research and Forecasting (HWRF) model to assimilate aircraft data in the inner-core of TCs (Zawislak et al., 2021). For ECMWF, a feasible test would be to assimilate the flight level data, if the data can be acquired in real-time. For evaluation, a more regular use of SFMR horizontal surface wind profiles can be explored to diagnose the size of the TCs.

Operational estimates of TC position and intensity, referred to as Best Track, are regularly used for evaluation (see following sections) and is tested in Section 6 for assimilation. The estimates are based on all available observations, such as infrared and visible satellite images via the Dvorak technique, satellite surface winds and reconnaissance flights. However, the practices and availability of observations differ between different parts of the world, which leads to inconsistencies. Additionally, there are several databases that contain the data. From the ECMWF side, we need to discuss the most feasible source for the forecast production and the evaluation, to fulfil future requirements for verification.

3. Tracking tropical cyclones

3.1. Introduction

In this section, ECMWF's cyclone tracking algorithms are described. For TCs, ECMWF uses the same tracker for all timescales based on Vitart et al. (1997) and van der Grijn et al. (2005), described in Vitart et al. (2012). For extra-tropical cyclones, the tracker described in Hewson and Titley (2010) is used. As there is an overlap between the two trackers during extra-tropical transitions, both trackers are described in Sections 3.2 and 3.3 respectively. Next, future possibilities to utilise machine learning in TC tracking are outlined in Section 3.4, followed by a short discussion in Section 3.5.

3.2. ECMWF operational tropical cyclone tracker

The main three steps of ECMWF's TC tracker are illustrated in Figure 3. In Step 1, the warm-core TCs are identified using 6-hourly fields of vorticity, temperature, and mean sea level pressure (MSLP) fields at a low resolution (T_L159, about 100 km). The use of low-resolution data is motivated by a reduction of the computing cost and the need to filter small-scale vorticity fields.

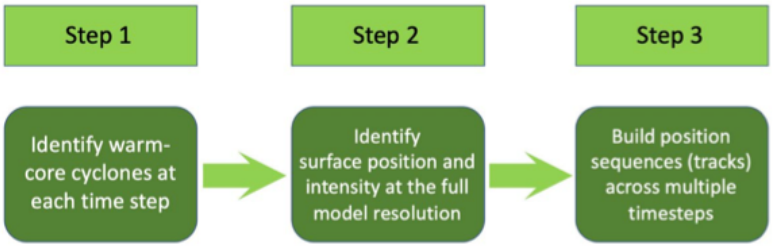


Figure 3: The main three steps of the ECMWF TC tracker.

Following Vitart et al. (1997), a series of criteria illustrated in Figure 4 then need to be met to satisfy Step 1. Although Criterion 3 on the maximum temperature anomaly and Criterion 4 on the maximum 200-1000 hPa thickness seem redundant as they both characterize the presence of a warm core, the use of both criteria is helpful to eliminate some extratropical cyclones that accidentally satisfy one of these criteria. To be defined as a TC, criteria 1, 2a, 3a, and 4a must be verified at all time steps, while criteria 3b and 4b need to be verified at least once in the lifetime of a TC (see step 3 in Figure 3).

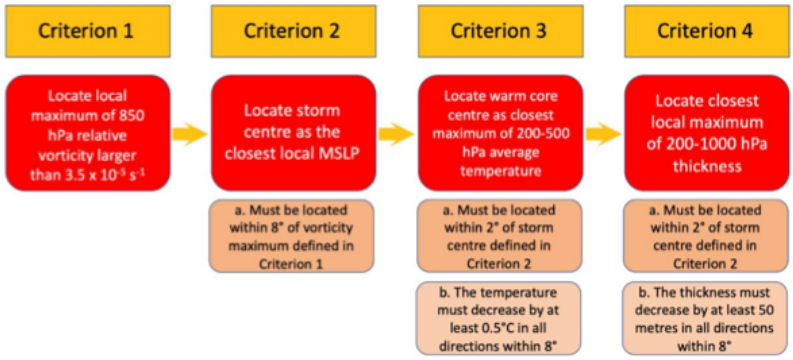


Figure 4: The criteria that need to be met to satisfy the identification of a warm core cyclone (Step 1 in Figure 3).

In Step 2 of Figure 3, the warm core TCs identified in Step 1 are detected using the full model resolution to retrieve the most accurate position and intensity. Since Cycle 47r1, the maximum radius to 34, 50 and 64 knots are also calculated for the four different quadrants (often referred to as wind radii), using a module provided from the Vortex Tracker package (Biswas et al., 2018) developed at the Geophysical Fluid Dynamics Laboratory (GFDL).

Step 3 of Figure 3 builds 6-hourly position sequences (tracks) from the warm core TCs identified in Steps 1 and 2. At each time step, the next position of the storm is calculated using a first guess position based on the steering wind between 200-850 hPa. The warm core TC identified in Steps 1 and 2 that is the closest to the first guess and within 350 km is considered as belonging to the same track. The difference in position between the first guess and the chosen position will be used in a correction term in the following step. A TC may be allowed to "disappear" for a maximum period of 24 hours (a TC may weaken for a short period of time when crossing an island for instance). All the criteria of Figure 4 defined in Step 1 need to be met a least once during the TC lifetime to be included in the tracker database. This means that the track contains both the early part of the life of a TC and the later parts (extra-tropical transition; after landfall). The TC also needs to be present for a minimum of two time-steps and to be located over a sea point at least once or being present in the initial conditions. Finally, the resulting track points are filtered into each of the TC basins. The tracker only searches for TCs over the basins recognised by WMO.

The final tracks are used for ECMWF forecast products and verification. They are also disseminated in BUFR format via FTP, on GTS and to the TIGGE-XML archive.

3.3. Extratropical cyclone tracking in Cyclone DataBase

For extra-tropical cyclone tracking, ECMWF uses the package described in Hewson and Titley (2010) to deliver operational products. This tracker is also able to pick up tropical cyclones that approach the extra-tropics. Shorthand for this facility and its output is "CDB" - i.e. the "Cyclone Data Base". The CDB is currently developed and maintained in collaboration with the Met Office.

During identification the CDB first uses objectively-defined lower tropospheric fronts, based on wet-bulb potential temperature, and then ostensibly looks for local maxima, on those fronts, in the vorticity of the cross-front geostrophic wind, which help define front-related cyclonic features of various types. These are complemented by a set of more simply identified low pressure centres. Together these all form the full "cyclone set" for a given time, albeit with close/co-location avoided by iterative combination, using an all-feature separation matrix, of cyclone pairs that are closer together in space than a pre-defined threshold. The second step of association - i.e. joining the dots, between consecutive times, employs the "half-time tracking" technique, whereby features from timesteps n and n+1 are respectively, and nominally, moved forwards and backwards in time, by half of the timestep interval, using an upper-level steering wind, and previous movement (for continuity) if available, in order to assess the credible and the most probable feature associations. In addition to separation at half-time, the CDB also uses cyclone type change probability (related to training data and a cyclone life-cycle conceptual model), and thickness changes in the algorithm to decide which feature moves where, in an iterative process.

The CDB uses for input standard gridded pressure level data (u, v, T, q) vertically interpolated to a height 1km above the model topography, with an upscaled grid length of ~50km used to retain a synoptic scale focus befitting cyclones. Many diagnostics are computed from the input fields before graphics-based post-processing of these pinpoints the fronts and cyclonic features. This graphical post-processing approach facilitates off-grid cyclone positions. Many thresholds are inevitably employed; these were refined over 2 years of working with real time data with forecasters.

Operationally at ECMWF CDB runs for a domain focusing over the north Atlantic and Europe. The low centre identification algorithm serves to successfully identify the TCs in tropical regions, and during extra-tropical transition (ET) these features are ordinarily seen to convert into the frontal wave cyclone type, as they ingest a semi-linear (i.e. non-axisymmetric) lower tropospheric thermal gradient.

The results from CDB processing are currently used in graphical products but the tracks are not distributed in an alphanumeric format.

3.4. Machine Learning approaches to tropical cyclone detection

In recent years, the problem of identifying TCs (described as Step 1 above) has become a popular application for machine learning solutions. In 2020, ECMWF has started a collaboration with NOAA and NVIDIA for exploring machine learning solutions for this problem. NOAA has initially provided a pre-developed machine learning tool (described in Kumler-Bonfanti et al., 2020) based on Convolutional Neural Networks, and within this collaboration ECMWF are optimizing selected parts of the algorithm. A similar application of machine learning was also explored in the ECMWF Summer of Weather Code 2020.

This project aims at exploring scientific challenges of TC tracking through machine learning, and more general technical aspects related to deploying a machine learning prediction system in operations (e.g., system requirements, reliability, maintainability, etc.). It aims to facilitate the integration of pre-trained machine learning models in high-performance computing environments for inference purposes, inside the NWP models and its post-processing workflows.

The algorithm under development can be used to detect the presence and location of TCs from fields of selected meteorological parameters. The network is currently trained using input fields from ERA5 reanalysis and ground-truth output fields, made by labelling the TC's areas from positional data, as provided by Best Track from the IBTrACS Database (see Section 2). During training the weights of the network are updated to minimize a pre-defined prediction error.

A Neural Network is used to scan multiple global fields and produce a single-channel output field where the TC areas are numerically labelled. This type of algorithm falls into the more general Machine Learning class of Image segmentation, where an image is labelled on a per-pixel basis according to pre-defined categories.

In this works there are many parameters to explore. The network used for this work has a U-NET architecture. The algorithm is currently trained on precipitable water content to mimic satellite images, but the variables in the training data set can be extended (e.g., to the same set as in the operational tracker described above) to include surface pressure, wind speeds, etc. The detection rate vs. computational performance can also be explored by using higher resolution in the input field such

as from ECMWF HRES analyses, or by applying cropping of the input data during the training phase. However, pairing analysis data (ERA5 or HRES) with Best Track can introduce inconsistencies if large errors are present for the TC position in the analysis or Best Track (see Section 5 for analysis errors). The network configuration itself also includes several settings to be further explored.

The future plans include to use this work as a basis to further understand processes behind tropical cyclone genesis and shortcomings in the ECMWF forecast model. This work will be undertaken in the upcoming EU- funded project H2020-CLINT.

3.5. Discussion

In this section we present details about tracking algorithms related to tropical cyclones that are used at ECMWF. We also present the first steps to explore machine learning for this application.

We are aware of the existence of several other tracking software packages, both in ECMWF Member States and elsewhere. One open question here is if it is possible to obtain future synergies in the development and maintenance of the software. ECMWF aims to make the tracking code open to the community in the future.

All TC trackers possess challenges, which are partially due to the difficulty for a model to accurately represent TCs and their vital statistics, and because of the complexity of TC structures in nature. A calibrated tracking algorithm may need updating as models are upgraded, and TCs in nature may not follow straightforward physical relationships that allow for easy calibration. This is especially true during the early stage, when the centre is not well-defined, and there may be significant vertical variation in the central location. The latter also holds for strongly sheared TCs. In contrast, for mature hurricanes, the central position is expected to be tracked well. Although the central pressure (Pmin) is a stable quantity to track, the maximum wind speed (Vmax) and radius of maximum wind are difficult, since they are dominated by small scales and can shift in space and time. Another major challenge in tracking is in distinguishing between tropical, subtropical, and extratropical cyclones. Accordingly, especially for these types of storms as they enter the extratropics, forecast verification statistics may be affected depending on the type of tracker used.

Although the use of machine learning for meteorological applications is still in its infancy, a particularly promising area for use is TC tracking. This is partly due to the presence of a well-defined observation dataset (Best Track, Section 2) to be used as labels in supervised training of a model. While the results are not yet operationally viable, the technique has potential to better mimic the TC properties in the Best Track than a manually calibrated tracking algorithm.

4. Verification Methods for Tropical Cyclones

4.1. Introduction

In the section, we discuss verification practices for TCs that are applied at ECMWF and elsewhere and highlight ongoing developments of the verification at ECMWF. Verification methods for TCs were comprehensively outlined in WMO (2013). The methodologies discussed here will be applied in the subsequent sections about predictability and forecast experiment evaluation.

Figure 5 shows the surface wind structure in Hurricane Laura (2020). When it comes to quantifying their predictive skill, TCs possess an advantage over other weather systems since they are self-contained features with distinct, measurable characteristics, a few of which are tabulated in the Best Track estimates. Such characteristics (highlighted in bold) include the mere **existence of a TC** (to measure TC activity), the latitudinal and longitudinal **position of the centre** of the TC (often referred to as the "track"), and various measures of the TC surface wind structure. The most common measure of the structure is the "intensity", which is usually represented by either the **minimum central pressure** ("Pmin") or the **maximum sustained surface wind speed** anywhere within the TC ("Vmax"). The distance between the points for Pmin and Vmax is referred to as the **radius of maximum wind.** All these metrics are routinely produced in the ECMWF TC tracker.

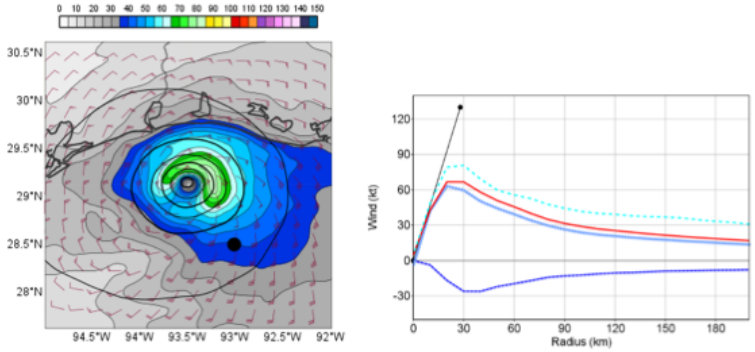


Figure 5: Visualisation of the surface wind structure in Hurricane Laura, for a 60 h HRES forecast initialized on 1200 UTC 24 August 2020. Left: MSLP (black contours), 10-metre winds (wind barbs, in kt), 10-metre wind speed (shading, in kt). Black dot: NHC Best Track position. Right: Radial mean of total (red), tangential (light blue, dotted) and radial (blue) 10-metre wind speed, and maximum wind speed at each radius (cyan, dashed). Black dot: Best Track value of Vmax and the radius to Vmax.

Since 2020, the **radial extent of specified surface wind speeds** (34 kt or 17 m/s, 50 kt or 26 m/s, 64 kt or 33 m/s) in the northeast, southeast, southwest, and northwest quadrants have also been produced in the tracker output and an example is given in Figure 6 for a 60 h forecast of TC Laura. Examples of strong hurricanes (e.g., HRES) to less intense hurricanes with an asymmetric wind distribution are illustrated. The tropical storm force winds in each of the illustrated forecast members extend further to the east than to the west, raising the possibility of broader wind damage and storm surges on the eastern side of the hurricane. The HRES and one of the ensemble members produced accurate 60 h forecasts of the position and wind distribution compared with the NHC Best Track (Figure 6, right panel), whereas other members were weaker or even not tracked.

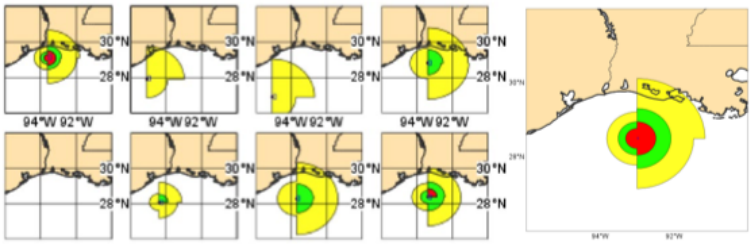


Figure 6: HRES (top left), 47r1-ENS-CF (second from left), and 6 of the 50 ensemble members' predictions of the radii of (yellow) 34 kt, (green) 50 kt, and (red) 64 kt surface winds, for the 60 h forecast of Hurricane Laura initialized at 24 August 12UTC 2020. Right: NHC Best Track of position and surface wind radii at 27 August 00UTC 2020.

To characterise the movement of the TC, the **propagation speed** can be calculated using the current and previous position (usually 6 hours earlier). Forecast track errors can also be decomposed into an **along-track part and a cross-track part** (see Figure 3 in WMO13). Other relevant aspects of the position are to correctly predict **genesis** (both location and timing) and to capture risk for **landfall**.

The key verification measures of TCs include both the predictive skill (via measures such as the Mean Absolute Error or MAE; Wilks, 2006) and bias (via the Mean Error or ME). For many TC characteristics, a probabilistic framework for prediction and verification is desirable.

The predictive skill of these TC metrics, and potentially others, depends not only on the intrinsic predictability of the Earth system, but the ability of the full NWP system to accurately simulate the processes on multiple scales. Traditionally, the focus at ECMWF (and other global NWP centres) has been on track prediction, which is largely (but not entirely) driven by synoptic-scale processes, which are well represented in the ECMWF data assimilation and modelling system. More recently, with the model grid spacing improving to less than 10 km, the model can capture mesoscale features such as the TC eyewall and rainbands. Accordingly, attention is now also being focused on structural characteristics, especially intensity. The intensity is defined in two ways by the TC community: Vmax is used by the NHC and other agencies, while Pmin is commonly used by Met Office and ECMWF, for example. It is generally viewed that Pmin is a more stable and robust metric by which to evaluate TC intensity in global models (Davis, 2018). The Vmax metric, while more volatile than Pmin in its predictability and more difficult to observe, was designed half a century ago with the worst-case scenarios for housing damage in mind. Vmax is largely dependent on convective-scale (O(1 km)) processes that are not resolved in the global models, and it is therefore expected to be regularly underestimated. Another difficulty with wind speed is that different reporting centres use different criteria to estimate the maximum wind speed. While the WMO recommends a 10-minute averaging period, in some centres the Vmax is measured over a 1-minute interval. This inconsistency is not taken into consideration in the verification of Vmax or wind radii in this report. The inclusion of both Pmin and Vmax in Best Track data allows for the evaluation of numerical predictions of both metrics, and they are also considered by other centres for use in assimilation and initialization. In this paper, "intensity" can refer to both Pmin and Vmax.

4.2. Operational verification and recent developments at ECMWF

Based on the verification against the Best Track dataset described in Section 2, statistics on position and Pmin errors are routinely produced at ECMWF. It is common practice to only include TCs that are present at the initial time of the forecast, which excludes forecasts of genesis. When comparing two (or more) models, homogeneous samples are based on the same pairs of forecast events shared by the different models.

As discussed in the Introduction, ECMWF has adopted the 3-day HRES forecast position error as one of its headline scores (see Figure 1). That score is included in ECMWF annual verification report (e.g., Haiden et al., 2021). The report also contains the RMSE of position errors at day 5, mean error and mean errors (bias) of propagation speed and intensity (in terms of central pressure) of HRES and the ENS control forecast. The metrics are also compared to ERA5 forecasts as a reference. For the ensemble, results of RMSE of the position error for ensemble mean and the ensemble standard deviation, as well as probabilistic verification of ensemble track forecasts are shown. Finally, the report includes results from basin-wide TC activity in the seasonal forecast.

Recently, ECMWF began an inter-comparison of the TC forecast performance between ECMWF and other centres. The track forecasts used are from the TIGGE-TC exchange, which provides tracks for both deterministic and ensemble models. All TCs that exist at the time of the forecast initialisation are included.

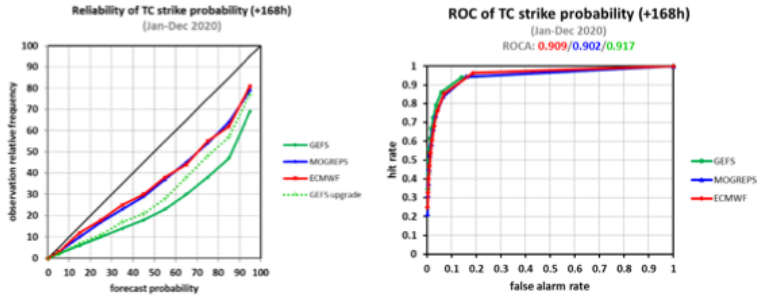


Figure 7: Probabilistic verification of ensemble TC forecasts at day 7 for January-December 2020; ENS (red), MOGREPS-G (blue) and GEFS (green). The dashed line (green) includes only the upgraded GEFS from 23 September 2020. Left: reliability diagram; Right: standard ROC diagram.

Evaluations of the 7-day strike probabilities are presented in Figure 7 for the ECMWF ENS, Met Office MOGREPS-G and NCEP GEFS ensemble forecast systems. This product is defined as the probability that the central position of an existing TC will pass within 120 km during the next 168 hours. All models are found to be overconfident, as shown by the curves below the diagonal line (0,0) to (1,1) in the reliability diagram (left panel) with the GEFS showing the lowest reliability. A significant model upgrade was announced by NCEP on 23 September 2020; the number of perturbed members increased from 21 to 31 and the resolution increased from 33 to 25 km. Preliminary results indicate an improvement of the NCEP GEFS reliability curve based on a smaller sample (from end of

September until December 2020). Using the same ensemble models, Titley et al. (2019) obtained better reliability curves based on a two-year period. Their results considered TCs with Best Track intensities greater than 34 knots (thus excluding tropical depressions) whereas the strike probabilities included times up to the last time that a matching observation of tropical storm intensity or higher was available, thus reducing the false alarms.

In contrast to the reliability diagrams, the Relative Operating characteristic (ROC) curves are almost identical between the three ensembles despite the differences of hit rates for larger probability thresholds for very small false alarm rates. GEFS has the best (largest) area under ROC curve amongst the three ensemble systems (0.917 against 0.909 and 0.902 of ENS and MOGREPS-G respectively).

When comparing the performance of TC forecasts from different centres, one faces several challenges. One is whether to use tracks calculated at each centre or to obtain the meteorological fields and apply the same tracker. The advantage of the former is that it reduces the amount of data to transfer but has the disadvantage of the properties of the different trackers affecting the forecasts (see Horn et al., 2014 for comparison of different tracking schemes). However, if the aim is to score the final product this is not a problem. One difficulty to obtain the meteorological fields is to make sure that they are on a resolution as close as possible to the native model resolution. In the TIGGE archive for example, the resolution is reduced for most centres, which affect the intensity of the TCs. To simplify the exchange of TC track data, there is a dedicated TIGGE archive (TIGGE-TC) for ensemble forecasts of TC tracks (NCAR, 2021), based on the trackers from each centre.

ECMWF is also exploring intensity verification based on wind speed, using Best Track files from other sources than those routinely used at the Centre (see Section 2). ECMWF occasionally produces statistics of Vmax verification, wind-pressure relations (see Section 5) and lately verification of wind radii as illustrated in Figure 8.

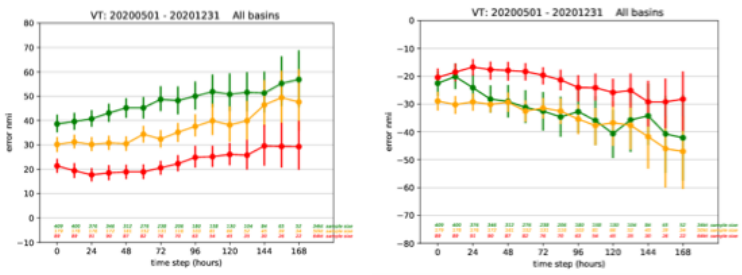


Figure 8: Mean absolute error (left) and mean error (right) for wind radii of 34- (green), 50- (yellow), and 64-knot (red) thresholds, for HRES operational forecasts between 1 May – 31 December 2020. Units in nautical miles (1 n mi = 1.852 km). Vertical bars represent 95% confidence intervals.

The verification of wind radii forecasts is difficult due to the scarcity of surface wind observations, which are critical for obtaining accurate wind structures of storms (Cangialosi and Landsea, 2016).

Because of this, TC forecast centres are still reluctant to publicly release verification on a regular basis. Figure 8 illustrates an example of wind radii verification following the implementation in operations of wind radii product in July 2020. A systematic negative bias is seen for the 34-, 50- and 64-knot thresholds up to 168 h. Overall, the HRES under-estimates the TC size in terms of the wind speed. This could partly be related to an underestimation of the wind-pressure relation discussed in Section 5. While the 34-knot wind radii observations are reported in all basins (south and north hemispheres), the 50- and 64-knot wind radii are only limited to Eastern Pacific and North Atlantic basins. For each wind speed threshold, the mean absolute radii error increases with the forecast lead time. Errors are slightly larger in the ENS control forecast than the HRES suggesting that model resolution has an impact on the TC size (not shown). The amplitude of the error of the 64-knot wind threshold is smaller comparatively with the 50- and 34-knot thresholds. This is expected as 34-knot wind radii is remarkably higher than the other two thresholds (Cangialosi and Landsea, 2016). A future development of this verification would therefore be to scale the error with the radius.

4.3. Future aspects of verification at ECMWF

The previous two sub-sections have summarized the progress in verification of TC position and surface wind, with the increasing ability to simulate TC structure enabling new metrics such as Vmax and the wind radii. Forecasting challenges related to TC structure will be further discussed in Section 5.

One future goal is to evaluate other variables that convey the impact, such as rainfall in connection with TCs. However, it is problematic to find reliable ground measurements from rain gauges and radar, due to the harsh environment. During a visit to ECMWF in 2016-2017 by Tung Nguyen, sponsored by WMO, the skill for rainfall during TC landfalls in Vietnam was evaluated over a ten-year period in ECMWF forecasts against high-density rainfall observations provided by the Vietnamese Meteorological Service. The skill was found to be on a higher level compared to other type of events. Another option for verification is to use satellite derived rainfall from the GPM dataset (Skofronick-Jackson et al, 2017; Huffman et al., 2014) along the TC tracks (Omranian et al., 2018). An example for such comparison is shown in Section 9 (64) for TC Idai (2019). As further discussed in Section 9, accurate representations of precipitation structure and duration are also necessary for flood forecasting and hazard mitigation.

The surface and 3-dimensional wind structure and precipitation distribution are important for capturing rapid intensification, for example. The TC structure is also important on its own, to provide improved forecasts and warnings of hazards associated with TCs. Examples include the broader wind field outside Vmax (which is only at one point), which contributes to storm surges and large-scale wind or wave damage as discussed above. However, many of the structural aspects of the hurricane are poorly observed and predicted, such as the fluxes at the air-sea interface, the structure and processes within the boundary layer, the organization and consequences of updrafts and downdrafts, and eyewall processes such as eyewall replacement cycles and intense mesovortices inside the eyewall. Some quantities can be derived such as outflow from atmospheric motion vectors and rain-structure from radar (airborne or ground-based). But it is not straightforward to set up regular verification of these quantities, and is expected to be time-consuming. It is therefore an area where ECMWF can seek collaboration for the wider community.

Another future goal is to verify variables that characterise the vertical structure of TCs. While these variables do not directly convey the impact, they are useful to evaluate the fidelity of the model analyses and forecasts. Characteristics include the vertical change in wind structure above the boundary layer, the warm core, and the upper-tropospheric outflow. However, it is difficult to verify these quantities routinely, given the limited availability and quality of observations in TCs.

As will be demonstrated in Section 5, the characteristics of the forecast errors and biases depend on the characteristics of the TCs themselves, such as the size or intensity. Hence, there is an ongoing need to perform conditional verification to understand the regional differences in errors and how they differ during different phases of the TC life cycle. In the past, this has helped to understand the impact on ocean coupling.

In operational ensemble verification of field variables, observation uncertainty is taking into account for the spread-error relation (metric discussed in Section 5). For TC verification, ECMWF could in the future explore to take the Best Track estimated error into account for this type of verification.

One uncertainty in the intensity verification is the inconsistency in the Best Track estimates due to different practices in different basins and availability of observations. Since 2016, ECMWF operationally produce simulated satellite images based on infrared channels on Meteosat (see Figure 54). This gives the opportunity to track TCs and estimate intensity in the model with the Advanced Dvorak Technique (Olander and Velden, 2007) in a consistent way as done with real satellite images. This was documented in Magnusson et al. (2017). The result was successful for most of the cases. The main shortcoming that was identified was in situations when cirrus clouds obscure the eye of the cyclones.

Direct evaluation based on IR-images as shown in Section 7 should also be considered in the region around tropical cyclones. Such targeted evaluation could also be applied based on data assimilation output for other assimilated observations. We could also consider making direct evaluations of non-assimilated observations discussed in Section 2, such as SFMR surface winds from reconnaissance flights.

ECMWF does not currently perform a regular verification of TC activity/genesis for medium-range forecasts. However, it is evaluated for extended-range forecasts. In Yamaguchi et al. (2015), the TC activity of ECMWF forecasts was compared with other forecasts in the TIGGE archive. One difficulty here is the use of specific wind speed threshold to detect cyclones, which makes the number of the TCs sensitive to biases in maximum wind speed. The solution applied in Yamaguchi et al. (2015) was to find the intensity threshold that maximised the skill for each system. As similar method was applied in Bergman et al. (2019) for seasonal forecasts. The results in Yamaguchi et et. (2015), performed during a sabbatical at ECMWF by Munehiko Yamaguchi (JMA), showed an advantage for ECMWF forecasts in most of the basins. However, as this study is based on forecasts that are about ten years old, it is time to redo this type of verification. The probability of TC genesis in the ENS forecasts is discussed in Section 5.

Focusing on seasonal timescales, Bergman et al. (2019) made verification of landfall risks in ECMWF forecasts. Such verification could also be adapted to shorter timescales such as extended-range.

5. Forecast challenges at different stages of the TC life cycle

5.1. Introduction

As for all weather systems, predictions of TCs pose challenges on multiple scales, spanning from correctly initialising the system on the relevant scales, to accurately modelling the processes on multiple scales, to accurately representing the climatology of their occurrence. In this section, the characteristics of predictions and their errors on different timescales are presented. To begin, the forecast and analysis errors based on ECMWF's standard overall verification are described in Sections 5.2 and 5.3 respectively. Following this, the characteristics of TC position (track) and intensity errors are provided in Section 5.4 and 5.5 respectively. Prediction of genesis, marking the beginning of the TC's lifetime, is addressed in Section 5.6. Extratropical transition and decay, marking the end of the TC's lifetime, is described in Section 5.7. Extended-range and seasonal predictions of TC activity are summarized in Section 5.8, followed by a summary of the key results in Section 5.9 and main challenges in Section 5.10.

5.2. Effect of model resolution on forecast errors

To give an overview of the errors in the operational forecast with different resolutions, Figure 9 shows the mean absolute error (MAE; Wilks, 2006) for position, Pmin, and Vmax, the mean error (or bias, ME) for Pmin and Vmax, and the wind-pressure relation for the following three sets of forecasts during July-November 2020:

(i) Operational HRES for Cycle 47r1;

(ii) Ensemble control for Cycle 47r1 (ENS-CF-47r1); and

(iii) Pre-operational control for Cycle 47r2 (ENS-CF-47r2).

During this period, the horizontal resolution for HRES was 9 km and ensemble control forecast (CF) 18 km. In Cycle 47r2 that became operational in May 2021, the HRES forecasts performed very similarly to the previously operational Cy47r1 HRES forecasts (not shown). In Cycle 47r2, the vertical resolution of the ensemble was increased from 91 to 137 levels to make it the same as HRES (Lang et al., 2021). The cycle also included a change to single precision that did not have any significant impact on TC performance.

The position errors in HRES, ENS-CF-47r1, and ENS-CF-47r2 are very similar up to day 3 (Figure 9(e)). Beyond day 3, the ENS-CF-47r1 errors become larger. On the other hand, the position error for ENS-CF-47r2 becomes closer to HRES. However, all differences are within the 95% (unpaired) confidence interval. One can note that the confidence interval is large for long lead times as the sample size decreases, and TCs are more likely to be interacting with extratropical systems where very large errors can occur.

For Pmin, the HRES ME (bias) remains at about +2 hPa during days 0-3, but then decreases and becomes negative after day 4 (Figure 9(b)). The change in ME is likely related to the stage of the TC lifecycle, as the sample in the verification of longer lead-times contains a higher fraction of TCs that were mature at the initial time (Rodwell et al., 2017, Figure 1). ENS-CF-47r1 develops a higher positive ME after the initialisation, given a more limited ability to intensify TCs due to the lower resolution. The difference in bias is reflected in the lower MAE for HRES than the ENS-CF-47r1

(Figure 9(a)). Stronger TCs are evident in ENS-CF-47r2. We believe that this is linked to a larger fraction of large-scale (resolved) precipitation with the higher vertical resolution, which helps to build up the vertical circulation in the TC (see discussion around explicit deep convection in Section 7). Finally, as will be discussed further down in this section, strong biases in HRES exist for Pmin when the initial TC is below hurricane strength, even if the average bias is small.

For Vmax, all configurations underestimate the maximum wind speed (Figure 9(d)). The relation between Pmin and Vmax, visualised as a wind-pressure plot in Figure 9(f), aggregated for 24-120 h forecasts, does not show any large differences between different lead-times. The lines are based on a quantile-quantile mapping between the two quantities. As expected, TCs with a lower Pmin possess a stronger Vmax. All configurations underestimate the winds for a given Pmin, and the underestimation is worse for both ENS-CF-47r1 and ENS-CF-47r2 compared with HRES. This indicates that increased horizontal resolution improves the relationship (see also Section 7).

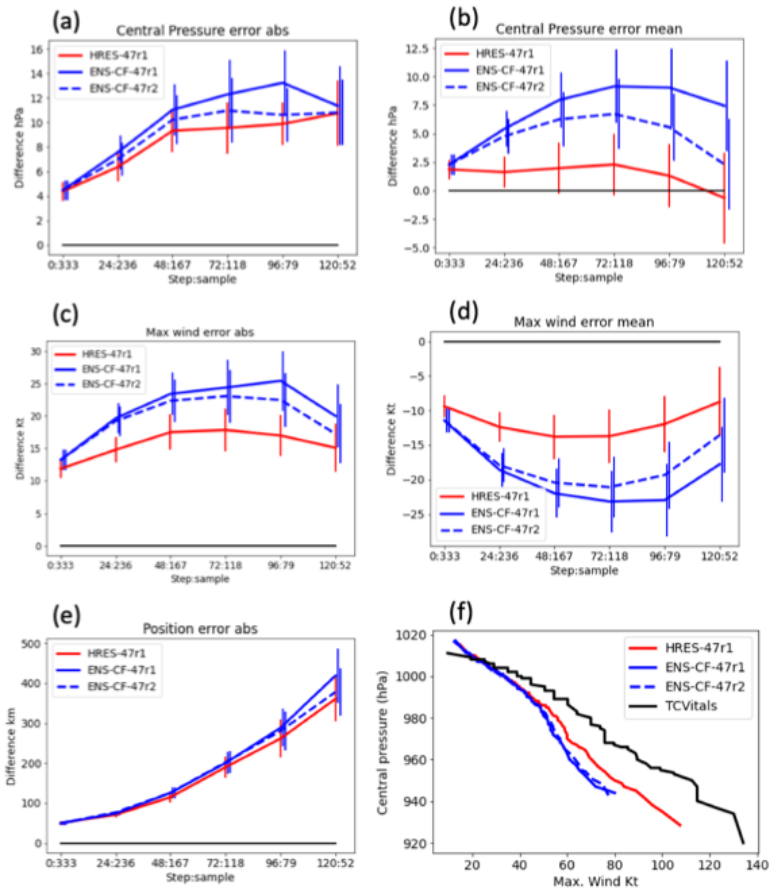


Figure 9: Position and intensity evaluations for (red) HRES, (solid blue) ENS Control, and (dashed blue) pre-Operational 47r2 ENS Control: (a) Mean Absolute Error and (b) Mean Error for Pmin. (c) Mean Absolute Error and (d) Mean Error for Vmax. (e) Mean Absolute Error for position. (f) Wind-pressure relation for 24–120-hour forecasts.

5.3. Analysis errors of tropical cyclones

Initializing an accurate physical structure of the TC in the data assimilation is a challenging task, due to the lack of observations within the TC to constrain the system and the limited resolution of the data assimilation. In this section, the analysis errors are presented. A more in-depth discussion of the data assimilation challenges will be provided in Section 6.

In Figure 10, the position error of the ECMWF analysis is illustrated for all Atlantic TCs after 1 July 2020 (cycle 47r1). The position is produced by the ECMWF tracker, and the verifying value is the Best Track from NHC. We note that there are uncertainties in the Best Track, as documented in Landsea and Franklin (2013), although we expect that these uncertainties are now lower in the 2020s due to more advanced satellite imagery. As an example, the 2013 NHC estimate of position error uncertainty for tropical storms (Vmax < 33 m/s) was 53 km if no aircraft reconnaissance data were available. For major hurricanes (Vmax > 49 m/s), the corresponding estimate was 26 km, given the relative ease in identifying the position of the eye. Overall, the majority of the 226 ECMWF analysis errors of TC position is 30 km or less. However, a small fraction of these errors is large (> 60 km), and most of these cases are within the subset of weak TCs including depressions (pink bars). On the other hand, no such large analysis errors of position are evident for hurricanes (light orange bars).

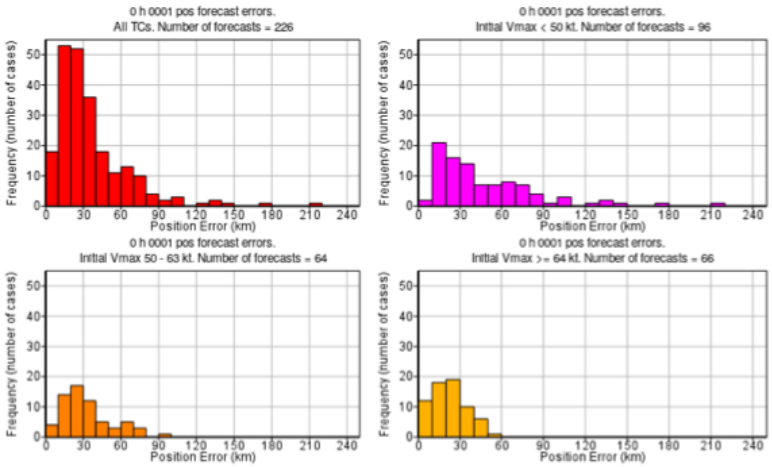


Figure 10: ECMWF analysis position error as a function of initial Vmax (Best Track value).

The corresponding ECMWF analysis error estimates for Pmin are shown in Figure 11. While over 100 of the 226 cases show a very small absolute analysis error (<2.5 hPa), there is approximately an equal number of cases where the analysis value of Pmin is too weak by at least 2.5 hPa. Several of these cases, especially the small number of extreme ones, correspond to hurricanes that are far too weak in the initial conditions (up to 40 hPa). Although not shown here, the majority of the analysis values of Vmax are at least 5 m/s weaker than the corresponding NHC Best Track values.

To conclude, the largest initial position errors usually occur in weak TCs while the largest initial intensity errors usually occur in strong TCs.

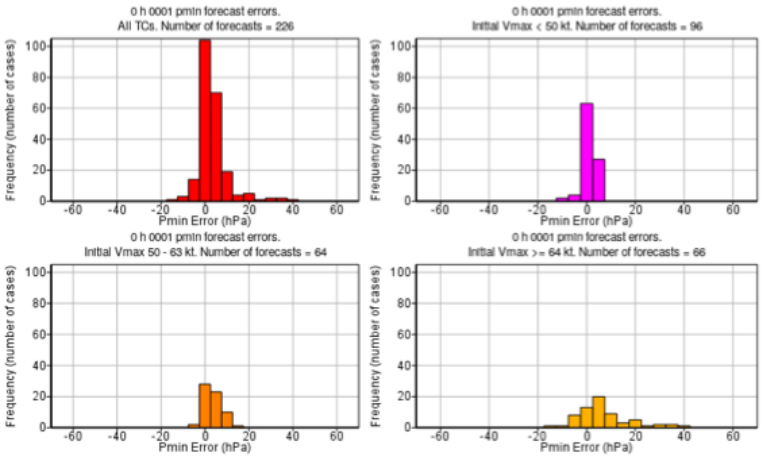


Figure 11: ECMWF analysis Pmin error as a function of initial Vmax (Best Track value).

To put these results in perspective, these values are compared against the US Navy's operational COAMPS-TC® system (Doyle et al. 2014) which uses a TC vortex-following grid with an innermost grid spacing of 4 km. Unlike ECMWF, COAMPS-TC employs the insertion of a synthetic, relocated and balanced Rankine vortex to initialize all numbered TCs. The value of Vmax is constructed to match the operational RSMC estimate, and the radial wind profile is adjusted to fit the RSMC estimate of the radius of maximum winds and radius of 17 m/s winds (Komaromi et al., 2021). Hence, by design, the initial vortex in COAMPS-TC is forced to contain much lower initial condition errors of position and intensity. The COAMPS-TC analysis errors of Pmin (evaluated versus the NHC Best Track) are shown in Figure 12. A comparison of this figure with Figure 11 shows a higher fraction of COAMPS-TC Pmin analysis errors nearer to zero. The contrasts between the COAMPS-TC and ECMWF analysis errors are more distinct in the position and especially the Vmax errors.

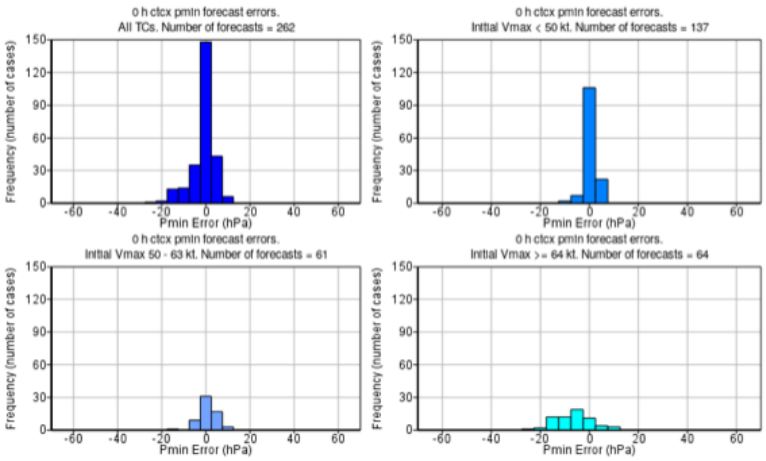


Figure 12: Same as Figure 11 but for COAMPS-TC.

While a homogeneous sample during the 2020 Atlantic hurricane season was used in this comparison between ECMWF and COAMPS-TC, the number of cases, even at 0 h, was substantially less in ECMWF (226) than in COAMPS-TC (262). The reason for this is that weak TCs were sometimes not trackable in the ECMWF analysis fields. In COAMPS-TC, the full sample of TCs exists by design of the synthetic vortex initialization.

5.4. Position (track) predictions

Arguably the most important aspect to predict for TCs in the medium range is its future position (track). As seen in Section 1, the skill of ECMWF TC position forecasts has improved over the past decades, but other centres have also improved at a similar pace. However, there exists a year-to-year variability in the errors. For example, as illustrated in Figure 1, the average 5-day HRES position errors in 2020 were larger than those of the previous five years for HRES. There are still several challenges in track prediction, some of which are described and illustrated below. In Magnusson et al. (2019b), the diagnostics of several challenging cases were discussed. Examples of tools used in such investigations include ensemble sensitivity (Torn et al., 2018), adjoint sensitivities (Doyle et al., 2012), steering flow diagnostics (Tang et al., 2021) and relaxation experiments (Magnusson et al., 2019b). Difficult cases to predict are often related to bifurcation points in the steering flow (Riemer and Jones, 2012), where a small error in the position of the TC can make a difference between a TC recurving into the mid-latitude flow, or instead continuing its westward propagation.

In Figure 10 we found that the analysis error and uncertainty for position is larger for weak TCs compared with stronger ones. From this finding, one can ask how the initial intensity of the TC affects the position (track) forecast. To answer this question, Figure 13 shows the position error for Atlantic TCs between 1 July – 18 November 2020 (Cy47r1), divided according to their initial intensity. It is evident that the average position error is over 70% larger for initially weaker TCs (magenta line) than

those for initially mature hurricanes (brown line) in the short range (24-48h), and this large discrepancy continues out to 84 h. This is an expected result, given that there is larger initial uncertainty in the structure of the weaker TCs and their environmental interactions.

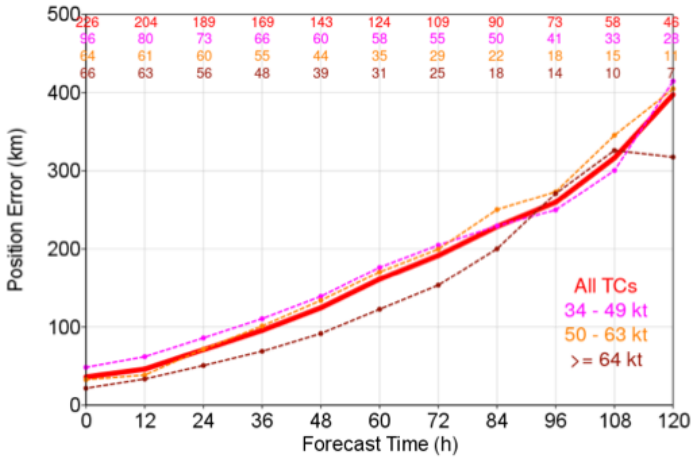


Figure 13: MAE of operational HRES forecasts for position, for all Atlantic basin TCs between 1 July – 18 November 2020 (AL05 Edouard – AL31 Iota), stratified by Vmax at the initial time. Red: all cases. Magenta: only TCs with Vmax between 34-49 kt (weak tropical storms). Orange: only TCs with intensity 50-63 kt (strong tropical storms). Brown: only TCs with Vmax> 64 kt (hurricane strength). The sample sizes of each subset are listed by their respective colour.

The key tool for predicting TC position in the medium range is ensemble forecasts, and we next focus the diagnostics on this system. A common metric in ensemble-based prediction and evaluation is the ensemble mean. For TC position, it refers to the Euclidean mean of the positions from each ensemble member (in contrast to other metrics where the ensemble mean refers to a field average of all members). As a start, the error of the ensemble mean and standard deviation (spread hereafter) of TC position are displayed in Figure 14 for all TCs in 2020 in all basins. In a well-tuned ensemble, these two quantities should coincide over many cases. On average, we find that the ensemble spread is lower than the average error of the ensemble mean. Figure 15 shows a scatterplot of ensemble spread and ensemble mean error for 3-day and 5-day forecasts from the Atlantic cases in 2020. The plot also includes the running mean based on 20 ensemble spread points sorted from low to high. For both lead-times, the running-mean line is above the 1-1 line, indicating a larger error than spread. However, it is also evident that the ensemble has some discriminatory skill to predict cases with larger error from the ensemble spread. For the 3-day lead-time, the most problematic forecasts were for Laura, Eta and Zeta. In the case of Laura, the forecast issued on 24 August 00UTC showed very large ensemble spread, while the forecast from 25 August 00UTC showed very large ensemble mean error, and this case will be discussed further down in this section.

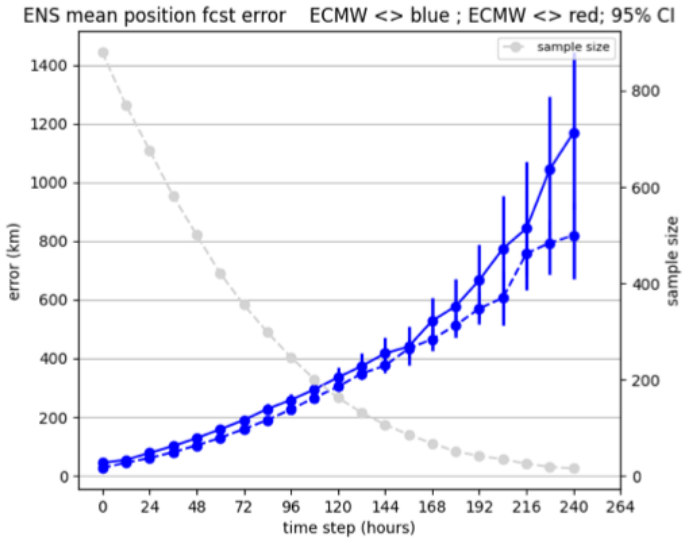


Figure 14: Mean position error of ensemble mean (solid) and ensemble spread (dashed) for cases during 2020 for all basins. Units in km. Vertical bars represent 95% CI.

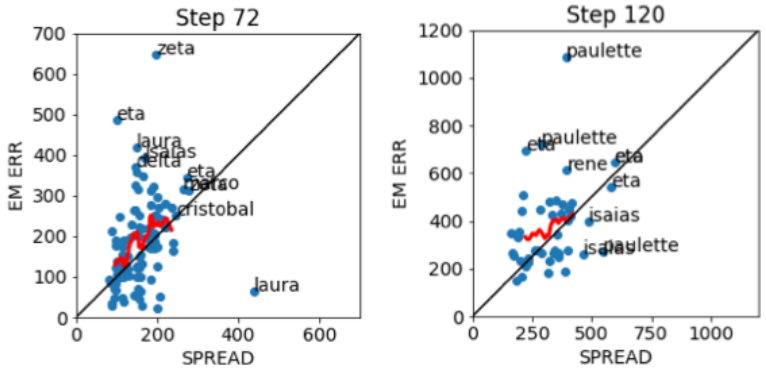


Figure 15: Scatterplot of 3-day (left) and 5-day (right) ensemble spread (x-axis) and ensemble mean position error (y-axis) in km for cyclones in the Atlantic during the 2020 season. A running mean based on 20 ensemble spread points are included as red line. The 1-1 line is included in black. The most extreme cases are annotated.

The four forecasts with the largest 5-day ensemble mean position errors in 2020 (each exceeding 600 km) are illustrated in Figure 16. In each case, the TC was gaining latitude during the forecast window as it moved towards the extra-tropics. The ensemble spread in each of these cases (dashed green line)

was sometimes substantially lower than the corresponding error of the ensemble mean (solid green line), consistent with the underprediction of the ensemble mean error shown in Figure 14. As is evident from a visual inspection of the ensemble forecast tracks (grey lines), the verification at five days lies either on the periphery of the ensemble, or outside the entire ensemble. Further investigation is required on the challenges of ensemble forecasts to encapsulate the actual TC position in a disproportionately large number of cases. These challenges include ensemble initialisation and model perturbation techniques, which will be discussed in Sections 6.9 and 7.4 respectively.

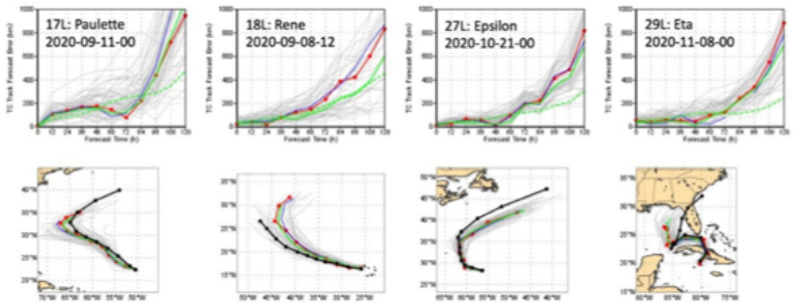


Figure 16: Operational HRES (red), ENS-CF-47r1 (blue), ensemble members (grey), ensemble mean (solid green) forecast errors (top) and forecasts (bottom) out to 5 days, evaluated against the NHC Best Track (black) for four TC cases. These cases possessed the largest 5-day ensemble mean position errors of all Atlantic TC forecasts in 2020. The ensemble spread is represented by the dashed green line in the top panels.

To further illustrate the forecasting issues with TC Laura, Figure 17 shows the longitude for the crossing of 30°N in each ensemble member (red) and HRES (blue), for forecasts initialised on different dates. The crossing of 30°N was chosen to be the latitude of the coast of Texas/Louisiana where the cyclone made landfall on 27 August. The forecasts issued from 24 August 12 UTC to 25 August 12 UTC had almost all ensemble members to the west of the actual crossing. For earlier forecasts, a large ensemble spread was present, but we can also see consecutive forecasts in which a majority of the ensemble members "flip-flop" between the east and west of the actual crossing, which is undesirable from a forecaster's perspective.

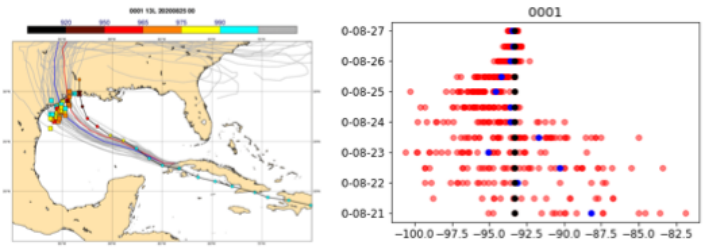


Figure 17: Left: Forecast for TC Laura from 25 August 00UTC. HRES (red), ENS Control (blue), ensemble members (grey) and BestTrack (black). Symbols indicate position and intensity on 27 August 00UTC. Right: Longitude of forecasts passing the latitude of 30N for HRES (blue) and ensemble members (red) for different initial times. Best Track longitude marked with black.

We can summarise the inconsistency (jumpiness) in a sequence of ensemble forecasts using the divergence index (DI) of Richardson et al. (2020), shown in Figure 18. Applying this to the cross-track errors for the 2020 Atlantic TCs shows that Laura stands out as the most inconsistent of all these cases. This is in agreement with subjective feedback from forecasters who were trying to assess the areas most at risk along the US Gulf coast. The inconsistency was especially large around the time of landfall (verification times between 26 August 12UTC and 27 August 12UTC).

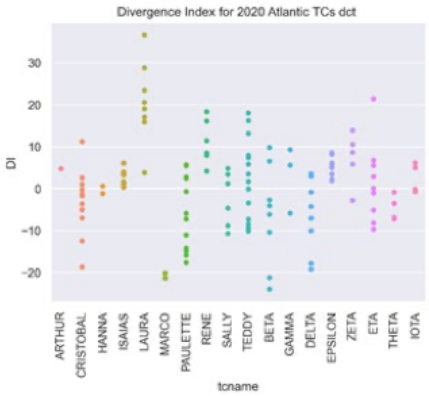


Figure 18: Inconsistency of ECMWF ensemble track forecasts (2020 Atlantic; all cases with at least 20 members for at least T+12,...T+60). Divergence index (DI) – large positive numbers indicate inconsistent cases (large negative numbers indicate high consistency).

While one can argue that the cross-track part of the error is most important to determine the position of the landfall, the along-track error can affect the timing of the landfall and hence the interaction between the storm surge and tides. Both can also be critical in determining interaction with extra-tropical features, such as an upper trough. Whether or not the TC ends up in the mid-latitudes, ahead

of the trough (where it may well be reinvigorated) or stays in the tropics can depend on whether it moves fast enough to catch the upper trough 'bus'. Some of the largest position errors can occur in such circumstances. Hurricane Leslie, which caused widespread damage when it hit Portugal as a re-intensified feature in October 2018, was one such case.

Figure 19 shows the distribution of along-track and cross-track errors for all 2-day forecasts of Atlantic TCs in 2020. For the cross-track biases (left panel), the westward propagating TCs (left panel) exhibit a slight overall northward (right-of-track) bias. For the subset of weak westward propagating TCs, this drift to the north is often evident, as illustrated in Figure 20. Most or all of the ensemble members (grey lines) are north of the actual track (black line), and this challenge warrants further investigation. On the other hand, TCs that propagated to the north, which often were strong TCs, in many cases showed a west (left-of-track) bias (Figure 19, right panel). For these northward moving storms, a strong negative (slow) along-track bias is also present.

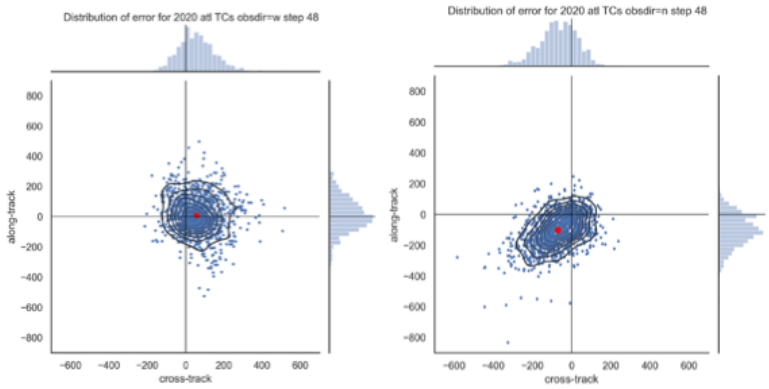


Figure 19: Cross-track and along track errors for westward moving (left) and northward moving (right) TCs (2020 Atlantic). T+48. The red circle shows the distribution mean. The sub setting of the cases is based on a movement direction +/- 20 degrees from west and north respectively. Positive cross-track bias indicates right-hand bias.

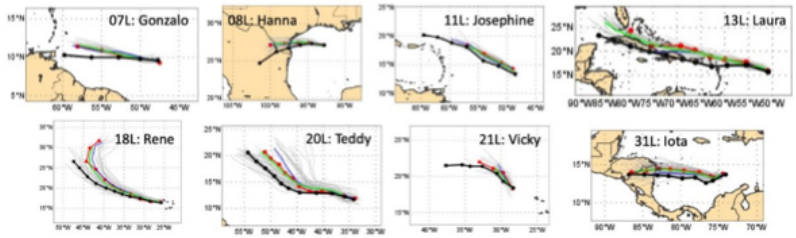


Figure 20: Examples of 2020 forecasts in which weak, westward-moving Atlantic TCs were predicted to the right (north) of the actual track. Black: NHC Best Track. Red: Operational HRES (47r1). Blue: ENS-CF-47r1. Grey: Ensemble members. Green: Ensemble mean. Dots: 12-hourly locations.

The along-track bias for short forecast ranges (48-hour) agrees with the propagation speed bias diagnostic that is routinely produced in the ECMWF verification report (Figure 32 in Haiden et al., 2021), where we find a long-term negative bias on the order of 1 km/h. This also agrees with other studies as Leonardo and Colle (2020), who used ECMWF ensemble forecasts from the TIGGE archive. They found the largest contribution to the bias from cases approaching the mid-latitudes. However, in their data a (smaller) bias was also present in cases that dissipated in the tropics. One could speculate that the stronger signal for the high-latitude cases is caused by the acceleration of the TC when interacting with the mid-latitude flow, which stretches the differences. In Chen et al. (2019), the GFDL-FV3 model and IFS with the same (ECMWF) initial conditions were compared. FV3 showed much less of a slow bias than IFS and had lower track errors for longer lead times.

In 2018 scientists at ECMWF made a range of sensitivity experiments to target the slow propagation bias. Even if some results indicated a sensitivity to the model time-step, the results were mainly inconclusive. This topic will be further explored in Section 7.

5.5. Intensity predictions

In contrast to track forecasts, operational forecasts of intensity did not improve at the same pace in the 1990s-early 2000s (e.g., Fig. 6 of Rappaport et al., 2009). In response to this concern, together with multiple hurricanes that made landfall in the United States in 2004 and 2005, NOAA initiated the Hurricane Forecast Improvement Program in 2009 (Gall et al., 2013). In Titley et al. (2019), intensity forecasts from ensembles were highlighted by users as an area where they want to see progress. The approach at many weather services has been to use regional NWP systems with TC-following nested grids with grid spacings of 4 km or less such as AROME Overseas, HWRF and COAMPS-TC, to resolve the convective-scale processes that produce high wind speeds in the eyewall. These regional systems have also been employed over the past decade in hundreds of research articles that shed new light on TC structure and intensity evolution, and their predictability. As is now evident in regional systems such as HWRF and COAMPS-TC, predictions of Vmax have improved substantially during the late 2010s (Cangialosi et al., 2020). It is worth noting that these models have their own vortex initialization schemes, with HWRF recently introducing regional data assimilation of inner-core data including aircraft data (Zawislak et al., 2021; Christophersen et al., 2021).

With the horizontal grid spacing of the ECMWF HRES now smaller than 10 km, it is finally within reach to accurately predict the intensity of larger TCs with eyewalls that can be resolved in the model, and an improvement in Pmin errors has been achieved in the past decade (Figure 1).

The evaluation of Pmin and Vmax analyses and forecasts for all Atlantic TCs between 1 July – 18 November 2020 reveals some important although unsurprising results for consideration. First, the MAE in intensity analyses and forecasts vary based on the initial intensity. Especially for forecasts of three days or less, the initially weaker TCs have lower forecast errors (Figure 21(a)). This is largely due to the presence of a high proportion of weaker TCs that do not intensify significantly, thereby producing several low-error cases. On the other hand, for TCs of hurricane intensity, the initial MAE of Pmin is 7.5 hPa (Figure 21(a)), indicating limitations in the data assimilation system to deepen strong TCs to their estimated values in the Best Track. This MAE increases to approximately 13 hPa for 36-48 h forecasts of TCs of hurricane strength. Similar conclusions are drawn for Vmax forecasts (Figure 21(b)). For weaker TCs, the initial Vmax errors are small, and they amplify gradually out to five days. For TCs of hurricane strength, the average initial Vmax errors are very large, exceeding 25 kt. The Vmax error decreases with longer lead times, partly because of the storms becoming weaker and partly as a compensation of errors with the developing Pmin bias.

The corresponding biases in the intensity analyses and forecasts provide further insights. For initially weak TCs, the Pmin forecasts are consistently too weak (Figure 21(c)). In this sample, several cases exist in which the forecast erroneously does not intensify an initially weak TC. For TCs of hurricane strength, the forecasts beyond one day yield overly strong TCs. An examination of the individual cases reveals that this strong bias reflects a generally limited ability of the forecasts to increase the Pmin during the weakening phase of the TC. The biases in Vmax demonstrate a systematic underestimation of the maximum surface wind at all intensities (Figure 21(d)). The bias at longer lead times is lowest for initially strong cyclones, which should be related to the too low Pmin in these cases.

To summarise, the stratification by initial intensity reveals several challenges in analysing and predicting TCs, with contrasting challenges for weak versus strong TCs.

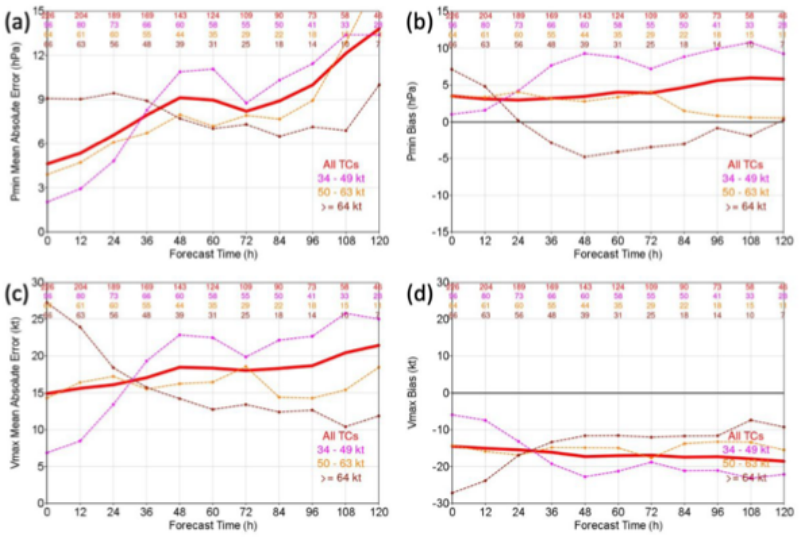


Figure 21: MAE of operational HRES forecasts for (a) Pmin and (c) Vmax, and respective mean error or bias (average of forecast value – verification value) for (b) Pmin and (d) Vmax, for all Atlantic TCs between 1 July – 18 November 2020, stratified by Vmax at the initial time. Red: all cases. Magenta: only TCs with Vmax between 34-49 kt (weak tropical storms). Orange: only TCs with intensity 50-63 kt (strong tropical storms). Brown: only TCs with Vmax> 64 kt (hurricane strength). The sample sizes of each subset are listed by their respective colour.

TC intensification and weakening processes need to be captured more accurately in the NWP system. The intensity change depends on a variety of processes on multiple scales, which may act non-linearly to positively reinforce each other or compete against each other. First, the coupling with the ocean is important to ensure that the fluxes of temperature and moisture at the ocean surface are appropriately represented. Further details of ocean coupling are provided in Section 7. Second, the environmental wind field, commonly represented by the 200-850 hPa vertical wind shear, is known to be influential on intensity change. However, although the general assumption is that moderate to strong wind shear (5-15 m/s) is usually detrimental to intensification, there are situations in which TCs can intensify, even rapidly, in these conditions. Third, the presence of low-humidity air is also usually assumed to be detrimental to intensification. Other environmental factors include interactions with subtropical or extratropical jets, which may control the outflow and thereby the intensity change, and interactions with land which modifies the supply of thermal energy and moisture through the TC surface layer and boundary layer.

Additionally, and of equal importance, are inner-core processes on the mesoscale and convective scales, which are governed in part by the environmental interactions. These inner-core processes include the organization of initially disorganized convection, rainband formation, and eyewall formation leading to intensification. The reverse of these processes can lead to TC weakening.

An especially challenging and societally important aspect of intensity change is Rapid Intensification (RI), in which Vmax increases by at least 30 kt (15 m/s) during a 24-hour period. The statistical and dynamical (global and regional) models to date have not exhibited high skill in predicting RI (Cangialosi et al., 2020), and it remains a very active research topic in the field of tropical meteorology. It is very difficult to discriminate between RI and steady intensification (intensification at a rate of < 15 m/s in 24 hours). RI can often begin early in the life cycle of a TC, even before it is a named tropical storm, which presents an additional challenge in the accurate initialization of a weak, disorganized TC. The abrupt axisymmetrization of convection and corresponding sharp increase in vorticity remains one of the toughest analysis and forecasting challenges in all NWP, even in regional models with a 1 km inner grid. While the environmental interactions are important, it is the precise nature of the O(1 km) inner-core processes, including eyewall formation and replacement, that needs to be captured physically. Nevertheless, some recent forecast cases (COAMPS-TC for Hurricanes Eta and Iota in 2020) demonstrated an ability to capture the RI process, albeit imperfectly, suggesting that there is potential for improvement in future global NWP systems when reaching 4-5 km resolution. A probabilistic approach to RI, and intensity change in general, would likely be necessary.

Figure 22 shows the Pmin change rate over six hours (48-42h) in BestTrack and HRES for 2020 cases in all basins. The slope of the quantile-quantile matching line shows that the model is both too slow for the intensification and too slow for the weakening, compared with the Best Track. This result is consistent with the discussion above based on the conditional bias for initially weak storms that develop a weak intensity bias, while initially strong storms develop a strong intensity bias. We can also see three cases that stand out for missed intensifications: Goni (NW Pac), Iota (Atl), Eta(Atl). These three were small TCs that underwent a very RI that was clearly missed in HRES. But as mentioned above, the RI was captured by the COAMPS-TC regional hurricane model, illustrating the future prospects for predicting RI.

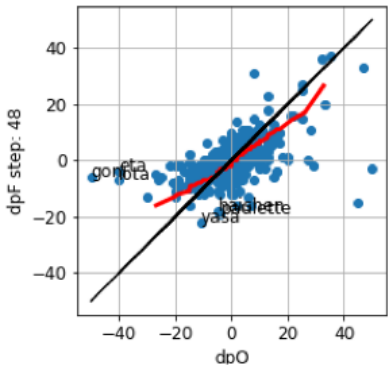


Figure 22: Intensification rate (hPa/6h) in BestTrack (x-axis) and 48-42 hour forecasts (y-axis) from 2020. Red line shows quantile-quantile matching between predicted and BestTrack.

5.6. Genesis

Tropical cyclogenesis (frequently referred to as "genesis") refers to the instant at which a TC forms. In its basic state, the tropical atmosphere is benign. For example, in the tropical Atlantic, the relative humidity above the boundary layer is typically 60-70%, and weak subsidence serves to impede the development of new clouds. Hence, the middle troposphere needs to be moistened by repeated convection for a TC to develop. In addition to this enhanced moisture, it is conventionally thought that low vertical wind shear and high thermodynamic instability (aided by a warm ocean) are favourable environmental conditions.

We first evaluate the capability of the model to simulate the frequency of genesis occurrences. However, this is dependent on the wind threshold set to count as a genesis event. Given the low Vmax bias discussed in Section 5.2, we expect the model to underpredict the number of cases if we apply the same threshold for a tropical storm (17 m/s) as in the Best Track. We illustrate the effect of this Vmax threshold on the number of forecast TCs in Figure 23 . In general, the total number of TCs (including tropical depressions and higher) is overpredicted in the model compared with the observed number, while the number of tropical storms is underpredicted, consistent with the general low bias in Vmax. The number of TCs is increased in cycle 47r2, as expected from the increased intensities shown in Section 5.2. This cycle increased the vertical resolution in the ENS to be the same as that of the HRES, and the increase in the number of TCs is consistent with the increased intensity seen in Figure 9.

We can evaluate the impacts of the wind speed threshold and the cycle change using the performance diagram (Halperin et al., 2013). Figure 24 shows for five-day forecasts that adjusting the maximum wind threshold to reduce overall bias is a trade-off between number of hits and false alarms (success ratio is equal to one-false alarm ratio), while the overall skill measured by the threat score (curved lines – closer to top-right corner indicates higher skill). Cycle 47r2 improves the skill in forecasting the genesis of tropical storms, especially at the longer forecast ranges.

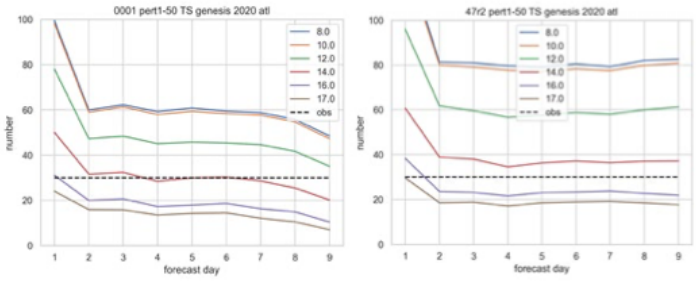


Figure 23: Genesis of TCs in the ENS. Average number of TCs that develop in the perturbed ENS members on forecast days 1 to 9 in the Atlantic basin in 2020 (10 May – 30 November). Left: operational ENS (47r1), right: pre-operational cycle 47r2. Coloured lines show the average number of TCs defined using different thresholds of maximum wind speed from 8 m/s (threshold for tropical depression in operational tracker) to 17 m/s (threshold for tropical storm). Black dashed line shows the observed number of tropical storms (30 cases; 31 TCs altogether including tropical depression).

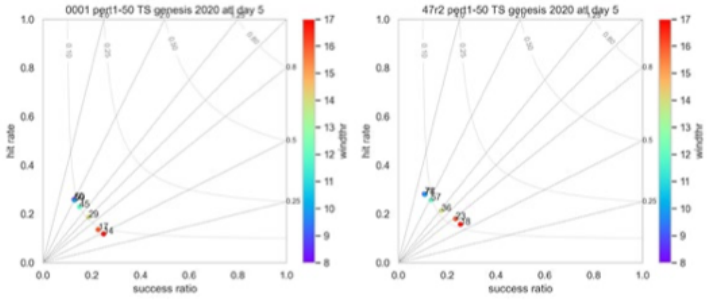


Figure 24: Performance diagram for genesis of tropical storms in the ENS at forecast day 5 the Atlantic basin in 2020 (10 May – 30 November) for operational ENS (left) and pre-operational cycle 47r2 (right). Different points are for TCs defined using different thresholds of maximum wind speed from 8m/s (blue) to 17 m/s (red); average number of TCs shown for each case. Dashed straight lines show bias; curved lines show threat score.

For genesis to occur, a precursor disturbance is necessary. These disturbances vary widely among different basins, and even in each individual basin. In the Atlantic basin, the majority of hurricanes initially developed from African Easterly Waves (AEWs). Other precursor disturbances include tropical waves of non-African origin, wave-ITCZ mergers, low pressure systems that did not originate from waves, broad gyres of low pressure over Central and South America, and extratropical baroclinic zones and even cyclones that drift into lower latitudes. In the north-western Pacific basin, TCs can develop from precursors such as a large monsoon trough, or easterly wave disturbances. In some instances, the wave structure and activity can be explained by shallow-wave theory (Kelvin waves, inertia-gravity waves and equational Rossby waves). Genesis can also be enhanced or hindered by large-scale oscillations, depending on their amplitudes and phases, and how these modulate the aforementioned environmental conditions such as wind shear and moisture. Such oscillations include El Niño - Southern Oscillation (ENSO), the Madden-Julian Oscillation (MJO), and Convectively Coupled Kelvin Waves. These and other oscillations may be concurrent, and their combined influence on genesis may be nonlinear, adding to the complications of predicting genesis on timescales of weeks to seasons. For example, in the south-west Indian Ocean, where Météo-France is the responsible RSMC, TCs most frequently form during phases 2 – 4 of the MJO, when MJO convection is active over the Indian Ocean. Landfall in south-east Africa has also been shown to be more common during La Niña, the cool phase of ENSO (Vitart et al., 2003). The potential for introducing diagnostics for tropical waves has recently been discussed with Météo-France, who demonstrated the utility of their tropical wave tracking software, and such diagnostics could provide important insights regarding TCs and genesis. Tropical wave diagnostics are also a part of the German Waves2Weather project. ECMWF aims to explore the utility of tropical wave diagnostics further in collaboration with these groups and in the project H2020-CLINT.

How far in advance are we able to predict genesis of tropical cyclones? It is a difficult question to answer as the variability is very large from case to case. For example, for Marco (Figure 25(a)) and Teddy (Figure 25(b)), the probability steadily grew from near 0% to near 100% as the forecast time

was shortened from 240 h to 0 h. On the other hand, for TCs such as Nana (Figure 25(c)), the corresponding probabilities were very small even three days prior to the TC being named. Additionally, isolated cases, especially pre-Laura (Figure 25(d)), exhibited "jumpiness" in the probabilities, even one-three days prior to genesis. Given the reliance on ensemble-based probabilities in products such as the NHC's Tropical Weather Outlook, and the rapid development or landfall (or both) in some TCs shortly after genesis, the predictive skill and predictability of genesis (and subsequent development) are important challenges to address.

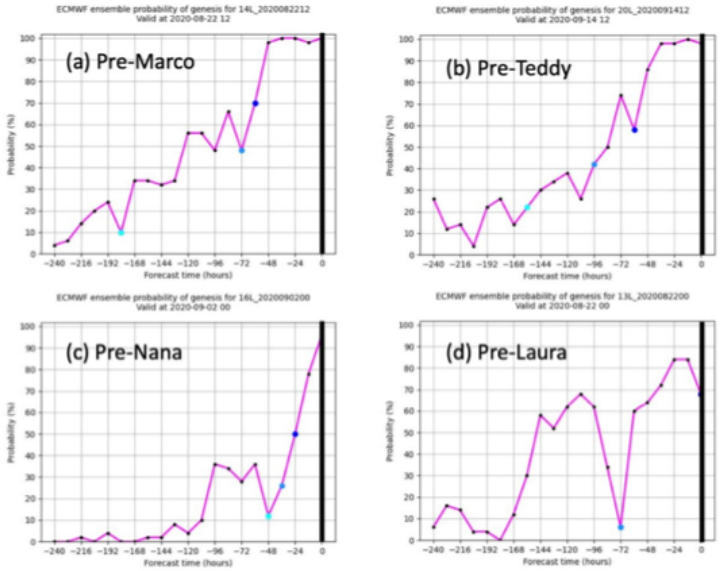


Figure 25: ECMWF ensemble-based probability of the existence of a TC within 500 km of the actual location of the named TC, at the fixed verification time that it became a named Tropical Storm. The x-axis shows the time (in hours) between the initial time of the ensemble and the fixed verification time.

Through a review of NHC Tropical Cyclone Reports and investigations of ERA5 reanalysis fields together with ECMWF forecasts, multiple pathways to genesis are evident, and these influence the predictability and ensemble forecast probabilities. The most straightforward cases in 2020 (Isaias, Marco, Rene, Teddy) were isolated AEWs with an amplitude stronger than most waves (measured, for example, by the 700 hPa relative vorticity averaged within a disk of radius 500 km). These robust waves developed in favourable environmental conditions, and we consider them to be the types of disturbances with the highest predictability. In other cases, such as Laura and Paulette, AEWs interacted with each other and/or low-pressure systems, with the probability of genesis being influenced strongly by the characteristics of these complex interactions. The ability of the analysis scheme to accurately represent these interactions is important in providing an accurate forecast. However, this is not straightforward, given the lack of observations in cloudy, mid-lower tropospheric

regions where the areas of weak vorticity are potentially beginning to amalgamate. Another substantial challenge in the theory and prediction of genesis is the initiation of convection, and how it organizes around a wave or broad region of low pressure. This challenge may be particularly important to address in the cases of very weak waves and disturbances, such as those that led to Gonzalo, Hanna, Josephine, Nana, Sally, Gamma, and Delta in 2020. In each of these cases, the NHC predicted 20% or lower chance of genesis five days before the TC was named. In many of these cases, the NHC probability was still low (20% or less) even just two days before. The ECMWF ensemble also predicted similarly low probabilities at five-day lead times. In several instances, the ECMWF ensemble probability rose sharply between three days and two days lead time. We suggest that accurate probabilistic predictions from weak waves or disturbances are the primary challenge for genesis prediction, to reduce the number of "misses" just two-three days out.

It is also important to account for non-developing disturbances in evaluating genesis probabilities. Using a wave tracker (developed by Quinton Lawton, PhD student at the University of Miami, based on Brammer and Thorncroft 2015 and Elless and Torn, 2018), many non-developing waves in the Atlantic basin were objectively identified. Some of these disturbances were weak waves in an unfavourable environment for genesis, whereas others were stronger waves that seemed more likely (but not certain) to develop. The ECMWF ensemble provided what we subjectively think are reasonable probabilities of development for these disturbances that ultimately did not develop. For the weak waves, these forecast probabilities were less than 20% for all lead times. For the stronger waves, the probabilities rose to 30-60% in some instances, but there were no cases in 2020 where there were obvious "false alarms". The main challenge in genesis prediction is to discriminate between weaker waves that develop into TCs, and weaker waves that do not.

5.7. Extratropical transitions and tropical cyclone decay

At the end of the lifecycle, some TCs curve towards the extratropics and start to interact with the waveguide in the mid-latitudes. During the extratropical transition, the cyclone becomes asymmetric with a frontal structure and the core changes from warm to cold. For a fundamental overview of the processes, see Jones et al (2003). Extra-tropical transitions can cause substantial impact in the mid-latitudes, both if the cyclones that directly (Evans et al.,2017; Baker et al., 2021) hit in a sub-tropical stage or soon after extra-tropical transition (e.g., Sandy, 2012; Leslie, 2018; Lorenzo, 2019) or indirectly (Keller et al., 2019) as the extratropical transitions can lead to downstream development (e.g., after TC Karl, 2016; Schäfler et al., 2018). Even if the cyclones do not hit land, the ocean waves can propagate long distances and hit the coasts of Europe.

Whether or not a TC will approach the extra-tropics is determined by the steering flow. If a TC is close to a bifurcation point in the flow (Riemer and Jones, 2013), very large track forecast uncertainties and track errors can occur. It is therefore critical to correctly predict bifurcations in the steering flow and the TC track towards these points. An example of such sensitivity is discussed in Magnusson et al. (2014) for TC Sandy (2012) and in Magnusson et al. (2019b) for TC Joaquin (2015), where small changes in the sub-tropical ridge caused very large differences in the future track of these TCs.

A related uncertainty is the phasing with the mid-latitude wave guide, where an upstream trough favours a northward propagation into the extratropics. Correctly predicting the mid-latitude wave guide is crucial to capture the extratropical transitions. Such a sensitivity was highlighted in McNally et al. (2014) where they found that satellite data over the northern Pacific influenced the predictions of the landfall of TC Sandy. It is difficult to determine if the forecast error in extratropical transitions is due to incorrect predictions of the TC or of the mid-latitude wave guide, as the latter can affect the track of the TC. More studies are needed to compare situations with an active waveguide with and without extratropical transitions to investigate the impact on the forecast skill from ET on the mid-latitudes. During the extra-tropical transition, the propagation speed of the cyclone can increase a lot, which also can lead to very large position errors. Large uncertainties also lie in the possibility for the cyclone to re-intensify when entering the extra-tropical stage.

The propagation speed of the TC in the tropics determines the phasing with bifurcation points and also the mid-latitude flow, while the propagation speed after the curving towards the extratropics determines the phase-lock with the movement of the mid-latitude waveguide. Both these aspects can cause substantial errors in the predictions of the extratropical transitions. As discussed above, the ECMWF forecasts have a negative bias in the propagation speed. Such bias is present both in cases that did and did not undergo extratropical transitions but is larger in the transition cases (Leonardo and Colle, 2020). However, further diagnostics are needed to understand the impact from this bias on the mid-latitude skill.

The divergent outflow from the TC can also modify the large-scale flow and contribute to the outcome of the extra-tropical transition (e.g., Agusti-Panareda et al., 2004; Keller et al., 2019). The modification of the potential vorticity in the outflow is governed partially by condensational warming in the TC. Therefore, both the strength of the secondary circulation (connected to intensity) in the TC and the precipitation rate can impact the transition. As diabatic processes are parametrized in the model this is a potential source for forecast errors. Leonardo and Colle (2020) found a sensitivity to the rain rate in TCs bound for extra-tropical transition to the propagation speed of the TCs.

It has been suggested that these transitions decrease the medium-range predictability over, for example, Europe (Keller et al., 2019). Lillo and Parsons (2017) investigated the climatology of forecast bust cases over Europe in the ERA-Interim forecasts and found that the busts are most frequent in September and October, which coincides with the most active period for tropical cyclones in the Atlantic. In a recent (unpublished) update of the annual forecast bust frequency, ECMWF and other NWP centres saw an anomalous high number of busts in the summer-autumn 2017. But even if the worst period of low skill coincided with TC Harvey, the forecast error was tracked to a tropical depression east of Florida at the same time, which shows that it is not straightforward to link the low skill to the most severe hurricanes. In recent years the evaluation at ECMWF has shown that the medium-range performance relative to other centres is worst during the autumn period, but the relation to tropical cyclones is still to be understood.

While the TCs that undergo extra-tropical transitions may create substantial impacts downstream over Europe, the majority of TCs do not undergo extra-tropical transition. As was especially evident in 2020, several TCs can make landfall in the deep tropics or subtropics, spinning down quickly into a remnant low pressure system that can bring substantial flooding rainfall for several days. Other TCs

weaken as they encounter high vertical wind shear or substantial low-humidity air, which may occur in the tropics and especially the extratropics. As a TC moves into the extratropics, it also encounters much colder waters, removing the supply of thermal energy and moisture from the ocean that is necessary to maintain the TC. As is implicit within the results of Figure 21, the HRES forecast often does not weaken strong TCs as fast as the actual weakening rate, which contributes to large errors in the intensity forecasts. Overall, intensity forecast errors are dominated by difficulties in strengthening initially weak TCs (and some stronger TCs that intensify explosively), and difficulties in weakening initially strong TCs.

5.8. Extended-range and seasonal predictions

This sub-section will discuss aspects of predicting TCs on extended (sub-seasonal) and seasonal ranges. Predicting genesis of TCs is an essential part of successful extended-range forecasts, and the subject was covered in Section 5.6.

The World Weather Research programme (WWRP) / World Climate Research Programme (WCRP) Sub-seasonal to Seasonal Prediction project (S2S) was established in 2013. Its main goals are to improve skill and our understanding of sub-seasonal to seasonal predictability and promote the uptake of S2S forecasts by the application community. A main delivery of this project has been the creation of a multi-model database containing 3-week behind real-time forecasts and re-forecasts from 11 operational centres. The S2S database represents a unique opportunity to investigate the skill of S2S models to predict TC activity up to week 6. Lee et al. (2018) produced an inter-comparison of the skill of the S2S database models to predict weekly TC genesis probabilities. Results show that the ECMWF model has the largest skill over the Atlantic, western North Pacific, eastern North Pacific and South Pacific, compared with the other models. The ECMWF extended-range forecasts display significant skill up to week 5 over the North Atlantic and western North Pacific and week 2 over the eastern North Pacific and South Pacific. Over the other basins, the skill is limited to week 1.

Important sources of extended-range predictability include ENSO and SST, and the MJO discussed in Maloney and Hartmann (2000). Vitart (2009) showed that this modulation of TCs by the MJO is well simulated in the ECMWF extended-range forecasts, except over the eastern Pacific and the Atlantic where the MJO teleconnections are too weak. Lee et al. (2018) showed that all the S2S models reproduce well this modulation, and there is a clear relationship between the ability of S2S models to represent the MJO and its impact on the TC activity and their skill in predicting extended-range TC activity. The model TC climatology also influences the performance in sub-seasonal prediction. S2S models are generally more skilful at predicting the probability of TC occurrence during the favourable phases of the MJO (Lee et al., 2020).

Lee et al. (2020) explored the impact of three different calibration methods to remove the mean TC genesis and occurrence biases, as well as the impact of a linear regression technique (van den Dool, 2017) on the probabilistic forecast skill scores. The linear regression method performed the best.

In recent years, and especially in 2020, the seasonal forecasts from ECMWF (SEAS5) of TC activity clearly missed the anomalies in the Atlantic TC activity, as seen in Figure 26. To follow up this issue, an internal working group met three times during the winter 2020/2021. The paragraphs below will therefore have a focus on seasonal forecasts for the Atlantic.

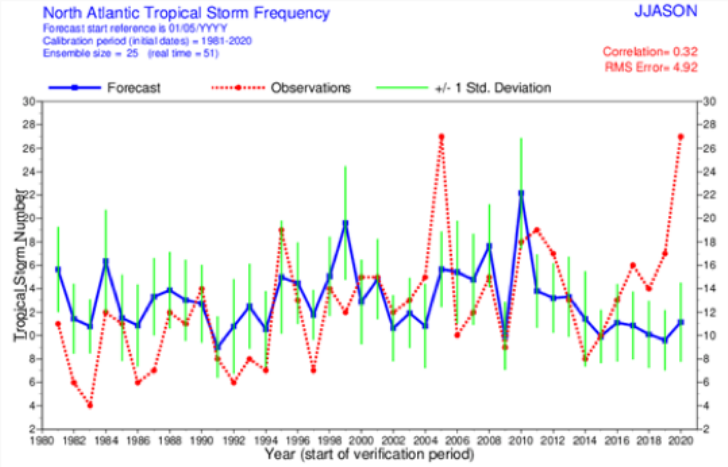


Figure 26: TC activity in seasonal forecasts from SEAS5 (blue) and observations (red) before introducing the revised calibration.

For longer lead times, the challenge is to identify drivers of predictability and the modulation of (basin-wide) anomalies in TC activity. Figure 27 shows the correlation between SST indices and TC frequency in the Atlantic, both based on observations and in SEAS5. While the SEAS5 has a comparable correlation from Nino3.4 with the observations, they are clearly missing the correlation with the local SST in the main development region (MDR) of TCs in the Atlantic, and also underestimate the link from Nino1.2 (easternmost part of the Pacific). Another important factor for the TC activity in the Atlantic is the vertical wind shear (difference of horizontal wind between 200 and 850 hPa) in the MDR, where large wind shear decreases the activity. The wind shear is partly modulated by the ENSO state as El Niño tends to increase the shear (Gray, 1979).

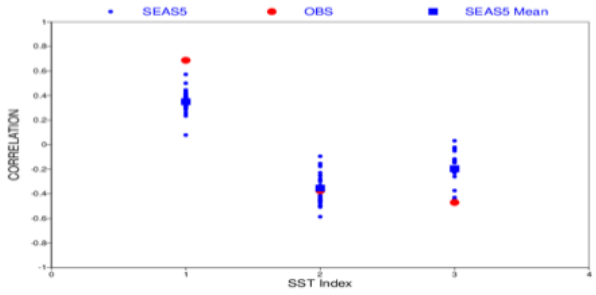


Figure 27: Correlation between Atlantic TC activity and area averaged SST in the Atlantic main development region (left), Nino3.4 (middle) and Nino1.2 (right).

In a recent investigation at ECMWF, the too weak link to the local SST was pointed out as one candidate, as 2020 saw a strong positive anomaly in the local SST. A positive trend in wind shear over the MDR was also found in SEAS5. This positive trend, which is not present in ERA5 reanalysis, could help explain the underprediction of TCs in the Atlantic in recent years. The underlying mechanism for this erroneous trend is under investigation.

Figure 28 highlights another problem for TCs in the Atlantic: the model climatology of TCs. The seasonal forecasts produce too many TCs in the Main Development Region and far too few in the Gulf in Mexico. This could have been an additional factor for the bad performance in 2020 as we saw a high activity in the Gulf of Mexico in the observations. Ultimately, the seasonal forecasts should be able to predict seasonal anomalies in landfalling TCs. This capability was investigated in Bergman et al. (2019) for ECMWF System 4, and the main result was that the skill of predicting landfalls was lower than the basin-wide activity. Marginal skill was found for landfall on the Atlantic coast, while the forecasts were not skilful for landfalls around the Gulf of Mexico.

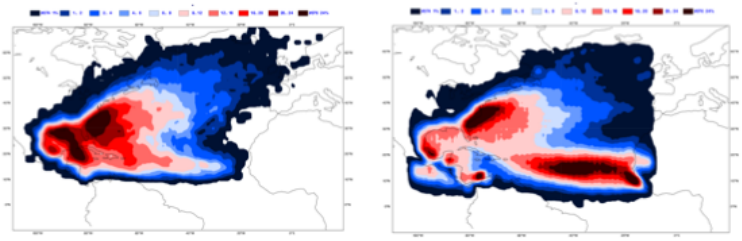


Figure 28: Climatology of tropical storm density from observations (left) and Seas5 for month 2-6 from May start dates.

To summarize the challenges for the seasonal forecasts, the important aspects to understand are the weak link between the local SST and TC activity, the suspicious trend in vertical wind shear, and the biased geographical distribution of TC activity. To temporarily minimise the effect of the erroneous trend the calibration period for TC activity in the seasonal forecast products was changed in 2021 from a fixed period (1993-2016) to the previous ten years.

5.9. Summary of challenges

The main findings from this section for all facets of predictive skill, related both to the intrinsic predictability of the atmosphere and the ECMWF system, are as follows:

Analysis errors:

- Initial errors and uncertainties of TC position are largest for weak TCs, where the centre is diffuse and difficult to determine from satellite observations. This applies to both ECMWF analyses and forecasters' Best Track estimates of position.
- Initial errors and uncertainties of TC intensity are largest for strong TCs, as the data assimilation struggles to replicate the convective-scale and mesoscale processes central to intense TCs. Further discussion of data assimilation in TCs will be provided in Section 6.

Position (track) predictions:

- Forecast errors of position are substantially larger for initially weak TCs than for strong TCs on average, especially in the short range.
- ECMWF forecasts have a long-standing systematic error with the propagation speed being too slow. There are indications that the main contribution to the error is during the propagation into the mid-latitudes, which could impact the extra-tropical transitions. This bias will be further discussed in Section 7.
- ECMWF forecasts of westward-moving TCs, especially weak TCs, often drift anomalously northward.
- ECMWF forecasts of TCs with a more northward component often drift anomalously westward.
- There can be run-to-run inconsistency ("jumpiness") in HRES and ensemble predictions.

Pmin predictions:

- Forecast errors are generally lowest for initially weak TCs, although there is a positive bias in Pmin forecasts which indicates that intensification rate is being under-predicted (too slow).
- Short-range forecast errors are generally highest for initially strong TCs, with a large positive bias. Later on, this bias for the initially strong cases changes sign as the weakening rate is under-predicted (too slow).
- Intensification in larger-scale TCs is captured better than intensification in smaller TCs.

Vmax predictions:

- There is a strong negative bias of Vmax, especially for initially strong TCs in the short-range. This is expected, as the definition of Vmax requires the simulation of convective-scale processes (O(1 km)).
- The under-prediction of the wind speed also affects the wind radii prediction (see Section 4).

Wind-pressure (Vmax vs Pmin) predictions:

- The wind-pressure relation is underestimated (see Vmax point above), and the underestimation is larger in the 18 km resolution than in 9 km resolution. The relation will be further discussed in Section 7.

Genesis predictions:

- There is a large variability in probabilistic predictions of genesis, and hence predictability, due to the variety of genesis mechanisms.
- For stronger, solitary African Easterly Waves, the pathway to genesis, and therefore the probabilistic predictions, are more straightforward and evident even over a week before genesis.
- For weaker African Easterly Waves, or wave/ITCZ/low pressure interactions, the pathway to genesis is more complex, and the probabilities of genesis are often very low less than three days prior to genesis. It is usually not straightforward to discriminate between developers and non-developers for these systems until one-two days prior to genesis.
- For non-developers (African Easterly Waves that do not develop into TCs), the probabilities of genesis do not appear to be unrealistically high.

- The precise timing of genesis is a challenge for the operational forecaster, and for the model, given the weak and disorganized structure of the TC at the time of genesis.

Extratropical transition predictions:

- Extratropical transitions are known to insert uncertainties in the mid-latitude flow. However, it is not fully understood whether the TC is the main source of the uncertainty, or if it acts as an amplifier of existing uncertainties from the upstream mid-latitude waveguide, or both.
- The effect on the slow propagation speed bias on extratropical transitions needs to be further understood.
- The effect from the intensity prediction to TC outflow and further on the extratropical transition needs to be explored.

Decay predictions:

- Decay is often under-predicted, due to the model maintaining too robust a TC as it moves towards the extratropics and encounters cooler water, stronger wind shear, and drier air.

Extended-range predictions of TC activity:

- Large biases in the model climatology of TCs were identified, especially between sub-basins in the Atlantic where the model produced too many TCs west of the coast of Africa and too few in the Gulf of Mexico.
- For the activity in the Atlantic, the model underestimated the link from the SST in the western tropical Pacific (related to ENSO) and the link from local SST in the Atlantic.
- The erroneous trend in SEAS5 wind shear over the northern tropical Atlantic need to be further understood, for example in the relation to global warming patterns in the model and reanalysis.

This section summarizes ECMWF's key challenges in TC prediction, for several aspects of TCs including their formation, motion, intensity, decay, and overall activity. These challenges arise in part due to the breakdown of intrinsic predictability of relevant atmospheric processes on the relevant timescales, and to limitations in the data assimilation (including observational availability) and modelling. In Sections 6 and 7, progress and challenges in ECMWF's data assimilation and modelling respectively are presented in detail, in the context of TC analysis and prediction.

6. Data assimilation

6.1. Introduction

As for any other weather event, the accuracy of TC forecasts is dependent on the accuracy of the initial conditions (Chen et al., 2019). The skill of a TC forecast is sensitive to the initial conditions in both the TC itself and the environmental dynamic and thermodynamic conditions that influence the track and structure (and hence the intensity).

Some of the fundamental challenges in creating accurate initial conditions of the TC structure are connected to the presence of sharp spatial gradients in the meteorological variables, together with the scarcity of observations inside the TC and of its active environment. Available in-situ observations (from buoys, aircraft, and ships) are limited to a few surface pressure measurements. Dropsondes

from aircraft reconnaissance and surveillance flights are only typically available in the Atlantic basin. It is therefore important to assimilate all observations, direct and remote, to define the best initialisation of TCs and their environment. The evolution of the observing system and the continuous improvement of data assimilation methods are key in improving TC analyses and therefore forecasts. Substantial efforts at ECMWF and other agencies are continuously dedicated to improving data assimilation methodologies, quality control and data selection. In this context, observing system experiments are useful for the following reasons: (i) to identify existing data sources with more impact and for which additional resources need to be allocated to improve their usage; and (ii) to promote the need for extra sources of data to be assimilated.

In contrast to many other NWP centres, ECMWF does not apply any specific data assimilation strategy targeted for TCs. For example, the Met Office assimilates the Best Track estimates as surface pressure observations and JMA applies a similar strategy. In the regional hurricane models, more attention is paid to the initialisation of the TCs. COAMPS-TC applies a special vortex initialisation scheme (Komaromi et al., 2021). For HWRF, an initialisation scheme for TCs is used. Until recently, increments from the HWRF assimilation in the TC core were masked, and the initialisation solely relied on vortex initialisation. Since 2019, instead a wave-number filtering of increments is applied (Christophersen et al., 2021). Importantly, HWRF now also assimilates flight-level data, airborne Doppler radar data, and dropsonde data from reconnaissance flights, and their TC forecasts have accordingly improved substantially in recent years (Cangialosi et al., 2020; Zawislak et al., 2021). The open question here is whether ECMWF should adopt any of these initialisation methods.

The observation systems used at ECMWF, described in Section 2, are continuously evolving, due to the launching and decommissioning of satellites and instruments. As instruments become older, the changes in the bias structures also need attention. Furthermore, the conventional observation system is non-stationary, and continuous monitoring is needed to minimise the risks in the data assimilation of problematic observations. For satellite observations, there is an ongoing move from only assimilating data in clear-sky conditions to all-sky (Geer et al., 2018). Another ongoing change is the increased availability of GNSS-RO data, further enhanced by the recent COSMIC-2 satellites. Since January 2020, ECMWF is also assimilating wind observations from the exploratory mission Aeolus. However, due to computational limitations together with the risk of correlated observation errors, satellite data are usually thinned before assimilation. In features with sharp gradients like TCs, it is necessary to ask if more information could be extracted from the datasets.

In recent years, coupled ocean-atmosphere assimilation has gained momentum. Given that some of the fastest change rates in SST occur under TCs, we expect that coupled data assimilation will play an important role in TC structure and intensity change prediction, and the representation of the ocean surface and subsurface currents, temperature, and salinity.

For this report, we have completed a range of parallel observation impact and coupled data assimilation experiments to explore the impact from recent and ongoing activities. The experiments were performed for the period of 15 August - 21 September 2020, to cover the most active period of the Atlantic hurricane season and a few key typhoons in the western Pacific (listed in Appendix B). For forecasts initialised every 12 hours, it resulted in more than 200 verified cases at the initialisation time and close to 100 cases in 3-day forecasts. We note that ECMWF regularly runs observing system

experiments (OSE); however, these are usually with reduced resolution (e.g., Lawrence et al., 2019). For TCs, we expect the resolution to have a significant impact, and so the experiments for this report are run at the current operational resolution of approximately 9 km grid spacing. To reduce the cost, the experiments only include the 12-hour long-window cycling (LWDA analyses), and all experiments use the background errors derived from the operational EDA.

This section is organised as follows. In Section 6.2, we will give a summary of progress at ECMWF in 4D-var and ensemble of data assimilation. Following this, we examine the observational coverage in and around TC Laura (2020) in Section 6.3, and the impact of individual observation types on TC performance in Sections 6.4-6.6. We then explore the possibility to assimilate the Best Track estimates in Section 6.7. The impact of variants of coupled data assimilation is reported in Section 6.8. In Section 6.9, we review the sampling of initial uncertainties around TCs in the ECMWF ensemble. Finally, a summary of the key points is provided in Section 6.10.

6.2. Progress and challenges in 4D-Var and ensemble data assimilation

During the past few years, incremental upgrades to the ECMWF system have gradually improved the realism of TC structure and intensity predictions and allowed better usage of available high-quality observations. Despite these advances, the initialisation of TCs remains challenging. A few challenges are described in this sub-section.

Since 2011, ECMWF has been using an ensemble of data assimilations (EDA) to estimate background errors and provide initial perturbations to the ensemble forecasts (Bonavita et al., 2016). Since 2016 the resolution of the EDA is TCo639 (Holm et al., 2015), which improved the intensity of TCs, and since 2019 the system contains 51 ensemble members (Lang et al., 2019). An example of the evolution of the position and Pmin in the HRES analysis and the first ten EDA members is illustrated in Figure 29 for TC Laura. During the first phase when the Laura was a weak Tropical Storm, relatively large spread in the analysis position is evident. This example is consistent with the increased initial position errors for weak TCs reported in Section 5. An additional uncertainty for this case is the effect from the terrain to find a minimum in the pressure field, resulting in large errors while passing Puerto Rico, Hispaniola, and Cuba. Once Laura started to intensify into a hurricane west of Cuba, the position among the members converged. For the intensity (both Pmin and Vmax), some differences among the analysis members are evident during the most intense phase (Figure 30). The difference in resolution between HRES-LWDA (long-window analysis) and EDA is evident, with a TC that is much more intense in HRES-LWDA but still weaker than the Best Track estimate.

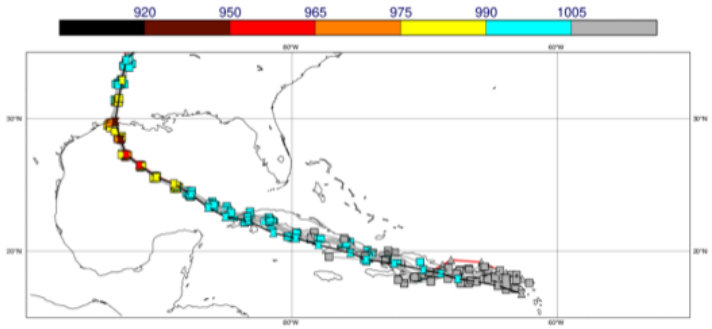


Figure 29: Position and intensity for TC Laura in concatenated HRES LWDA analyses (red line, triangle), the 10 first EDA members (grey lines, square) and Best Track (black, hourglass). The colour of the symbol shows Pmin.

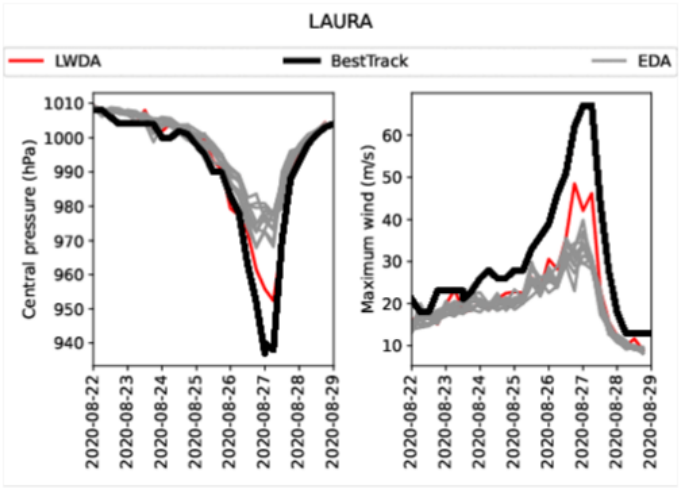


Figure 30: Intensity in terms of Pmin (left) and Vmax (right) for TC Laura in HRES LWDA analysis (red), the 10 first EDA members (grey) and BestTrack (black). The colour of the symbol shows Pmin.

Even if the non-linear model used in the 4D-Var data assimilation update trajectories (outer-loop trajectories) is run with 9 km resolution, the increments are calculated at a lower resolution (~50 km in the last minimisation). This implies that the sharp gradients and detailed structures present in the analysed TC in the vicinity of its centre are largely the product of the forecast model. The incremental 4D-Var approach aims to adjust the large scale whilst retaining finer scale gradients, but if the TC is mislocated in the background the increments may smooth the gradients. Another candidate for the

smoothing here is that the background error covariance structures may be broader than the real covariance structures in the core of the TCs.

Another consequence of sharp gradients near TCs is that small discrepancies between the observed and predicted structure can result in very large discrepancies between available observations and their model counterparts. As discussed in Bonavita et al. (2017), the interplay between extreme observations, large background forecast errors and linear assumptions in incremental 4D-Var can be problematic near TCs. With increases in the resolution in the EDA, the intensity of the TCs is more realistic (Holm et al., 2015). This leads to larger (and more realistic) ensemble spread/background errors. However, this also makes the analysis update more exposed to extreme and unrepresentative observations, e.g., of wind speed from dropsondes in the eyewall of a TC. The adaptive first guess quality control (as discussed in Bonavita et al., 2017) mitigated the issue, in particular for dropsonde winds, but occasional problems still occur. Further improvements need to be explored (e.g., improved QC, activation of VarQC in the first minimisation, etc.). Discrepancies between the predicted and observed position can also lead to wrong selection of observations to be used by the analysis (e.g., selection of the wrong ambiguity in the case of scatterometer observations, as discussed in Section 2).

Another challenge was the issue of multiple local minima in the analysis and first few hours of the forecast of surface pressure in TCs, caused by the EDA and the increased model resolution (Bonavita et al., 2017). This issue was mostly cured by a reduction of the resolution of the EDA-derived background errors used in 4D-Var and a non-homogeneous noise filtering technique together with an adaptive, stricter quality control technique applied to dropsonde wind observations.

An occasional issue is the assimilation of SYNOP/BUOY observations of surface pressure. As there are normally only small day-to-day fluctuations in surface pressure in the tropics, faulty (or misplaced) observations could go unnoticed or cause the bias correction algorithm to converge to erroneous values. Once the TC approaches the observed position, the faulty surface pressure observation starts affecting the assimilation of the TC with undesirable consequences (typically filling up the TC). One example was for the TC Leslie (2017) east of Australia, where a faulty position of a ship observation had a severe impact on the data assimilation. Another example is TC Surigae (2021) where a SYNOP station had the wrong position. One forecast just before the cyclone reached the location of the mis-placed SYNOP (17 April 12UTC) and just after (18 April 00UTC) is shown in Figure 31. For the latter forecast, the intensity is much reduced, and the track forecast had large errors. In the experiment without the SYNOP station, both the intensity and track are much better. In this type of case, the observations were given low weight in the quality control but were still allowed to have a significant impact during the first minimisation in 4D-Var. More generally, it is an open question whether tighter QC should be applied in the first 4D-Var minimisation, and it should be investigated.

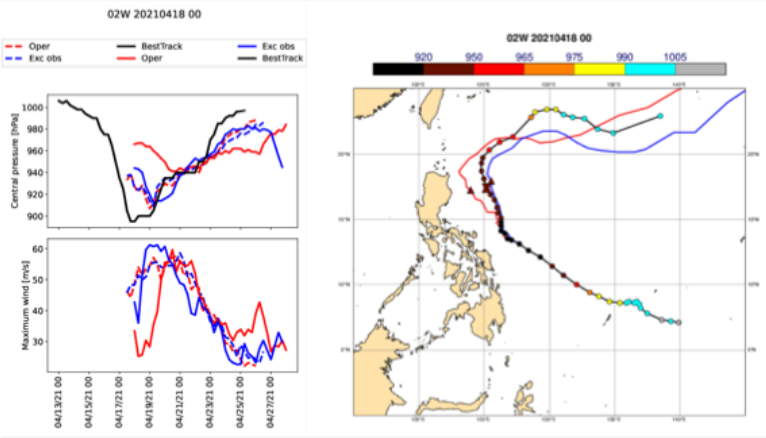


Figure 31: Pmin (left-top) and Vmax (left-bottom) forecasts from 17 April 2021 12UTC (dashed) and 18 April 2021 00UTC (solid), and track forecast from 18 April 00UTC (right) for TC Surigae. Operational forecasts (red), experiment without misplaced SYNOP (blue) and Best Track (black). Symbols in right panel indicate Pmin for Best Track every 6th hours and at 21 April 00UTC in experiments (triangles) and BestTrack (hourglass).

Another problem with the assimilation of surface pressure is the lack of stiffness of the current adaptive bias correction algorithm, which can react too quickly to large departures. If a cyclone passes a buoy with a too weak intensity in the first guess forecast, a bias correction will be applied to the buoy in the next assimilation cycle, and it can then have an impact for several days. A potentially more robust VarBC approach to the surface bias correction problem is currently being assessed.

6.3. Observations and their impact on Hurricane Laura

Figure 32 shows the observation coverage from one assimilation cycle over Hurricane Laura (27 August 00UTC) for a selection of observation platforms described in Section 2. The analysis time coincided with the most intense phase of Laura just before landfall. It was a busy period with reconnaissance flights as seen in the coverage of dropsondes (Figure 32(a)). For ASCAT, both Metop-B and Metop-C had direct overpasses of the TC (Figure 32(b)). With the operational thinning of 100 km, it resulted in just 6 wind vectors inside the 1000hPa isobar of the TC. Figure 32(c) and Figure 32(d) respectively show AMSU-A and microwave humidity sounders (MHS) for one tropospheric channel each. Operationally, AMSU-A is assimilated in clear-sky mode while MHS is in all-sky. As the contributing satellites are almost the same in the plots (NOAA-15 has a functioning AMSU-A but not an MHS instrument), the impact on the coverage of all-sky assimilation is clearly seen with cloud-masked observations for AMSU-A close to Laura (only two observations inside 1000 hPa). However, also for MHS the number of observations is relatively sparse due to thinning to approximately 130 km. For radiances from geostationary satellites, ECMWF currently only assimilate observations in clear-sky mode. This results in very few used observations near TCs (Figure 32(e)). From the same instruments Atmospheric Motion Vectors (AMV) are also derived from cloud and water-vapour

features. However, even if TCs contains a lot of clouds to trace, the number of used observations in the case of Laura is very few (none within the 1000hPa isobar, Figure 32(f), due to cloud tops above the limit for GOES-16 AMV and/or lack of contrasts in cloud signatures (see Section 2.4). In the bottom row GNSS-RO (Figure 32(g)) and Aeolus (Figure 32(h)) are presented. For GNSS-RO there are a few observations present in the outskirts of the TC. Aeolus measures the wind in one line tilted under the satellite, and one such passage was close but not immediately over the TC. All these panels together show that there are very different types of coverage from different observation platforms, with only a limited number directly inside or over the TC.

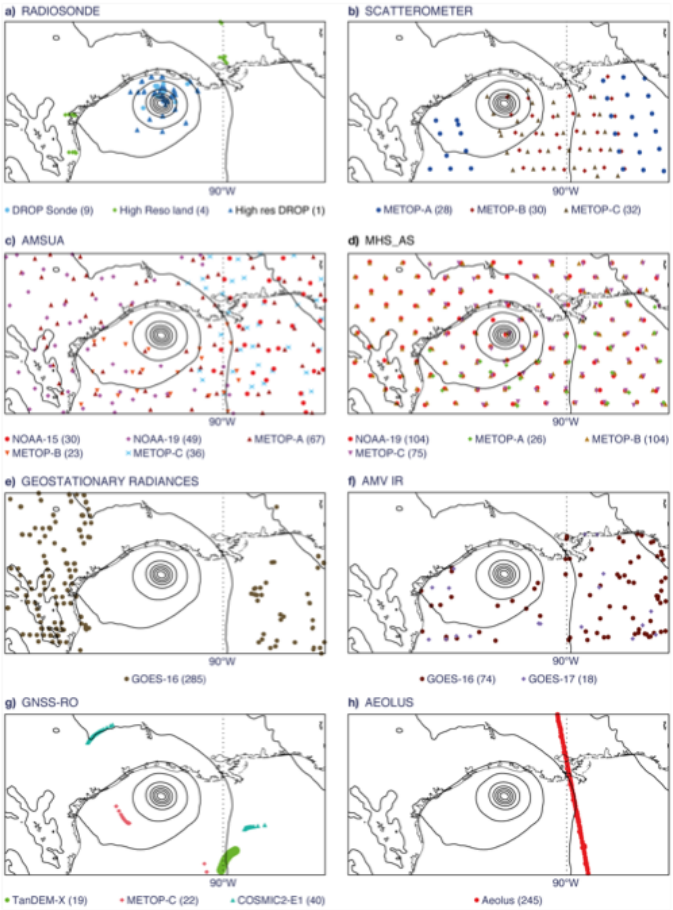


Figure 32: Observation coverage for the operational (6-hour window) assimilation on 27 August 00UTC for TC Laura for (a) radiosondes and dropsondes, (b) ASCAT, (c) AMSU-A channel 5, (d) microwave humidity sounder channel 4 in all-sky mode, (e) IR radiances from geostationary, (f) atmospheric motion vectors from geostationary, (g) GNSS-RO and (h) Aeolus.

To illustrate the impact of assimilating the dropsonde observations, Figure 33 shows the intensity forecasts for TC Laura initialised at four consecutive times: 25 August 12UTC and 26 August 00UTC just before the RI (left), and 26 August 12UTC and 27 August 00UTC during the intense phase of Laura (right). The plots include forecasts initialised from the operational configuration of LWDA (blue) and an experiment without dropsonde data (green). For the forecasts initialised just before the intensification, the intensity was best predicted by the operational configuration, indicating that the dropsondes provided useful information. This result is in line with a similar experiment reported in Magnusson et al. (2019a). However, for the initialisation on 27 August 00UTC, the intensity is

degraded in the LWDA control analysis, with Pmin changing from 941hPa in the 12 h background forecast to 951hPa in the analysis, while the Best Track estimate was 940hPa. At the same time, the experiment without dropsondes only gave a slight degradation. The dropsonde impact will be further discussed later in this section.

Degradation of the intensity for the analysis during intense phases been noted on several occasions by analysts at ECMWF. One could therefore ask the question how the intensity would look if no observations were assimilated in the vicinity of the TC. A parallel assimilation experiment was run for Laura between 22-27 Aug, during which all observations within a 4 x 4-degree box centred on Laura were withheld (Figure 33, orange line). It is evident that the denial of these data local to the moving TC substantially degraded the intensity forecasts. For the 25 August 12UTC and 26 August 00UTC initialisations, the track forecast was also worse (not shown). These results illustrate for this TC that the observations close to the TC that are assimilated are beneficial to the forecasts, but that there are potential problems during the most intense phase.

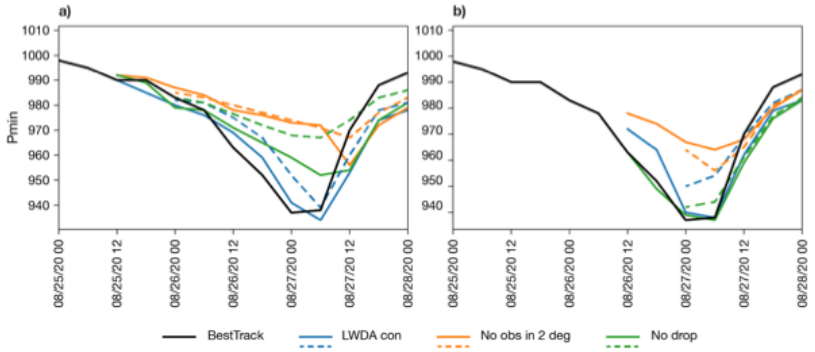


Figure 33: Pmin in forecasts for Hurricane Laura in experiments initialised from LWDA control experiment (blue), assimilation without observations inside a 4 x 4-degree box around the cyclone (orange) and without dropsonde observations (green). Forecasts initialised 25 Aug 12UTC (solid, left), 26 Aug 00UTC (left, dashed) and 26 Aug 12UTC (right, solid) and 27 Aug 00UTC (right, dashed).

6.4. Impact of assimilating dropsondes from reconnaissance flights

As introduced in Section 2, dropsondes are regularly deployed from reconnaissance missions, both inside the core of the TC and the surrounding environment. A case example of the impact was presented in Section 6.3, and here we briefly provide an overview of progress at ECMWF and results from the 2020 season.

ECMWF assimilates temperature, wind and humidity from dropsondes. In recent years, several experiments at ECMWF have aimed to quantify the impact from dropsondes on TC forecasts. While Magnusson et al. (2019a) reported a positive impact of dropsondes during the intensification phases of TC Harvey (2017) and Maria (2017), it was difficult to find any statistically significant improvement averaged over the full 2017 season. Since 2019, some of the observations are reported in the BUFR

format that contains information about the position of each reading and makes it possible to account for the drift (Ingleby et al., 2020). At the same time, the impact of accounting for the drift showed a slight positive impact on the intensity forecast.

Despite the difficulties due to COVID-19, the 2020 Atlantic hurricane season resulted in a record number of reconnaissance missions. Figure 34 provides results from an experiment in which all dropsonde observations were withheld from the assimilation during the 2020 test period. Although the impact on position error is neutral (not shown), an improvement to the average Pmin absolute error is found for the experiment with dropsonde observations, especially for a small number of intensifying TCs of initially moderate strength (50-64 kt, e.g., Laura prior to 25 August 12UTC). This is consistent with the results of Magnusson et al. (2019a) where positive impacts were found for two rapidly intensifying TCs (Harvey and Maria in 2017). Positive impacts were usually not found for initially very weak TCs (< 50 kt), or TCs of hurricane strength (>64 kt). Further research is needed into the corrections to the analysis fields due to the dropsondes, and whether this impact is carried over into the forecasts. One could also speculate that additional impact can be achieved in the future with higher resolution and sharper covariance structures in the data assimilation, which would permit analysis corrections that are closer to the observed data from the dropsondes.

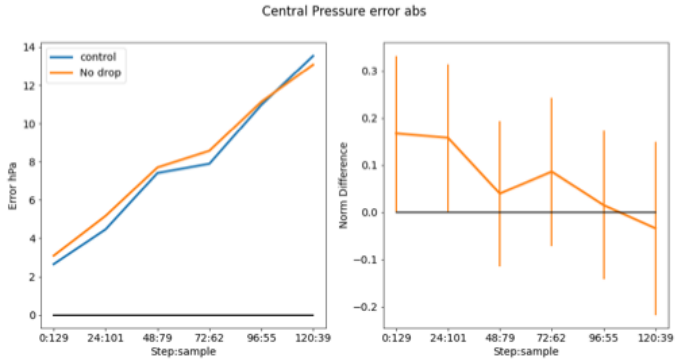


Figure 34: Mean absolute errors for Pmin for experiment without dropsondes for cyclones in the Atlantic. Normalised difference to operational configuration in the right panel with 95% confidence interval plotted as error bars.

6.5. Impact of satellite observations: recent progress at ECMWF

As discussed in Section 2, ECMWF is using a range of different satellite observations in the data assimilation. In McNally et al. (2014), the impact of satellite data on the track for Sandy (2012) was discussed, and for that case the satellite data were found to be crucial for the prediction of Sandy's landfall over the United States east coast. Similar results for Irma (2017) were demonstrated in Magnusson et al. (2019a). In both experiments, large volumes of available satellite data were removed from the data assimilation system for a short period before the genesis of the TCs.

The impact of a selection of satellite observing systems on TC forecasts has also been assessed for a recent set of global observing system experiments (Bormann et al., 2019). While these experiments were run at a lower spatial resolution, the observation usage reflects that of the operational system, including the quality control and data thinning applied to satellite data. The experiments covered a period of eight months over two seasons (1 June - 30 Sept 2016, and 1 Dec 2017 - 31 March 2018), and TC statistics were calculated over all basins. Several aspects are noteworthy in the impact on position and intensity error (Figure 35): The experiment denying all microwave radiances shows the clearest impact, with statistically significant increase in error up to day 3 without these observations. This observing system was also found to have one of the largest impacts on general headline scores in Bormann et al. (2019). It is notable that MW radiances provide regular observations near the core of the TC through the all-sky assimilation of humidity-sensitive MW radiances (cf. Figure 32(d)). While it is not clear that the impact is due to the near-core observations, it provides a plausible mechanism, and the impact of near-core observations has also been noted in assessments of FSOI statistics (not shown). Other observing systems also show statistically significant benefits for specific forecast ranges, with some of the most consistent improvements for position errors found for AMVs. We note that scatterometer data also provide near-core observations, and while they were not considered in Bormann et al. (2019), experiments with scatterometer data together with microwave data are presented for 2020 in Section 6.6.

Another aspect is also apparent from these experiments: the ECMWF system shows robustness to denials of single observing systems for TC forecasts. Even removing sizeable contributions to the global observing system such as all microwave observations translates to a loss of only around 6 h for TC forecasting. In contrast, the overall strong benefit of all satellite data combined has previously been demonstrated, for instance, in McNally et al. (2014). The results for single observing systems shown here also highlight the challenge of verifying improvements in TCs.

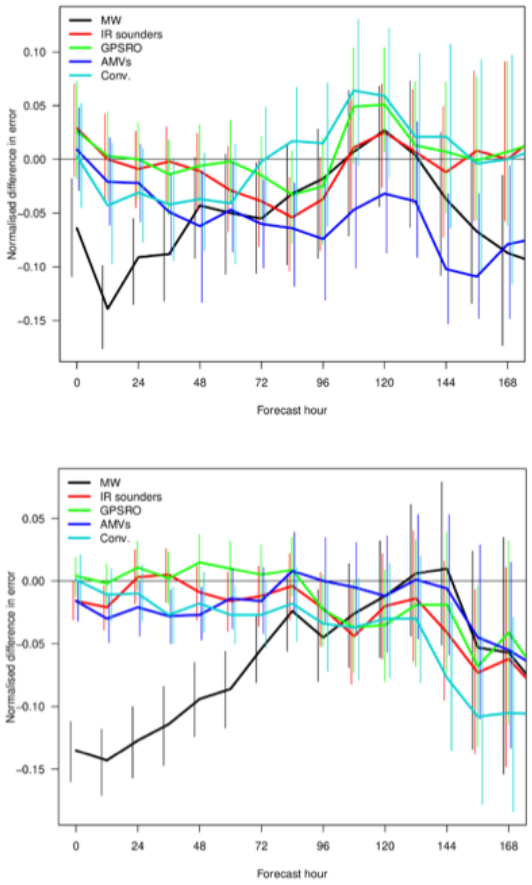


Figure 35: Normalised difference(control minus experiment) in the position (top) and absolute intensity error (bottom) from the observing system experiments considered in Bormann et al. (2019). These measure the impact from denying selected observations from an otherwise full observing system. Considered are all satellite MW radiances (MW, black), all hyperspectral IR sounders (red), bending angles from GPS radio occultation (GPSRO, green), AMVs (blue), and in-situ observations, encompassing all aircraft data, radio- and drop-sondes, synops, etc (Conv., cyan). Statistics are based on experimentation over 8 months (1 June - 30 Sept 2016, and 1 Dec 2017 - 31 March 2018), with the number of TC forecast cases considered ranging from 700 in the short range to over 110 at day 7. Error bars indicate statistical significance at the 95 % confidence level.

6.6. Impact of satellite observations: 2020 experiments

Several new observing system experiments with satellite observations were performed for the special 37-day period in Autumn 2020. A control experiment ("LWDA Con") based on the operational observation usage is used as a reference for all the observing system experiments in this section and paper. In Figure 36, the denial of the following observation types which were assimilated operationally in 2020 is evaluated: (i) satellite wind observations from Aeolus; and (ii) GNSS-RO data from COSMIC-2. To put these results in perspective, these impacts are compared against an experiment in which all microwave observations are denied, similar to the experiment of Bormann et al. (2019) illustrated above in Figure 35. As we remove observations in these experiments, we expect the error to increase for observation types with large impact. A "positive" impact for the denial experiments refer to increased errors when the observations are removed from the experiment. In addition to the denial experiments, Figure 36 also includes an experiment with the expected improvements from all-sky assimilation of AMSU-A temperature sounders upon the operational configuration. This change to the all-sky assimilation is planned for implementation in Cycle 47r3 in late 2021. However, as all the individual systems by themselves are a small part of the full observation set, one should expect relatively small impacts. While the focus in this section is on TC predictions, results for the impact over the full tropics are provided in Appendix C. As the results in Appendix C are based on verification against the operational analysis, the results are not straightforward to interpret for short lead times. However, one can note that all instruments discussed here have a statistically significant positive impact for 700hPa winds in the tropics for 3-day forecast, as an example.

As expected, the largest impacts on the TC position and Pmin are found when all microwave observations are removed (Figure 36) and the results are in line with the low-resolution experiment over a longer period presented in Figure 35. Since the humidity sounders and microwave imagers are assimilated in all-sky mode, including near the TC (as seen in Figure 32), this could be one reason for the positive impact on Pmin. An examination of the impact of adding all-sky assimilation of AMSU-A temperature sounders reveals a slight improvement in the position error for one-day predictions, and no statistically significant benefit for the Pmin error. However, a similar experiment during a longer period of the 2019 TC season at lower resolution provided a positive impact on both position and intensity forecasts (Duncan et al., 2021).

For COSMIC-2, statistically significant improvements at the 95% level are found for Pmin for the first two days and still on the positive side out to five days (Figure 36, right). However, the magnitude of the average improvement compared with microwave is small, while the impact on position forecasts is neutral (Figure 36, left).

For Aeolus, the results are neutral for both position and Pmin. However, as seen in Figure 32, with the line measurement from one instrument, only a few crossings over a centre of a TC are expected in a 40-day sample, which makes it difficult to reach a statistically robust impact. For TC Teddy, the system was fortunate to achieve three straight passages over the TC. This provides an opportunity for a further evaluation of this case in the future.

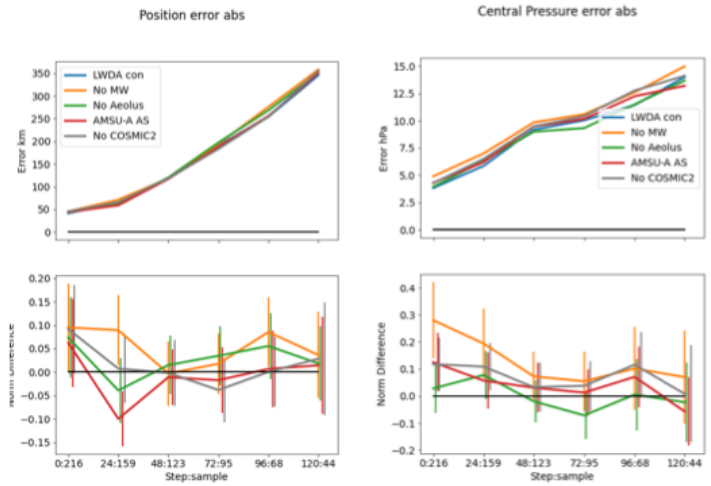


Figure 36: Mean absolute errors for position (left) and Pmin (right) for different satellite observation system experiments (experiment minus control). Normalised difference to operational configuration in the bottom panels with 95% confidence interval plotted as error bars.

As discussed in Section 2, surface winds from scatterometer instruments have been assimilated operationally for many years. Since 2020, ECMWF has been assimilating ASCAT winds from three Metop satellites, which are provided with an original grid spacing of 25 km and thinned to 100 km. An observation error of 1.5 m/s is assigned to the wind components. While providing high quality observations, sometimes a wrong direction is presented. This can be due to observation errors, the retrieval process or ambiguity removal issues, with the latter being the most common source of wind direction problems. Since the ambiguity removal process is based on the short-range forecast (background) wind fields, in case of TCs, this issue usually happens when we have large first guess errors in wind direction. An example of erroneous wind vectors presented to the 4D-Var is shown in Figure 30 for TC Paulette where a couple of wrong directions can be found north-east of the TC. However, these wind vectors are the closest to the background, and the alternative solutions do not appear to be better. Most likely, in this specific case, the problem could be due to either observation error in an area with a quite chaotic ocean surface, or an issue in the retrieval process. The incorrect selections lead to very large differences between the observations and background forecast, and the variational quality control (VarQC, Andersson and Jarvinen, 1999) then correctly assigns zero weight to these observations, as seen for this case in Figure 38, where the problematic vectors are not among the active observations). However, we also see that other (correct) observations of winds above 20 m/s were rejected. To reduce the number of rejected ASCAT observations the usage of Huber Norm instead of the VarQC has undergone some preliminary investigations (De Chiara et al., 2018) and this could be a promising line of future research.

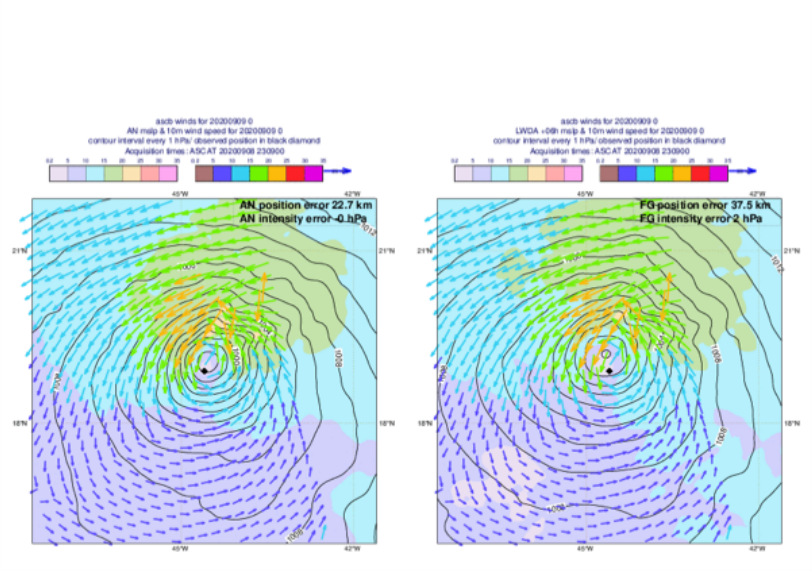


*Figure 37: Analysis (left) and first-guess (right) of MSLP (contour) and wind speed in m/s (shading), and **all** ASCAT observations (wind vectors) coloured by wind speed for TC Paulette on 9 September 00UTC in the operational analysis. Best Track position indicated as diamond symbol.*

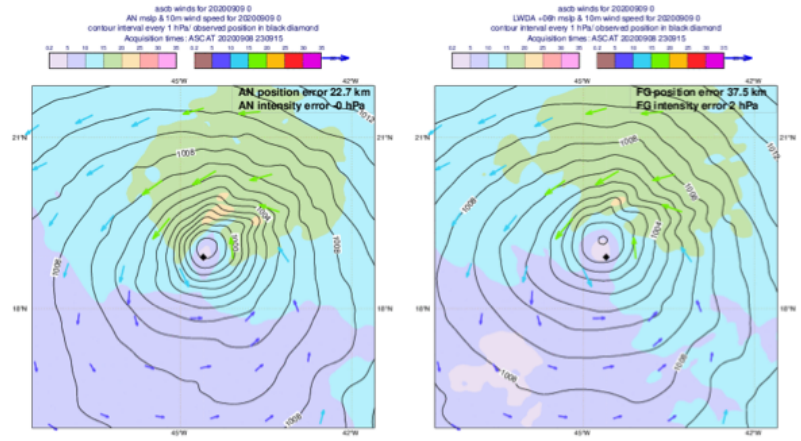


*Figure 38: Same as Figure 37 but for **active** observations.*

Due to the challenges (ambiguity selection, thinning, wind saturation) in using ASCAT winds close to TCs, one can ask the question of whether they are beneficial in TC analysis and forecasting. In operations, ASCAT observations are thinned, with one in every four observations assimilated along and across track, as seen when comparing Figure 37 (all observations) and Figure 38 (active observations). This results in a final horizontal sampling of 100 km. To explore the sensitivity to the assimilation of ASCAT winds, two experiments with a different thinning of the observations have been completed. As the reduced thinning increases the correlation of observation errors, this is compensated for by increasing the observation errors.

One experiment ("Scat Thin2") is thinning the observations by a factor of 2 (instead of 4 in operations) and uses an observation error of 2.25 m/s. The other experiment ("Scat No thin") has no thinning and an observation error of 3.75 m/s. Note that increasing the observation error also leads to a relaxed quality control. Another experiment with all ASCAT winds removed was also conducted. The results from all experiments are presented in Figure 39 and Figure 40. First, the removal of all scatterometer observations is found to degrade the forecasts of Pmin out to 48 h (Figure 39, orange line). Next, while the experiment "Scat Thin2" only shows a slight (non-significant) reduction in Pmin error (Figure 39, green line), the biases in Pmin are reduced (Figure 40, green line), strengthening analyses and forecasts of hurricane-force TCs out to two days (not shown). This is more clearly illustrated in the scatterplot in Figure 41, where the analysis Pmin values for deep TCs is usually lower than in the experiment analysis ("Scat Thin2") than in the LWDA Control analysis ("LWDA con"). The mechanisms behind this result will be further explored. Similar results, albeit more modest, are found in the experiment with no thinning and higher observation errors ("Scat No thin").

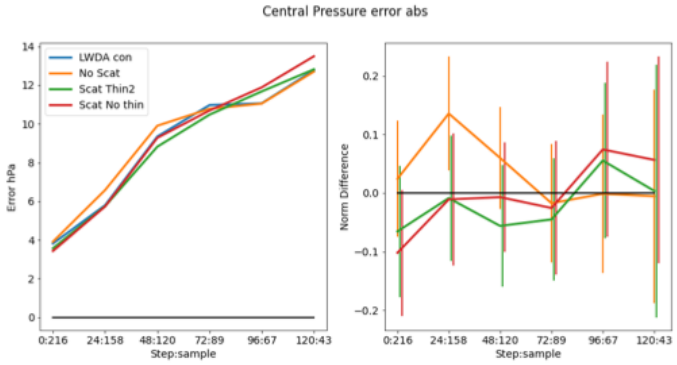


Figure 39: Mean absolute error for Pmin for experiment with different ASCAT configurations. Normalised difference to operational configuration in the right panel with 95% confidence interval plotted as error bars.

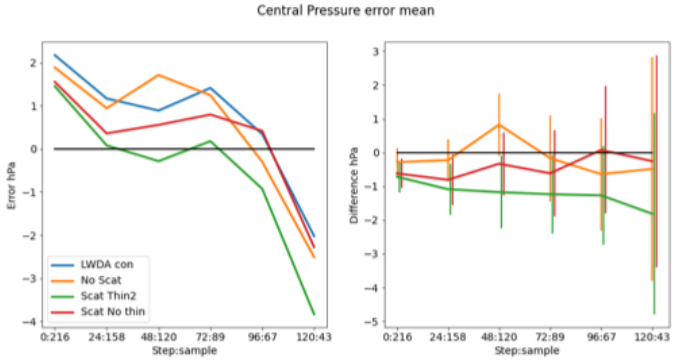


Figure 40: Mean error for Pmin for experiment with different ASCAT configurations. Difference to operational configuration in the right panel with 95% confidence interval plotted as error bars.

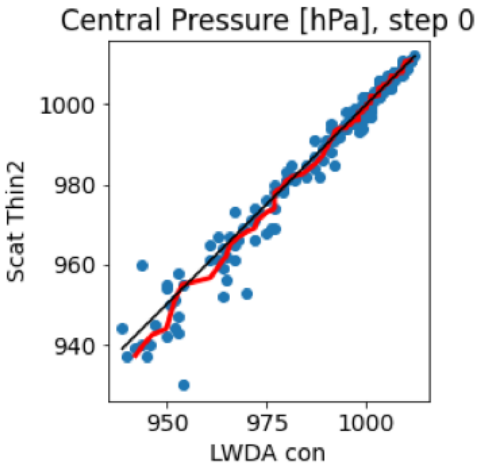


Figure 41: Scatterplot of Pmin in control vs reduced thinning of ASCAT with a factor of 2 ("Scat Thin2"). Quantile-quantile matching is shown in red.

6.7. Assimilation of Best Track

Most of the other NWP centres make use of the information in the Best Track in the data assimilation or initialization. Some centres apply a vortex relocation scheme to move the TC according to the

estimated position. Other centres like the Met Office (Heming, 2016) infer the estimate as a surface observation of pressure in the data assimilation. A similar approach was also undertaken in the ERA5 back-extension to 1950. However, severe issues were detected during this approach in cases where the background forecast lacked a TC or was far from the truth. This led to the development and implementation of a safer Best Track observations assimilation for the pre-satellite era and a renewal of the production of the ERA5 back extension.

In the case of the Met Office, they interpolate the values between the six-hourly estimates to produce hourly observations to assimilate, to increase the impact. To test the impact on assimilating Best Track, we have conducted experiments that follow the operational approach at the Met Office. The Best Track position and intensity has been linearly interpolated from six-hourly (from "operational" Best Track) and three-hourly data (from IBTRACS files) into one-hourly sampling and is used as hourly surface pressure observations. There is also a potential difference in the TC life-length spanned by these two sources. The results are also compared with an experiment only assimilating the six-hourly estimations. Observation errors have been set to 3.5 hPa for Atlantic TCs (which are usually observed by both satellites and reconnaissance aircraft) and 6 hPa for TCs in the Pacific. For hourly interpolated values, the observation errors are inflated by 1.2 to account for interpolation errors. When the TC intensity is constant for more than 12 hours, the errors are inflated by a factor of 1.5.

Figure 42 illustrates the position and Pmin mean absolute errors for these three experiments. For the position error, all three experiments lead to an improvement of 3-5% for one to two-day forecast, but do not pass the 95% significance level. The magnitude is somewhat lower compared to the improvement found at the Met Office (Heming, 2016, table 3). For the Pmin, the impact on MAE is positive but non-significant, while the bias is significantly reduced at the analysis time for both experiments using 1-hourly interpolation (not shown).The positive impact for the bias is found in making strong TCs more intense at the analysis time, as seen in the scatterplot of Pmin values in Figure 43. The main contributions here are TC Maysak and TC Haishen in the north-western Pacific. While the impact is positive in the northwest Pacific, the experiment seems to degrade the forecasts in the Atlantic (not shown). One could speculate that the difference in intensity distribution, the estimate of observation error, or availability of dropsondes in the Atlantic may cause this difference. However, the variability from case to case is large and none of the results passed the 95% significance level.

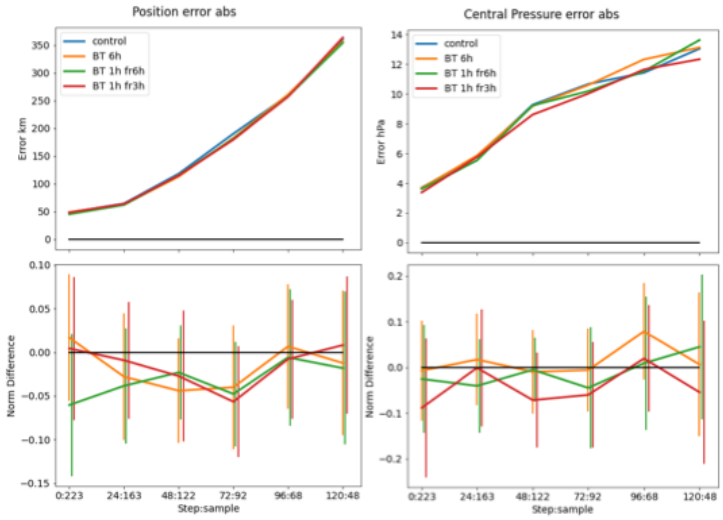


Figure 42: Mean absolute errors for position (left) and Pmin (right) for experiments with assimilating BestTrack estimates. Normalised difference to operational configuration in the bottom panels with 95% confidence interval plotted as error bars.

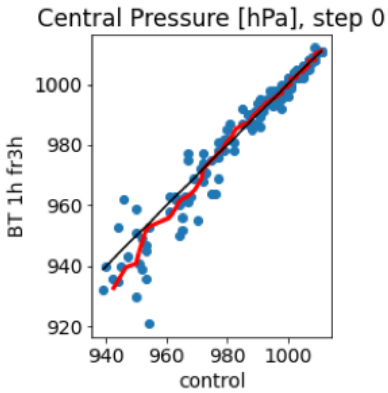


Figure 43: Scatterplot of Pmin in control vs assimilation BT from 3h estimates. Quantile-quantile matching is shown in red.

Applying this procedure poses several difficulties. One lies in the estimation of the observation error. This is expected to vary according to the availability of reconnaissance flights etc. (Torn and Snyder, 2012; Landsea and Franklin, 2013). The estimate of the observation error also has an impact on the

VarQC. If the Best Track observation is too far from the first guess forecast value, the observation will have very little or no impact due to the quality control. Another lies in the exposure to human errors and technical faults in the reporting chain, that can lead to assimilating erroneous values. As noted above, non-human observations can also suffer from difficulties with estimating observation errors, biases, and occasional technical faults. The variability in the results from the three different experiments also indicates a sensitivity to interpolation and selection of cyclones to include (pre- and post-stages of the TCs). All these aspects need a careful consideration before this type of observation is put into operations, which could be a time-consuming exercise.

6.8. Ocean Aspects: Coupled Data Assimilation

An important aspect is the treatment of the SST, given that rapid upper-ocean cooling may occur in the wake of TCs. Since 2013 (ensemble) and 2018 (HRES), the ECMWF atmospheric forecasts have been coupled to a dynamical ocean model (see Section 7). Since 2019, the SST boundary conditions for the atmospheric analysis have been taken in the tropics from the ocean data assimilation system, providing a weakly coupled Atmosphere-Ocean data assimilation system (Browne et al., 2019). In the extra-tropics the SST boundary conditions continue to come from the OSTIA (Good et al., 2020) observational product. For observations of SST, ECMWF currently relies on the OSTIA product (see Section 2), which is nudged into the ocean data assimilation at the top level in the ocean model. However, the latency in the availability of the observational product could lead to errors in the initial SST in the wake of TCs.

Figure 44 illustrates the problem with OSTIA just after the passage of Hurricane Teddy (2020). During 21 September, Teddy passed over a cluster of buoys, launched from aircraft ahead of the storm. Around the same time, TC Teddy passed over the area where TC Paulette passed a week before. The figure shows SST from a 2-day forecast valid 22 September 00UTC (top-left) and the OSTIA analysis (top-right) valid at the same time. Both panels include buoy observations. The figure also includes panels with SST from satellite products valid around the same time. The 2-day forecast agrees well with the buoys about the strength of the cold wake created by the TC. However, the used OSTIA analysis lacks the trace of the TC and it is much warmer than the buoys, although the wake after Paulette is still visible. There are two reasons behind the lack of signal from TC Teddy. First, as a daily averaged product it is necessarily received one day behind real time. The other issue is seen in the panels with the SST from the satellites. Due to cloudiness the information is missing from the regions where we expect to see the cooling. However, this case shows that the information from the model can be valuable to fill in the missing parts in the wake of the TC.

The solution to both the observation latency and difficulty of satellites to see through the optically deep clouds around a TC is to use the ocean model as part of the analysis system. This is a cornerstone of the benefits of the weakly coupled ocean-atmosphere data assimilation system which is currently operational. To show the impact of ocean atmosphere coupled assimilation configurations we show in Figure 44 and Figure 45 a number of experiments: an LWDA control (LWDA con) with SST data from OCEAN5 used between +-20° and OSTIA elsewhere, an experiment where this region is extended to +-30° (Partial 30), an experiment (OSTIA) where we do not use any SST from OCEAN5 but rather use OSTIA globally, and finally an experiment where we have outer loop coupling active within the +-30° region (LWDA Outloop con). The middle right panel of Figure 44

shows that when the cold wake of a TC occurs within the region of coupled SST analysis, the ocean model is able to capture signals which can be seen in the in situ observations that remain hidden from the satellites.

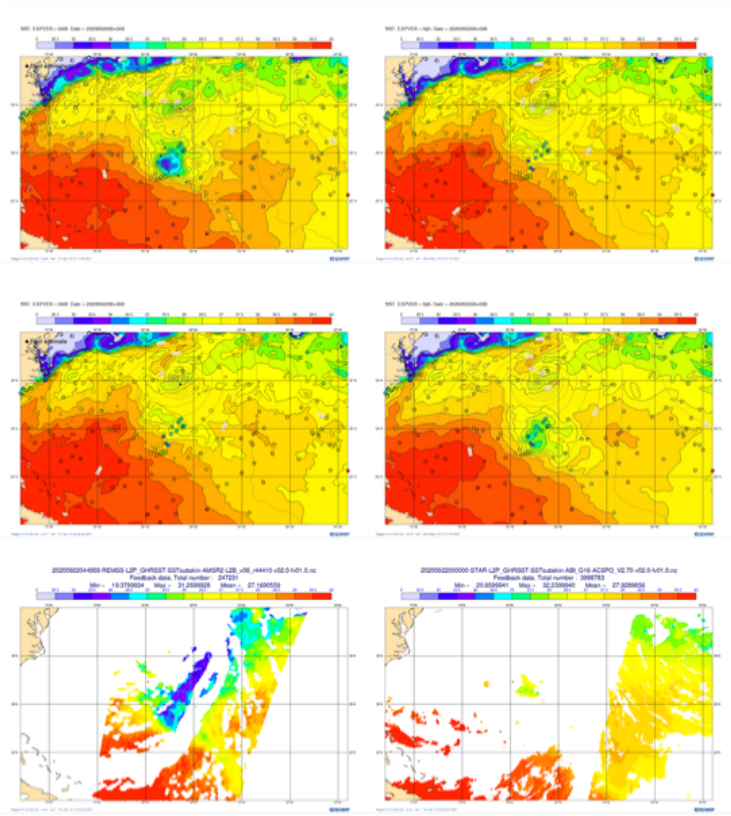


Figure 44: SST valid 22 September 00UTC from a 2-day forecast (top-left) and from OSTIA (top-right). SST initial conditions for 22 September 00UTC based on the operational configuration with OSTIA SST mainly used ("partial coupling") north of 20N (middle-left) and an experiment with the the boundary moved to 30N (middle-right). SST observations from buoys are marked in circles. SST from AMSR2 valid 22 Sept 04UTC (bottom-left) and from GOES-16 valid 00UTC (bottom-right)

Figure 45 shows the mean absolute error for position and Pmin for various experiments in different configurations of coupled assimilation and hence different treatments of SST. The impact of the changes on the future of the TC is small on average. However, as the TC mainly affects the wake, other parameters could be affected, as may future TCs if they pass over the cooler wake. The similarity in TC performance with different coupled DA systems is encouraging as it helps show that

coupled DA is not passing ocean model biases to the atmosphere which would show up as changes to the steering flow and therefore degradations to track and intensity forecasts.

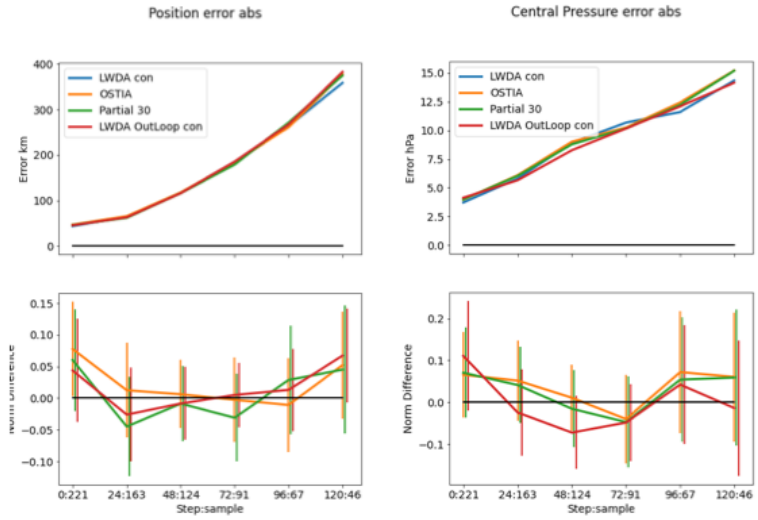


Figure 45: Mean absolute errors for position (left) and Pmin (right) for various coupled DA experiments.

6.9. Sampling initial uncertainties for tropical cyclones

The perturbed initial conditions for the ensemble members are generated by adding perturbations based on a combination of EDA and singular vector perturbations (SVs, e.g. Leutbecher and Palmer, 2008) to the deterministic high-resolution analysis (Buizza et al. 2008, Lang et al. 2021). To increase the spread of the ENS TC track forecasts, SVs are targeted on each reported TC (Barkmeijer et al., 2001, Puri et al., 2001). At ECMWF, SVs are computed at relatively low (T42) horizontal resolution. Hence, perturbations are more focused on the large-scale steering flow around the TC (Yamaguchi and Majumdar, 2010). For a detailed discussion of the impact of the different perturbation methods and their relative contributions to the ENS TC track and intensity forecasts spread, see Lang et al. (2012).

To determine the SV TC target regions, the reported position together with the TC tracks of the previous ensemble run are taken into account (see Leutbecher and Palmer, 2008). The maximum number of targeted SVs for TCs is set to six. If there are more TCs active at the same time than the maximum number of targeted SV sets, target regions that are close together are combined.

This led to problems during the very active 2020 season. Because of the large number of active TCs, very big target regions were created, which encompassed multiple TCs. This is exemplified by the 17 September 2020 in Figure 46, when ten different tropical systems were active. In the Atlantic, Teddy,

Vicky and Sally was active together with three INVEST systems which a day later became Wilfred, Alpha and Beta. As a result, one SV calculation region was created for Teddy, Wilfred, Alpha and Beta, and perturbation amplitudes associated with each individual system were markedly reduced for several dates (Figure 47, left panel). Increasing the maximum number of allowed target regions to 12 prevents the creation of overly large target regions (comparing the panels in Figure 46) and increases the SV perturbation amplitude in the vicinity of the TCs that would otherwise have been assigned the same target region (comparing the panels in). In the case of Teddy (the most intense system), the initial perturbations were increased in the southern edge of the sub-tropical anti-cyclone to the north-east of Teddy, a structure the future track for Teddy is sensitive to. As expected, the increased perturbation amplitude leads to an increase in TC track spread in these cases, while otherwise the track spread would have been artificially reduced (not shown).

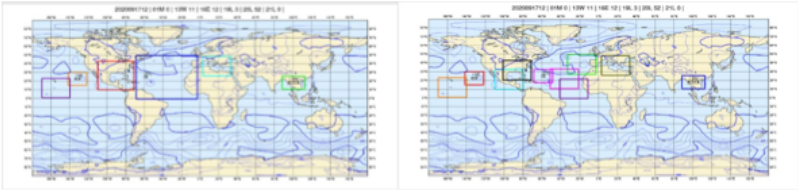


Figure 46: Tropical cyclone target regions (20200917 12UTC); a) maximum number of 6 regions and b) maximum number of 12 regions.

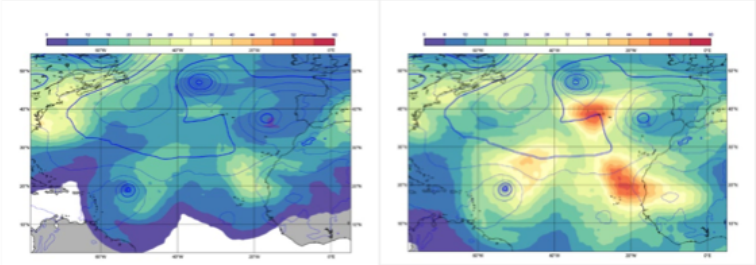


Figure 47: Vertically integrated total energy of the singular vector initial perturbation and MSLP of control forecast at initial time (2020-09-17-12 00 UTC); a) maximum of 6 target regions and b) maximum number of 12 target regions.

6.10. Summary

In this section, recent progress at ECMWF in data assimilation in the context of TC analysis and forecasting has been reviewed. Several data assimilation experiments were completed during a special 37-day period in Autumn 2020, run at the same resolution as the operational system at the time (47r1). These parallel experiments provided a unified framework to evaluate and understand the roles of different types of observational data, in the context of the same TCs under investigation. These were conducted with the recognition that any single observation system has a small impact, and that all observing systems combined in the operational system have a substantial collective impact. The goal

of the individual experiments was thus to provide insights on how to make the best use of the observations that we already have, and to identify the context behind which any specific datasets are especially useful. Based on the 2020 experiments, the results are summarized as follows:

- Analysis values of the position are generally accurate (within 50 km of the Best Track), except for a few weaker TCs in which there is uncertainty in the actual position (Section 5.3).
- Analysis values of the intensity (Pmin) can be substantially weaker than the observed or estimated values, especially in intensifying TCs.
- Faulty surface pressure observations occasionally deteriorate the ECMWF analysis of TCs. A similar risk also exists for ASCAT observations with wrong selection of ambiguity.
- Targeted experiments for TC Laura showed that observations within a 4x4 degree box centred on the TC are important for predictions of both position and intensity.
- The impact of assimilating dropsonde data on position forecasts is neutral while being beneficial for Pmin forecasts. The gains are evident in a small number of cases of intensifying TCs, consistent with the results of Magnusson et al. (2019a).
- Some satellite datasets in the assimilation (Atmospheric Motion Vectors; scatterometers) are sparse in the vicinity of TCs, due to thinning of the data, cloud-top height restrictions (AMV) and saturation of signal for wind speeds above 30 m/s (ASCAT).
- Withholding all satellite-based microwave observations significantly degrades the TC forecasts out to two days.
- The individual impact of each other satellite dataset by itself is marginal and of lower magnitude than the denial of all microwave data, as expected since it comprises a very large fraction of the observational network. Some positive impacts are evident for recently added COSMIC-2.
- The nature of sampling from Aeolus (being only one instrument on one satellite) makes it difficult to reach statistically significant results, and a sub-sampling of cases with good passes might give more insights.
- The assimilation of interpolated hourly central pressure and position data from the Best Track, following the practice at the Met Office, has a small but overall positive impact on intensity for stronger TCs in the north-western Pacific basin, but not in the Atlantic basin.
- An increase in the number of TC target regions for ensemble initialization leads to larger and more relevant initial perturbations for each individual TC during particularly active periods.
- The impact usually varies depending on the initial intensity of the TC.
- The results can differ from basin to basin.

6.11. Discussion and future directions

TCs comprise processes on multiple spatial scales ranging from O(1)-O(100) km, and these scale decompositions also vary depending on the stage of the TC. Questions arise on how effectively the observational data can capture these processes; how the data assimilation scheme is able to exploit these data and provide realistic initial conditions for TCs; how meaningful insights can be derived on observation impact; and how initial conditions and their uncertainty can be appropriately represented. The remainder of this final section discusses these facets and suggests ways forward with observations and data assimilation within this context.

Although the experiments reported in this section have shown that several observing platforms have contributed to improving TC intensity analyses and forecasts, it is not yet known whether the signal is from the assimilation of observations near the core of the TC, or from the surrounding conditions (e.g., environmental humidity and vertical wind shear) important for intensity changes. In parallel, the question of representativeness of the observations, especially in the inner core of the TC where the processes are small-scale and vary rapidly, needs addressing. Similarly, the volume of observations that are of high quality but are rejected from the assimilation warrant investigation.

For dropsonde data, the intriguing result of why they improved intensification forecasts, albeit for a small sample, warrants further investigation, together with the more general question of how they are contributing to the inner-core structure. One open question here is if the horizontal scale of the data assimilation is still too coarse to take the full advantage of this fine-scale information in areas with sharp gradients in the wind fields. It is also worth noting that other aircraft data, such as flight-level data from the aircraft itself, and remotely sensed data such as airborne Doppler radar are now being routinely assimilated in NOAA's operational systems and are demonstrating benefit. An initial step at ECMWF is under way to explore the impact of flight-level data or additional surface wind products that can help to anchor the position and low-level wind field. In the longer term, the assimilation of airborne Doppler radar data may be feasible, subject to the availability and maintenance of not only the data but the observation operator and associated error characteristics.

The NHC deploys the NOAA G-IV aircraft around TCs for synoptic surveillance, and they continuously seek to improve their strategies for targeting the aircraft, to optimise the use of dropsonde data for assimilation at operational centres including ECMWF (Michael Brennan, personal communication, 2020). Additional studies to investigate the useful distances and effectiveness of synoptic-scale sampling around the TC in the ECMWF system will help guide NHC's sampling strategy.

The global network of satellite observations is large and complex, with the backbone comprising geostationary satellites and low-Earth orbit satellites. Given that TCs spend most of their lifetimes over the ocean, and aircraft data are usually only available (and are limited in volume) in the Atlantic basin, the combined assimilation of many types of satellite data is especially important for TC predictions.

The positive impact of microwave observations on TC forecasts is due in part to the large volume of observations from several instruments and satellites sampling different times of the day. We also expect that the positive impact on the intensity is coming from the effort in recent years to assimilate observations in all-sky conditions and not only clear sky (Geer et al., 2018). This effort continues with the introduction of all-sky assimilation of AMSU-A temperature sounding radiances, and further investigation on how the all-sky assimilation can be further improved is warranted.

It is important to ask what more can be done with current observations. Most satellite datasets, including scatterometer winds, radiances, and AMVs, are thinned substantially in the operational assimilation. For example, ASCAT winds, whose native grid of wind retrievals is 25 km, are thinned to 100 km, thereby yielding only very sparse data near the TC. The experiment that reduced the thinning before the assimilation further elevated the positive impact of ASCAT. In fact, we now plan to implement the 50 km scatterometer thinning configuration in Cy48r1 based on these results. There

is also more spatial information available in MW radiances, including on warm core structures (Tian and Zou, 2016) down to scales of around 20-50 km. Further work is recommended to establish the optimal spacing of satellite data in which mesoscale features within the TC (including its outflow) are captured, while accounting for correlated observation errors; ambiguities and bias correction would also likely need to be accounted for. One avenue for exploration would be to conduct experiments in which the spatial and temporal density of satellite data in the immediate environment of a TC is increased in the assimilation.

For microwave imagers, some sensors are only passively monitored because in its current configuration the assimilation system is not able to successfully use the additional information. With the addition of some of the sensors that are currently passive (for example, Windsat, DMSP-F15 SSMI and DMSP-F18 SSMIS), temporal coverage of TCs from these instruments could be substantially improved, provided the issues preventing us from using more of these instruments can be overcome.

Infrared radiances are thinned on 100 km scales and are currently only assimilated in clear-sky and fully overcast situations (a very limited number of scenes, particularly around TCs). All-sky infrared assimilation is looking increasingly feasible both in the IFS (e.g., Geer et al., 2019) and in the convective assimilation context (e.g., Sawada, et al. 2019). The combination of geostationary infrared observations at fine time and space scales, and all-sky assimilation, could be very helpful to TC initialisation; these are some of the main observations used in deriving best track estimates. Current satellite observations could therefore provide far more detail on the temperature, humidity and cloud and precipitation structure of TCs than is currently being assimilated.

The GNSS-RO measurements have a limb geometry, and this limits the horizontal resolution of the measurement to hundreds of km, which is clearly not ideal in the context of TCs. In addition, the sampling of the GNSS-RO observations is still quite sparse despite the availability of COSMIC-2 (see Figure 32(g)). Some new GNSS-RO measurements will become available from Sentinel-6a and Spire in 2021-22, but we will also lose Metop-A GRAS. The new operational processing of Metop GRAS measurements, scheduled for implementation in September 2021, should improve the quality of these measurements in the troposphere. In addition, ECMWF will investigate updating the uncertainty model used to weight the GNSS-RO data in the 4D-Var system.

Aeolus has been demonstrated to provide positive impact on forecasts of tropical wind and temperature (Rennie et al. 2021), however showing a positive impact on TCs is proving more difficult. This is probably due to the length of the OSEs and the limited spatial coverage of one instrument. A much longer OSE using reprocessed Aeolus data is underway to try to determine the impact of Aeolus on TCs and extreme weather (an ESA funded project). There are some residual biases for Aeolus winds varying with altitude and wind-speed which are expected to be improved with future reprocessed Level-2B winds. It is possible that the reduced biases will increase the impact, making TC impact more detectable. The early phases of a possible operational EUMETSAT Doppler Wind Lidar mission are underway; so any improvements in the way we assimilate Aeolus or in the data processing may benefit NWP in the future.

Additional satellite platforms targeting TCs, which are being tested in limited-area assimilation, could be considered for testing the ECMWF system. One example is SAR images of surface winds, which

are similar to ASCAT images. Additionally, new small satellites are being developed and deployed, at a small fraction of the cost of the more conventional satellites. Although they are in their infancy, these new types of satellites may hold promise for further improving analyses and forecasts at ECMWF, if the data are timely and of high quality. One example is the TROPICS (Time-Resolved Observations of Precipitation structure and storm Intensity with a Constellation of Smallsats) constellation (Blackwell et al., 2018), which will be launched in 2022, following successful deployment of a pathfinder satellite in June 2021. TROPICS contains microwave soundings with a higher temporal sampling of TCs than is presently achievable. Another promising technology is CYGNSS, whose mission was recently extended to 2023. However, the CYGNSS measurements are less accurate at high windspeeds and are not yet available in near-real-time. The introduction of new observations brings additional costs for the system and needs to be motivated.

The investigations of observation impact in this report have highlighted the question of how to evaluate the impact of new observations in the most efficient way for TCs. There are several caveats in evaluating observation impact in TCs. Firstly, with the huge number of observations present in the operational system, it is difficult to reach a statistically significant impact from a single observation type. There is also the impact of previous assimilation cycles, remote versus local impact, and nonlinear effects in the weights given to the observations by the quality control. Given that the results are often a mixture of improvements and degradations for specific cases, it is important to consider a sufficiently large sample to achieve robust overall results. However, since TCs are not common events, a long integration period would be necessary. Such experiments are often not affordable computationally with the operational resolution, which was motivated in this special study by the fact that TCs are dominated by convective-scale and mesoscale processes. It remains an open question as to how experiments at lower resolution affect the observation impact on analyses and forecasts of TC position, intensity, and structure. For example, the effect of reduced thinning of observations might be under-played in low resolution experiments. As an alternative, the Forecast Sensitivity Observation Impact (FSOI) method can be utilised. Examples of the use of FSOI for TC cases can be found in Duncan et al. (2021). Even if no results from FSOI have been presented in this report, an exploratory study suggested that the method needs to be further adapted to reach reliable results for tropical cyclone impact when aggregated over many cases.

There are several aspects of the data assimilation methodology that could be further explored. With processes on multiple spatial scales, the fidelity of the data assimilation scheme in spatially and temporally resolving TCs in its outer and inner loops requires investigation. The scales also vary with the life-cycle of the cyclone. For example, the onset of RI usually occurs when a TC is a disorganised depression or weak tropical storm, with individual convective bursts dominating until a coherent vortex is formed and begins to axisymmetrize and intensify. These convective bursts and intensification processes occur on timescales of a few hours, within or on the edges of the 12-hour optimisation window in 4D-Var. On the other hand, a mature TC that is not undergoing major fluctuations (such as eyewall replacement cycles) may be resolved both spatially and temporally in the model, with lower wavenumbers dominating. The consequences of deficient resolution that may preclude the effective use of observations on finer time and space scales include: a) the inability to correctly represent spatial correlations in the observation errors; b) overly broad scales in the background errors, and the necessity to further tailor the background errors to TC conditions; and c)

the time and space scales of TC evolution may in some instances be too fast and small for the current system, with its 12 h window and around 50 km increment scale. Experiments to increase both the horizontal resolution of the data assimilation and reduce the length of the assimilation window (and/or to introduce a weak-constraint approach enabling increments to be generated on closer to 3 h timescales) may shed new light on the potential benefits to TC analysis and forecasting, especially for important cases of small, intensifying TCs which occurred several times in 2020 and whose intensity forecasts were poor.

Another assumption of the data assimilation methodology is the tangent linear model, whose fidelity requires investigation especially for small-scale features whose evolution is highly nonlinear. The physical tangent linear structures and subsequent increments in TCs warrant further analysis.

A separate but similarly important issue is the handling of erroneous observations that can severely impact the assimilation. Although several updates have been made in recent years to make the system more stable around TCs, occasional problems with erroneous observations still occur.

Another aspect to further investigate is the quality control. When the difference between an observation and the background is very large, the observation is not used in the analysis so as to avoid any erroneous outliers that degrade the analysis. VarQC was implemented at ECMWF in 1999 (Andersson and Jarvinen, 1999) and is currently applied to most satellite observations and to some conventional ones. But sometimes the large departure could be due to errors in the background, not in the observations. This is more commonly the case for rejected scatterometer wind observations close to the centre of a TC: if the location in the background is misplaced, the scatterometer observations close to the TC's centre have very large departures and are therefore rejected (the observations are given no weight in the assimilation). Usage of the Huber Norm (Tavolato and Isaksen, 2015) as alternative approach to VarQC is a good candidate for ASCAT winds; the Huber Norm allows observations with large departures to be included in the analysis even if with a reduced weight (which is a function of the magnitude of the departure). Preliminary investigations on the ASCAT winds with a few previous model cycles gave mixed results (from neutral to positive; De Chiara et al., 2018). It may be worth testing this again in combination with reduced thinning.

We have also discussed the aspect of coupled data assimilation and the generation of SST analysis. This aspect is most important in TCs as the temperature cooling rate is strong, and yet clouds limit satellite observations of the surface temperature.

The SVs targeted at TCs are important for the reliability of ensemble forecasts of TC track, and increasing the number of SV target regions prevents spread collapse in crucial periods during which many TCs are active at the same time. Future work will assess the impact of increasing the horizontal resolution of the SV computations, and how to improve the generation of the perturbed ensemble members in the context of ensemble data assimilation. Currently, EDA perturbations are constructed from short-range forecasts. The perturbations are then re-centred on the high-resolution deterministic analysis, which can lead to spurious structures in case of tropical cyclones (Lang et. al., 2015). The impact of alternative re-centring options in the EDA and multi-level Monte Carlo methods will also be evaluated.

7. Modelling of tropical cyclones

7.1. Introduction

In this section, recent and ongoing developments in the ECMWF atmospheric and ocean modelling system are presented in the context of TCs. During the 2020 Atlantic hurricane season, model Cy47r1 was operational. While working on this report, Cy47r2 was implemented with an increase of vertical levels for the ensemble from 91 to 137, and the move from double precision to single precision in both HRES and ENS. At the same time, model Cy47r3 was in preparation.

While the large-scale features of TCs can be captured by global models, the core of TCs is dominated by convective-scale and mesoscale processes with sharp gradients, which are at (and beyond) the limit of what a global model can accurately simulate operationally in 2021. To account for these small-scale processes, km-scale limited area models are used as the main tool for TC intensity forecasts. Examples include AROME run operationally at Météo-France with 2.5 km resolution over five different domains in the tropics, and the HWRF (2 km) and COAMPS-TC (4 km) regional models that are run globally at NOAA and the United States Navy respectively. The benefits of this higher resolution are to further resolve processes in the boundary layer, eyewall and rainbands, yielding improved predictions of Pmin and Vmax, and potentially the track due to more realistic TC structures interacting with the environment. However, it should be noted that improvements to the forecast quality are not automatically gained due to an increase in resolution, but go hand in hand with updates to the model physics and the dynamical core of the model.

With increased resolution, the distinction between the model dynamics (the resolved transport) and physical parametrizations (the subgrid-scale transport and mixing) is becoming more blurred, a challenge often referred to as the 'grey-zone'-problem. This is a wide topic for research, for example on the need of parametrized convection, and in Section 7.2 some relevant ongoing developments at ECMWF and their impacts on TC forecasts are highlighted. The recent development of the new moist physics package for the forthcoming Cy47r3 is also presented, as are the effects of a possible future resolution of 4 km and how convection is handled on this scale. Improvements to the model dynamics and the physics-dynamics coupling are ongoing, and the impact of this on TC forecasts is presented. Finally, in Section 7.2, an update on the future developments of the dynamical core towards a non-hydrostatic finite volume model is provided.

TCs develop and are maintained by heat and moisture fluxes from the ocean and are sensitive to the momentum flux at the air-sea interface. It is therefore important to accurately represent these processes with the coupling to the dynamical wave and ocean models, and relevant work is discussed in Section 7.3.

With the imperfections in modelling the evolution of TCs, it is important to also simulate the model uncertainties in the ensemble system to obtain a more reliable ensemble forecast. Some recent and ongoing developments in this field at ECMWF are highlighted in Section 7.4.

In Section 5 we highlighted challenges in simulating the observed relation between Pmin and Vmax, as the model has too slow propagation speeds and too slow intensification and weakening rates. In this section, we examine average errors (biases) and evaluate if any of these aspects are improved. One has to keep in mind that the compensation of errors can lead to a small bias for the wrong reason. For

example, we can ask if the current version of HRES with 9 km grid-spacing and effective resolution of roughly 30 km should have a bias in TC intensity close to zero as seen in Figure 9.

As in Section 5, most experiments reported on in this Section were conducted for the 37-day period 15 August-21 September 2020. The baseline experiment is a 9 km (TCo1279) coupled forecast with Cy47r1, as used operationally in autumn 2020. All experiments are initialised from the operational analysis.

7.2. Model developments in physics and dynamics

As TCs are one of the phenomena with the strongest near-surface winds and gradients on the globe, they are one of the most challenging test cases for the numerical aspects of the model. The strong horizontal and vertical winds push some of the theoretical assumptions, which guarantee stable and efficient performance of the model numerical algorithms, to their limits. For example, it was found that the numerical algorithm used to compute the departure points of the semi-Lagrangian (SL) advection scheme needs more iterations to converge in the high wind speed and high wind shear area of TCs than elsewhere (Diamantakis and Magnusson, 2016). Analysis showed that in such regions, the Lipschitz condition (deformational Courant number smaller than 1) which needs to be satisfied to guarantee convergence of the departure-point iterative procedure is close to its limit. A failure to converge results in mislocated air parcels and large errors in the predicted flow. For the long timesteps that are permitted by the good stability and dispersion properties of the semi-Lagrangian method, convergence of this iterative procedure is slow, and five iterations were found to be required. The numerical convergence can also be improved with a shorter timestep but at the expense of a much larger increase in the overall computational cost.

A reduction in the **model timestep** affects not only dynamical but also physical processes. Therefore, the model is sensitive to the timestep choice, which may have an impact on the TC core pressure and track. Figure 48 includes results from an experiment with the model time-step reduced from 450s ("9km-oper") to 300s ("300s"). In previous experiments we have found occasionally deeper TCs with smaller timesteps, but in this sample the mean difference is small and slightly to the weaker side (not shown). In the past we have also seen indications of a sensitivity for propagation speed to the time-step going from 450s to 225s resulting in ~0.1 m/s faster TCs. For this experimentation, evaluating the average speed for TCs between step 24 and 48 hours, we find a weaker sensitivity (decrease in bias from -0.30 to -0.28 m/s). As seen in Appendix A, the tropical RMSE for 700hPa temperature and winds are slightly degraded with a shorter time-step. There is currently no other systematic improvement known in the overall model forecast skill to justify shortening the timestep and raising the model cost.

There are several ongoing dynamical core developments that could potentially affect TCs. The first one concerns the issue of **slow convergence of the departure-point iterations** that was mentioned earlier in this section. A reformulation of this algorithm (Diamantakis and Vana, 2021), in a geocentric Cartesian framework, allows one to simplify these computations and to accelerate their convergence with no impact on TC track and forecast error. This is illustrated in Figure 48 where a single-precision version of the IFS (as operational in Cy47r2) that uses the "new SETTLS" scheme is compared with a double-precision version of the operational scheme ("9km-oper") for departure points. Both experiments use model Cy47r1 and identical physics, but the former is almost twice as

fast as the latter (due to single precision and use of less iterations) and performs equally well in terms of predictive skill.

In the IFS the **Coriolis force** is only included for horizontal motions. However, in Liang and Chan (2005) it was argued that including the Coriolis force due to vertical motion in the model should lead to a shift of the TC propagation towards the south-west. This result is relevant as the IFS currently has a propagation speed that is slow in comparison with observations, and with a frequently observed drift to the right for weak westward propagating TCs in the Atlantic (as illustrated in Figure 20). In Tort and Dubos (2014), a reformulation of the shallow-atmosphere equations that retains both the vertical and horizontal components of the Coriolis force is proposed. This reformulation is currently being tested in the IFS and the results are presented in Figure 48. Whilst the impact on position and Pmin absolute error is small, we find an impact on the propagation speed bias from -0.30 m/s in the 9 km control experiment ("9km-oper") to -0.24 m/s in the experiment with the change to the Coriolis force ("Coriolis"), and the difference seems to be larger for low-latitude storms. As the magnitude of the bias is small compared with the variability in propagation speed, further investigation is needed to understand the robustness of this result.

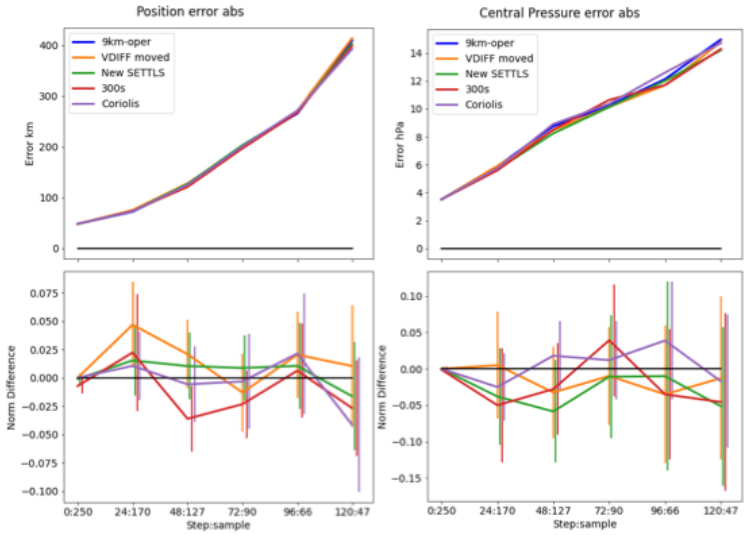


Figure 48: Mean absolute error for position (left) and Pmin mean absolute error (right) for experiments with changes to the model dynamics. The bottom plots show the normalised difference in error to the 9km-oper experiment.

A **major upgrade of the moist physics** has been prepared for Cy47r3 (Bechtold et al., 2020). One main driver of this development is the future grid resolution upgrade of the ensemble to O(5 km). With this in mind, the complicated interactions between the turbulence in the lowest part of the atmosphere, convective motions and the cloud physics are parameterised as simply, efficiently, accurately and scale-independently as possible. An important component of this upgrade is a revision

to the convective closure that takes into account, beyond the convective instability, the total moisture advection, allowing a more realistic representation of mesoscale convective systems in terms of rainfall statistics and propagation (Becker et al., 2021).

However, during initial evaluations of the physics upgrade, a large sensitivity of TC core pressure to the convective closure became apparent. TCs and especially the eyewall constitutes an extreme case of total moisture advection, and an excessive increase in convective stabilisation will lead to TCs that are too shallow. A reasonable fix to this problem in Cy47r3 was to exclude the moisture advection term in regions where the vertically integrated saturation fraction exceeds a threshold of 0.94 marking the transition to resolved moist overturning.

The forecast position errors (Figure 49), Pmin errors (Figure 50), Vmax errors (Figure 51) and wind-pressure relationship (Pmin versus Vmax, Figure 52) are illustrated for model runs with the operational Cy47r1 moist physics ("9km-oper") and the new moist physics ("9km-newMP") that forms the basis for Cy47r3. The forecasts have also been run with a resolution of 4 km (TCo2559) representative of a possible future resolution upgrade, for both the operational moist physics ("4km-oper") and the new moist physics ("4km-newMP").

The lowest position error is produced by the "4km-newMP", while both the 9 and 4 km runs of the newMP reduce the position error compared to the respective operational runs, in particular during the first three days (Figure 49(a),(b)). This result might be an effect of the improved large-scale winds in the tropics with the new moist physics, as shown in Appendix C.

The improved position forecasts for the 4 km resolution experiment with new moist physics ("4km-newMP") are encouraging. Next, the corresponding evaluations of the intensity are considered. An illustration of the improvement to the surface wind structure is provided in panels (c) and (d) of Figure 49, for the same 60 h forecast of Hurricane Laura for which the corresponding operational 9 km forecast was illustrated in Figure 5. From Figure 49(c), the TC is much stronger in the "4km-newMP" experiment than in the operational forecast, with a sharper eyewall. This is an encouraging result, as it demonstrates the ability of the 4 km model to realistically simulate the most intense hurricanes. Comparing Figure 49(d) against Figure 5(right), the maximum wind speed in the "4 km-newMP" simulation is accordingly closer to the NHC Best Track value of Vmax, and there is a stronger inflow (blue line). Overall, this encouraging example suggests the potential for overall improvement in intensity forecasts for the entire sample, which will be described next.

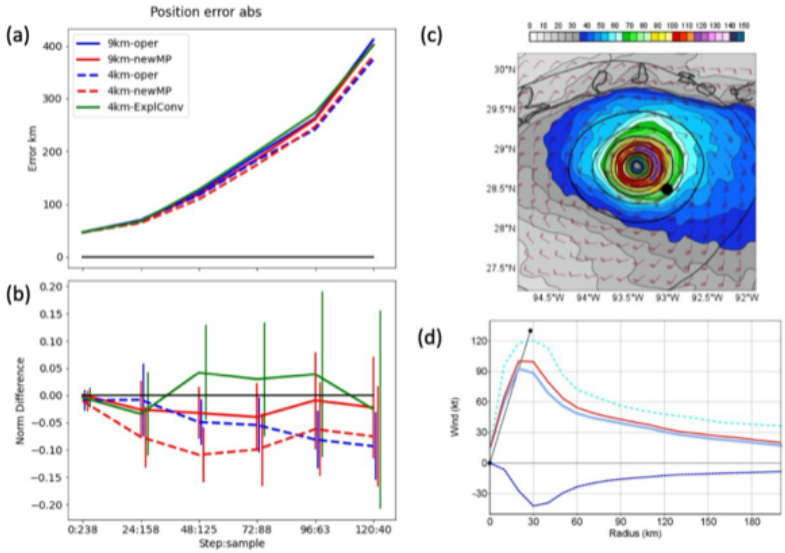


Figure 49: (a) Mean absolute error for position for experiments with the new moist physics package and experiments with 4-km resolution. (b) Normalised difference in error to the 9km-oper experiment. (c) Surface wind structure in Hurricane Laura, for a 60 h 4km-newMP experimental forecast initialized from the operational analysis on 1200 UTC 24 August 2020. Black contours: Forecast MSLP. Wind barbs: 10 m winds (kt). Shading: 10 m wind speed (kt). Black dot: NHC Best Track position. (d) Radial mean of total (red), tangential (light blue, dotted) and radial (blue) 10-metre wind speed, and maximum wind speed at each radius (cyan, dashed) for 4km-newMP experiment. Black dot: NHC Best Track value of Vmax and the radius to Vmax.

The operational configuration of HRES ("9km-oper") has a bias in Pmin close to zero (Figure 50) although, as presented in Section 5.5, this is due to the cancelling out between a low Pmin bias for initially strong TCs and a high Pmin bias for initially weak TCs (Figure 21). The newMP with 9 km resolution ("9km-newMP") produces weaker TCs on average due to the moisture convergence change. The mean absolute error in Pmin is slightly lower for this configuration out to 3 days, and these improvements are mostly evident in initially strong TCs (in which the operational forecasts have a low - i.e., too deep - Pmin bias, and the new moist physics reduces this bias). When the resolution is increased to 4 km, the TCs are deeper in both physics configurations. The operational moist physics ("4km-oper") produces TCs that are too deep, which increases the mean absolute Pmin error with respect to the operational 9 km simulations. In contrast, the 4 km simulations with new moist physics ("4km-newMP") brings the overall bias back near zero and produces the lowest mean absolute error of all configurations presented in this paper.

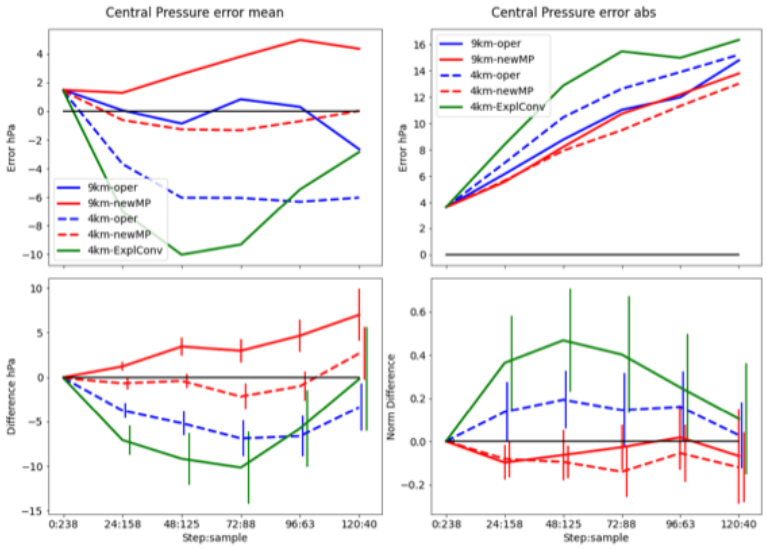


Figure 50: Mean error (left) and mean absolute error for position (right) for Pmin for experiments with the new moist physics package and experiments with 4-km resolution. The bottom plots show the (normalised in right) difference in error to the 9km-oper experiment.

The corresponding results for Vmax are presented in Figure 51. Given that Vmax is the key intensity metric used by most operational TC forecasting centres, and it is generally recognized that models of resolution 4 km or less are required for meaningful predictions of this metric, the 4 km experiments here are of particular interest. The results are impressive. Not only does the increase in resolution to 4 km substantially reduce the too weak Vmax bias, but the mean absolute error in Vmax for two-four-day forecasts is reduced by ~7 kt. For the Atlantic TCs in the experiment sample, these improvements are statistically significant (compared with "9km-oper", using a one-tailed t-test) at the 85% level for all forecast times, and usually the 90% or 95% levels. The improvements are most evident and significant for the cases in the sample that were initially weak TCs (Vmax < 50 kt), for all forecast times out to five days (not shown). These improvements due to the 4 km resolution corroborate those for Pmin, although unlike for Pmin, no clear distinction in MAE between "4km-oper" and "4km-newMP" is evident for Vmax probably due to a lower bias for 4km-oper. Overall, the level and consistency of improvement in both Pmin and Vmax when the resolution was increased to 4 km with newMP was not evident in any of the other modelling or data assimilation experiments.

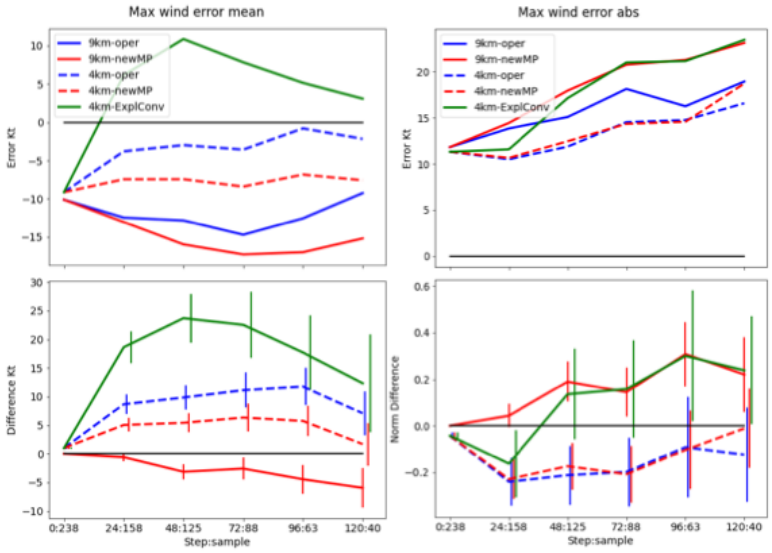


Figure 51: Mean error (left) and mean absolute error (right) for Vmax for experiments with the new moist physics package and experiments with 4-km resolution. The bottom plots show the (normalised in right) difference in error to the 9km-oper experiment.

For the pressure-wind relation (Pmin vs Vmax, Figure 52), all configurations underestimate Vmax on average compared with the Best Track. However, one needs to keep in mind the difference in temporal and spatial sampling for the maximum wind in the Best Track, as the model represents a grid box average over a full time-step. As will be discussed later in this section, at ECMWF the pressure-wind relation has improved substantially since the introduction of cycle 47r1. Increasing the model resolution from 9 km to 4 km improves the pressure-wind further for Vmax > 50 kt, and substantial improvements are especially evident at high intensities, where the dashed lines are significantly closer to the black line in Figure 52. This suggests an improved gradient wind balance relation, which is the dominant horizontal balance in intense TCs.

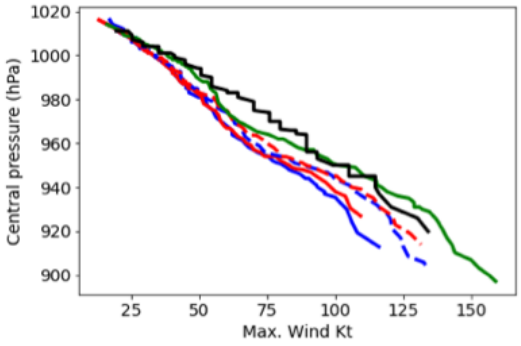


Figure 52: Wind-pressure relation for experiments with the new moist physics package and experiments with 4-km resolution evaluated for time-steps 24-120h. For legend, see Figure 51 (Best Track in black).

Table 2 shows the number of tracked feature points for all basins relative to the 9 km control experiment. (Note that these numbers are not compared with the Best Track as the ECMWF tracker includes pre-TC and post-TC stages.) Here we find that the runs with newMP (both 4 km and 9 km) produce ~10% fewer features compared with their old physics counterparts. Comparing 4 km resolution with 9 km resolution, we find a similar number of features, but about 30% more cases of hurricane intensity with the higher resolution. Further diagnostics of the TC activity is needed to validate the benefit of the changes (see Section 4.3).

Table 2: Number of features relative to 9km control experiment for different cyclone strengths.

	>8 m/s	>17 m/s	>32 m/s
9km	1	1	1
9km-newMP	0.87	0.89	0.93
4km	1.02	1.24	1.31
4km-newMP	0.90	1.06	1.21
4km-ExplConv	1.24	1.60	1.69

Going towards km-scale model resolutions, one open question is the handling of deep convection. In the DYAMOND project (Judt et al., 2021) a 4 km version of the **IFS without parameterised** convection was compared with other global models with similar resolution. We have therefore also run here the newMP with the deep convection switched off, to let the model explicitly resolve the convection ("4km-ExplConv"). This required a reduction of the model time step from 300 to 180 s to

avoid model failures due to excessive vertical velocities. The experiment with explicitly resolved deep convection provided slightly higher errors in the position (Figure 49), and far too intense TCs (Figure 50 and Figure 51). However, the wind-pressure relation is improved with 4km-ExplConv compared with 4km-newMP (Figure 52). This result is in line with experiments for Polar Lows where simulations with 5 km resolution and explicit convection produced stronger and more localised wind maxima compared with simulations with parameterised convection (Hallerstig et al., 2021).

The impacts of increased model resolution and deep-convection parameterisation are illustrated for one forecast for TC Laura (Figure 53). For the 4 km experiment with explicit deep convection, Pmin almost reached 900 hPa when the Best Track was around 940 hPa. Among the other experiments, the best results are seen for 4km-newMP, in line with the verification results for the full sample. This is further illustrated by the comparison of simulated infrared satellite images at the most intense time of TC Laura, with a satellite image from GOES-16 (Figure 54). Here the 4km-newMP has a clear eye of the cyclone, in line with the real image. For 4km with explicit convection, the eye is hidden by thick clouds and the TC is smaller in scale, resulting in too small a radius of maximum wind (not shown).

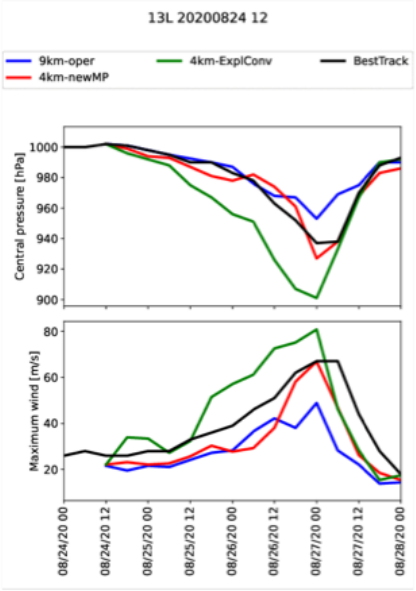


Figure 53: Intensity (Pmin) forecasts for TC Laura in forecasts from 24 August 12UTC.

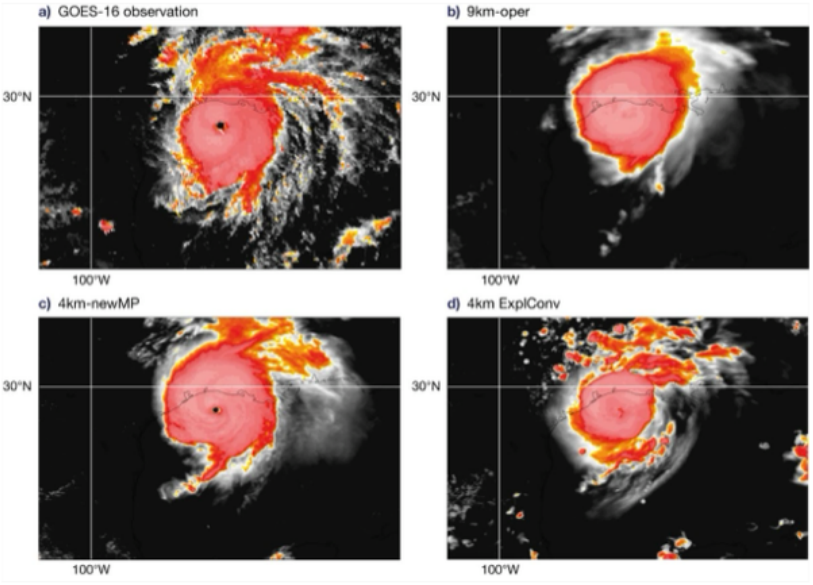


Figure 54: Satellite images for TC Laura on 27 Aug 00UTC from GOES-16 (a), and simulated images from forecasts initialised 24 Aug 12UTC in 9km-Oper (b), 4km-newMP (c) and 4-km with explicit convection (d).

For the propagation speed, averaged between step 24 and 48 hours, we find for this sample a bias of -0.28 m/s for the control experiment (9 km). Note that the numbers are slightly different to the ones from the dynamical core experimentation due to homogenisation over a different set of experiments. 9km-newMP increases this bias to -0.32 m/s. Both 4 km experiments decrease this bias to -0.20 m/s (4km) and -0.26 m/s (4km-newMP). However, as documented in Appendix C, both newMP experiments have the lowest RMSE for 700hPa winds in the tropics. Furthermore, 4km-ExplConv results in a lowest propagation speed bias (-0.09 m/s) while being the experiment by far with the highest RMSE for wind errors at 700hPa in the tropics. One can note that the two experiments with the strongest intensity bias also have the fastest propagation of the TCs. As discussed above, we did see some sensitivity to introducing the vertical component of the Coriolis force (reduction from -0.30 to -0.24 m/s). Further investigations with a larger forecast sample are needed to better understand these differences, and how it relates to different types of TCs (strong vs. weak, low-latitude vs high-latitude).

Experiments have also been undertaken to investigate sensitivities in the **coupling between the dynamics and physics** in the model. Results with the new moist physics show that the quality of the TC forecast is sensitive to the coupling of the diabatic processes parametrization to the model dynamical core. Following suggestions made by Beljaars (2016) and results of the theoretical study of Termonia and Hamdi (2007), the vertical diffusion scheme has been moved to the very end of the sequential physics chain. This allows the convection scheme to directly respond to the explicit

tendencies from dynamics, while having the vertical diffusion scheme moved at the end stabilizes the whole physics. This theoretically justified option currently represents only an experimental design requiring a double call of the vertical diffusion scheme – before convection, to provide BL fluxes to convective closure and then second time at the end of physics. Furthermore, changes in the coupling between the vertical diffusion scheme and the radiation, showed promising results in terms of forecast skill were also tested on TCs. Including both coupling changes shows a small positive impact with respect to the new moist physics (see VDIFF experiment in Figure 48). More optimal alternatives in terms of computational cost for the coupling between the convection and vertical diffusion scheme may be investigated in future.

Another open question is the impact of moving to a **non-hydrostatic dynamical core**. However, tests with the IFS at 9 km horizontal resolution indicate much larger sensitivities to the special numerical algorithms, which need to be activated to integrate in a stable fashion the more demanding non-hydrostatic equation set, than to relaxing the hydrostatic approximation itself. For example, recent experimentation shows that the physics-dynamics coupling of the so called Iterative Centred Implicit scheme (ICI) used by the non-hydrostatic model needs improvement to be competitive with the scheme used in the hydrostatic dynamical core. The TC performance is degraded also in a hydrostatic experiment with the ICI activated to a similar level as the non-hydrostatic experiments, compared to the operational hydrostatic model without the ICI scheme (not shown). Interestingly, while the ICI scheme deepens the TCs, the maximum wind in the TCs is almost unchanged, leading to a significant degradation in the maximum wind vs. minimum pressure relationship.

As a further development, the **IFS-FVM non-hydrostatic dynamical core** employs a fundamentally different finite-volume design with inherent conservation and robust semi-implicit integration (Kühnlein et al., 2019). The IFS-FVM foundations lie in cloud-scale large-eddy simulation, potentially offering numerical techniques for NWP at kilometre-scale resolutions and beyond. Finite-volume methods are common among the newest generation of NWP models such as ICON of DWD, but formulations differ significantly in details. An experimental IFS-FVM forecast configuration using the IFS Cy43r3 physical parametrizations has been developed, which enables re-forecast studies (as yet uncoupled) with this new model formulation, including tropical cyclone cases. In 2020, a first single case study of TC Irma (2017) verified the IFS-FVM against the established IFS as well as the best track data (see Kühnlein et al., 2020). The TC case study has been very useful to demonstrate the basic skill, as well as to identify open issues and specific aspects for further refinement of the IFS-FVM forecast configuration, particularly in initialization and physics coupling. Further systematic comparison studies against the hydrostatic and non-hydrostatic spectral-transform IFS will follow. To enable the highest possible future resolutions, the IFS-FVM is currently undergoing revision in terms of its software infrastructure towards emerging computing technologies (Bauer et al., 2020).

7.3. Ocean and wave modelling

Since 2018, all ECMWF forecasts are produced with a **coupling to the NEMO dynamical ocean model**. One of the main motivations for making this step was improved intensity in TCs. Strong TCs lead to SST cooling due to vertical mixing, enhanced upwelling, and heat and moisture flow to the atmosphere. Historically, when ECMWF forecasts predicted too weak TCs (see Section 1) due to the lower horizontal resolution, the lack of ocean coupling was not crucial due to absence of very strong

wind. Also, the lack of SST cooling might have partly compensated intensity errors due to low resolution. However, with increasing model resolution, cases of too deep TCs (Pmin too low) started to appear, most pronounced in the north-western corner of the North-western Pacific basin. Mogensen et al. (2017) investigated this issue and related it to the lack of ocean coupling, which had the strongest effect on the shallow, warm layers in the ocean such as south of Japan. The positive effect of ocean coupling has also been found at other global modelling centres, such as the Met Office (Vellinga et al., 2020).

Figure 55 includes results for experiments with ("9km-oper") and without ocean coupling ("FC uncoup"). On average, the experiment without ocean coupling led to too strong TCs and increased absolute errors. The clearest example in this sample from Aug-Sept 2020 was TC Haishen, whose path took it over an area south of the Japanese island Kyushu with low ocean heat content, but initial warm SST of around 30°C. This is illustrated in Figure 56 where we have plotted the mean temperature in the uppermost 300 m as a measure of upper ocean heat content. The bottom panels of Figure 56 show the evolution of SST in the coupled experiment (left) and the uncoupled experiments (right) overlayed with SST observations from drifters and ships. The coupled model was able to simulate the cold wake after the TC even though the cooling was less than what is observed. As the uncoupled model lacked the physics needed for the air-sea interaction, it continued to be driven by the prescribed warm SSTs leading to the over-intensification of the forecast. It is worth noting that TC Haishen behaves in a very similar way to TC Neoguri 2014 which was discussed in detail in Mogensen et al. (2017).

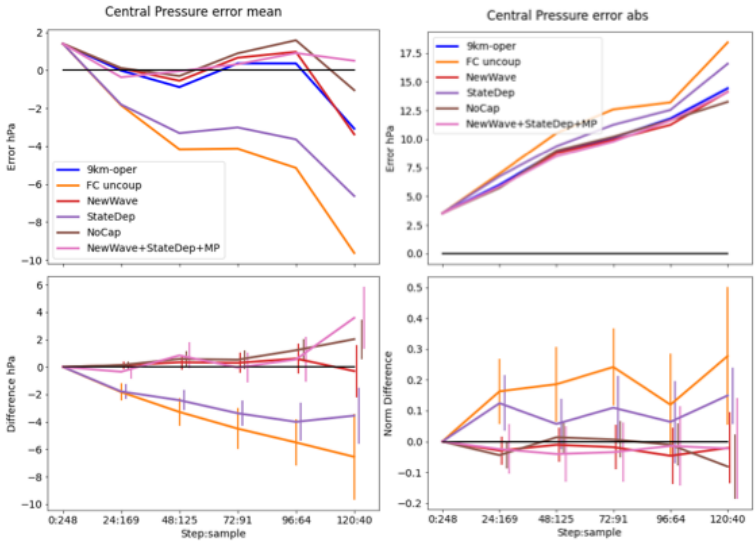


Figure 55: Mean error (left) and mean absolute error (right) for Pmin for various experiments related to ocean and waves. The bottom plots show the (normalised in right) difference in error to the 9km-oper experiment.

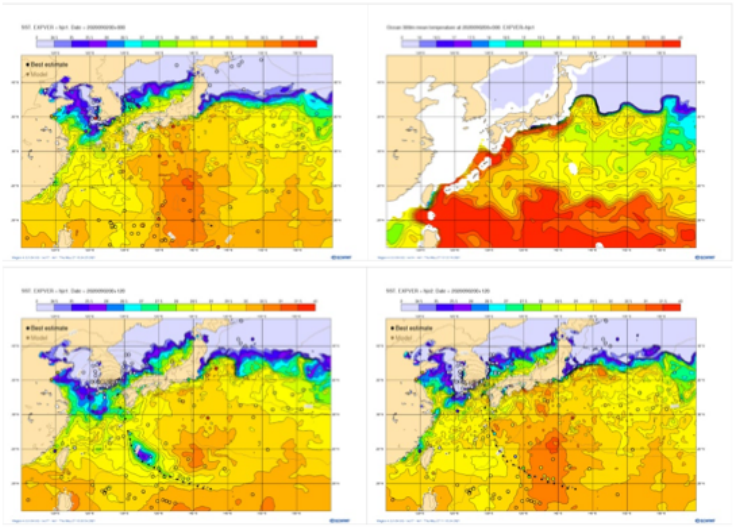


Figure 56: Sea surface temperature (top left) and mean upper 300 m temperature (top right) both valid at 20200902 00z and SST from coupled (bottom left) and uncoupled forecasts (bottom right) valid 120 hours later. Black dots indicate Best Track for TC Haishen and coloured dots are SST from observations.

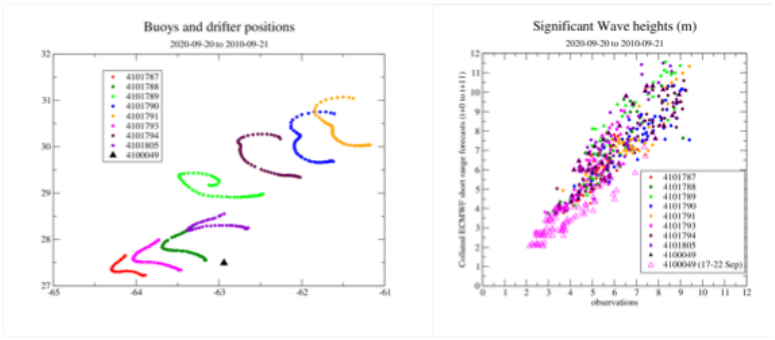


Figure 57: Observation positions for buoy and drifter after passage of TC Teddy (left) and significant wave height co-located in short HRES forecasts and observations.

All ECMWF forecasts are coupled to a dynamic ocean wave model. Besides providing users with sea state parameters, as detailed as the full two-dimensional wave spectrum, the two-way coupling to the wave model ensures a physically more consistent representation of air-sea interactions, linking the lower atmosphere to the upper ocean. As an example of wave evaluation for one cyclone, Figure 57 shows significant wave height from HRES and observations under TC Teddy, from an array of buoys seen in Figure 44, around 28°N, 62°W. For this case there is a general good match between modelled

and observed wave height. For the buoys closest to the centre of the cyclone centre path (purple and light green), the wave height was somewhat over-estimated in HRES.

Prior to IFS Cy47r1, ECMWF forecasts underestimated maximum wind speed for intense TCs even given the correct central pressure. While there are many different factors that could account for this behaviour, a strong candidate is that it could be linked to the parametrization of **momentum exchange at the ocean surface.** This momentum exchange is dependent on the state of the waves (sea state) and not just the surface winds, resulting in a range of values for the drag coefficient over the oceans for similar wind speeds. It is achieved by coupling the atmosphere with an ocean wave model.

Over the last decade, it has been suggested that the drag coefficient should tail off for strong winds. In the IFS, the momentum exchange with the sea surface is modelled via a dependency of the roughness length (z_0) on the surface stress. This expression accounts for both low and high wind regimes. At low wind speed, the sea surface becomes aerodynamically smooth and z_0 is determined by viscosity. At high wind speed, Charnock's relation is used, in which z_0 is expressed as a function of surface stress, air density, gravitational acceleration and a sea-state-dependent Charnock parameter. In ECMWF's wave model, the Charnock parameter depends on the state of development of the resolved waves and a tuneable parameter (α_b) which represents the impact of unresolved short waves (background roughness beyond the highest frequency resolved by the model) on the overall surface stress. Until Cy47r1, this parameter had a constant value.

Observational evidence that the drag coefficient should be much lower for high winds suggests that the coupling between the ocean surface and the wind above becomes less efficient at transferring momentum for high winds. For this to happen, it was realised that the Charnock parameter should be considerably smaller in the case of high winds (above 35 m/s). This was achieved by reducing α_b for strong wind speeds in Cy47r1. This development is the result of internal ECMWF work informed by discussions with scientists at Météo-France and the US National Centers for Environmental Prediction (NCEP).

The effect of the drag change in Cy47r1 is evident in the pressure-wind (Pmin versus Vmax) relation in Figure 58, where the experiment "NoCap" has the formulation from the previous cycle without the cap of the drag. The wind-pressure relation is much less consistent with observations for the experiment "NoCap" for wind speeds above 70 Kt, compared to 9km-oper. However, it could also be seen that all experiments still underestimate the winds in a similar way for speed between 60-70 Kt. This indicates that more than just a capping of the surface drag is necessary to solve this issue.

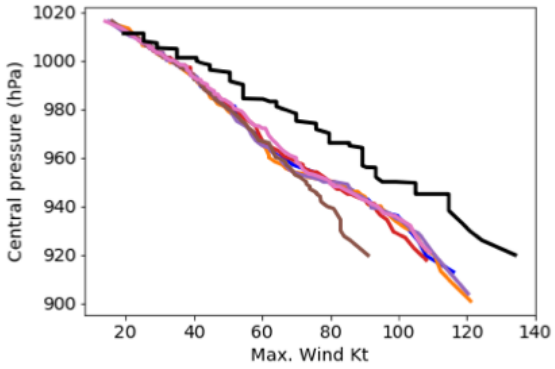


Figure 58: Wind-pressure relation for experiments with the different ocean configurations evaluated for time-steps 24-120h. For legend, see Figure 55.

We are currently exploring an alternative approach to the simple capping of the Charnock parameter. The current parametrization for the wind input into waves is not properly accounting for the impact of the short gravity-capillary waves, nor is the nonlinear effect of the wave spectrum on the growth rate of wind waves accounted for. Recent work as presented in Janssen and Bidlot (2021), is proposing a simple model for the impact of gravity-capillary waves on the wave growth and is extending the current parametrization to include the nonlinear effect. Both effects yield a reduction of the drag that is function of the sea state. Because the impact of the new approach affects all wind regimes, it will require full meteorological testing. However, for the current experimentation, we see that the "NewWave" simulation gives similar results as the reference (Figure 55, Figure 58).

The sea state also has an impact on the latent and sensible heat fluxes via the exchange coefficients for heat and moisture. From recent observation campaigns (Brut et al., 2004; Cook and Renfrew, 2015), there are indications that the exchange coefficients for heat and moisture should have a stronger dependency on the wind speed (i.e., the sea state) than currently modelled. Recent sensitivity experiments have shown the potential impact on the atmosphere (Jansen and Bidlot, 2018). In the ongoing work, we have quantified the impact of this new parameterization on TCs ("StateDep"). With this change, more heat and moisture can be extracted from the ocean, the TC gets deeper, and surface winds stronger. It is worth stressing that such a change would not have been feasible without an interactive ocean, as the TC-induced cooling of the ocean is needed to avoid further over deepening of TCs. There still appears to be some over intensification. However, when we combine "NewWave" and "StateDep" with the new moist physics (see Figure 55), the mean error for Pmin is close to zero (and similar to control), as the new moist physics reduces the intensity. This illustrates that there are several configurations that can give a close to zero bias for Pmin, and highlights the care needed when interpreting the results due to compensating errors.

We also note that the effect of spray as generated by whitecaps and breaking waves is still ignored and should considered in future work (Wu et al. 2015).

7.4. Sampling model uncertainties

At ECMWF, the Stochastically Perturbed Parametrisation Tendency scheme (SPPT) is used to simulate model uncertainties in ensemble forecasts (Buizza et al., 1999). The idea behind SPPT is to perturb the total physics tendency, that is, the sum of the tendencies from all physical parametrization schemes, such as convection, radiation, etc. The perturbations are made by multiplying the total tendency with a factor determined from a two-dimensional (2D) Gaussian random field with prescribed space and time decorrelation scales. The currently operational implementation of SPPT at ECMWF is described in Leutbecher et al. (2017) and Lock et al. (2019).

Figure 59 shows the mean absolute error of Pmin during July-November 2020 for operational HRES (the performance of Cy47r2 HRES was very similar, not shown), ensemble control (ENS-CF) and one ensemble member (ENS-PF1), from the pre-operational ensemble experimentation for Cy47r2. For the Pmin, the ENS-CF and ENS-PF1 has a larger positive bias (too weak TCs) than HRES as expected from the lower horizontal resolution. We find the ENS-PF1 has somewhat stronger TCs than the ENS-CF for all lead-times. We suspect this difference is coming from the use of the SPPT scheme to simulate model uncertainties. The wind-pressure relation is similar for the ENS-PF1 and ENS-CF (not shown).

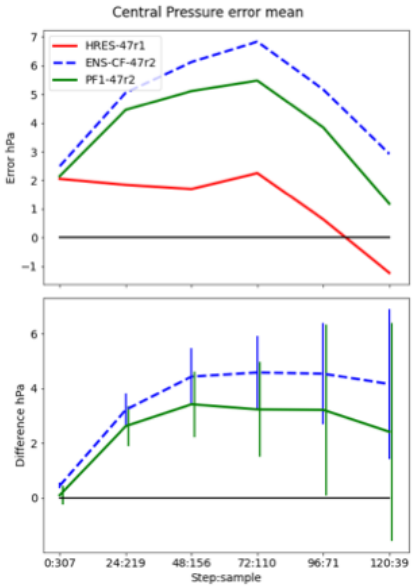


Figure 59: Pmin mean error for HRES and control forecast and one perturbed member from cy47r2. The bottom plots show the difference in error to the HRES.

Table 3 shows the number of cyclonic features relative to operational HRES. With the lower horizontal resolution for the ensemble we expect fewer features, especially for the stronger categories. This is confirmed and the ENS-CF has only 45% of the number of hurricanes in HRES. Comparing ENS-CF and ENS-PF, we find more features in the perturbed members and the signal is stronger for the weaker categories. This result is in line with results from System 5 seasonal forecasts (Stockdale et al., 2018) and long simulations presented in Vidale et al. (2021).

Table 3: Number of cyclonic features relative to 9km HRES forecasts for different cyclone strengths for step 24-120h.

	>8 m/s	>17 m/s	>32 m/s
HRES (9km)	1	1	1
ENS-CF (18km)	0.92	0.80	0.45
ENS-PF (18km)	1.13	0.95	0.56

ECMWF is currently developing a Stochastically Perturbed Parametrizations scheme (SPP), which represents model uncertainty by introducing stochastic perturbations directly into the physical parametrization schemes. In contrast to SPPT, SPP preserves conservation properties of the parametrization schemes within the model column. While the first implementation of SPP produces less skilful medium-range probabilistic forecasts than SPPT, a recently revised SPP configuration is now competitive with SPPT (Lang et al., 2021). The revised version from the scheme introduces perturbations to additional quantities and modifies the probability distributions sampled by the scheme compared to the original implementation of SPP documented in Ollinaho et al. (2017) and Leutbecher et al. (2017).

Lang et al. (2021) compared the ensemble spread for TCs in lower-resolution ensemble forecasts (TCo399, approx. 30 km horizontal grid spacing) and found that SPP results in larger TC intensity (Pmin) ensemble spread compared to SPPT, while intensity forecast errors are very similar, leading to a better spread-error relation. TC track errors and ensemble spread are also very similar for SPPT and SPP.

A further scheme that is under development explores how to represent model uncertainty arising from the IFS dynamical core: STOCHDP introduces stochastic perturbations to the departure point (DP) calculation in the semi-Lagrangian transport scheme (Leutbecher et al., 2017). As discussed in Section 7.2, in Diamantakis and Magnusson (2016) the iterative DP calculation was shown to converge most slowly (indicating greatest uncertainty) in regions associated with more complex flow. In particular, the authors presented a case study of TC Neoguri (2014), which demonstrated that where the windspeeds and Lipschitz numbers are largest, the DP calculation is slowest (and in places even fails) to converge.

In STOCHDP, the DP convergence rate is used to scale the stochastic perturbations that are added to the DP calculation. An examination of the TC Neoguri case with STOCHDP active has shown that the scheme successfully generates ensemble spread which tracks and develops with the TC (see Lock, ECMWF Annual Seminar 2020). Work is ongoing to explore the scheme sensitivities to the choice of perturbations and model resolution.

7.5. Summary of results

In this section, several recent and ongoing developments to the IFS model have been discussed, both in the physics and dynamics of the atmosphere, and aspects related to the ocean. The simulation of model uncertainties has also been discussed. All these aspects have largely been discussed in the context of processes in the vicinity of TCs, which occur mostly on the convective-scale and mesoscale (of order 1 km up to a few hundred km).

Particular attention has focused on a range of modelling experiments during the special 37-day period in autumn 2020, which were conducted using the same initial conditions as in operations. We therefore note that the challenges described in Section 5 in creating accurate operational analyses of TCs carry over to the modelling experiments, as the modelled TC is dependent on the realism of the initial conditions. Cognizant of this caveat, these parallel experiments have nevertheless provided a unified framework to understand the roles of the different modelling tests, with the same TCs under investigation. The intent in some of the tests is not necessarily to improve the forecast, but instead to understand the behaviour of a particular modification in several cases. A summary of the finding of the 2020 experiments is provided here, with further discussion and future directions in the final sub-section.

Generally, as was the case for the data assimilation experiments, the modelling experiments did not produce dramatically different forecasts from the (near) operational version that was used as the control. This is as desired, since it demonstrates robustness and a lack of volatility in the model.

The reduction of the timestep (in the 9 km configuration) did not result in a systematic improvement to the TC forecast skill, and hence these results do not provide a justification to shorten the timestep and raise the model cost.

The single-precision test performed as well (with new SETTLS scheme) as the (47r1) double-precision test, encouragingly suggesting that overall degradations are not likely when single-precision is used. This is consistent with Cy47r2 that included the change to single-precision in HRES.

The introduction of the contribution to the Coriolis force from the vertical motion provided some modest improvements to the track forecasts of weak TCs in the Atlantic basin. The experiment showed a decreased propagation speed bias, a result that will be followed up with a longer experiment period to increase the robustness.

The change to the moist physics package at 9 km resolution, which will be included in the 47r3 cycle, provided improved TC position forecasts. Statistically significant improvements to Pmin forecasts were found in the Atlantic basin, especially for cases being hurricanes (> 64 kt) at the initialisation time.

The experiments in which the resolution was increased to 4 km provided the most significant combined improvements to Pmin, Vmax, and the wind-pressure relation out of all the 2020 experiments. The usage of the new moist physics package in the 4 km experiment yielded further improvements in Pmin, with a reduction in the strong bias for initially strong TCs, and a reduction in the weak bias for initially weak TCs. As we expect the intensity bias to asymptotically approach zero with increasing resolution, we see this as a positive result for the new moist physics. The experiment with explicitly resolved deep convection provided slightly higher position errors, and the TCs were far too intense.

The experiment with no coupling of the atmospheric model to the ocean yielded significantly degraded results, especially for the minimum pressure of hurricane-strength TCs. This result confirms the importance of the coupling in the operational system. Since Cy47r1, the wave drag has been modified for extreme wind speeds to improve the wind-pressure relation for tropical cyclones, a result that also was confirmed in this report.

For the ensemble we find on average stronger TCs with the use of the SPPT scheme and a higher number of cyclonic features, something that needs further investigation to understand. Experiments with the SPP scheme shows promising impact on the TC prediction in terms of generating ensemble spread.

8. Discussion and future directions

The benefits of increasing the model resolution from 9 km to 4 km are evident in the Pmin and Vmax forecasts, and the wind-pressure relation. This is especially clear when the new moist physics package is used. However, as the model resolution is increased, the traditional boundary between model physics and dynamics becomes blurred. As the resolution approaches 4 km, it opens the question of whether parametrized deep convection should continue to be used, or whether to let the dynamics explicitly resolve the convection. The explicit convection experiment in this section suggests that parametrized convection is still beneficial for TC simulation at this resolution, as the simulations without the parametrization created too strong TCs. Related to the research on **slow convergence of the departure-point iterations**, the impact of a new IFS development that introduces a more accurate algorithm to find the departure points based on the 4th order Runge-Kutta Lobatto IIIA scheme will be investigated. Future model resolutions also bring up the question of using a non-hydrostatic dynamical core. However, this choice has implications for the coupling to the model physics that are currently being explored. In the longer term, the aim is to move the dynamical core to a finite-volume model, which is non-hydrostatic.

As part of the ongoing INCITE20 project, 1 km global simulations (Wedi et al., 2020) have also been conducted for two seasons covering Nov-Feb 2019 and Aug-Oct 2019. The TC statistics of these seasonal simulations are currently being evaluated in collaboration with Reading University and compared with the 9 km resolution equivalents. A special case has also been simulated at 1 km with higher temporal resolution information as a reference example (Hurricane Dorian in late August 2019). Although only a single simulation, it indicates that the track forecast was good when started from the interpolated 9 km operational analysis six days earlier, and the storm had an increased

intensity compared to the operational run, but it also shares features of the described 4 km simulations with deep convection simply switched off such as a cloud filling of the cyclone eye.

Regarding ocean processes, the ECMWF HRES forecast saw an improvement in the intensity forecast with the introduction of ocean coupling in 2018, a result that is confirmed by the no-coupling experiment for our special 37-day period in 2020. In 2020, a change to the ocean drag in the wave model led to improved forecasts of Vmax, and therefore a better pressure-wind relation. Further developments are under way to improve the heat and momentum exchange under extreme conditions, such as TCs. One needs to be aware of that these types of developments require a coupling to an ocean model to give realistic results in terms of energy exchanges.

The current ocean model has a resolution of around 0.25° globally, which puts it in the eddy permitting regime. To have a fully eddy resolving model would require a resolution of 1/12° or better, which is unrealistic to implement operationally until the next HPC after the Atos upgrade. However, to gain a better understanding of the ocean response of coupled models an informal collaboration with Météo-France, the Met Office and NRL has been established to compare different models with the observational ALAMO float data deployed by the US Naval Academy and Woods Hole Oceanographic Institute. By comparing very high-resolution limited area coupled models with global coarser resolution coupled models, we hope to be able to quantify the importance of horizontal atmospheric and oceanographic resolution for the ocean response in TC conditions.

We have investigated sensitivities to some of the key challenges found in Section 5, such as the slow propagation speed bias, wind pressure-relation and issues in capturing intensification and weakening rates. For the propagation speed, we found improved biases in the 4 km experiment with explicit deep convection. However, this experiment had at the same time the strongest intensity bias, and the worst tropical winds in general. But we also saw positive results from the experiment considering the Coriolis force due to vertical motion, a result that we will explore further.

The wind-pressure relation diagnostic helps to understand model deficiencies in simulating the extreme winds under TC conditions. The change in ocean drag provided significant improvements to the relation in Cy47r1. In the experiments with 4 km resolution, we found further improvement. Together with the result for the ENS control forecast presented in Section 5, this suggests a resolution dependence on capturing the relation. We also found an even stronger relation in the experiment with explicit convection, which could be related to convective wind structures. Finally, in all experiments, there is an underestimation of the relation for TCs in the range 25-40 m/s, something that needs further investigation.

All the steps discussed above are important to explore in order to deliver a km-scale ensemble forecast system with high quality for extreme events such as TCs. Although we have so far mainly investigated standard TC metrics such as position error, Pmin and Vmax, the modelling of fields that contribute to useful probabilistic forecasts of TC storm surge and flooding will become more important, as will be discussed in Section 9.

9. Forecast products

9.1. Current and future forecast products for tropical cyclones

Based on the tracker output (described in Section 3), ECMWF produces different graphical products. For each active TC reported at the initial forecast time by RSMC/TCWC, a combined product with the strike probability or track plume map up to ten days (or out to six days for 06/18UTC forecast runs introduced operationally in May 2021) together with Lagrangian metgrams for intensity (Pmin and Vmax) is computed. One example is shown in Figure 60 (with the track plume). Recently, the TC products from 06/18UTC forecast runs were introduced and can be accessed in the web open-charts or via dissemination together with 00/12UTC forecasts.

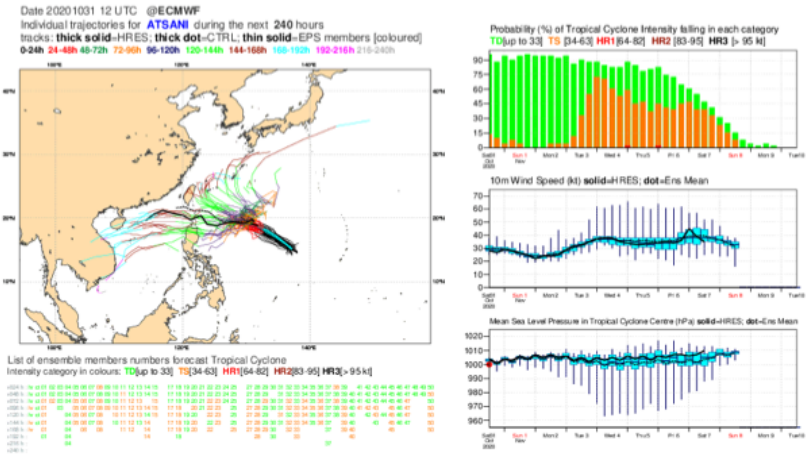


Figure 60: Lagrangian metgram for TC Atsani from 31 October 12UTC 2020.

ECMWF is also producing TC activity maps that include genesis (probability maps) both for medium-range forecasts (with a 48-hour time window) and extended-range forecasts (with 7-day accumulation period).

For the extended-range and seasonal forecasts, statistics of basin-wide weekly/seasonal Tropical Storm frequency and ACE (up to 30 days, only Mon and Thu) are also produced.

The ECMWF TC tracking and products are today restricted to the WMO official TC basins. This means that the system does not produce tracks for occasional cyclones in the southern Atlantic, or for "Medicanes" in the Mediterranean. For the Medicanes, the products from the extra-tropical cyclone database (CDB, Section 3.3) can be used, but it could also be considered to extend the TC infrastructure to the Mediterranean.

In the 12-24 h or so spanning an extra-tropical (ET) transition event, Lagrangian products from both the tropical and extratropical cyclone trackers (CDB) are normally available (see Section 3). If users want to continue seeing plots in a similar (tropical) format, they should continue with the TC tracker

output, but must recognise that at some point after ET it will look like the cyclone has decayed, when in fact it may have intensified a lot, as an extratropical feature. If, conversely, users want to see the full life cycle and try to identify a nominal "time of ET" they are recommended to use the CDB plots to see when the TC, denoted as a barotropic low (black dot), changes to a frontal wave cyclone (orange dot). That feature should then continue on as a trackable feature right through its extratropical phase, encompassing any re-intensification (likely reverting to barotropic low form) and also subsequent decay. The main CDB limitation is that it currently only covers part of the tropical north Atlantic, so there are many TCs for which it cannot be used.

As discussed in previous sections, the TC size based on the wind radii forecasts of the 34, 50 and 64 kt became available to users in July 2020. One example of a possible future product for ensembles was shown in Figure 6, and one example following the TC in HRES is shown for hurricane Dorian (2019) in Figure 61. The chart displays the 34 kt wind radii forecast of HRES every 12 h. The product is available from HRES and ENS forecasts in BUFR for dissemination, but not yet in the chart catalogue.

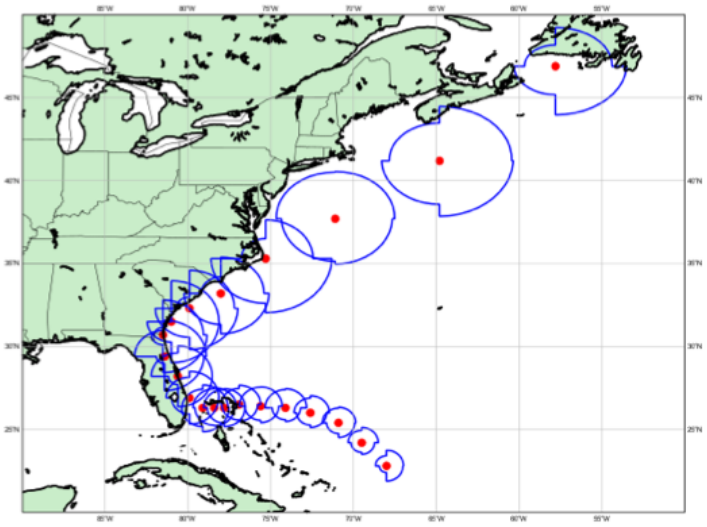


Figure 61: HRES wind radii forecast for the 34 kt wind threshold out to 240 h, initialised at on 30 August 00UTC 2019. The red dots indicate the predicted centre of hurricane Dorian at 12 h intervals.

While the TC products only comprise the position and intensity (core pressure, maximum surface wind speed and wind radii) no hazard products associated with TCs are currently available. Work is planned to explore ways to better use forecasts that will help users assess TC hazards. Figure 62 shows an example of the probability of 24 h maximum wind gusts above 20 m/s within a 250 km radius, centred on the position forecast of TC Tauktae (2021) as it progressed northwards, parallel to the west coast of India before landfall near Diu in the state of Gujarat. This highlights the coastal

regions exposed to the impact of cyclone TC Tauktae. Rainfall is also a significant TC-related hazard (discussed further in Section 9), and future forecast products may include charts specifically for TC rainfall.

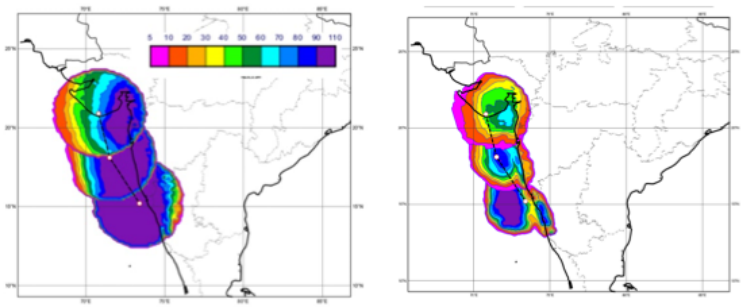


Figure 62: Probability (%) of 24 h maximum wind gust exceeding 20 m/s (above Tropical Storm strength threshold – 17 m/s) centred at 0000 UTC on 16, 17 and 18 May 2021, based on the forecast initiated at 1200 UTC on 15 May (left) and probability of total precipitation exceeding 100 mm/24 h (right) valid for the same dates. Symbols connected by a dashed line correspond to the HRES position forecast (close to the ENS mean track) valid at the same dates. Probability circles are produced for a radius of 300 km.

9.2. Future clustering of tropical cyclone tracks

Despite the significant progress made in position forecasts of TCs in recent years as a result of better NWP models, cases of large uncertainty regarding the track still occur (e.g., Magnusson et al., 2019b), especially when the TC passes close to a bifurcation point in the steering flow. Figure 60 shows such a case for TC Atsani (2020). While some ENS members depict the TC moving towards the Philippines, other members had a curvature to the north. In these situations, cluster analysis techniques can provide better guidance to forecasters under multi-modality track scenarios (Tsai and Elsberry, 2013; Kowaleski and Evans, 2020), and help to depict intensity uncertainty for example.

For extended-range TC prediction, clustering of TC tracks also has potential. Such a product would help address some of the issues linked to model biases (e.g., too rare TC landfall in extended-range and seasonal forecasts because of a too early TC track curvature) by using the dynamical model to predict the probability of a given cluster and a statistical method to predict its impact (e.g., probability of landfall). This dynamical-statistical method would be similar to the use of weather regimes in the Euro-Atlantic sector to predict the probability of extreme events in Europe.

Camargo et al. (2021) applied a clustering technique, described in Gaffney et al. (2007), to observed TC tracks and the ECMWF extended-range reforecasts produced in 2018 over the North Atlantic. This analysis indicates that there are clear spatial differences; the ECMWF model having an additional cluster with recurving tracks close to the African coast, with characteristics that do not correspond to observations. In addition, the relative population of the clusters is not realistic in the ECMWF extended-range forecasts, with, for instance, too few TCs in the cluster containing Caribbean TCs. This study also showed that the clusters have different levels of predictability since

they are differently modulated by sources of predictability such as the MJO, ENSO, AMM (Atlantic Meridional mode) and the NAO.

9.3. Discussion

Based on TC tracks, ECMWF provides a set of products freely available on the ECMWF website for different time-ranges. However, there are a lot of other products related to tropical cyclones that are currently not produced by ECMWF. As the tracks are also provided to external users, via GTS or FTP, there are many other sites that display ECMWF TC forecasts, either separate or as a part of a multi-model ensemble. The open question here is how much resource ECMWF should spend on increasing the graphical product catalogue for TCs.

10. Applications: Impact Forecasting

10.1. Introduction

Often, the first thing that comes to mind when considering TC impacts is the winds, and this has typically been the focus of intensity forecasting. However, the most dangerous impacts of TCs are often those related to water, including various types of flood hazard: flash flooding, river flooding, and flooding from storm surge and extreme ocean waves. These water-related hazards require additional modelling capabilities to provide forecasts of storm surge and rainfall-induced inundation. As discussed in Section 4, traditionally the focus in terms of TC predictability and predictive skill has been on track prediction, and is increasingly moving towards structural characteristics and intensity as model resolution has increased and it has become possible to capture mesoscale features. Increasingly, emphasis is also moving towards hazard- and impact-based forecasting, aiming to narrow the gap between science and decision-making for early action based on forecast information. For TCs, this means incorporating forecasts of not only track and intensity, but also wind fields, precipitation and flooding, and another step further, risk information such as populations and infrastructure exposed to these hazards.

The downstream hazard models required for predicting water-related TC hazards rely on accurate forecasts of rainfall and other meteorological variables from numerical weather prediction models and highlight the importance of accurate rainfall prediction for TCs. It is known that rainfall amounts are not directly related to the intensity of TCs, but the translation speed and size of the TC alongside topography and geography of the landfall region are instead key. Therefore, predicting TC rainfall relies on several factors: TC track, intensity, size, structure, and interactions with land and the wider atmospheric circulation (Titley et al., 2021). Furthermore, flood severity from TCs depends on additional factors including the TC duration, total precipitation, and catchment characteristics such as antecedent soil moisture and orography height and gradient (Titley et al., 2021). An important avenue for future work is to assess the predictability and predictive skill of TC precipitation and flood forecasts, in order to fully understand whether forecast skill is enough for accurate and useful impact-based forecasting.

Current capabilities, ongoing work and future directions for hazard forecasting and impact modelling in the context of TCs are discussed in the following subsections, and Section 9.5 provides examples and case studies of the use of ECMWF forecasts for such applications and decision-making.

10.2. River Flooding

One aspect of hazard forecasting for water-related TC impacts is forecasting river flow, including increased river flow and flooding from TCs. As part of the Copernicus Emergency Management Service (CEMS), ECMWF produces global flood forecasts, using a hydrological model forced by the ECMWF ensemble forecasts (The Global Flood Awareness System, GloFAS). GloFAS v3.1 (Copernicus, 2021; www.globalfloods.eu), implemented on 26 May 2021, uses a hydrological modelling system based on the open source global LISFLOOD model (Van der Knijff et al., 2010). LISFLOOD is a spatially distributed rainfall-runoff-routing model, developed at the European Commission's Joint Research Centre (JRC). ECMWF's 00UTC ensemble forecasts (for precipitation, temperature, dew point temperature, 10-metre u and v wind components and surface net solar and thermal radiations) are used to drive the LISFLOOD model, which calculates a complete water balance and produces forecasts of runoff at each grid cell (0.1° resolution). It then routes this runoff through the river network, producing forecasts of river flow. The initial conditions for GloFAS are primarily taken from the GloFAS-ERA5 river flow reanalysis, produced using the GloFAS modelling chain forced with ERA5 data. Harrigan et al. (2020) provide an overview of GloFAS-ERA5 production and evaluation, based on the GloFAS v2.1 modelling chain, and the GloFAS Wiki (Copernicus, 2021) provides details of the different GloFAS model versions.

Each new GloFAS forecast is compared against flood thresholds, calculated from the GloFAS-ERA5 reanalysis for various return periods, to provide the probability of exceeding various flood severity thresholds, out to 30 days ahead. These probabilities are provided through different products including maps, hydrographs and persistence plots, amongst various other hydrological and meteorological forecast layers. This approach based on exceedance probabilities limits the influence of systematic biases, which are expected in regions where the model remains uncalibrated. At present, GloFAS does not provide forecasts for coastal or pluvial flooding, which are also of significant concern during TC landfalls.

GloFAS forecasts have been used operationally for forecasting river flooding from TCs, in combination with a detailed flood inundation model in order to predict population exposure (see Section 9.5 below). Further research as part of an ongoing collaboration between ECMWF and the University of Reading, is evaluating the ability of GloFAS to forecast fluvial flooding from TCs by examining all parts of the forecast chain, including the track, intensity, precipitation and hydrological components. Recent research published in Titley et al. (2021) has identified key factors that influence the severity of TC flooding, such as translation speed and river catchment characteristics, and ongoing work aims to verify ensemble precipitation forecasts in TCs and highlight the main influences on predictability of fluvial flooding from TCs, in order to improve flood forecasts and their use for decision-making during TCs. Initial work based on a case study of Hurricane Iota (2020) indicated that in one river, the key driver of good flood forecasts was the accurate prediction of the broader onshore atmospheric flow into the mountains and was less sensitive to the prediction of landfall location. In another river, successful flood forecasts relied on the accurate prediction of landfall location and of heavy rainfall close to the TC centre, which is less predictable at longer lead times (Figure 63).

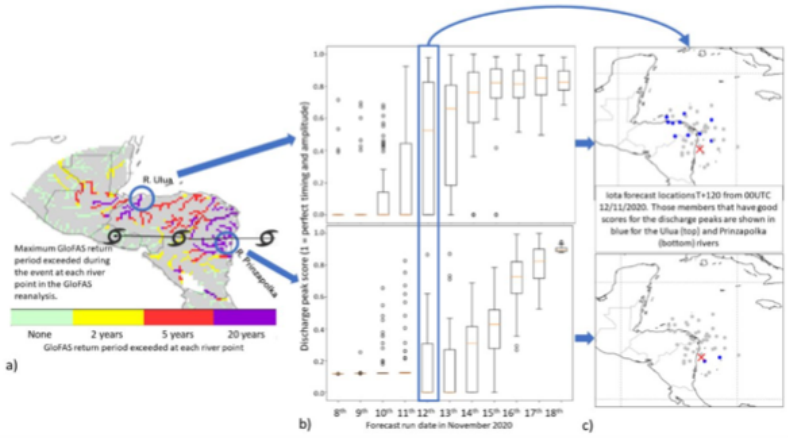


Figure 63: Initial analysis of the GloFAS forecast predictability for Hurricane Iota: a) the GloFAS return period exceeded at each river point in the Iota flood event, using the GloFAS-ERA5 reanalysis; b) box-whisker plot showing the range of discharge peak scores across the ensemble for two river points (circled in blue in (a)) for all forecast runs in the lead up to the flood event; c) Iota track forecast centre locations at T+120 from 0000 UTC 12th November 2020, with the ensemble members that displayed good forecast skill highlighted in blue.

As part of a wider evaluation and discussion of flood forecast bulletins for Cyclones Idai and Kenneth in Mozambique in 2019 (discussed further in Section 9.5), Emerton et al. (2020) evaluated the chain of forecasts from TC tracks and rainfall to river flow and flood inundation forecasts. The location of the storm is key for both precipitation and flood forecasts, and it is important to consider that the track forecasts indicate only the centre of the TC, but winds and rain associated with the storm extend much further. This was a consideration when these TCs made landfall, as in both cases, track forecasts indicated that the storms would continue to move further inland before dissipating. However, both Idai and Kenneth stalled after making landfall, resulting in sustained periods of heavy rainfall over the same region rather than smaller rainfall amounts spread over a larger region. This stalling was only picked up ~one day ahead and resulted in uncertainty in the rainfall and flood forecasts, in terms of the probability of severe flooding and the rivers which were likely to flood. This is seen in Figure 64(b), which shows under-estimation of rainfall at two days ahead (red shading) nearer the coast where Idai stalled (and over the Mozambique Channel), and an over-estimation further west due to track forecasts indicating the storm would move further inland. This had implications for decision-making based on the associated flood forecasts, and again highlights the importance of more evaluation and diagnostics accounting for various aspects of TC forecasts including hazards beyond the track and intensity.

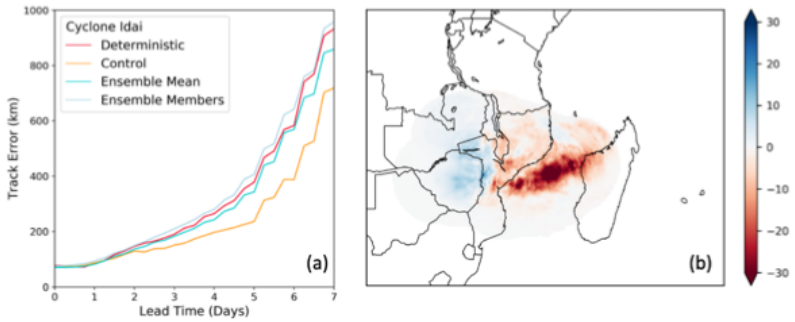


Figure 64: (a) Track location errors (km) with lead time for Cyclone Idai, Mozambique March 2019 and (b) TC-related precipitation errors (mm/day) for all 2-day ensemble mean precipitation forecasts throughout the lifecycle of Cyclone Idai, where red indicates an under-estimation of rainfall, and blue an over-estimation, compared to GPM IMERG satellite rainfall data. Figure adapted from Emerton et al. (2020).

10.3. Inundation Forecasting/Impact Modelling

Beyond forecasts of river flow, in order to move towards hazard and impact forecasting, it is important to provide information for decision-making purposes on the inundation area expected from the predicted river flows, and on the associated risks, such as population and critical infrastructure that may be exposed to flooding.

In addition to forecast products indicating predicted river flow and threshold exceedance probabilities, GloFAS has a risk mapping component showing the estimated maximum predicted inundation extent at a 1 km scale, and associated level of risk based on the exposed population (using the Global Human Settlement Layer dataset) and timing of maximum flood hazard. These products are available for river points with an upstream area greater than 5000km^2, where the flood hazard is predicted to exceed the 10-year return period threshold during the 30-day forecast horizon. Flood risk is estimated per global administrative region and provided through the 'rapid impact assessment' forecast layer of the GloFAS interface (Figure 65). Additional information on land cover types and critical infrastructure (using data from e.g., the European Space Agency and OpenStreetMap) is also provided for each region with an expected impact. The risk is assessed to be higher for events predicted with a shorter lead time, and with larger exposed populations. Further details are provided via the GloFAS Wiki (Copernicus, 2021), and limitations of this approach and the global datasets used are discussed on the GloFAS website (Copernicus EMS, 2021). While these products are not specific to TCs, they provide added information at the global scale that is applicable for predicting river flooding impacts from landfalling TCs.

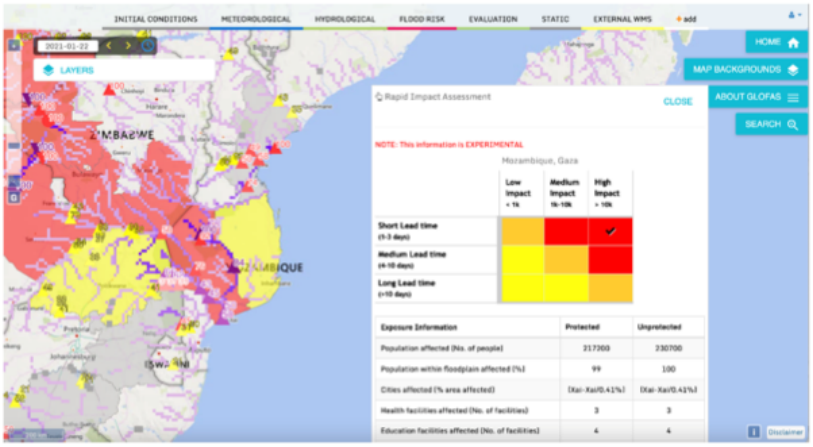


Figure 65: Screenshot of the GloFAS interface for the forecast from 22 January 2021, 1 day ahead of Cyclone Eloise's landfall in Mozambique. The map indicates rivers where flooding is expected to exceed the 20-year return period (purple shading, where darker colours indicate higher probabilities), and the rapid impact assessment layer (red and yellow shading). The box provides additional rapid impact assessment information, shown here for the Gaza administration region in Mozambique, indicating a high potential impact at a short lead time on the impact matrix, and estimates of the population affected.

In the future, it may be possible to incorporate flood inundation mapping at much higher resolutions e.g., through assimilation of satellite derived inundation extents, or use of river flow forecasts to drive a dynamical hydraulic model and provide inundation forecasts that directly relate to the probabilistic river flow forecasts. It will also be important to incorporate further information on critical infrastructure and vulnerability, which can be significantly impacted by both winds and flooding during TCs and post longer-term risks beyond the duration of the storm. A similar approach to that described here has been used operationally to forecast flood inundation and population exposure from TCs, in collaboration with the University of Bristol, for early humanitarian action and response. This is discussed in section 9.5.

10.4. Storm Surge Forecasting

With rising sea levels and increasing populations in coastal regions (Neumann et al., 2015), flooding from storm surge, which can be destructive and dangerous, is a key consideration for TC hazard and impact forecasting. The inundation risk from storm surge is also affected by other factors such as astronomical tides, waves, river flow and precipitation (Kohno et al., 2018). As discussed in Section 5, the ECMWF forecasts include an ocean wave model. ECMWF does not produce any storm surge forecasts, but other institutions run storm surge models using forcing from ECMWF meteorological and ocean forecasts. There are also steps to build full inundation model systems based on forcing from ECMWF forecasts (Zhou et al., 2020). This is important for risk communication, in order to convert

forecasts of storm surge into estimates of total sea level and inundation, but it requires detailed information on land elevation which is not always available (Kohno et al., 2018).

There are two distinct approaches to the use of atmospheric forcing for storm surge modelling. The first is to use only the forecasts of TC position and create an artificial wind field around the TC, and the second is to use more realistic wind fields directly from the forecasts.

The first approach uses the track forecasts for TCs and derives wind fields based on parametric models, for example using empirical pressure profiles such as Holland (1980) or Fujita (1952) to define the sea level pressure in the region of the TC. Gradient wind relations can then be used to estimate the wind fields, and adding the forward motion of the TC gives an asymmetric wind field. This approach has been used historically because NWP systems were unable to provide a realistic enough wind field to produce useful storm surge forecasts. While TC forecast skill for track and intensity has been continuously improving, it is often insufficient for accurate storm surge predictions (Kohno et al., 2018), and as such, these parametric modelling approaches are popular. A similar approach is being trialled by HR Wallingford using ECMWF's forecasts as part of a pilot study by the UK government for early humanitarian action, which is discussed in Section 9.5. This trial looks at moving from using a deterministic TC track forecast from the relevant RSMC, to additionally using ensemble TC track forecasts from ECMWF to provide probabilistic storm surge forecasts. For this application, the Holland (1980) profile model is used to determine the wind field, and the forward motion of the storm is added to give an asymmetric wind field, with the wind directions corrected based on the inflow angel. The TELEMAC2D model is then used to simulate storm surge. There is further potential to use ECMWF's new wind radii forecast products to provide a more accurate TC structure.

However, NWP systems moving to higher resolutions and providing more realistic simulations of TC winds provides new opportunities to use the wind fields directly from NWP models to force storm surge models. This second approach is particularly desirable in cases when TC structures vary from the standard structure used in parametric models. Moreover, it is known that the state of development of the wave field can have quite some impact on the actual storm surge intensity (Bertin et al., 2015). Therefore, ECMWF coupled forecasts can also provide consistent information on the sea state and surface fluxes for better storm surge forecasts.

Development of a storm surge forecasting system based on NWP, with global coverage (GLOSSIS, the Global Storm Surge Forecasting and Information System, (Deltares, 2021)), is also ongoing. This configuration uses meteorological forcing and the Global Tide and Surge Model (GTSM), alongside Delft-FEWS, to produce 10-day water level and storm surge forecasts for ~16,000 coastal segments around the globe. The model uses an unstructured grid, with coastal areas represented at ~5 km resolution and open oceans at 50 km resolution, and provides return period exceedances based on return periods derived from reanalysis data. GLOSSIS currently makes use of NOAA's GFS forecasts for the meteorological forcing, but also has functionality to use ECMWF's forecasts alongside track forecasts from the JTWC.

ECMWF has received a request by the Norwegian met service to explore the feasibility of using NEMO capabilities to provide storm surge information derived from the ensemble runs. This activity is currently at a very early stage but might prove useful in deriving storm surge early warnings at a

global scale. Additionally, the JRC has been developing a new unstructured mesh model to simulate storm surge globally, forced with ECMWF's operational HRES forecasts and ERA5 for testing. While this new model is currently under development, in future it could be linked with GloFAS to provide information on coastal flooding, which, as mentioned in Section 9.2, is not currently provided by GloFAS but would be further beneficial for forecasting flood impacts from TCs.

10.5. Examples of forecast uptake for decision-making

ECMWF's medium-range TC forecast products are used by national meteorological services and the RSMCs around the globe for operational TC forecasting, warning and decision-making, alongside seasonal forecasts of TC activity in each ocean basin. As mentioned in Section 9.2, forecasts of rainfall and flooding from ECMWF and GloFAS are also used for hazard and impact forecasting ahead of tropical cyclones. Here we provide some recent examples of the combined use of ECMWF TC and associated hazard forecasts for decision-making purposes.

10.5.1. Forecast-based Financing using GloFAS flood forecasts

GloFAS forecasts have been used to take early action ahead of flood events since 2015, when humanitarian action was triggered by a GloFAS flood forecast for the first time, ahead of flooding in Uganda, allowing aid to be distributed before flooding began. Typically, humanitarian aid reaches communities after a disaster has occurred, but the forecast-based financing programme (IFRC, 2021) enables access to humanitarian funding based on forecast information and risk analysis, using specific forecast thresholds that trigger the release of pre-agreed financial resources set out in an Early Action Protocol (EAP). In some regions, it is possible to trigger early humanitarian action for flooding from tropical cyclones based on forecasts. Research into the use of GloFAS forecasts for early humanitarian action is carried out at the University of Reading through the FATHUM project, in collaboration with CEMS and ECMWF.

10.5.2. Emergency flood bulletins for tropical cyclones

Researchers at the Universities of Reading and Bristol in the UK collaborated with ECMWF in 2019 to assist in providing flood forecast information for TCs Idai and Kenneth, which made landfall in Mozambique with devastating impacts. After TC Idai made landfall in March 2019, the President of Mozambique declared a state of emergency and requested international assistance. The UK government's Foreign, Commonwealth and Development Office (FCDO, previously the Department for International Development, DFID) tasked a team of scientists from the two Universities to produce real-time flood forecast bulletins to support humanitarian decision-making during the flooding that followed Idai's landfall. Using ECMWF TC track and rainfall forecasts and GloFAS flood forecasts, with the support of ECMWF, the Universities produced daily bulletins detailing the expected location and timing of landfall, alongside the rivers most at risk of flooding and the predicted timing of the flood peak and recession. This was based on the forecasts and methods described in Section 9.3. ECMWF also provided GloFAS river flow forecast data, which was used to produce inundation and population exposure information using a model framework similar to that described in Section 9.3, based on a hydrodynamic model with the LISFLOOD-FP code, and global scale gridded population data from the High-Resolution Settlement Layer (HSRL) dataset.

The objective of the bulletins was to facilitate decision-making, such as the distribution of resources ahead of flooding, and increase understanding of the situation and the areas most at risk of flooding and when, to be used alongside forecast and warning information from the mandated regional and national authorities. The flood bulletins, modelling framework, and an evaluation of the forecasts and the bulletins themselves, are described in detail in Emerton et al. (2020).

Six weeks after TC Idai struck Mozambique, the UN Office for the Coordination of Humanitarian Affairs (UN OCHA) and the UK government requested re-activation of these emergency flood bulletins ahead of TC Kenneth, which was forecast to impact northern Mozambique. This allowed forecast information to be used for humanitarian decision-making ahead of the TC and consequent flooding, and further led to the introduction of a pilot project by FCDO to operationalise the production of emergency flood bulletins by scientists at the Universities of Reading and Bristol, and HR Wallingford, with the support of ECMWF. This pilot project was activated a number of times during 2020 and 2021, including for TCs Iota, Amphan and Eloise. ECMWF provided forecasts, data, expertise and training in relation to this pilot project. This pilot project also allowed for the inclusion of storm surge forecasting by HR Wallingford, based on official track forecasts from the RSMC in La Réunion. As discussed in Section 9.3, this storm surge forecasting is being tested using ECMWF TC track forecasts, with the potential to further include the new wind radii products. The inclusion of storm surge information allows a more complete overview of the hazards associated with TCs, from both riverine and coastal flooding. At present, the outcomes of the pilot project are being evaluated in consideration of the next steps for the production of such bulletins.

10.5.3. The ARISTOTLE Consortium

ECMWF has also provided emergency reports for six TCs that included flood impacts since October 2020, using GloFAS flood forecasts and a range of TC forecast products for track and winds, including from ECMWF and the Met Office. These reports were produced as part of the ARISTOTLE (All Risk Integrated System TOwards Transboundary hoListic Early-warning) consortium, of which ECMWF is a partner alongside 17 other institutions. ARISTOTLE is a multi-hazard partnership covering expertise across multiple hazard groups (severe weather, flooding, forest fires, volcanoes, earthquakes and tsunamis). The institutions in each hazard group work in rotation to provide routine monitoring reports (three times per week) and emergency reports (within three hours) when activated by the European Commission's Emergency Response Coordination Centre (ERCC). ARISTOTLE is funded by the EU Directorate General for Humanitarian Aid and Civil Protection (DG ECHO) and is set to continue for the next five years.

10.6. Summary

Current capabilities, ongoing work and future directions for TC hazard forecasting and impact modelling have been discussed in this section. Often, the focus for TC forecasting is on the prediction of TC track and intensity, which are key in terms of decision-making and in terms of the accuracy of other forecast products and models. However, water-related hazards due to intense and prolonged rainfall, and flooding including flash flood, river flood and coastal flood hazards, can be the most dangerous aspects of TCs. It is therefore important to move towards provision of forecast information that can be used for these hazards. Hazard forecasting and impact modelling often requires the use of

additional models such as hydrological, hydraulic and storm surge models, typically driven with the output of NWP models such as ECMWF's IFS. In this section, we provided an overview of the forecasting capabilities for river flow and river flood forecasting, and inundation modelling and exposure estimates, through the Copernicus Emergency Management Service (CEMS) Global Flood Awareness System (GloFAS), in the context of TCs. ECMWF does not currently have capabilities in storm surge forecasting, but other organisations make use of ECMWF forecasts to drive storm surge models. Some recent examples of the application of ECMWF forecasts for decision-making were discussed further, with a focus on hazard and impact forecasting for TC flooding. One aim of the WMO/WWRP project HIWeather is to evaluate the full forecast production chain for extreme events, including TCs, and ECMWF is participating in this work.

In order to provide useful input to downstream models, ECMWF needs to make sure that both the forecast quality and the forecast output is accurate. For the forecast quality, more focus on the precipitation and wind extent around TCs is needed in the evaluation, and model development is necessary. Regarding the model output, close communication with forecast users is necessary to make sure that the model output frequency, parameters and products meet their needs.

11. Concluding remarks and avenues for improvement

11.1. Current progress and challenges

The performance of forecasts for TCs needs to be viewed from the perspective of the **overall performance of the forecasting system**. The performance of tropical winds and the extra-tropical waveguide will have a large impact on TC propagation and the humidity over the tropical oceans. However, there are some aspects that are more germane to TCs than other systems. Expanding on and updating two recent articles on ECMWF activities (a report on data assimilation by Bonavita et al., 2017 and an overview article by Magnusson et al., 2019), the aspects of the overall forecasting system most relevant to TCs have been presented in this report and are summarised below.

Observations are used for multiple purposes, including the "Best Track" estimation of TC position and structure (provided by RSMCs) which is used for verification; case study diagnostics; and of most importance to ECMWF's forecasts, data assimilation (summarised further below). For TCs, there is an especially large reliance on satellite data. Additionally, specially tasked aircraft are sometimes deployed, mostly in the Atlantic basin, to provide data on the TC and its environment. Ongoing challenges with observational data include the timely processing of incoming data, quality control schemes, handling faulty observations, managing observation operators and specifying their errors, prescribing errors and biases of the observations (which may vary as a platform ages), and accounting for correlated observation errors. Moreover, old observation platforms can cease to exist, sometimes suddenly, and new observation platforms are introduced each year.

Data assimilation at ECMWF involves the ingestion of over 800 million observations each day, of which more than 60 million are processed by the data assimilation. Given that forecasts of TC **position** are determined by the steering flow over a large region, several million observations have an influence on these forecasts. It is therefore difficult to quantify the impact from single observation systems on forecasts of TC position. For short-range forecasts of TC intensity (Pmin or Vmax), an

accurate initial representation of the conditions within and near the TC is especially necessary. However, the hostile environment within a TC and the cloudy conditions around the TC substantially limit the coverage of both conventional and satellite data. Here, the impact of single observation systems might be more easily detected. From the special observing system experiments conducted for the August-September 2020 period, improvements to the short-range intensity forecast errors were evident due to the assimilation of all-sky microwave observations, COSMIC2 radio occultation, and surface wind scatterometer observations.

Several recent improvements to the **data assimilation methodology** at ECMWF have provided benefits for TC analyses and forecasts. In the vicinity of TCs, data assimilation is challenging due to the sharp gradient and lack of observations. One recent improvement has been to increase the **resolution of the Ensemble of Data Assimilations** (EDA), which improved the background error statistics for TCs (Holm et al, 2015). However, this also led to larger weights for observations close to the TC. As the drift of dropsondes was not accounted for, very large observation errors occurred that occasionally had a severe effect on the analysis. This issue was mitigated with **adaptive first guess quality control.** Also, since the introduction of the BUFR format for dropsondes has made it possible to report the position for each observation, the position was updated during the ascent. Even if it is difficult to measure the impact on the overall skill, these improvements are expected to avoid damaging situations. A remaining danger for assimilation of data within TCs is occasionally **faulty observations**, from surface synoptic observations (SYNOP), buoys, or incorrect positions of ships or land stations. On a few occasions in recent years, the assimilation of these faulty observations has had a severe impact on TC initial conditions.

Data assimilation schemes in the ECMWF system and other global NWP systems all possess substantial **limitations in TCs**. These limitations include the resolution of the data assimilation scheme, some components (such as tangent linear models) that assume linear dynamics, and background error covariances that may not reflect the main dynamic and thermodynamic balances in TCs. This is further compounded by the relative paucity of observations within the TC. These limitations lead to several consequences. For an intensifying TC, especially one that is small and/or intensifying rapidly, small-scale processes are especially important. The successive analysis cycles are often found to be **unable to keep up with the actual intensification of a TC**. As a result, the short-range forecast that provides the background field for assimilation is too weak, and the subsequent analysis TC structure (and intensity) drifts even further behind the actual TC structure (and intensity). These difficulties are not so severe for large, relatively symmetric TCs with eyewalls that are resolvable. Some similar challenges exist in weakening TCs, where the weakening rate in successive analyses is not as sharp as the actual weakening rate.

Challenges also exist in the **use of observations** within the TC. In the assimilation experiments, we have not yet investigated specifically how observations served to correct the analysis structure in the TC. Equally importantly, we would need to investigate those observations that did not pass the quality control. Given the very large discrepancies between observed values and those in the background field, especially for an intense hurricane, several high-quality observations may fail the quality control process or not be weighted substantially in the assimilation. However, this is likely a necessary practice to prevent overly drastic changes to the analysis fields, which could cause more harm than if these extreme observations were excluded.

To help provide initial conditions of TCs that are more consistent with forecasters' estimates of the (working) Best Track, other NWP centres directly **assimilate Best Track estimates**. These estimates include uncertainties, and as they are subjective, different practices can lead to different error characteristics. Following the method at the Met Office (Heming, 2016), new assimilation experiments were conducted here to use the Best Track estimation of central pressure as a regular surface observation. From these experiments, no statistically significant results were consistently evident, although improvements did occur for a small sample of strong TCs in the north-west Pacific basin.

Recent **model developments** have led to the improvements of TC forecasts. The introduction of **ocean coupling** has improved intensity forecasts, via avoiding over-deepening of TCs. The introduction of a **cap on the surface drag coefficient** during very high wind speed regimes has helped improve the forecast of the maximum wind speed, and the wind-pressure relation. With these recent improvements, it is also time to raise the bar and evaluate the **structure of the surface wind speed**. Since Cy47r1, the tracker output includes the radius of different wind speed thresholds (including tropical storm force and hurricane force).

In the 2020 operational configuration of HRES with 9 km resolution (Cy47r1), the bias of the central pressure (Pmin) appeared close to zero. However, there were strong conditional biases where initially weak TCs developed a bias toward too shallow TCs (underpredicting intensification rates), and initially strong TCs developed a bias towards too deep TCs (underpredicting weakening rates). As expected, the bias in the maximum wind speed (Vmax) was consistently low, regardless of the initial intensity of the TC. We note that Vmax can be dominated by small-scale (O(1 km)) processes.

The question of whether an increase in the **model resolution** improves TC intensity forecasts has been widely investigated in the NWP community. In this study, experiments with 4 km resolution have been tested over the special 37-day period in August-September 2020. Using the 2020 operational physics package, the 4 km experiments provided TCs that were too deep on average. However, when the 9 km and 4 km experiments were repeated with the **new physics package** intended for Cy47r3, the intensity was on average reduced, leading to an increased bias at 9 km resolution but a reduced bias for 4 km. The intriguing result was that the mean absolute error improved for both resolutions, by reducing the conditional biases discussed above. We should therefore be ready for an increase in intensity bias with Cy47r3 in the knowledge that it is probably due to unresolved structures in the 9 km resolution model and compensation of errors. However, we still believe that there are several compensating errors in the model that need to be further explored.

The wind-pressure relation, which has been a long-standing problem in ECMWF forecasts and other global models, was improved in Cy47r1 as mentioned above. In the 4 km experiment the relation is further improved, demonstrating that the model resolution is important to predict not only the maximum wind speed (Vmax) and central pressure (Pmin), but also the relationship between the two. However, there is still an underestimation for wind speeds of 30-40 m/s, which will need to be further investigated in the context of modelling the drag from ocean waves.

A long-standing **systematic error in ECMWF TC track forecasts** is a too slow propagation speed on average (Figure 32 in Haiden et al., 2021) and a tendency to curve too much to the right of the observed track in the tropics. This bias is difficult to target as it depends on the sample of test cases,

and it is relatively small in magnitude compared with the random component of the error. The same argument holds in the context of a visual inspection of weak, westward-moving TCs in the Atlantic basin that often drift slightly to the right (north) of the actual track, even at early forecast times. Although there are difficulties in finding statistical evidence, one can expect these biases to affect forecasts of not only TCs but extra-tropical transitions and downstream impacts in the mid-latitudes. In our experimentation, we found encouraging results by including the Coriolis effect due to vertical motion. This is expected from theoretical work (Liang and Chan, 2005), and seems to reduce the bias in the IFS but more experimentation is needed to confirm the results.

For **long-range predictions** of TC activity, the forecasts do not reproduce the pattern over the Atlantic. They overestimate the activity over the central Atlantic, and underestimate the number of TCs in the Gulf of Mexico. This is an obstacle to the goal of predicting the risk of landfalling TCs in the coming season. An ECMWF team that was recently tasked with investigating the seasonal forecasts found a trend of increased wind shear over the tropical Atlantic in the model that is not present in the reanalysis. This has led to an underestimation of the TC activity in recent years. This finding about the wind structure also opens the much wider question about the global warming response in the model and might also have implications for climate models.

Predicting TC **genesis** is important both in the medium- and extended-range. However, it is difficult to draw general conclusions about the performance, as the variability in multi-scale mechanisms and predictability between individual cases is very high. The predictability varies from more than a week in advance to not even capturing the TC at the formation time. In situations of higher predictability, there is often a relatively large-scale, coherent precursor disturbance. In situations of lower predictability, the precursor disturbance is often less defined, and there may be a greater dependence on bursts of convection on short time scales.

11.2. Avenues for future improvement

Based on evolving user needs, the challenges described in the individual sections and summarised above, and the continuous progress in TC forecasting at other NWP centres around the world, several recommendations for improvement are provided in this closing sub-section. Some of the items below are related to planned activities at ECMWF and some are to consider for future plans.

First, a **structured evaluation framework** for TCs that advances beyond the current framework is proposed, to monitor progress and identify further challenges:

- For **position**, evaluate errors for paired and homogeneous samples, including metrics for along-track and cross-track forecast biases, and propagation speed.
- For **intensity**, evaluate the minimum central pressure, the maximum surface wind, and the wind-pressure relation.
- For **surface wind structure**, continue the recent development of evaluation techniques for wind radii (34 kt, 50 kt, 64 kt), and explore the radius of maximum wind.
- For **other aspects of TC structure**, explore verification against infrared and other satellite images by using simulated images.
- Include **TC activity** in the routine verification, investigating both predictive skill and bias, and accounting for environmental variables (such as wind) that affect the detection threshold.

- Probabilistic verification of **TC genesis**, using related tools to those used for TC activity.
- For **impact related metrics**, investigate the possibility of rainfall verification.

New methods to **evaluate model structures** require investigation, given that the Best Track values only represent a small number of characteristics related to the surface pressure and wind. Examples include comparisons of simulated infrared images versus actual images, comparisons against dropsonde and other aircraft data, and comparisons against specialised satellite data such as SAR. For the evaluation of both future assimilation and model improvements, TCs put special constraints on the testing procedure since the resolution is expected to be especially important.

For **impact forecasting**, accurate modelling of the structure of the wind and precipitation fields is necessary. An additional requirement in **storm surge** forecasting is for accurate predictions of the timing of landfall and the corresponding wind fields, given that the storm surge is superposed on the astronomical tide. Ultimately, for **flood** forecasts based on both storm surge and TC precipitation, the different model components will form a basis for inundation models that aim to predict water levels down to street resolution. However, this will not only demand high-quality TC forecasts, but also the processing of ensemble data at high spatial and temporal resolutions.

For the **data assimilation**, several aspects were not tested here, due to the substantial effort and research required. It includes **modifications to the data assimilation** scheme itself, such as the resolution of the loops in the 4D-Var algorithm, error covariance matrices that account for the dynamics of the TC structures and the observation quality control. We believe these aspects could have a relatively strong impact on TC analyses and forecasts, and should therefore be evaluated with special attention.

For **observation usage in the data assimilation**, the following priorities have been identified:

- Increase the resilience to faulty surface observations in the tropics.
    - The avoidance of such cases will be targeted via the continued development of the observation monitoring and quality control, with help from machine learning algorithms.
- Increase the use of observations close to the TC.
    - Examples include adaptive thinning of satellite observations to capture finer-scale structures in TCs, a continued move to all-sky assimilation of satellite radiances (e.g. infrared channels) and exploring the increased use of atmospheric motion vectors near the TC.
- Explore new satellite products.
    - Examples with marine wind information at high windspeed include passive L-band (SMOS, SMAP, CIMR), building on work carried out in the SMOS collaboration, and Synthetic Aperture Radar (SAR), building on the work already done at Météo-France.
- Continue the initial exploration of assimilation of Best Track.
    - The work should be conducted with the awareness that finding the “correct” way to use the data can be a time-consuming task.

- Establish contact with Met Office and other centres to exchange knowledge on the topic

Several other new and future satellite observation types and algorithms may enhance analyses and therefore forecasts of TCs. Examples include new LEO satellites, targeted assimilation of fine-scale structures now discernible in geostationary satellites, rapid-scan AMVs, future Lidars, ocean surface wind retrievals, small satellite constellations such as TROPICS and CYGNSS, and new remotely sensed and in situ ocean observations. The effective assimilation of all these observation types not only depends on their accuracy, but their availability on the GTS in real-time together with the provision of error characteristics and observation operators.

Further investigation is required on how reconnaissance and surveillance aircraft data are enhancing the analysis, and whether new methodologies would make better use of these specialised, high-quality observations. In addition to the in-flight data mentioned in the bullet point above, the large volume of aircraft data that are not presently assimilated, including airborne Doppler radar data, may substantially improve analysis structures of TCs. Future datasets to be considered for assimilation include Stepped Frequency Microwave Radiometer (SFMR) data at the ocean surface, and data from a variety of unmanned aircraft.

The ongoing **model development and experimentation** at ECMWF is expected to provide long-term improvements to TC forecasts. These include not only the resolution and physics, but also continued improvements to ocean coupling, wave modelling, and representation of the atmospheric boundary layer. Continued experimentation with nonhydrostatic modelling, a finite-volume dynamical core, and explicit convection, is strongly encouraged in the context of improving TC forecasting. Some priorities to address include:

- Improved modelling of the propagation speed
- Further investigation of the tendency to a right-hand track bias
- Understand the causes behind the too slow intensification and decay of TCs
- Understand the deficiencies in simulating the TC climatology for Atlantic sub-basins in seasonal forecasts

For **seasonal forecasting**, it is also important to further investigate the weak teleconnections from ENSO and local SST. It is also important to further understand the global warming trend in TCs in the model compared with reanalysis data, a topic that is of importance to the wider climate modelling community. All these aspects need to be worked on in collaboration with the model developers.

Ensemble perturbation methods require evaluation in the context of TC prediction. These include the EDA method, moist singular vectors that are targeted over TCs, and model perturbation methods. Results here indicate that quite often the best track falls outside the ensemble plume, something that can be related to the track biases mentioned above. It is therefore important to further assess the reliability of the ensemble track forecasts and how they can be improved in the future.

The ensemble prediction system benefits from all model and data assimilation improvements. However, in the current configuration the lower resolution for the ensemble affects the TC predictions, especially for intensity. The current plan is to bring the ensemble resolution to 9-11 km in 2022-2023. Figure 66 illustrates the impact of increased ensemble resolution to 9 km for TC Laura.

The prediction of the RI is much improved. Such resolution increase is also expected to improve the ensemble spread for the intensity.

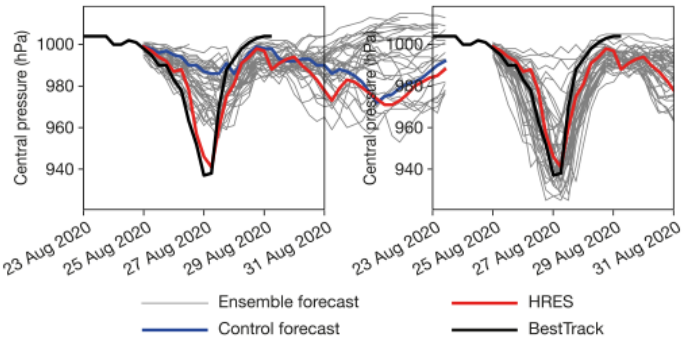


Figure 66: Forecasts for TC Laura initialised 25 August 00UTC of Pmin comprising the operational ensemble forecast, control forecast, high-resolution forecast (HRES) and Best Track data (left) and HRES and preliminary. Best Track data with the experimental ensemble with the same resolution as HRES (right).

For **usage of forecast output**, the following recommendations are provided:

- Increase collaborations across multiple centres on **TC tracking methods**
- Introduction of forecast products targeted towards the **impact** of the TCs
  - Rainfall and wind gusts in TCs are examples
  - The flooding impact from TCs needs to be validated
- Work together with the groups that use ECMWF forecasts for **storm surge modelling**
  - This is necessary to ensure the ECMWF forecasts are used optimally

Machine learning may help address some of the challenges described in this section. For example, a new project will explore machine learning to find drivers of predictability related to TC genesis, and also to be used as a forecasting aid.

All these points are important to address in collaboration with ECMWF Member States, other NWP centres and the wider research community. Examples include exploring high-resolution ocean modelling together with Météo-France and the Met Office. The Destination Earth programme would also provide opportunities to explore some of the listed aspects.

Although this report is extensive, there are inevitably parts that have not been covered. An example is the impact from aerosols on TCs. With the forecasting in Copernicus Atmospheric Monitoring Service (CAMS), this is an area that could be explored in the future. We have also not covered reanalysis aspects of TCs, something that is challenging especially for the pre-satellite era.

To conclude, substantial progress continues to be made in ECMWF's forecasts of TCs. As global user demands increase and diversify, and global and regional modelling of TCs at other centres continues to improve quickly, several avenues require investigation and improvement. These involve more

advanced evaluation and diagnostic methods, improvements to observational usage, and continuous advances in the data assimilation and modelling systems in several directions.

12. Acknowledgements

This report is a result of cross-departmental contributions from more than 40 scientists at ECMWF. We would also like to give a special acknowledgement to Helen Titley (Met Office) and Sylvie Malardel (Météo-France) for input to this report. Parts of the report were funded by the EUMETSAT fellowship programme. The one-year visit by Prof Sharan Majumdar was funded by the University of Miami, the Office of Naval Research, the National Science Foundation, and ECMWF. We also gratefully acknowledge the provision of COAMPS-TC model forecast data by James Doyle (Naval Research Laboratory, USA) and verification statistics by Michael Brennan (National Hurricane Center, USA), and helpful feedback from many colleagues around the world.

13. References

Agusti-Panareda, A., Thorncroft, C. D., Craig, G. C., & Gray, S. L. (2004). The extratropical transition of hurricane *Irene* (1999): A potential-vorticity perspective. Quarterly Journal of the Royal Meteorological Society, 130(598), 1047–1074. https://doi.org/10.1256/qj.02.140

Andersson, E., & Järvinen, H. (1999). Variational quality control. Quarterly Journal of the Royal Meteorological Society, 125(554), 697–722. https://doi.org/10.1002/qj.49712555416

Baker, A. J., Hodges, K. I., Schiemann, R. K. H., & Vidale, P. L. (2021). Historical variability and lifecycles of North Atlantic midlatitude cyclones originating in the tropics. Journal of Geophysical Research: Atmospheres, 126, e2020JD033924. https://doi.org/10.1029/2020JD033924

Barkmeijer, J., Buizza, R., Palmer, T.N., Puri, K. and Mahfouf, J.-F. (2001). Tropical singular vectors computed with linearized diabatic physics. Q.J.R. Meteorol. Soc., 127: 685-708. https://doi.org/10.1002/qj.49712757221

Bauer, P, Quintino, T, Wedi, N, Bonanni, A, Chrust, M, Deconinck, W, Diamantakis, M, Düben, P, English, S, Flemming, J, Gillies, P, Hadade, I, Hawkes, J, Hawkins, M, Iffrig, O, Kühnlein, C, Lange, M, Lean, P, Marsden, O, Müller, A, Saarinen, S, Sarmany, D, Sleigh, M, Smart, S, Smolarkiewicz, P, Thiemert, D, Tumolo, G, Weihrauch, C, Zanna, C, Maciel, P. (2020). The ECMWF Scalability Programme: Progress and Plans. ECMWF Tech Memo 857

Bechtold, P., Forbes, R., Sandu, I., Lang, STK., Ahlgrimm, M. (2020). A major moist physics upgrade for the IFS. ECMWF Newsletter 164. https://www.ecmwf.int/en/newsletter/164/meteorology/major-moist-physics-upgrade-ifs

Becker, T., Bechtold, P. & Sandu, I. (2021). Characteristics of convective precipitation over tropical Africa in storm-resolving global simulations. Q.J.R. Meteorol. Soc., Accepted. https://doi.org/10.1002/qj.4185

Beljaars, A. (2015). Time step dependence of wind errors in storm conditions, ECMWF RD memo RD16-041.

Bergman, D. L., Magnusson, L., Nilsson, J., Vitart, F. (2019). Seasonal forecasting of tropical cyclone landfall using ECMWF's System 4, Weather and Forecasting, 34 (5), 1239-125, https://doi.org/10.1175/WAF-D-18-0032.1

Bertin, X., Li, K., Roland, A., & Bidlot, J.-R. (2015). The contribution of short-waves in storm surges: Two case studies in the Bay of Biscay. *Continental Shelf Research*, *96*, 1–15. https://doi.org/10.1016/j.csr.2015.01.005

Bidlot, J.-R., Prates, F., Ribas R., Mueller-Quintino, A., Crepulja, M., Vitart, F. (2020). Enhancing tropical cyclone wind forecasts, ECMWF Newsletter 164. https://www.ecmwf.int/en/newsletter/164/meteorology/enhancing-tropical-cyclone-wind-forecasts

Biswas, M.K., D. Stark & L. Carson, (2018). GFDL Vortex Tracker Users's Guide V3.9a.

Blackwell, WJ, Braun, S, Bennartz, R, et al. (2018). An overview of the TROPICS NASA Earth Venture Mission. Q J R Meteorol Soc.144 (Suppl. 1): 16– 26. https://doi.org/10.1002/qj.3290

Blake, E. S., and D. A. Zelinsky, (2018). Hurricane Harvey (AL092017), National Hurricane Center Tropical Cyclone Report, https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf

Bloemendaal, N. et al. (2021). Adequately reflecting the severity of tropical cyclones using the new Tropical Cyclone Severity Scale, Environ. Res. Lett. 16 014048, https://iopscience.iop.org/article/10.1088/1748-9326/abd131/meta

Brammer, A., & Thorncroft, C. D. (2015). Variability and evolution of african easterly wave structures and their relationship with tropical cyclogenesis over the eastern atlantic. Monthly Weather Review, 143(12), 4975–4995. https://doi.org/10.1175/MWR-D-15-0106.1

Brut, A., Butet, A., Durand, P., Caniaux, G., & Planton, S. (2005). Air–sea exchanges in the equatorial area from the EQUALANT99 dataset: Bulk parametrizations of turbulent fluxes corrected for airflow distortion. Quarterly Journal of the Royal Meteorological Society, 131(610), 2497–2538. https://doi.org/10.1256/qj.03.185

Bonavita, M., Hólm, E., Isaksen, L., & Fisher, M. (2016). The evolution of the ECMWF hybrid data assimilation system. Quarterly Journal of the Royal Meteorological Society, 142(694), 287–303. https://doi.org/https://doi.org/10.1002/qj.2652

Bonavita, M., Dahoui, M., Lopez, P., Prates, F., Hólm, E., De Chiara, G., Geer, A., Isaksen, L., & Ingleby, B. (2017). On the initialization of tropical cyclones. ECMWF Tech Memo 810

Bormann, N., Lawrence, H., Farnan, J. (2019). Global observing system experiments in the ECMWF assimilation system. ECMWF Tech Memo 839

Brammer, A. and C. D. Thorncroft, 2015: Variability and evolution of African easterly wave structure and the relationship with tropical cyclogenesis over the eastern Atlantic. *Mon. Wea. Rev.*, **143**, 4975-4995.

Browne, P. A., de Rosnay, P., Zuo, H., Bennett, A., & Dawson, A. (2019). Weakly Coupled Ocean-Atmosphere Data Assimilation in the ECMWF NWP System. Remote Sensing, 11(234), 1–24. https://doi.org/10.3390/rs11030234

Buizza, R., Leutbecher, M. and Isaksen, L. (2008). Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 134, 2051–2066. URL: https://doi.org/10.1002/qj.346.

Camargo, S., F. Vitart, C.-Y. Lee and M. Tippett. (2021). Skill, predictability, and cluster analysis of Atlantic hurricanes in the ECMWF monthly forecasts. Submitted to Monthly Weather Review.

Cangialosi, J. P., and C. W. Landsea, 2016: An examination of model and official National Hurricane Center tropical cyclone size forecasts. *Wea. Forecasting*, **31**, 1293–1300

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. N. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. Wea. Forecasting, 35, 1913–1922

Chen, J. H., Lin, S. J., Magnusson, L., Bender, M., Chen, X., Zhou, L., Xiang, B. (2019). Advancements in Hurricane Prediction With NOAA's Next-Generation Forecast System, Geophysical Research Letters 46 (8), 4495-4501

Christophersen, H., J. Sippel, A. Aksoy, and N. L. Baker, 2021: Tropical Cyclone Data Assimilation. AGU Geophysical Monograph Series: "Earth's Climate and Weather: Dominant Variability and Disastrous Extremes". Under Review.

Cione, J. J. and Coauthors, 2020: Eye of the Storm: Observing Hurricanes with a Small Unmanned Aircraft System. *Bull. Amer. Meteor. Soc.*, **101**, E186-E205.

Cook, P. A., & Renfrew, I. A. (2015). Aircraft-based observations of air-sea turbulent fluxes around the British Isles: Observations of Air-Sea Fluxes. Quarterly Journal of the Royal Meteorological Society, 141(686), 139–152. https://doi.org/10.1002/qj.2345

Copernicus (2021), *Global Flood Awareness System - Copernicus Services - ECMWF Confluence Wiki*. Retrieved 23 June 2021, from https://confluence.ecmwf.int/display/COPSRV/Global+Flood+Awareness+System

Copernicus EMS. (2021). *GloFAS Rapid Risk Assessment*. Global Flood Awareness System; Copernicus. Retrieved 23 June 2021, from https://www.globalfloods.eu/technical-information/glofas-impact-forecasts

Cotton, J., Francis, P., Heming, J., Forsythe, M., Reul, N., & Donlon, C. (2018). Assimilation of SMOS L-band wind speeds: Impact on Met Office global NWP and tropical cyclone predictions. Quarterly Journal of the Royal Meteorological Society, 144(711), 614–629. https://doi.org/https://doi.org/10.1002/qj.3237

Coughlan de Perez, E., van den Hurk, B., van Aalst, M. K., Jongman, B., Klose, T., & Suarez, P. (2015). Forecast-based financing: An approach for catalyzing humanitarian action based on extreme weather and climate forecasts. Natural Hazards and Earth System Sciences, 15(4), 895–904. https://doi.org/https://doi.org/10.5194/nhess-15-895-2015

Davis, C. A. (2018). Resolving tropical cyclone intensity in models. Geophysical Research Letters, 45(4), 2082–2087. https://doi.org/10.1002/2017GL076966

De Chiara, G., English S., (2016). "SMOS Hurricane wind speed analysis", Report for ESA contract 4000101703/10/NL/FF/fk CCN5.

De Chiara, G, Isaksen, L., English, S. (2018). ECMWF support for EPS/ASCAT ocean wind assessment", Final Report Contract No. EUM/CO/15/4600001497/JF

Deltares (2021). Global storm surge information system(GLOSSIS). Retrieved 23 June 2021, from https://www.deltares.nl/en/projects/global-storm-surge-information-system-glossis/

DeMaria, M., Franklin, J. L., Onderlinde, M. J., & Kaplan, J. (2021). Operational forecasting of tropical cyclone rapid intensification at the national hurricane center. Atmosphere, 12(6), 683. https://doi.org/10.3390/atmos12060683

Diamantakis, M. and Magnusson, L. (2016). Sensitivity of the ECMWF model to Semi-Lagrangian departure point iterations, Monthly Weather Review, 144 (9), 3233-3250

Diamantakis, M. and F. Vana, (2021). A fast converging and concise algorithm for computing the departure points in semi-Lagrangian weather and climate models, submitted in Q. J. R. Meteor. Soc.

Duncan, D. I., Bormann, N., Geer, A. J., and Weston, P. (2021). Assimilation of AMSU-A in All-sky Conditions. 57, EUMETSAT/ECMWF Fellowship Programme Research Report.

Duong, Q.-P., Langlade, S., Payan, C., Husson, R., Mouche, A., & Malardel, S. (2021). C-band sar winds for tropical cyclone monitoring and forecast in the south-west indian ocean. Atmosphere, 12(5), 576. https://doi.org/10.3390/atmos12050576

Doyle, J.D., C.A. Reynolds, C. Amerault, J. Moskaitis, (2012). Adjoint sensitivity and predictability of tropical cyclogenesis. J. Atmos. Sci., 69, 3535-3557.

Doyle, J. D. and Coauthors, 2014: Tropical cyclone prediction using COAMPS-TC. *Oceanography*, **27**, 104–115

van den Dool, H., Becker, E., Chen, L.-C., & Zhang, Q. (2017). The probability anomaly correlation and calibration of probabilistic forecasts. Weather and Forecasting, 32(1), 199–206. https://doi.org/10.1175/WAF-D-16-0115.1

Dvorak, V. F., (1984). Tropical cyclone intensity analysis using satellite data. NOAA Tech. Rep. NESDIS 11, 47 pp.

Elless, T. J., & Torn, R. D. (2018). African easterly wave forecast verification and its relation to convective errors within the ecmwf ensemble prediction system. Weather and Forecasting, 33(2), 461–477. https://doi.org/10.1175/WAF-D-17-0130.1

Emerton, R., Cloke, H., Ficchi, A., Hawker, L., de Wit, S., Speight, L., Prudhomme, C., Rundell, P., West, R., Neal, J., Cuna, J., Harrigan, S., Titley, H., Magnusson, L., Pappenberger, F., Klingaman, N., & Stephens, E. (2020). Emergency flood bulletins for Cyclones Idai and Kenneth: A critical evaluation of the use of global flood forecasts for international humanitarian preparedness and response. International Journal of Disaster Risk Reduction, 50, 101811. https://doi.org/10.1016/j.ijdrr.2020.101811

Evans, C., et al. (2017). The Extratropical Transition of Tropical Cyclones. Part I: Cyclone Evolution and Direct Impacts, Monthly Weather Review, 145(11), 4317-4344. https://doi.org/10.1175/MWR-D-17-0027.1

Fujita, T. (1952). Pressure distribution within typhoon. Geophys. Mag., 23, 437–451.

Gaffney, S. J., A. W. Robertson, P. Smyth, S. J. Camargo, and M. Ghil, (2007). Probabilistic clustering of extratropical cyclones using regression mixture models. Clim. Dyn., 29, 423–440, 643 doi:10.1007/s00382-007-0235-z

Geer, A. J., Lonitz, K., Weston, P., Kazumori, M., Okamoto, K., Zhu, Y., Liu, E. H., Collard, A., Bell, W., Migliorini, S., Chambon, P., Fourrié, N., Kim, M., Köpken-Watts, C., & Schraff, C. (2018). All-sky satellite data assimilation at operational weather forecasting centres. Quarterly Journal of the Royal Meteorological Society, 144(713), 1191–1217. https://doi.org/10.1002/qj.3202

Geer, A. J., Migliorini, S., & Matricardi, M. (2019). All-sky assimilation of infrared radiances sensitive to mid- and upper-tropospheric moisture and cloud. Atmospheric Measurement Techniques, 12(9), 4903–4929. https://doi.org/10.5194/amt-12-4903-2019

Gall, R., Franklin, J., Marks, F., Rappaport, E. N., & Toepfer, F. (2013). The hurricane forecast improvement project. Bulletin of the American Meteorological Society, 94(3), 329–343. https://doi.org/10.1175/BAMS-D-12-00071.1

Good S, Fiedler E, Mao C, Martin MJ, Maycock A, Reid R, Roberts-Jones J, Searle T, Waters J, While J, Worsfold M. (2020). The Current Configuration of the OSTIA System for Operational Production of Foundation Sea Surface Temperature and Ice Concentration Analyses. Remote Sensing. 2020; 12(4):720. https://doi.org/10.3390/rs12040720

Gray, W. (1979). Hurricanes: Their formation, structure and likely role in the tropical circulation, Meteorology over the Tropical Oceans, D. B. Shaw, Ed., Royal Meteorological Society, 155– 218.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Ferranti, L., Prates, C., & Richardson, D. (2021). Evaluation of ECMWF forecasts, including the 2020 upgrade. ECMWF Tech Memo 880 https://doi.org/10.21957/6NJP8BYZ4

Hallerstig, M., Magnusson, L., Kolstad, E. W., & Mayer, S. (2021). How grid-spacing and convection representation affected the wind speed forecasts of four polar lows. Quarterly Journal of the Royal Meteorological Society, 147(734), 150–165. https://doi.org/10.1002/qj.3911

Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., &

Pappenberger, F. (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present [Preprint]. Hydrology and Soil Science – Hydrology. https://doi.org/10.5194/essd-2019-232

Heming, J. T. (2016). Met Office Unified Model Tropical Cyclone Performance Following Major Changes to the Initialization Scheme and a Model Upgrade. Weather and Forecasting 31.5, 1433-1449, https://doi.org/10.1175/WAF-D-16-0040.1.

Heming, J. T., Prates, F., Bender, M. A., Bowyer, R., Cangialosi, J., Caroff, P., Coleman, T., Doyle, J. D., Dube, A., Faure, G., Fraser, J., Howell, B. C., Igarashi, Y., McTaggart-Cowan, R., Mohapatra, M.,

Moskaitis, J. R., Murtha, J., Rivett, R., Sharma, M., … Xiao, Y. (2019). Review of recent progress in tropical cyclone track forecasting and expression of uncertainties. Tropical Cyclone Research and Review, 8(4), 181–218. https://doi.org/10.1016/j.tcrr.2020.01.001

Hewson, T. D., & Titley, H. A. (2010). Objective identification, typing and tracking of the complete life-cycles of cyclonic features at high spatial resolution: Objective identification, typing and tracking of the complete life-cycles of cyclonic features. Meteorological Applications, 17(3), 355–381. https://doi.org/10.1002/met.204

Hodges, K. I. (1995). Feature tracking on the unit sphere. Monthly Weather Review, 123(12), 3458–3465. https://doi.org/10.1175/1520-0493(1995)123<3458:FTOTUS>2.0.CO;2

Hodges, K. I. (1999). Adaptive constraints for feature tracking. Monthly Weather Review, 127(6), 1362–1373. https://doi.org/10.1175/1520-0493(1999)127<1362:ACFFT>2.0.CO;2

Hodges, K., Cobb, A., & Vidale, P. L. (2017). How well are tropical cyclones represented in reanalysis datasets? Journal of Climate, 30(14), 5243–5264. https://doi.org/10.1175/JCLI-D-16-0557.1

Hodges, K. I., Klingaman, N. P. (2019). Prediction Errors of Tropical Cyclones in the Western North Pacific in the Met Office Global Forecast Model, Weather and Forecasting, 34(5), 1189-1209. https://doi.org/10.1175/WAF-D-19-0005.1

Holland, G. J. (1980). An analytic model of the wind and pressure profiles in hurricanes. *Monthly Weather Review*, *108*(8), 1212–1218. https://doi.org/10.1175/1520-0493(1980)108<1212:AAMOTW>2.0.CO;2

Holm, E. V., Bonavita, M., Magnusson, L., (2015). Improved spread and accuracy in higher-resolution Ensemble of Data Assimilations. ECMWF Newsletter 145, https://www.ecmwf.int/sites/default/files/elibrary/2015/14589-newsletter-no145-autumn-2015.pdf

Horn, M., Walsh, K., Zhao, M., Camargo, S. J., Scoccimarro, E., Murakami, H., Wang, H., Ballinger, A., Kumar, A., Shaevitz, D. A., Jonas, J. A., Oouchi, K. (2014). Tracking Scheme Dependence of Simulated Tropical Cyclone Response to Idealized Climate Simulations, Journal of Climate, 27(24), 9197-9213. https://doi.org/10.1175/JCLI-D-14-00200.1

Huffman, G., Bolvin, D., Braithwaite, D., Hsu, K., Joyce, R., Xie, P. (2014), Integrated Multi-satellitE Retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center, accessed 31 March, 2015, ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/

IFRC (2021). *Forecast-Based Financing*. Retrieved 23 June 2021, from https://www.forecast-based-financing.org/about/

Ingleby, B., Prates, F., Isaksen, L., Bonavita, M. (2020). Recent BUFR dropsonde data improved forecasts. ECMWF Newsletter 162.

Janssen, P. (1997). Effect of surface gravity waves on the heat flux. ECMWF Tech Memo 239, https://doi.org/10.21957/80E0INRGX

Janssen, P. A. E. M., & Bidlot, J.-R. (2018). Progress in operational wave forecasting. Procedia IUTAM, 26, 14–29. https://doi.org/10.1016/j.piutam.2018.03.003

Janssen, P., & Bidlot, J.-R. (2021). On the consequences of nonlinearity and gravity-capillary waves on wind-wave interaction. ECMWF Tech Memo 882, https://doi.org/10.21957/B88TQJD6Q

Jones, S. C., Harr, P. A., Abraham, J., Bosart, L. F., Bowyer, P. J., Evans, J. L., Hanley, D. E., Hanstrum, B. N., Hart, R. E., Lalaurette, F., Sinclair, M. R., Smith, R. K., & Thorncroft, C. (2003). The Extratropical Transition of Tropical Cyclones: Forecast Challenges, Current Understanding, and Future Directions, Weather and Forecasting, 18(6), 1052-1092. https://doi.org/10.1175/1520-0434(2003)018<1052:TETOTC>2.0.CO;2

Judt, F., Klocke, D., Rios-Berrios, R., Vanniere, B., Ziemen, F., Auger, L., Biercamp, J., Bretherton, C., Chen, X., Düben, P., Hohenegger, C., Khairoutdinov, M., Kodama, C., Kornblueh, L., Lin, S.-J., Nakano, M., Neumann, P., Putman, W., Röber, N., … Zhou, L. (2021). Tropical cyclones in global storm-resolving models. Journal of the Meteorological Society of Japan. Ser. II, 99(3), 579–602. https://doi.org/10.2151/jmsj.2021-029

Jurlina, T., Baugh, C., Pappenberger, F., Prudhomme, C. (2020). Flood hazard risk forecasting index (FHRFI) for urban areas: The Hurricane Harvey case study. Meteorol Appl.; 27:e1845. https://doi.org/10.1002/met.1845

Keller, J. H., et al. (2019). The Extratropical Transition of Tropical Cyclones. Part II: Interaction with the Midlatitude Flow, Downstream Impacts, and Implications for Predictability, Monthly Weather Review, 147(4), 1077-1106. https://doi.org/10.1175/MWR-D-17-0329.1

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The international best track archive for climate stewardship (Ibtracs): Unifying tropical cyclone data. Bulletin of the American Meteorological Society, 91(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kohno, N. Dube, S. K., Entel, M., Fakhruddin, S.H.M., Greenslade, D., Leroux, M.-D., Rhome, J., Ba Thuy, N. (2018). Recent Progress in Storm Surge Forecasting, TCRR, Vol 7, 128-139, https://www.sciencedirect.com/science/article/pii/S2225603219300207

Komaromi, W. A., P. A. Reinecke, J. D. Doyle, and J. R. Moskaitis, 2021: The Naval Research Laboratory's Coupled Ocean-Atmosphere Mesoscale Prediction System – Tropical Cyclone Ensemble (COAMPS-TC Ensemble). Wea. Forecasting, 36, 499-517.

Kowaleski, A. M., & Evans, J. L. (2020). Use of multiensemble track clustering to inform medium-range tropical cyclone forecasts. Weather and Forecasting, 35(4), 1407–1426. https://doi.org/10.1175/WAF-D-20-0003.1

Kowaleski, A. M., Morss, R. E., Ahijevych, D., Fossell, K. R. (2020). Using a WRF-ADCIRC Ensemble and Track Clustering to Investigate Storm Surge Hazards and Inundation Scenarios Associated with Hurricane Irma, Weather and Forecasting, 35(4), 1289-1315, https://doi.org/10.1175/WAF-D-19-0169.1

Kühnlein C., W. Deconinck, R. Klein, S. Malardel, Z. P. Piotrowski, P. K. Smolarkiewicz, J. Szmelter, and N. P. Wedi, 2019. FVM 1.0: a nonhydrostatic finite-volume dynamical core for the IFS. Geosci. Model Dev., 12, 651–676.

Kühnlein C. et al., Towards global weather forecasting with IFS-FVM, ECMWF Annual Seminar 2020, https://events.ecmwf.int/event/167/contributions/1367/attachments/799/1410/AS2020-Kuehnlein.pdf

Kumler-Bonfanti, C., Stewart, J., Hall, D., Govett, M. (2020). Tropical and Extratropical Cyclone Detection Using Deep Learning, Journal of Applied Meteorology and Climatology, 59(12), 1971-1985. https://doi.org/10.1175/JAMC-D-20-0117.1

Landsea, C. W., & Franklin, J. L. (2013). Atlantic hurricane database uncertainty and presentation of a new database format. Monthly Weather Review, 141(10), 3576–3592. https://doi.org/10.1175/MWR-D-12-00254.1

Lang, S.T.K., Leutbecher, M. and Jones, S.C. (2012), Impact of perturbation methods in the ECMWF ensemble prediction system on tropical cyclone forecasts. Q.J.R. Meteorol. Soc., 138: 2030-2046. https://doi.org/10.1002/qj.1942

Lang, S.T.K., Bonavita, M. and Leutbecher, M. (2015), On the impact of re-centring initial conditions for ensemble forecasts. Q.J.R. Meteorol. Soc., 141: 2571-2581. https://doi.org/10.1002/qj.2543

Lang, STK, Lock, S-J, Leutbecher, M, Bechtold, P, Forbes, RM. (2021), Revision of the Stochastically Perturbed Parametrisations model uncertainty scheme in the Integrated Forecasting System. Q J R Meteorol Soc. 2021; 147: 1364– 1381. https://doi.org/10.1002/qj.3978

Lang, S.T.K., Dawson, A., Diamantakis, M., Dueben, P., Hatfield, S., Leutbecher, M., Palmer, T., Prates, F., Roberts, C.D., Sandu, I., Wedi, N. (2021), More Accuracy with Less Precision, submitted to Q.J.R. Meteorol. Soc.

Lawrence, H., Bormann, N., Sandu, I., Day, J., Farnan, J., & Bauer, P. (2019). Use and impact of arctic observations in the ecmwf numerical weather prediction system. Quarterly Journal of the Royal Meteorological Society, 145(725), 3432–3454. https://doi.org/10.1002/qj.3628

Lee, C., Camargo, S. J., Vitart, F., Sobel, A. H., & Tippett, M. K. (2018). Subseasonal Tropical Cyclone Genesis Prediction and MJO in the S2S Dataset, Weather and Forecasting, 33(4), 967-988, https://doi.org/10.1175/WAF-D-17-0165.1

Lee, C., Camargo, S. J., Vitart, F., Sobel, A. H., Camp, J., Wang, S., Tippett, M. K., & Yang, Q. (2020). Subseasonal Predictions of Tropical Cyclone Occurrence and ACE in the S2S Dataset, Weather and Forecasting, 35(3), 921-938. https://doi.org/10.1175/WAF-D-19-0217.1

Liang, X., Chan, J. C. L. (2005). The Effects of the Full Coriolis Force on the Structure and Motion of a Tropical Cyclone. Part I: Effects due to Vertical Motion, Journal of the Atmospheric Sciences, 62(10), 3825-3830, https://doi.org/10.1175/JAS3545.1

Lillo, S.P., Parsons, D.B. (2017). Investigating the dynamics of error growth in ECMWF medium-range forecast busts. Q.J.R. Meteorol. Soc., 143: 1211-1226. https://doi.org/10.1002/qj.2938

Liu, Z., Wang, H., Zhang, Y. J., Magnusson, L., Loftis, J. D., Forrest, D. (2020). Cross-scale modeling of storm surge, tide, and inundation in Mid-Atlantic Bight and New York City during hurricane Sandy, 2012, Estuarine, Coastal and Shelf Science 233, 106544

Leonardo, N. M., & Colle, B. A. (2020). An Investigation of Large Along-Track Errors in Extratropical Transitioning North Atlantic Tropical Cyclones in the ECMWF Ensemble, Monthly Weather Review, 148(1), 457-476. https://doi.org/10.1175/MWR-D-19-0044.1

Leonardo, N. M., & Colle, B. A. (2021). An Investigation of Large Cross-Track Errors in North Atlantic Tropical Cyclones in the GEFS and ECMWF Ensembles, Monthly Weather Review, 149(2), 395-417. https://doi.org/10.1175/MWR-D-20-0035.1

Leutbecher, M. and Palmer, T. N. (2008) Ensemble forecasting. J. Comput. Phys., 227, 3515–3539. URL: https://doi.org/10.1016/ j.jcp.2007.02.014.

Leutbecher, M., et al. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision, QJRMS, 143, 2315-2339

MacLeod, D., Easton-Calabria, E., de Perez, E. C., & Jaime, C. (2021). Verification of forecasts for extreme rainfall, tropical cyclones, flood and storm surge over Myanmar and the Philippines. Weather and Climate Extremes, 33, 100325. https://doi.org/10.1016/j.wace.2021.100325

Magnusson, L., Bidlot, J. R., Lang, S. T. K., Thorpe, A., Wedi, N., & Yamaguchi, M. (2014). Evaluation of medium-range forecasts for hurricane Sandy. Monthly Weather Review, 142(5), 1962-1981.

Magnusson, L. , Dahoui, M., Velden, C. S., Olander, T. L., (2017). A fresh look at tropical cyclone intensity estimates. ECMWF Newsletter 152. https://www.ecmwf.int/en/newsletter/152/news/fresh-look-tropical-cyclone-intensity-estimates

Magnusson, L., Bidlot, J.-R., Bonavita, M., Brown, A., Browne, P., De Chiara, G., Dahoui, M., Lang, S. T. K., McNally, T., Mogensen, K. S., Pappenberger, F., Prates, F., Rabier, F., Richardson, D. S., Vitart, F. Malardel, S. (2019a). ECMWF activities for improved hurrivane forecasts, BAMS, 100 (3), 445-458

Magnusson, L., Doyle, J. D., Komaromi, W. A., Zhang, F., Torn, R., Tang, C. K., Chan, C. L., Yamaguchi, M. (2019b). Advances in understanding difficult cases of track forecasts, Tropical Cyclone Research and Review, 8(3), 109-122. https://doi.org/10.1016/j.tcrr.2019.10.001

Magnusson, L. and Bidlot, J.-R. (2020). Challenges in forecasting Hurricane Lorenzo. ECMWF Newsletter, 162, 2-3.

Maloney, E. D., & Hartmann, D. L. (2000). Modulation of Eastern North Pacific Hurricanes by the Madden–Julian Oscillation, Journal of Climate, 13(9), 1451-1460. https://doi.org/10.1175/1520-0442(2000)013<1451:MOENPH>2.0.CO;2

McNally, T., Bonavita, M., & Thépaut, J. (2014). The Role of Satellite Data in the Forecasting of Hurricane Sandy, Monthly Weather Review, 142(2), 634-646. https://doi.org/10.1175/MWR-D-13-00170.1

Mogensen, K. S., Magnusson, L. and Bidlot, J.-R. (2017). Tropical cyclone sensitivity to ocean coupling in the ECMWF coupled model, Journal of Geophysical Research: Oceans, 122, 4392-4412,

Mouche, A., Chapron, B., Knaff, J., Zhao, Y., Zhang, B., Combot, C. (2019). Copolarized and cross-polarized SAR measurements for high-resolution description of major hurricane wind structures:

Application to Irma category 5 hurricane. Journal of Geophysical Research: Oceans, 124, 3905– 3922. https://doi.org/10.1029/2019JC015056

NCAR. (2021). THORPEX Interactive Grand Global Ensemble (TIGGE) Model Tropical Cyclone Track Data, https://doi.org/10.5065/D6GH9GSZ.

Neumann, B., Vafeidis, A. T., Zimmermann, J., & Nicholls, R. J. (2015). Future coastal population growth and exposure to sea-level rise and coastal flooding—A global assessment. PLOS ONE, 10(3), e0118571. https://doi.org/10.1371/journal.pone.0118571

Olander, T. L., and C. S. Velden, 2007: The advanced Dvorak technique: Continued development of an objective scheme to estimated tropical cyclone intensity using geostationary infrared satellite imagery. *Wea. Forecasting*, **22**, 287—298.

Omranian, E., Sharif, H., & Tavakoly, A. (2018). How well can global precipitation measurement (Gpm) capture hurricanes? Case study: hurricane harvey. Remote Sensing, 10(7), 1150. https://doi.org/10.3390/rs10071150

Puri, K., Barkmeijer, J. and Palmer, T.N. (2001), Ensemble prediction of tropical cyclones using targeted diabatic singular vectors. Q.J.R. Meteorol. Soc., 127: 709-731. https://doi.org/10.1002/qj.49712757222

Rappaport, E. N., Franklin, J. L., Avila, L. A., Baig, S. R., Beven, J. L., Blake, E. S., Burr, C. A., Jiing, J.-G., Juckins, C. A., Knabb, R. D., Landsea, C. W., Mainelli, M., Mayfield, M., McAdie, C. J., Pasch, R. J., Sisko, C., Stewart, S. R., & Tribble, A. N. (2009). Advances and challenges at the national hurricane center. Weather and Forecasting, 24(2), 395–419. https://doi.org/10.1175/2008WAF2222128.1

Rennie, M.P., Isaksen, L., Weiler, F., de Kloe, J., Kanitz, T. & Reitebuch, O. (2021) The impact of Aeolus wind retrievals on ECMWF global weather forecasts. Quarterly Journal of the Royal Meteorological Society, 1–32. Available from: https://doi.org/10.1002/qj.4142

Reul, N., Chapron, B., Zabolotskikh, E., Donlon, C., Mouche, A., Tenerelli, J., Collard, F., Piolle, J. F., Fore, A., Yueh, S., Cotton, J., Francis, P., Quilfen, Y., & Kudryavtsev, V. (2017). A new generation of tropical cyclone size measurements from space. Bulletin of the American Meteorological Society, 98(11), 2367–2385. https://doi.org/10.1175/BAMS-D-15-00291.1

Richardson, D. S., Cloke, H. L., & Pappenberger, F. (2020). Evaluation of the consistency of ecmwf ensemble forecasts. Geophysical Research Letters, 47(11). https://doi.org/10.1029/2020GL087934

Riemer, M. and Jones, S.C. (2014). Interaction of a tropical cyclone with a high-amplitude, midlatitude wave pattern: Waviness analysis, trough deformation and track bifurcation. Q.J.R. Meteorol. Soc., 140: 1362-1376. https://doi.org/10.1002/qj.2221

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, B., Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Roberts, C., Senan, R., Zarzycki, C., & Ullrich, P. (2020). Impact of model resolution on tropical cyclone simulation using the highresmip–primavera multimodel ensemble. Journal of Climate, 33(7), 2557–2583. https://doi.org/10.1175/JCLI-D-19-0639.1

Rodwell, M., Ferranti, L., Haiden, T., Magnusson, L. et al. (2015). New developments in the diagnosis and verification of high-impact weather forecasts. ECMWF Tech Memo 759, https://doi.org/10.21957/36B0T3MKE

Ruf, C. S., Atlas, R., Chang, P. S., Clarizia, M. P., Garrison, J. L., Gleason, S., Katzberg, S. J., Jelenak, Z., Johnson, J. T., Majumdar, S. J., O'Brien, A., Posselt, D. J., Ridley, A. J., Rose, R. J., & Zavorotny, V. U. (2016). New ocean winds satellite mission to probe hurricanes and tropical convection. Bulletin of the American Meteorological Society, 97(3), 385–395. https://doi.org/10.1175/BAMS-D-14-00218.1

Ruston, B., & Healy, S. (2021). Forecast Impact of FORMOSAT -7/ COSMIC -2 GNSS Radio Occultation Measurements. Atmospheric Science Letters, 22(3). https://doi.org/10.1002/asl.1019

Sanabia, E. R., & Jayne, S. R. (2020). Ocean observations under two major hurricanes: Evolution of the response across the storm wakes. AGU Advances, 1, e2019AV000161. https://doi.org/10.1029/2019AV000161

Sawada, Y., Okamoto, K., Kunii, M., & Miyoshi, T. (2019). Assimilating every-10-minute himawari-8 infrared radiances to improve convective predictability. Journal of Geophysical Research: Atmospheres, 124(5), 2546–2561. https://doi.org/10.1029/2018JD029643

Schäfler, A., et al. (2018). The north Atlantic waveguide and downstream impact experiment, BAMS, 99 (8), 1607-1637, https://doi.org/10.1175/BAMS-D-17-0003.1

Skofronick-Jackson, G., Petersen, W. A., Berg, W., Kidd, C., Stocker, E. F., Kirschbaum, D. B., Kakar, R., Braun, S. A., Huffman, G. J., Iguchi, T., Kirstetter, P. E., Kummerow, C., Meneghini, R., Oki, R., Olson, W. S., Takayabu, Y. N., Furukawa, K., & Wilheit, T. (2017). The global precipitation measurement (Gpm) mission for science and society. Bulletin of the American Meteorological Society, 98(8), 1679–1695. https://doi.org/10.1175/BAMS-D-15-00306.1

Stockdale, T. and co-authors. (2018). SEAS5 and the future evolution of the long-range forecast system. ECMWF Tech Memo 835

Tang, C. K., Chan, J. C. L., & Yamaguchi, M. (2021). Large tropical cyclone track forecast errors of global numerical weather prediction models in western north pacific basin. Tropical Cyclone Research and Review, S2225603221000242. https://doi.org/10.1016/j.tcrr.2021.07.001

Tavolato, C., & Isaksen, L. (2015). On the use of a Huber norm for observation quality control in the ECMWF 4D-Var. Quarterly Journal of the Royal Meteorological Society, 141(690), 1514–1527. https://doi.org/10.1002/qj.2440

Termonia, P., & Hamdi, R. (2007). Stability and accuracy of the physics—Dynamics coupling in spectral models. Quarterly Journal of the Royal Meteorological Society. https://doi.org/10.1002/qj.119

Tian, X., & Zou, X. (2016). ATMS- and AMSU-A-derived hurricane warm core structures using a modified retrieval algorithm. Journal of Geophysical Research: Atmospheres, 121(21), 12,630-12,646. https://doi.org/10.1002/2016JD025042

Titley, H., Yamaguchi, M., Magnusson, L. (2019). Current and potential use of ensemble forecasts in operational TC forecasting: results from a global forecaster survey, Tropical Cyclone Research and Review, 8(3), 166-180.

Titley, H. A., Bowyer, R. L. Cloke, H. L. (2020). A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. Q J R Meteorol Soc., 146: 531– 545. https://doi.org/10.1002/qj.3712

Titley, H. A., Cloke, H. L., Harrigan, S., Pappenberger, F., Prudhomme, C., Robbins, J. C., Stephens, E. M., & Zsoter, E. (2021). Key factors influencing the severity of fluvial flood hazard from tropical cyclones. Journal of Hydrometeorology, 1(aop). https://doi.org/10.1175/JHM-D-20-0250.1

Torn, R. D., & Snyder, C. (2012). Uncertainty of tropical cyclone best-track information. Weather and Forecasting, 27(3), 715–729. https://doi.org/10.1175/WAF-D-11-00085.1

Torn, R. D., T. J. Elless, P. Papin, C. A. Davis, (2018). The sensitivity of TC track forecasts within deformation steering flows. Mon. Wea. Rev., 146, 3183-3201

Tort, M., & Dubos, T. (2014). Dynamically consistent shallow-atmosphere equations with a complete Coriolis force: Non-Traditional Shallow-Atmosphere Equations. Quarterly Journal of the Royal Meteorological Society, 140(684), 2388–2392. https://doi.org/10.1002/qj.2274

Tsai, H.-C., & Elsberry, R. L. (2013). Detection of tropical cyclone track changes from the ECMWF ensemble prediction system: TC TRACK CHANGES IN ECMWF ENSEMBLE. Geophysical Research Letters, 40(4), 797–801. https://doi.org/10.1002/grl.50172

Van der Grijn, G., Paulsen, J. E., Lalaurette, F., Leutbecher, M. (2005). Early medium-range forecasts of tropical cyclones. ECMWF Newsletter 102, https://www.ecmwf.int/sites/default/files/elibrary/2004/14623-newsletter-no102-winter-200405.pdf

Van Der Knijff, J. M., Younis, J., & De Roo, A. P. J. (2010). LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation. *International Journal of Geographical Information Science*, *24*(2), 189–212. https://doi.org/10.1080/13658810802549154

Vellinga, M., Copsey, D., Graham, T., Milton, S., Johns, T. (2020). Evaluating Benefits of Two-Way Ocean–Atmosphere Coupling for Global NWP Forecasts, Weather and Forecasting, 35(5), 2127-2144. https://doi.org/10.1175/WAF-D-20-0035.1

Vidale, P. L., Hodges, K., Vannière, B., Davini, P., Roberts, M. J., Strommen, K., Weisheimer, A., Plesca, E., & Corti, S. (2021). Impact of stochastic physics and model resolution on the simulation of tropical cyclones in climate gcms. Journal of Climate, 34(11), 4315–4341. https://doi.org/10.1175/JCLI-D-20-0507.1

Vitart, F., Anderson, J. L., Stern, W. F. (1997). Simulation of Interannual Variability of Tropical Storm Frequency in an Ensemble of GCM Integrations, Journal of Climate, 10(4), 745-760. https://doi.org/10.1175/1520-0442(1997)010<0745:SOIVOT>2.0.CO;2

Vitart, F., Anderson, D., & Stockdale, T. (2003). Seasonal forecasting of tropical cyclone landfall over mozambique. *Journal of Climate*, *16*(23), 3932–3945. https://doi.org/10.1175/1520-0442(2003)016<3932:SFOTCL>2.0.CO;2

Vitart, F. (2009). Impact of the Madden Julian oscillation on tropical storms and risk of landfall in the ECMWF forecast system: MJO composites of model tropical storms. Geophysical Research Letters, 36(15), n/a-n/a. https://doi.org/10.1029/2009GL039089

Vitart, F., Prates, F., Bonet, A., & Sahin, C. (2012). New tropical cyclone products on the web. ECMWF Newsletter 130, https://doi.org/10.21957/TI1191E2

Wedi, N. P., Polichtchouk, I., Dueben, P., Anantharaj, V. G., Bauer, P., Boussetta, S., Browne, P., Deconinck, W., Gaudin, W., Hadade, I., Hatfield, S., Iffrig, O., Lopez, P., Maciel, P., Mueller, A., Saarinen, S., Sandu, I., Quintino, T., & Vitart, F. (2020). A baseline for global weather and climate simulations at 1 km resolution. Journal of Advances in Modeling Earth Systems, 12(11). https://doi.org/10.1029/2020MS002192

Wilks, D. S. (2006), Statistical Methods in the Atmospheric Sciences.2nd Academic Press, 627 pp

WMO (2013), Verification methods for tropical cyclone forecasts, WWRP/WGNE Joint Working Group on Forecast Verification Research, November 2013

Wu, L., Rutgersson, A., Sahlée, E., & Larsén, X. G. (2015). The impact of waves and sea spray on modelling storm track and development. Tellus A: Dynamic Meteorology and Oceanography, 67(1), 27967. https://doi.org/10.3402/tellusa.v67.27967

Yamaguchi, M., Majumdar, S. J. (2010). Using TIGGE Data to Diagnose Initial Perturbations and Their Growth for Tropical Cyclone Ensemble Forecasts, Monthly Weather Review, 138(9), 3634-3655. https://doi.org/10.1175/2010MWR3176.1

Yamaguchi, M., Vitart, F., Lang, S. T. K., Magnusson, L., Elsberry, R. L., Elliott, G., Kyouda, M. and Nakazawa, T. (2015). Global Distribution of the Skill of Tropical Cyclone Activity Forecasts on Short- to Medium-Range Time Scales. Wea. Forecasting, 30, 1695–1709.

Yamaguchi, M., Ishida, J., Sato, H., Nakagawa, M. (2017). WGNE Intercomparison of Tropical Cyclone Forecasts by Operational NWP Models: A Quarter Century and Beyond, Bulletin of the American Meteorological Society, 98(11), 2337-2349. https://doi.org/10.1175/BAMS-D-16-0133.1

Zawislak, J., Rogers, R. F., Aberson, S. D., Alaka, G. J., Alvey, G., Aksoy, A., Bucci, L., Cione, J., Dorst, N., Dunion, J., Fischer, M., Gamache, J., Gopalakrishnan, S., Hazelton, A., Holbach, H. M., Kaplan, J., Leighton, H., Marks, F., Murillo, S. T., … Zhang, J. A. (2021). Accomplishments of noaa's airborne hurricane field program and a broader future approach to forecast improvement. Bulletin of the American Meteorological Society, 1–79. https://doi.org/10.1175/BAMS-D-20-0174.1

Appendix A: List of Abbreviations

COAMPS-TC	Coupled Ocean/Atmosphere Mesoscale Prediction System - Tropical Cyclones (USA)
ENS	Operational ECMWF ensemble forecast
ENSO	El Niño and the Southern Oscillation
GTS	Global Telecommunications System
HRES	Operational high-resolution ECMWF forecast
HWRF	Hurricane Weather Research and Forecasting model (USA)
IBTrACS	International Best Track Archive for Climate Stewardship
JMA	Japan Meteorological Agency
MDR	Main Development Region in Atlantic basin (10-20N, 20-80W)
MHS	Microwave Humidity Sounder
MJO	Madden-Julian Oscillation
MSLP	Mean Sea Level Pressure
NASA	National Aeronautics and Space Administration (USA)
NCEP	National Centers for Environmental Prediction (USA)
NHC	National Hurricane Center (USA / NOAA)
NOAA	National Oceanographic and Atmospheric Administration (USA)
NWP	Numerical Weather Prediction
Pmin	Minimum central surface pressure in the tropical cyclone
RSMC	WMO Regional Specialized Meteorological Center
RI	Rapid Intensification
SEAS5	ECMWF's Seasonal Forecasting System
SFMR	Stepped Frequency Microwave Radiometer

SST	Sea Surface Temperature
TC	Tropical Cyclone
TCWC	Tropical Cyclone Warning Centre
UKMO	UK Met Office
Vmax	Maximum sustained surface wind speed at any location in the tropical cyclone
WGNE	WMO Working Group on Numerical Experimentation
WMO	World Meteorological Organisation
WWRP	World Weather Research Program

Appendix B: List of Tropical Cyclones in Special Experiment Period

The cases highlighted in grey refer to tropical cyclones with a Pmin lower than 980 hPa. Each of these cases lasted for at least 108 hours and provided the most substantial contributions to the statistics.

Vmax refers to the maximum sustained surface wind speed averaged over 1 minute, except for the western north Pacific tropical cyclones (marked by *) in which the averaging is over 10 minutes.

TC Code	TC Name	Genesis Time	Final Time	Duration (h)	Maximum Vmax (kt)	Minimum Pmin (hPa)
13L	Laura	2020-08-20-00	2020-08-29-00	216	130	937
14L	Marco	2020-08-20-12	2020-08-25-00	108	65	991
15L	Omar	2020-09-02-00	2020-09-05-12	84	35	1003
16L	Nana	2020-09-01-12	2020-09-04-00	60	65	994
17L	Paulette	2020-09-07-12	2020-09-16-06	210	90	965
18L	Rene	2020-09-07-18	2020-09-14-12	162	40	1001
19L	Sally	2020-09-12-18	2020-09-17-06	108	95	966
20L	Teddy	2020-09-14-06	2020-09-23-00	210	120	945
21L	Vicky	2020-09-14-12	2020-09-17-18	78	45	1001
22L	Beta	2020-09-18-18	2020-09-22-18	96	55	993

23L	Wilfred	2020-09-18-18	2020-09-21-00	54	35	1006
11E	Fausto	2020-08-16-00	2020-08-17-12	36	35	1004
12E	Genevieve	2020-08-16-12	2020-08-21-12	120	115	950
13E	Hernan	2020-08-24-00	2020-08-28-18	114	40	1001
14E	Iselle	2020-08-24-12	2020-08-30-18	150	50	997
15E	Julio	2020-09-03-12	2020-09-07-06	90	40	1004
16E	Karina	2020-09-10-00	2020-09-17-00	168	50	996
08W	Higos	2020-08-17-18	2020-08-19-06	36	60*	992
09W	Bavi	2020-08-21-18	2020-08-27-12	138	85*	950
10W	Maysak	2020-08-28-00	2020-09-03-06	150	95*	935
11W	Haishen	2020-08-31-12	2020-09-07-12	168	105*	910
13W	Noul	2020-09-15-18	2020-09-18-12	66	45*	992

Appendix C. Special Experiments: Verification Scores for Temperature and Winds in the Tropics

The appendix presents root-mean-square (RMSE) errors for the tropics (20°N-20°S). The plots show normalised RMSE differences against the control experiment, where positive values mean that experiment performed better than the control. The verification is against the operational analysis. For data assimilation experiments, this will favour experiments similar to the operational configuration for short lead-times.

Figure 67 shows results for satellite observation experiments, corresponding to Figure 36. As the experiments introduce difference to the verifying analysis, a negative impact is seen for short lead times. We expect the impact to be negative from the experiments where we remove observations, and significant differences remain into day 2-3 for several experiments, indicating a benefit at 700hPa in the tropics of assimilating these.

Figure 68 shows the results for the scatterometer experiments from Figure 39. Again, we see a negative impact for the shortest lead-times as mentioned above. But here we see improvements after Day 1 with reduced thinning and increased error for the experiment without any ASCAT data.

Figure 69 includes results for model dynamics experiments and Figure 70 for model physics and resolution experiments. For the dynamics experiments (corresponding to Figure 48) the results are mainly neutral, but with an interesting improvement from the move of the vertical diffusion calculation for longer lead-times. For the new moist physics (Figure 70) we find a positive impact for the tropical winds and temperature for both 4 km and 9 km resolution, corresponding to Figure 49 to Figure 51. The experiment with explicit convection (no parameterised convection) and 4 km resolution is clearly worse than the control experiment.

Figure 71 shows the results for the experiment without ocean coupling and for different wave model experiments, corresponding to Figure 55. By removing the ocean coupling we see a negative impact on the forecast scores. For the new state dependent wave physics we see a degradation in the tropics, but together with the new moist physics is still results in a positive impact.

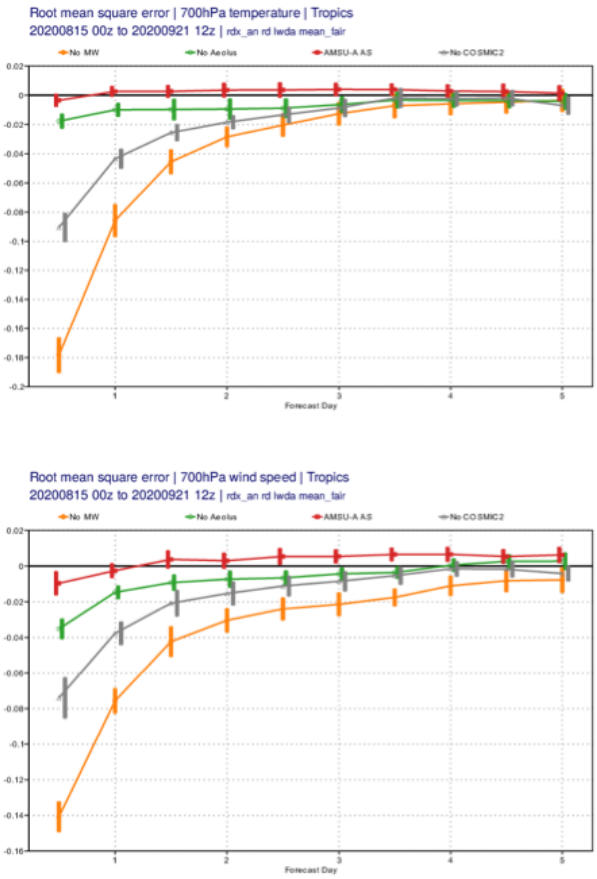


Figure 67: Normalised RMSE difference to control for satellite observation experiments for 700hPa temperature (top) and 700hPa wind speed (bottom) in the Tropics. (Positive means experiment better than control.

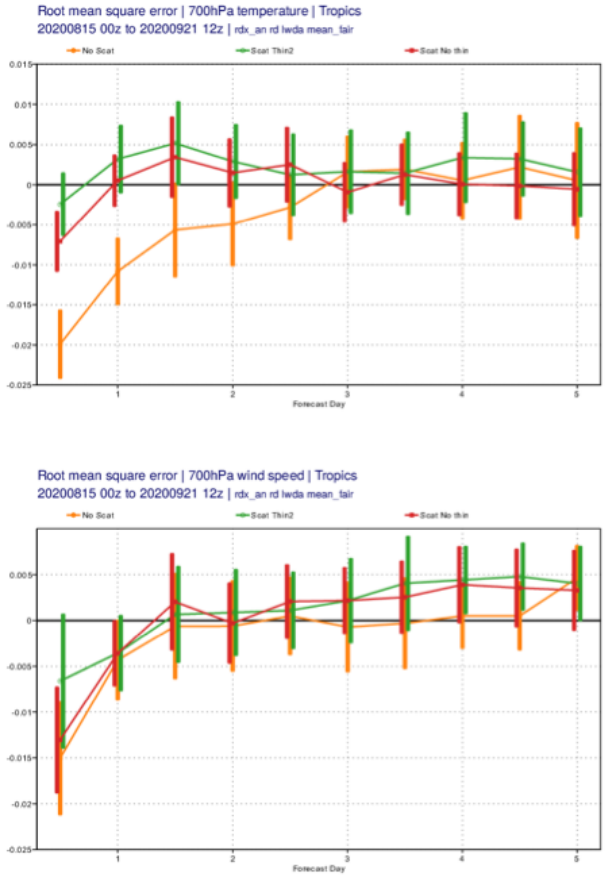


Figure 68: Normalised RMSE difference to control for scatterometer experiments for 700hPa temperature (top) and 700hPa wind speed (bottom) in the Tropics. (Positive means experiment better than control.

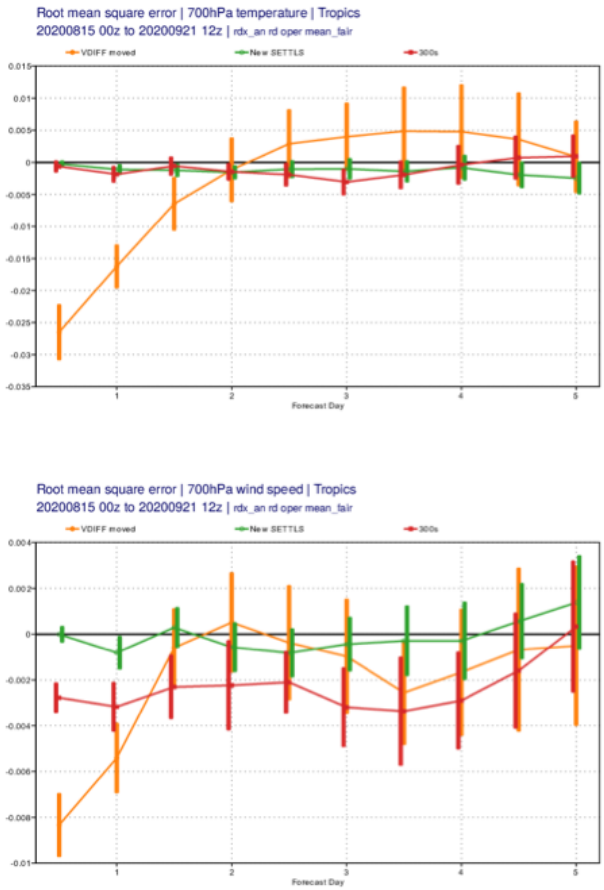


Figure 69: Normalised RMSE difference to 9km-oper for model dynamics experiments for 700hPa temperature (top) and 700hPa wind speed (bottom) in the Tropics. (Positive means experiment better than control.

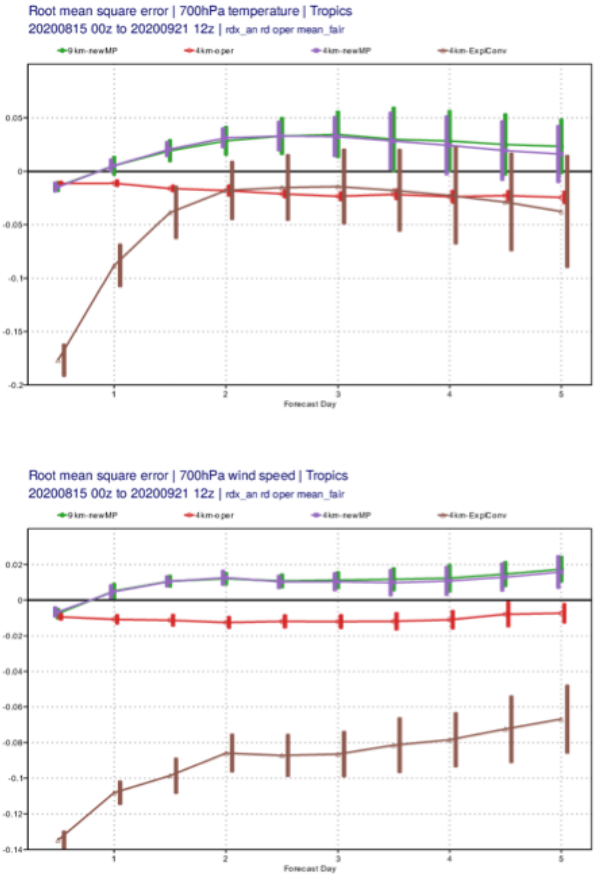


Figure 70: Normalised RMSE difference to 9km-oper for model physics and resolution experiments for 700hPa temperature (top) and 700hPa wind speed (bottom) in the Tropics. (Positive means experiment better than control.

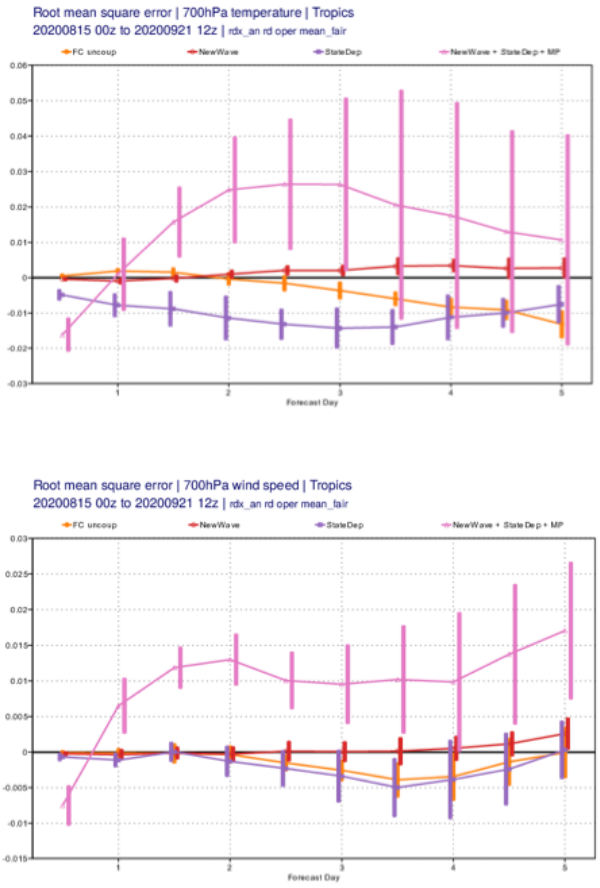


Figure 71: Normalised RMSE difference to 9km-oper for ocean-related model experiments for 700hPa temperature (top) and 700hPa wind speed (bottom) in the Tropics. (Positive means experiment better than control.

A5. Published Article: User decisions, and how these could guide developments in probabilistic forecasting

This appendix contains the published version of the following paper (summarised in Section 6.5.1):

Rodwell, M.J., Hammond, J., Thornton, S. and Richardson, D.S. (2020) 'User decisions, and how these could guide developments in probabilistic forecasting', *Quarterly Journal of the Royal Meteorological Society*, 146(732), pp. 3266–3284. Available at: https://doi.org/10.1002/qj.3845.

Received: 8 October 2019 | Revised: 21 May 2020 | Accepted: 1 June 2020 | Published on: 5 August 2020

DOI: 10.1002/qj.3845

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

User decisions, and how these could guide developments in probabilistic forecasting

M. J. Rodwell[1] | J. Hammond[2] | S. Thornton[3] | D. S. Richardson[1,4]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK

[2]weathertrending, High Wycombe, Buckinghamshire, UK

[3]weathertrending, Great Missenden, Buckinghamshire, UK

[4]Department of Geography and Environmental Science, University of Reading, Reading, UK

Correspondence
M.J. Rodwell, ECMWF, Shinfield Park, Reading RG2 9AX, UK.
Email: mark.rodwell@ecmwf.int

Abstract

We investigate how users combine objective probabilities with their own subjective feelings when deciding how to act on weather forecast information. Results are based on two scenarios investigated at a *Live Science* event held by the Royal Meteorological Society. When deciding whether to go to the beach with the possibility of warm, dry weather, we find that users attempt to identify their 'Bayes Action': the one which minimises their expected negative feeling or utility. Key factors are the 'thrill' of a nice day at the beach and the 'pain' of coping with, for example, children in wet weather, and the costs of travel. The users' threshold probabilities for deciding to go to the beach thus approximately define their distribution of cost/loss ratios. This is used to calculate a 'User Brier Score' (UBS): a measure of the overall utility to society, and which could be used to guide forecast system development. When applied to operational ensemble forecasts issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) over the period 1995–2018, the UBS tends to be higher (i.e., worse) than the Brier Score, largely because users tended not to exhibit high cost/loss ratios. When deciding whether to leave a campsite in the face of potentially dangerous gales, users try to find a balance between the 'regret' of serious injury and the 'pain' of spoiling an enjoyable holiday. Some users decide to stay even at high probabilities of serious consequences – partly due to a lack of experience. On the other hand, forecasts suffer from 'complete misses' – where probabilities of zero are accompanied by non-negligible outcome frequencies. These dominate the overall Brier Score. The frequency of complete misses halved over the period 1995–2018: a welcome improvement for users who do wish to avoid danger at low probabilities.

KEYWORDS

Bayes action, cost/loss ratio, ensemble forecast, proper score, refinement, reliability, User Brier Score, user decision

1 | INTRODUCTION

Weather forecasts have improved greatly over recent decades (Haiden *et al.*, 2018; Ben Bouallègue *et al.*, 2019), but does this improvement feed through to better weather-related decisions made by the forecast user community (Fundel *et al.*, 2019)? Such an assessment is complicated by the fact that almost every user decision has

TABLE 1 Scenarios, weather events and issued probabilities

Scenario	Weather event	Probabilities (%)	Action (Yes/No)
It is a Monday in summer, and you are thinking of going to the beach at the weekend with family, friends or on your own	"Warm and dry": greater than 20 °C and with less than 0.5 mm rain in 24 hr	0, 20, 40, 60, 80, 90, 95, 100	Go to the beach
You are camping with elderly and/or young relatives. There is a possibility of strong winds tomorrow, which could blow down your tent	"Strong winds": sustained wind-speeds of more than 25 mph (11 $m \cdot s^{-1}$), accompanied by stronger gusts	100, 80, 60, 40, 20, 10, 5, 0	Pack up and leave

a unique set of circumstances, costs and feelings associated with it, which the users themselves may not even be fully conscious of.

This study is based on a Royal Meteorological Society *Live Science* event, where the audience was given two weather-dependent scenarios (Table 1). The first was a potential trip to the beach in 5 days' time: something which might be enjoyable. The second represents the potential for catastrophic winds tomorrow while camping. The audience was asked to make binary no/yes decisions on the basis of a range of forecast probabilities. Through subsequent questioning, we sought to establish whether the users' cited feelings were consistent with their decisions (Savage, 1971). The distribution of users' threshold probabilities for a change in their actions was derived for each scenario. These hitherto elusive distributions (Murphy, 1966; 1969) were used to assess the last quarter-century of operational ensemble forecasts produced by the European Centre for Medium-Range Weather Forecasts (ECMWF).

Based on this *Live Science* event and related theory, we attempt to bridge the gap between the forecasting and user communities by addressing a few key questions:

- Can we determine the users' distribution of threshold probabilities for given weather-dependent decisions?
- To what extent do users make rational decisions on the basis of subjective feelings?
- Can forecast products be tailored to users, and how much direction should be given in terms of decision-making?
- How have ensemble forecasts improved from the users' perspective over the last quarter-century?
- Can forecast system development be guided by user-relevant scores?

In order to interpret the results of the *Live Science* event, we present some theory in section 2. Much of this is not new but reproduced here in a form where the user's no/yes decision – their "action" $\mathbb{I} \in \{0, 1\}$ – is more evident. Equations are presented in their numerical form so they can be readily applied. Later, we discuss salient terms in these equations. We discuss the importance of 'proper' scores for guiding both modelling and observational developments in operational ensemble forecasting. We also highlight the fact that, if the user takes the so-called 'Bayes Action', then the expense associated with their decision and the eventual outcome is a proper score of the forecast system (Dawid, 2007; Gneiting and Raftery, 2007). The derivation leads up to the proposal of a 'User Brier Score' (UBS): representing a minor development on the generalised score used by Richardson (2001) and Palmer (2002). This is a user-relevant score with potential to guide forecast system development. A reader less interested in the derivation might wish to go straight to Section 3 where we initially consider how a 'hyperrational decision maker' (Millner, 2009) would act in the presence of competing feelings or 'utilities' (Savage, 1971; Roebber and Bosart, 1996; Granger and Pesaran, 2000). The intention here is to then investigate how well a real decision-maker might approximate this hypothetical one. In Section 4 we give details of the forecast and verifying data used. More discussion of the *Live Science* event is presented in Section 5. Section 6 presents the results and our interpretations of these for both scenarios. In particular, we discuss our novel attempt to determine the users' distributions of threshold probabilities, and assess how well this relates to their stated feelings. Trends in the relative value of forecasts for individual users, and in the UBS are evaluated. We highlight key shortcomings and improvements in the forecast system of relevance to users and developers. A discussion and conclusions are presented in Section 7.

2 | THEORY

2.1 | Proper scores

Ensemble forecasts provide a probability for a given weather event which takes into account uncertainties in observations and in the forecast model itself. (The probability can be estimated as the fraction of ensemble members which predict the event). Scores of forecast performance compare the issued probabilities $p(t)$, where $t = 1, \ldots, N$ is an index of the forecast start time, with the observed outcomes $o(t)$, where $o(t) = 1$ if the event occurs and 0 if not. Key attributes of probabilistic forecasting systems are their 'reliability' and 'sharpness'.

Reliability measures how well the outcome frequency matches the forecast probability. If a large-enough sample of reliable forecast probabilities are partitioned into a set of probability bins $\{b_k\}$ with $k = 0, \ldots, K-1$ then, of the n_k occasions when the issued probability $p(t)$ is in bin b_k, the event should occur, on average, a fraction $\bar{o}_k = \bar{p}_k$ of the time, where

$$\begin{aligned} \bar{p}_k &= \frac{1}{n_k} \sum_{p(t) \in b_k} p(t), \\ \bar{o}_k &= \frac{1}{n_k} \sum_{p(t) \in b_k} o(t). \end{aligned} \tag{1}$$

Reliability is important to ensure users' decisions are unbiased. Sharpness relates to the propensity for the forecast system to issue probabilities close to 1 (definitely going to happen) and 0 (definitely not going to happen). The sharper the forecast is, the smaller its uncertainty.

Operational ensemble forecasting involves making a finite set of forecasts from a potentially unreliable set of initial conditions. However, to understand 'proper scores', such as the Brier score (Brier, 1950), in the context of operational weather forecasting, it is useful to think of a reliable distribution of initial conditions (with the truth lying somewhere within this distribution) being mapped into the future by a real, approximated model and by a hypothetical, perfect model. The mapping by the approximated model would become unreliable while that of the perfect model would remain reliable. Let p (q) be the probability for a given event based on the approximated (perfect) model. A proper score (Murphy and Epstein, 1967; Winkler and Murphy, 1968) is defined to be one for which the expected score of the unreliable forecast would never be better than that of the reliable forecast. In the 'divergence–entropy' decomposition of a proper score (Gneiting and Raftery, 2007), it is the 'divergence' component which penalises forecasts for deficiencies in reliability ($p \neq q$). Developments in the approximated model, which brought p closer to q, would therefore be rewarded. The 'entropy' component, for a (negatively oriented) proper score, is an upper-convex function of q (Schervish, 1989; Gneiting and Raftery, 2007) and thus leads to smaller penalties when q is close to 0 or 1. To obtain more 'refined' values of q closer to 0 or 1 (up to any limit imposed by chaotic error growth) would require a sharper initial distribution (and lead to a sharper forecast). Better observational information would be key to achieving this. Hence it can be argued that proper scores helpfully guide both modelling and observational developments in operational forecasting.

TABLE 2 The "Cost–Loss" matrix, representing the forecast-related costs associated with each action and outcome

Action	Outcome	
	No	Yes
No	0	L
Yes	C	C

2.2 | The Bayes Action

Although ensemble forecasts provide probabilities, users will often need to turn the issued probability p for a given weather event into a binary no/yes decision or "action" $\mathbb{I} \in \{0, 1\}$. The obvious way to do this is to set a threshold probability p_T and define

$$\mathbb{I}_p = [p > p_T] \quad , \tag{2}$$

where the expression $[\ \cdot\ ]$ is 0 or 1 depending on whether the statement it contains is untrue or true, respectively. A key focus of this study is to understand how each user chooses their own p_T.

In the simple cost–loss model (Thompson 1952; also discussion by Liljas and Murphy 1994 on earlier work by Ångström 1922), taking action (e.g., to protect a piece of infrastructure from the effects of a tropical storm) has a cost \$ C, while not taking action leads to a potential loss \$ $L > 0$, as shown in Table 2. A key parameter is the cost/loss ratio:

$$\alpha = \frac{C}{L}, \tag{3}$$

with the expense per unit L to the user being:

$$\begin{aligned} E_p &= \alpha \mathbb{I}_p + o(1 - \mathbb{I}_p) \\ &= (\alpha - o)\mathbb{I}_p + o. \end{aligned} \tag{4}$$

Using "expense per unit L" makes the presentation of the equations easier; from now on, we will simply refer to this as the "expense". In the first line of Equation (4), it

is clear that cost (per unit L) α is incurred if action is taken ($\mathbb{I}_p = 1$); and loss 1 is incurred if action is not taken ($\mathbb{I}_p = 0$) and the event happens ($o = 1$). The user's expected expense is

$$\mathbb{E}(E_p) = \{\alpha - o(p)\}\mathbb{I}_p + o(p), \tag{5}$$

where $o(p) \equiv \mathbb{E}(o|p)$, the expected outcome frequency given forecast probability p. Often the user will wish to take their "Bayes Action": the action which minimises their expected expense associated with the given forecast information. Clearly the expected expense in Equation (5) is minimised when their action $\mathbb{I}_p$ is equivalent to

$$\mathbb{I}_{o(p)} = [o(p) > \alpha]. \tag{6}$$

However in general $o(p)$ is not known (due to limitations on the number of past cases) and may not reflect the flow-dependence of reliability (Rodwell *et al.*, 2018), and so $\mathbb{I}_{o(p)}$ cannot be determined. For the given forecast probability p, and for the purposes of minimising their expected expense, the user can do little but assume that p is reliable and so their Bayes Action is

$$\mathbb{I}_p = [p > \alpha]. \tag{7}$$

This will lead to a larger (or at most equivalent) expected expense to that based on $\mathbb{I}_{o(p)}$. Importantly, this means that the user's actual expense E_p in Equation (4) represents a proper score of the forecast probability p (Dawid, 2007; Gneiting and Raftery, 2007). We will assume here that all users wish to take their Bayes Action. From Equations (7) and (2), this implies that

$$\alpha = p_T. \tag{8}$$

We will also assume that

$$0 < \alpha < 1, \tag{9}$$

since otherwise the user's Bayes Action is the same for all p, and so the user derives no benefit from the forecast. Such users need not be considered part of the user community for the given event. (Note that for the exceptional case of $p = 0$, $\alpha = 0$, the Bayes Action can also be $\mathbb{I}_p = 1$. Note also that we do not consider the possibility of $L < 0$).

Later we investigate whether the user's intention is to make their Bayes Action and, if so, how well they achieve it. The latter will depend on how well they are able to relate their threshold probabilities to their own costs – which might be subjective in nature.

2.3 | Relative value of forecasts

From Equation (7), the user's Bayes Actions associated with a set of operational forecast probabilities $\{p(t)\}$ will be

$$\mathbb{I}_{p(t)} = [p(t) > \alpha], \tag{10}$$

and, from Equation (4), their mean expense will be

$$\begin{aligned}\overline{E}_p &= \overline{\alpha\mathbb{I}}_p + \overline{o(1-\mathbb{I}_p)} \\ &= \overline{(\alpha - o)\mathbb{I}}_p + \overline{o},\end{aligned} \tag{11}$$

where an overbar $\overline{\cdot}$ indicates the mean over all t (or, in the continuous case, an integral over the distribution of values of p). Being a mean of proper scores, $\overline{E}_p$ is also a proper score of the forecast system.

To put the expense associated with the operational forecast into context, it is worth calculating the expense associated with poor and perfect forecast systems. For a set of forecasts based only the climatological (expected) event frequency $\mathbb{E}(o)$, the Bayes Action $\mathbb{I}_{\mathbb{E}(o)}$ will always be the same – either always take action or never take action. Following Equations (4) and (5), the expected expense is thus given by

$$\begin{aligned}\mathbb{E}(E_{\mathbb{E}(o)}) &= \alpha\mathbb{I}_{\mathbb{E}(o)} + \mathbb{E}(o)(1 - \mathbb{I}_{\mathbb{E}(o)}) \\ &= \{\alpha - \mathbb{E}(o)\}\mathbb{I}_{\mathbb{E}(o)} + \mathbb{E}(o),\end{aligned} \tag{12}$$

from which it is clear that the Bayes Action is

$$\mathbb{I}_{\mathbb{E}(o)} = [\mathbb{E}(o) > \alpha], \tag{13}$$

and the corresponding (minimum) expected expense is

$$\mathbb{E}(E_{\mathbb{E}(o)}) = \min\{\alpha, \mathbb{E}(o)\}. \tag{14}$$

In the asymptotic limit as $N \to \infty$, we have that the mean event frequency $\overline{o} \to \mathbb{E}(o)$. For 'sufficiently large' samples of forecasts, if the assumption is made that $\overline{o} = \mathbb{E}(o)$, then the Bayes Action becomes

$$\mathbb{I}_{\overline{o}} = [\overline{o} > \alpha], \tag{15}$$

and the mean expense becomes

$$\overline{E}_{\overline{o}} = \min(\alpha, \overline{o}). \tag{16}$$

As before, this can be interpreted as "either always take action at cost α or never take action and incur a loss 1 on the fraction $\overline{o}$ of occasions that the event occurs".

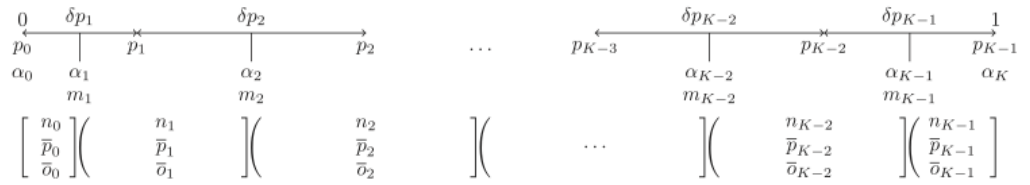


FIGURE 1 Numerical framework used for the calculation of Relative Value, and Brier and User Brier Scores. Variables are defined in the text

If we had access to a perfect (deterministic) forecast, the Bayes Action would be the outcome, $\mathbb{I}_o = o$, and we would incur the cost α only on the fraction of days that the event occurred:

$$\mathbb{E}(E_o) = \alpha \mathbb{E}(o), \tag{17}$$

and

$$\overline{E}_o = \alpha \overline{o}. \tag{18}$$

In Section 2.4, we will also make reference to the worst possible deterministic forecast where the Bayes Action is given by $\mathbb{I}_{1-o} = 1 - o$: the wrong decision is always made – so a cost α is incurred every time the event does not happen, and a loss 1 is incurred every time the event does happen. The mean expense is thus given by

$$\overline{E}_{1-o} = \alpha(1 - \overline{o}) + \overline{o}. \tag{19}$$

For $0 < \alpha < 1$ and $0 < \overline{o} < 1$, the relative value $V(\alpha)$ is defined as

$$V(\alpha) = \frac{\overline{E}_{\overline{o}} - \overline{E}_p}{\overline{E}_{\overline{o}} - \overline{E}_o}. \tag{20}$$

Note that $V = 0$ for forecasts based on the (sample) climate frequency, and $V = 1$ for a set of perfect forecasts.

Following Murphy (1977), the Appendix shows that the expected relative value $\mathbb{E}(V)$ for reliable forecast systems is non-negative. Hence, as $N \to \infty$, we have $V \geqslant 0$. Negative $V(\alpha)$ for any α may therefore be indicative of a lack of reliability. As also shown in the Appendix, the maximum expected relative value occurs for users who make their Bayes Action and have a cost/loss ratio equal to the expected event frequency. For a finite sample, following Murphy (1977), Richardson (2000) and Granger and Pesaran (2000), we also have

$$\begin{aligned} \max_{\alpha} V(\alpha) &= V(\overline{o}) \\ &= \frac{\overline{o\, \mathbb{I}_p}}{\overline{o}} - \frac{\overline{(1-o)\, \mathbb{I}_p}}{1 - \overline{o}} \\ &\equiv \mathrm{H} - \mathrm{F} \\ &\equiv \mathrm{PS}, \end{aligned} \tag{21}$$

where H is the Hit Rate, F is the False Alarm Rate and PS is the Peirce Score – all evaluated here at $\alpha = \overline{o}$. The Peirce Score was first discussed by Peirce (1884), and is used in the medical profession (Youden, 1950) and other disciplines as well as in weather forecasting; another name is Kuiper's Score. It has the useful property of being 'equitable' (Gandin and Murphy, 1992), so that unskilful forecasts (constant or random forecasts) are accorded the same expected score. Rodwell (2011) demonstrated an equivalence to Peirce's motivation, emphasizing that equitable scores are also useful for evaluating the performance of more skilful forecast systems.

In order to calculate the relative value for users with different α from a set of probabilistic forecasts and outcomes (as in the *Live Science* event), we use the numerical framework shown in Figure 1. For a given weather-related decision, users are presented with a range of exploratory probabilities for the weather event

$$0 = p_0 < p_1 < \ \ldots \ < p_{K-1} = 1, \tag{22}$$

and asked to make a decision 'no' or 'yes' for each probability. If a user first changes their action from 'no' to 'yes' between probabilities p_{j-1} and p_j (where $j \in \{1, \ldots, K-1\}$), then their cost/loss ratio α (and threshold probability p_T) is estimated as

$$\alpha_j = (p_{j-1} + p_j)/2. \tag{23}$$

Note that, if a user never changes their action, they are assumed to have cost/loss ratio of $\alpha = 0$ or 1 and so, as discussed above, they are disregarded here. Later we discuss other possible user choices.

To calculate the mean expense, $\overline{E}_p(\alpha_j)$, for this user from Equations (11) and (7), we need to identify the forecast probabilities $p(t)$ that are more than (or less than) α_j. Hence it is useful to base the forecast probability bins (as discussed in Section 2.1) on the possible values of α:

$$\begin{aligned} b_k &= (\alpha_k, \alpha_{k+1}], \qquad k = 1, \ \ldots \ , K-1, \\ b_0 &= [\alpha_0, \alpha_1], \end{aligned} \tag{24}$$

where $\alpha_0 = 0$ and $\alpha_K = 1$. As in Section (2.1), let n_k be the number of forecasts which fall in the probability bin b_k, and let $\bar{o}_k$ be as in Equation (1): the mean outcome frequency of the event for these forecasts. Integrating over the distribution of forecast probabilities, the user's mean expense (Equation (11)) can thus be calculated as

$$\begin{aligned}\overline{E}_p(\alpha_j) &= \frac{1}{N}\sum_{k=j}^{K-1} n_k\alpha_j + \frac{1}{N}\sum_{k=0}^{j-1} n_k\bar{o}_k \\ &= \frac{1}{N}\sum_{k=j}^{K-1} n_k(\alpha_j - \bar{o}_k) + \bar{o}. \end{aligned} \tag{25}$$

The values of $\overline{E}_{\bar{o}}$ and $\overline{E}_o$, also required to calculate the relative value $V(\alpha_j)$ in Equation (20), are readily obtained from Equations (16) and (18), respectively.

2.4 | The User Brier Score

In addition to establishing the relative value of forecasts for individual users, it is also useful to measure the utility of the forecasts to the user community as a whole. If we can ascertain the cost/loss ratio for each member of the user community for a given event, then we can calculate $\widetilde{\overline{E}}_p$, where a tilde $\widetilde{\ }$ indicates the mean over all users' cost/loss ratios (or, in the continuous case, the integral over the users' distribution of cost/loss ratios). Here, we propose a 'User Brier Score' which relates $\widetilde{\overline{E}}_p$ to the mean expense for the user community associated with the set of perfect forecasts, $\widetilde{\overline{E}}_o$:

$$\mathrm{UBS} = \frac{\widetilde{\overline{E}}_p - \widetilde{\overline{E}}_o}{\widetilde{\overline{E}}_{1-o} - \widetilde{\overline{E}}_o}, \tag{26}$$

Below, we show how this score can be calculated and then discuss why it might be thought of as a User Brier Score.

Referring again to Figure 1, suppose there are M users and m_j of these have cost/loss ratio α_j with $j \in \{1, \ldots, K-1\}$ and $\sum_{j=1}^{K-1} m_j = M$. If we write $\delta p_j = p_j - p_{j-1}$, then the distribution of cost/loss ratios can be estimated as

$$u(\alpha_j) \equiv u_j = \frac{m_j}{M\delta p_j}. \tag{27}$$

Notice that we again disregard users who never change their action and who thus have a cost/loss ratio of $\alpha_0 = 0$ or $\alpha_K = 1$.

Integrating the mean expense (Equation (25)) over this distribution, we obtain

$$\begin{aligned}\widetilde{\overline{E}}_p &= \sum_{j=1}^{K-1} u_j\delta p_j \overline{E}_p(\alpha_j) \\ &= \sum_{j=1}^{K-1} u_j\delta p_j \left\{ \frac{1}{N}\sum_{k=j}^{K-1} n_k(\alpha_j - \bar{o}_k) + \bar{o} \right\} \\ &= \frac{1}{N}\sum_{k=1}^{K-1} n_k \sum_{j=1}^{k} (\alpha_j - \bar{o}_k) u_j\delta p_j + \bar{o}. \end{aligned} \tag{28}$$

On the last line, we have used the fact that the integral of the distribution of cost/loss ratios $\sum_{j=1}^{K-1} u_j\delta p_j = 1$. The order of summation has also been swapped but still sums over the same triangle of (j,k) values. The mean of the users' distribution of cost/loss ratios itself is given by

$$\tilde{\alpha} = \sum_{j=1}^{K-1} u_j\delta p_j\alpha_j. \tag{29}$$

and integrals of the expense associated with perfect and worst deterministic forecasts can be written as

$$\widetilde{\overline{E}}_o = \tilde{\alpha}\ \bar{o}, \tag{30}$$

and

$$\widetilde{\overline{E}}_{1-o} = (1-\bar{o})\tilde{\alpha} + \bar{o}. \tag{31}$$

We now have all the components to calculate UBS in Equation (26).

By replacing $\mathbb{I}_p$ in Equation (4) with $\mathbb{I}_{1-o}$ and $\mathbb{I}_o$, it is straightforward to show that $1 \geqslant E_{1-o} \geqslant E_p \geqslant E_o \geqslant 0$ for any α or $\mathbb{I}_p$ (with $E_{1-o} > E_o$ for $\alpha \in (0,1)$), and so we have $0 \leqslant \mathrm{UBS} \leqslant 1$.

The numerator in Equation (26) is a proper score of the forecasts $\{p(t)\}$ since $\widetilde{\overline{E}}_p$ is proper and $\widetilde{\overline{E}}_o$ is only dependent on the outcome. (This relates to the notion of 'equivalence' discussed by Dawid, 2007.) The denominator in Equation (26) is $(1-2\tilde{\alpha})\bar{o} + \tilde{\alpha}$. If $\tilde{\alpha} = 1/2$, then this denominator is 1/2 (i.e., constant) so that UBS is also a proper score. If $\tilde{\alpha} \neq 1/2$ then UBS is asymptotically proper as $N \to \infty$ (Murphy, 1973; Gneiting and Raftery, 2007).

For of a uniform distribution of cost/loss ratios, we have $u_j = 1$ for all j, $\tilde{\alpha} = 1/2$ and, following Murphy (1966), UBS becomes the Brier Score (BS):

$$\begin{aligned}\mathrm{UBS}^{\mathrm{unif}} &= \frac{\widetilde{\overline{E}}^{\mathrm{unif}}_p - \widetilde{\overline{E}}^{\mathrm{unif}}_o}{\widetilde{\overline{E}}^{\mathrm{unif}}_{1-o} - \widetilde{\overline{E}}^{\mathrm{unif}}_o} \\ &= \frac{1}{N}\sum_{k=1}^{K-1} n_k\{p_k^2 - 2\bar{o}_k p_k\} + \bar{o} \\ &\approx \frac{1}{N}\sum_{k=0}^{K-1} n_k\{(\bar{p}_k - \bar{o}_k)^2 + \bar{o}_k(1-\bar{o}_k)\} \\ &\qquad\qquad \text{Reliability} \qquad \text{Refinement} \\ &= \frac{1}{N}\sum_{k=0}^{K-1} n_k\{\bar{o}_k(\bar{p}_k - 1)^2 + (1-\bar{o}_k)(\bar{p}_k - 0)^2\} \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\{p(t) - o(t)\}^2. \\ &\qquad\qquad \text{Brier Score} \end{aligned} \tag{32}$$

On the second line we have used the facts that $\alpha_j = (p_j + p_{j-1})/2$, $\delta p_j = (p_j - p_{j-1})$, $p_0 = 0$, $p_{K-1} = 1$, so that $\sum_{j=1}^{k} \alpha_j \delta p_j = p_k^2/2$ and $\tilde{\alpha} = 1/2$. On the third line, we have used the facts that $p_0 = 0$, $\overline{o} = \frac{1}{N} \sum_{k=0}^{K-1} n_k \overline{o}_k$ and the "≈" is because $p_k \approx \overline{p}_k$. To get to the last line, note that there are $n_k \overline{o}_k$ events, and $n_k(1 - \overline{o}_k)$ non-events associated with bin b_k: these contributions to the BS are evident in the fourth line. The "≈" on the last line is because $p(t) \in p_k \Rightarrow p(t) \approx \overline{p}_k$ (Stephenson *et al.*, 2008). In plots for the situations described here, UBS^{unif} and BS are virtually indistinguishable.

Since the possible range of UBS is the same as for the BS, since UBS is (asymptotically) proper, and since UBS converges to the BS in the case of a uniform distribution of cost/loss ratios, we argue that UBS in Equation (26) can be thought of as a 'User Brier Score'.

To ensure propriety for cases where $\tilde{\alpha} \neq 1/2$, then a modified version

$$\text{UBS}_{\text{prop}} = \frac{\overline{\overline{E}}_p - \overline{\overline{E}}_o}{\overline{\overline{E}}_{1-o} - \overline{\overline{E}}_o} \tag{33}$$

could be considered, where the denominator is given a fixed climatological value (the double overbar $\overline{\overline{\cdot}}$ indicates the mean over some large climatology which is not dependent on the sample of forecasts being scored in the numerator). This will be a proper score but might stray above 1 in some circumstances. In the cases highlighted here Equation (33) is almost indistinguishable from Equation (26). We might also imagine a "Continuous UBS" (CUBS) analogous to the Continuous Rank Probability Score (CRPS), perhaps weighted more heavily for extreme events (Gneiting and Ranjan, 2011) which the users might be more sensitive to. This is not pursued here. The propriety and user relevance suggest that the UBS or (UBS_{prop} or CUBS, where appropriate) should be useful in unifying the goals of forecasters and users (also Richardson, 2001; Palmer, 2002).

The third line in Equation (32) shows the 'Reliability–Refinement' decomposition (Degroot and Fienberg, 1983; Rodwell *et al.*, 2018) of the Brier Score. This appears to be a more useful decomposition for flow-dependent evaluation of forecast systems (Rodwell *et al.*, 2018) than the more traditional three-component decomposition, where the Refinement term is split into 'Resolution' and 'Uncertainty' components. Note that we find considerable cancellation in year-to-year variations of the Resolution and Uncertainty components associated with changes in the base-rate $\overline{o}$.

For a reliable forecast system $\overline{o}_k \approx \overline{p}_k$ and so the refinement term becomes $\approx \sum_{k=0}^{K-1} n_k \overline{p}_k (1 - \overline{p}_k)/N$: a measure of overall sharpness of the forecast system (the smaller the better) and this is improved by increasing the proportion of forecast probabilities close to 0 or 1 (large n_0 and/or large n_{K-1}).

What is of particular interest here is the users' distribution of cost/loss ratios $u(\alpha_j)$ as in Equation (27). Is there, for example, a particular cost/loss ratio that the majority of users share, or are the users' cost/loss ratios uniformly distributed between 0 and 1? Obtaining the 'true' distribution of cost/loss ratios is useful because it allows forecasters to determine the integrated expense $\overline{\overline{E}}_p$ for the user community as a whole (Richardson, 2001). Any differences from a uniform distribution will lead to differences between the UBS and the BS: do improvements to the forecast system as evidenced in the BS (or CRPS) feed through into the UBS (or CUBS), for example?

3 | COMBINING FEELINGS

In the example of planning of a trip to the beach (Table 1), cost C in Table 2 might represent the travel costs (in $) involved, but this hardly embodies the often intangible, subjective feelings associated with the various outcomes. Possible feelings associated with each action/outcome pair are listed in Table 3. For example, a user might feel 'Satisfaction' S if they missed a bad day at the beach. Conversely, missing a nice day at the beach might lead to some 'Regret' R. Experiencing a bad day at the beach might lead to some 'Pain' P. Last but not least, the 'Thrill' T would surely embody the feelings associated with a nice day at the beach. Being highly subjective, the quantification of these feelings is likely to be different for each user. Some might be thinking of taking small children, others might be wishing to swim in the sea or just walk on the beach. The feelings could also embody what alternative plans could be made. Note that, while P,R,S,T might not explicitly represent the cost in $ of travelling to the beach, they could represent the user's *feeling* about this travel cost – which is arguably more relevant. The key motivation of this study is to investigate how well users are able to quantify such feelings and act on them appropriately. A similar list of possible feelings associated with the camping scenario is presented in Table 4. Here, the feelings reflect the higher stakes, with the potential for injury and curtailment of a summer holiday. An insurance aspect is also included. The ways P,R,S,T are associated with each action and outcome pair are summarised in the generalised matrix shown in Table 5.

The names and corresponding letters P,R,S,T have been chosen in an attempt to relate to the likely feelings involved: they seem appropriate for the two scenarios

TABLE 3 Feelings associated with the beach event.

Action	Outcome	
	Not warm and dry	**Warm and dry**
Don't go to beach	Satisfaction S ↑ Missing 'bad' day at beach ↑ Doing something else ↑ No travel costs	Regret R ↑ Missing 'good' day at beach ↓ Doing something else ↓ No travel costs
Go to beach	Pain P ↑ 'Bad' day at beach ↑ Not doing something else ↑ Travel costs	Thrill T ↑ 'Good' day at beach ↓ Not doing something else ↓ Travel costs

Note: ↑ (↓) implies something which increases (decreases) the feeling.

TABLE 4 Feelings associated with the camping event.

Action	Outcome	
	No strong winds	**Strong winds**
Stay	Satisfaction S ↑ Continuing holiday ↑ Gamble pays off ↑ No expensive travel change	Regret R ↑ Major loss or injury ↑ Feeling of responsibility ↓ No expensive travel change ↑ Loss of no-claims discount
Pack up and go	Pain P ↑ Ending of holiday ↑ Feeling of being too cautious ↑ Expensive travel change	Thrill T ↑ Avoiding loss and injury ↑ Feeling right decision made ↓ Expensive travel change ↑ Maintain no-claims discount

Note: ↑ (↓) implies something which increases (decreases) the feeling.

TABLE 5 Generalised matrix of costs or feelings associated with each action and outcome

Action	Outcome	
	No	Yes
No	$-S$	R
Yes	P	$-T$

considered here, but it remains to be seen if they are appropriate more generally. As with any 'pros and cons' list, there is some symmetry between the rows and columns in Tables 3 and 4. The Satisfaction represents opposite feelings to the Pain, and might be similarly quantified. Similarly the Thrill represents opposite feelings to the Regret and might also be similarly quantified.

When making decisions, users will weigh up these various factors and feelings. It is not always clear how this is done but suppose that, for some hypothetical hyperrational user, these factors and feelings can be costed and manipulated like real numbers (or, in the terminology of Murphy, 1966, a user's "negative utilities" associated with each "consequence"). This is a major assumption and the extent to which it might be valid for *real* users will be explored later. With this assumption, our hypothetical user's decision can be based on the standard cost–loss model. This is because Table 5 can then be written as the sum of two terms: the standard cost–loss matrix (Table 2) which depends on both the user's action and the outcome, and a second matrix which is independent of the user's action:

$$\begin{Bmatrix} -S & R \\ P & -T \end{Bmatrix} = \begin{Bmatrix} 0 & L \\ C & C \end{Bmatrix} + \begin{Bmatrix} A & B \\ A & B \end{Bmatrix}, \tag{34}$$

where $C = P+S$, $L = R+T+P+S$, $A = -S$, and $B = -T-P-S$. Note that in the term on the right, if the event does not happen, the cost A is incurred regardless of the action. Similarly, if the event happens, the cost B is incurred regardless of the action. Hence the second matrix on the right of Equation (34) is indeed independent of the action and will have no bearing on our hypothetical user's decision. Evidently, even in this more generalised framework of costs, our hypothetical user's decision can be represented by the standard cost–loss model, with cost/loss ratio:

$$\alpha = \frac{C}{L} = \frac{P+S}{R+T+P+S}. \tag{35}$$

C and L may not look like the traditional cost and loss but, if used in the standard cost–loss model, they would define the appropriate Bayes Action (also Roebber and Bosart, 1996; Granger and Pesaran, 2000). As in Equation (4), the action-dependent expense (per unit L) is:

$$E = \alpha\mathbb{I} + o(1 - \mathbb{I}). \tag{36}$$

From Equation (35) the criterion for non-trivial decisions ($0 < \alpha < 1$, discussed earlier) implies that $-S < P$ and $-T < R$. Hence we see from Table 5 that, for any given outcome, an incorrect forecast will always have a higher expense than that of a correct forecast. In the discussion above, we implied that P,R,S,T were all positive and so these inequalities would clearly be true.

We aim to explore the extent to which Equation (35) and the assumption that a user wishes to make their Bayes Action (Equation (8)) are useful for relating *real* users' professed feelings P,R,S,T to their decisions p_T. If a user has feelings associated with all action/outcome pairs, can they really do the mental work to derive their effective C and L? Can, for example, the "=" in Equation (35) be replaced with "≈", or is Equation (35) simply not valid? We will clearly not be able to answer these questions in any detail, but hope to get a general sense of how the user moves mentally from feelings to decisions.

4 | FORECASTS AND OBSERVATIONS

The ensemble forecasts used within this study are those that were produced operationally at the European Centre for Medium-Range Weather Forecasts (ECMWF) from 1995 to 2018 (24 years in total). For consistency over this period, we are limited by the reduced forecasting and 12-hourly archiving practices in the earlier period. Hence we only consider ensemble forecasts initiated at 1200 UTC each day. There are 32 ensemble members until 9 December 1996, and 50 members thereafter.

In order to assess forecast performance and improvement from the users' perspective, the observations used are *in situ* point observations (known as SYNOP observations). Specifically, we use point observations of temperature 2 m above the ground, windspeed 10 m above the ground, and precipitation accumulated over a 24 hr period. The corresponding forecast probabilities are based on the same parameters predicted at the nearest model grid-point. Although forecasts can be calibrated through post-processing to reduce biases and improve scores, we use the raw forecast data and the probabilities are simply based on the fraction of ensemble members predicting a given event. This makes it easier to diagnose problems and monitor progress in the underlying forecast system – which is what, ultimately, we wish to improve. In addition, we note that the computed expense associated with decisions based on the operational forecast (Equation (11)) is likely to be inflated due to errors in the verifying observations. This means, for example, that the relative value of the operational forecast (Equation (20)) is likely to be underestimated.

Since there are relatively few SYNOP observing sites (e.g., around the coast of the UK), we use SYNOP observations from all sites in the northern midlatitudes (between 50° and 60°N) for both scenarios. This gives ~75,000 forecast/observation pairs per season for the beach scenario, and ~114,000 pairs for the camping scenario. Of course, the participants' threshold probabilities and cost/loss ratios for the beach scenario will still reflect their feelings about the specific beach they have in mind, but the scores may be somewhat more generally applicable to the Northern Hemisphere summer (June–August) season. Similarly, while the participants' decisions about the camping scenario will reflect their feelings about a campsite and time of year they have in mind, to ensure a sufficiently high frequency of strong wind situations, scores are based on the Autumn (September – November) season.

For the beach scenario, a warm dry event in an ensemble member requires the temperature to be $\geqslant 20$ °C at forecast lead time 5 days and the accumulated precipitation between lead times 4 and 5 days to be $\leqslant 0.5$ mm. A warm dry event in the observations is similarly configured. For the camping scenario, a strong wind event in an ensemble member requires the windspeed to be $\geqslant 11\ \mathrm{m{\cdot}s^{-1}}$ at a lead time of 1 day.

5 | THE *LIVE SCIENCE* SESSION

5.1 | Audience

The *Live Science* event was staged during a national meeting of the Royal Meteorological Society at ECMWF, Reading, UK. The audience was largely composed of professional and amateur meteorologists of all ages. Results from the on-line participation (four people), and from a smaller event (16 people) with a similar audience have been combined to give a total of 74 active participants. This is less than that of, for example, Morss *et al.* (2010), but it is hoped it is still meaningful for our purposes.

5.2 | Elaborating on the scenario

The audience was presented with each scenario and corresponding weather event, and asked to consider carefully

the factors and feelings which might influence their decision. For the beach scenario, they were asked to consider what they might do at the beach if the weather was warm and dry, how much they would enjoy (or not enjoy) this, how far they live from the beach and what would they do if the weather did not turn out to be warm and dry. They were also asked to consider feelings of regret if it turned out that they had missed a nice day at the beach. For the camping scenario, the audience was asked to consider who they would be with, whether they are experienced campers, whether they have been caught out before, how far they are into their holiday and how far from home they are, whether they have insurance, and how upset would they be to pack up and leave early. These questions required a participant to elaborate mentally on each scenario in a way which might realistically apply to them. While each participant could probably have elaborated in several ways, the hope was that the audience as a whole would sample reasonably well the full range of possible options.

5.3 | Forecasts and actions

The audience was then presented with a range of exploratory forecast probabilities $\{p_j\}$ as in Equation (22) and asked to indicate their decision ('no' or 'yes') to the given action (Table 1, column 4) through use of a "Plickers" card (the orientation of each barcode-like card indicates the participant' decision, which can be interpretted by a camera at the front of the lecture theatre: Chng and Gurvitch, 2018), or on-line. For the beach scenario, the presented probabilities (column 3) for 'nice' weather increased monotonically ($p_j : j = 0, \ldots, K-1$). The rationale here is that this was likely to be a non-critical decision, that users would go to the beach only at reasonably high probabilities, and thus they would have more time to weigh up the pros and cons as the probabilities rose. Note that results from Morss *et al.* (2010) suggest that a random presentation of probabilities is not necessary. For the camping scenario, the presented probabilities for strong winds decreased monotonically ($p_j : j = K-1, \ldots, 0$). The rationale here is that this is potentially a catastrophic event, where users might be expected to only stay at small probabilities and hence, again, they would have more time to consider the key costs involved before changing their decision. Moreover, probability intervals were reduced as the probability approached zero in order to better resolve the potentially complex decision processes which might take place for high-loss–low-probability situations. A user's threshold probability p_T was deduced as in Equation (23) where p_j is the lowest probability for which they took action 'yes'. Participants who always or never took action 'yes' were discounted as they are effectively indicating that the forecast is of no use to them for this particular decision (as discussed above). The distribution of users' threshold probabilities, along with the relative value to individual users and comparison with the Brier Score (as discussed in Section 1) were presented to the audience and discussed after all probability levels had been presented.

Immediately after the event (in most cases the next day) users were invited to provide more information via email about the key factors which influenced their decisions. The aim with this questioning was to determine how well each user's threshold probabilities relate to their own feelings associated with the two scenarios. We obtained 17 sets of responses – sufficient for a preliminary assessment.

5.4 | *Vox pop* of the general population

Since the audience at the *Live Science* event was largely composed of individuals with a specific interest in meteorology, a small *vox pop* of the public in the market town of Amersham in rural England was also conducted for the same two scenarios. It involved a small sample of 11 people and a more limited set of probability levels, but it nevertheless provides a useful means of assessing the general applicability of the results obtained within the event itself. Of course, we acknowledge that Amersham itself is not fully representative of society as a whole.

6 | RESULTS

6.1 | Beach scenario

The bar-chart in Figure 2a shows, for the beach scenario, the users' distribution of threshold probabilities. As anticipated, it can be seen that few participants would go to the beach at very low probabilities of nice weather. The most popular threshold is centred on 70%. There is then a fast drop-off towards 100%, indicating that most participants do not wait for very high probabilities to decide to go.

A key question is whether participants' threshold probabilities really do relate well to the factors they identify as important in their decision. Table 6 highlights, for a range of threshold probabilities, a representative sample of participants and the key factors and decision strategies they cited. It does appear that participants are weighing up their feelings with the aim of choosing their Bayes Action. An interpretation in terms of P,R,S,T values is presented in the last column. The key factors highlighted appear to be associated with the Pain and Thrill categories. Few participants highlighted feelings associated with Satisfaction and Regret (although we might infer that these are similar to

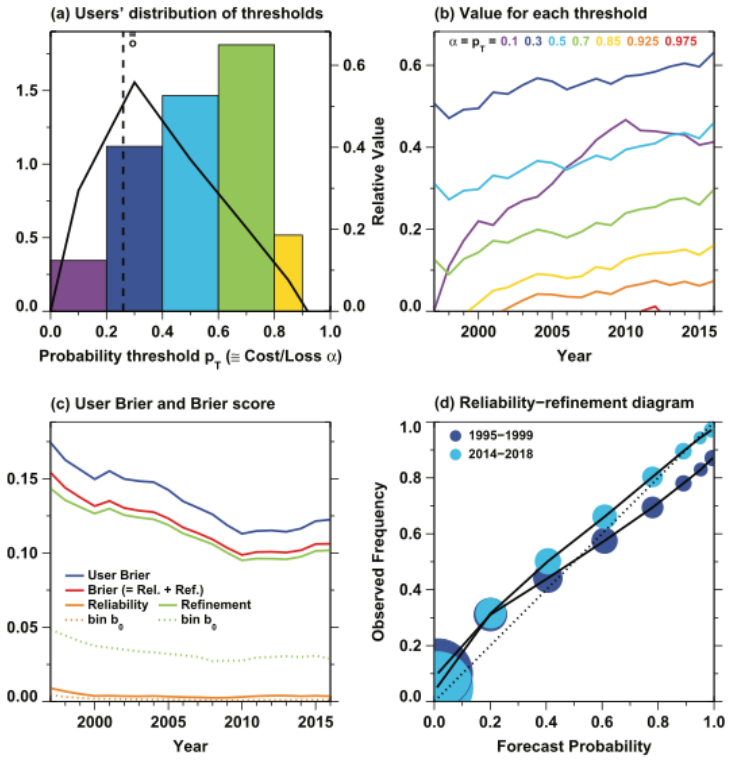


FIGURE 2 Results for the beach scenario. (a) The distribution of users' threshold probabilities p_T (bars, left axis) with the mean relative value over the 24 years (black curve, right axis); the 24-year event frequency $\bar{o}$ is indicated by the vertical dashed line. (b) Timeseries of the relative value of forecasts for each cost/loss ratio considered. A 5 year running-mean has been applied and the central year is indicated on the x-axis. (c) The User Brier Score based on the distribution of participants cost/loss ratios (blue), the Brier Score (red), and the Reliability (orange solid) and Refinement (green solid) components of the Brier Score. The contributions to the Reliability and Refinement associated with the lowest probability bin are shown dashed. A similar 5-year running mean has been applied. (d) The Reliability–Refinement diagram showing observed frequency against mean forecast probability for each bin. The areas of the circles indicate the fraction of forecasts within each bin for 1995–1999 (dark blue) and 2014–2018 (light blue)

the Pain and Thrill, respectively, as discussed above). Taking into account cancellations in Equation (35) of various feelings displayed in Table 3, we might then interpret the cost/loss ratio, for this (pleasurable) beach scenario, as

$$\alpha_{\text{pleasure}} \approx \frac{P}{T+P} \approx \frac{\text{bad beach day} + \text{miss something else} + \text{travel costs}}{\text{good beach day} + \text{bad beach day}}. \quad (37)$$

Table 6 might suggest a slight reticence of participants to go to the beach. For example, the comment for $p_T = 50\%$: "Going to the beach is about walking and watching the waves, so I don't mind as long as it's above 10 °C and not too wet". One might think that this participant would be prepared to go to the beach at probabilities lower than 50%. However, it could be the case that they actually require a higher probability of, say, 80% that the temperature will be above 10 °C and they are mentally equating this to a 50% probability of above 20 °C. This would argue for more tailorable forecast products. It might also be that this is a non-critical event and the participant is simply waiting for higher probabilities than they would actually settle for: "I could always go another day", for example. Subject to such qualifications, it would appear that users are able to make appropriate use of the forecast probabilities, and that feelings can be combined, semi-quantitatively, as in Equation (37). Hence we feel that the distribution of threshold probabilities is a reasonable indication of the users' true distribution of cost/loss ratios (Equation (27)). From the users' responses, it was also noted that eight participants reversed their decision at high probabilities to avoid a crowded beach. In addition, twelve participants indicated that they would always go to the beach, regardless of the forecast. Interestingly, all these aspects are represented in the *Vox pop* results (Table 7), suggesting that they are not specific to the meteorologically minded audience in the *Live Science* event.

The black curve in Figure 2a shows the relative value of all the forecasts (from all years). As discussed in Section 2.3, this should be maximum at a cost/loss ratio $\alpha = \bar{o}$ (≈0.26: the climatological frequency for this beach weather event). Bearing in mind the limited set of cost/loss ratios considered here, this is indeed the case.

TABLE 6 Users' replies when asked to identify the key factors that influenced their decision for the beach scenario

p_T	Participants' comments on key factors (paraphrased)	*P,R,S,T* interpretation
0%	I love being on beaches, whatever the weather. I live 75 miles from the beach so creating an opportunity is rare. Weather would never dissuade me.	Low pain despite the distance, high thrill. The forecast is of no use in decision to go to the beach, but might help them decide what to wear.
10%	Anything over 17 °C is perfectly fine for me. 0.5 mm rain is not much and could fall overnight anyway. However, 0% could imply potential for very unpleasant weather.	Low pain, moderate thrill.
30%	I live near a beach and go a lot. My decision is not often related to the forecast. For probabilities above 40%, I'd be more likely to go than do something else.	Low pain since lives close, small regret due to lack of alternatives. The forecast actually has its maximum relative value at this probability.
50%	Going to the beach is about walking and watching the waves, so I don't mind as long as it's above 10 °C and not too wet.	Moderate pain, moderate thrill.
70%	I hate sitting on the beach in the rain. We also live a long way from a beach and with three kids it's quite an expedition.	High pain, moderate-to-low thrill.
70%	For me, 20°C is the bare minimum to be able to enjoy the beach. I really like it warm.	Very high pain, moderate thrill. A warmer event definition would be more appropriate.

TABLE 7 *Vox pop* for the beach scenario

	Respondent										
p	1	2	3	4	5	6	7	8	9	10	11
20%	0	1	1	1	1	1	0	0	1	0	0
40%	1	1	1	1	1	1	0	0	1	1	1
80%	1	1	1	1	1	1	1	1	0	1	1
90%	1	1	1	1	1	1	1	1	0	1	1

Note: Shown are the respondents' chosen actions $\mathbb{I}_p$ when given forecast probability p for warm and dry weather. $\mathbb{I}_p = 0$ means 'don't go to the beach', $\mathbb{I}_p = 1$ means 'go to the beach'.

Figure 2b shows how the relative value (Equation (20)) of forecasts for each cost/loss ratio considered improved over the period 1995–2018. When computing the forecast relative value for each year, we use the 24-year mean event frequency $\overline{\overline{o}}$ instead of $\overline{o}$ in the calculation of the expense based on climatology. This avoids overfitting to the seasonal anomaly. To make the curves smoother, a 5-year running mean has been applied: for example, for 1997, the overline in Equation (20) corresponds here to the mean over June–August 1995–1999. Perhaps the most striking change has been the rise in relative value for users with the lowest non-zero cost/loss ratio considered for this scenario ($\alpha_1 = 0.10$). This is largely associated with missed events in the early years, and is discussed in more detail in relation to the camping scenario where the feature is more pronounced. Progress has also been made for all other cost/loss ratios. The relative value for $\alpha = 0.30$ is 0.6 at the end of the period: this means that the forecast, even at day 5 and for the bivariate (temperature and precipitation) weather event being considered here, is 60% as good as a perfect deterministic forecast. This is despite the fact that relative value is likely to be underestimated due to the existence of errors in the verifying observations. Unfortunately, for many users, the relative value at $\alpha = 0.70$ (the most popular cost/loss ratio) is somewhat less because this cost/loss ratio is far from the observed frequency. For the highest cost/loss ratios considered (e.g., α =0.925 and 0.975) strong feelings of 'pain' mean that the user rarely goes to the beach: a similar set of decisions to those based on the climatological forecast (never go to the beach) and thus with similar expense, and little relative value.

Figure 2c shows that the User Brier Score for the user community as a whole (blue curve) is somewhat larger than the Brier Score (red curve) derived as in the first line of Equation (32). This is partly because the distribution of users' cost/loss ratios for this scenario places less weight on high values of α (Figure 2a) than does the uniform distribution, and so does not benefit from the small relative expense near $\alpha = 1$, where $\mathbb{I}_p = 0$ and $\overline{E}_p - \overline{E}_o = 0$; Equations (11) and (18). Note that, somewhat fortuitously, for this scenario $\tilde{\alpha} = 0.52 \approx 1/2$ and the denominator in Equation (26) varies interannually within

the range (0.504, 0.511), so is approximately constant and does not contribute to the difference seen with the BS.

The orange and green curves in Figure 2c show the reliability and refinement components of the BS derived as in the third line of Equation (32). We see an improvement (i.e., a drop) in the reliability curve prior to about the year 2000, but the main improvement has been associated with refinement. In essence, this relates to better (sharper) initialisation of the ensemble in the latter years due to improvements in observations and data assimilation as well as to the underlying forecast model and representation of model uncertainty.

Figure 2d shows a Reliability–Refinement diagram for the first and last 5-year periods 1995–1999 (dark blue) and 2014–2018 (light blue), with observed frequency $\overline{o}_k$ plotted against the mean forecast probability in each bin $\overline{p}_k$ as in Equation (1). The area of the circles is proportional to n_k/N: the fraction of forecasts in each bin. The circles for the later period are generally closer to the diagonal than for the early period, indicating an improvement in reliability. The large size of the circle for the lowest probability bin (with $\overline{p}_0 \sim 0.009$) reflects the fact that it contains ~58% of all forecasts. Partly because of its size, it has the largest impact on the improvements to overall reliability and refinement. (The components for this bin alone are shown with the dotted orange and green curves in Figure 2c.) Other probability bins give similar improvements to refinement as that of bin 0 reliability.

The bin containing the forecasts with highest probabilities (with $\overline{p}_7 \sim 0.991$) has a much smaller bin-width, and there is actually a secondary peak in the distribution of issued forecast probabilities there. Since sharpness (a propensity for a forecast system to issue probabilities close to 0 and 1) is desirable, this secondary peak is welcome. However, the system clearly also issues probabilities between these two extremes. The extent to which these can be avoided in future depends on the ultimate limits of 5-day predictability set by the chaotic dynamics of the real world.

6.2 | Camping scenario

For the camping scenario, the distribution of threshold probabilities (Figure 3a) is more complicated than that of the beach scenario but, broadly speaking, it is backed up by the results of the *Vox pop* in Table 9. Participants interpreted the risk in very different ways. As Table 8 shows, those with high threshold probabilities were sometimes more concerned about leaving the holiday due to the distance home (high Pain) while others saw the risk as a challenge to overcome (and thus effectively had low Regret). For the central threshold probabilities, the consequences of the potential aftermath with young or old people in the group (high Regret) needed to be balanced against the responsibility of spoiling a nice holiday (high Pain). The low threshold probability users cited a dislike of camping (low Pain in leaving) along with high Regret if they were to get caught in strong winds. The option by some of imagining a car to jump into also led to different decision strategies. Again, we infer that participants are weighing up factors in an attempt to make their Bayes Action. Clearly Pain and Regret were the main decision drivers in this scenario. Taking into account cancellations in (35) of various feelings displayed in Table 4, and neglecting the insurance aspect which no participant highlighted, it is tempting to try to interpret the cost/loss ratio for this (dangerous) camping scenario as

$$\alpha_{\text{danger}} \approx \frac{P}{R+P}$$

$$\approx \frac{\text{end holiday} + \text{too cautious} + \text{travel costs}}{\text{injury} + \text{responsible} + \text{end holiday} + \text{too cautious}}. \tag{38}$$

However, as Table 9 suggests, some users struggle with decisions involving high potential losses, and do not take action even at high probabilities. A lack of first-hand experience in dealing with very rare and extreme events is probably both inevitable and problematic.

Being a more extreme weather situation, with a much lower climatological frequency $\overline{\overline{o}} = 0.039$ (and hence a smaller sample of events), the relative value curves shown in Figure 3b are not quite so monotonic as for the beach scenario, but they do generally increase over the 24-year period. The maximum relative value occurs for $\alpha_2 = 0.075$, which is close to $\overline{\overline{o}}$. This maximum relative value of ≈0.6 is similar to that for the beach scenario. However, here the forecast lead time is 1 day rather than 5 days.

There is a very pronounced rise in relative value for users with the lowest non-zero cost/loss ratio considered ($\alpha_1 = 0.025$). In the early 1995–1999 period, $\overline{E}_p(\alpha_1) > \overline{E}_{\overline{o}}(\alpha_1)$ and hence $V(\alpha_1) < 0$ (i.e., negative relative value which, from the Appendix, suggests a lack of reliability). The main cause is the high number of "complete misses". The mean probability for forecasts falling into bin b_0 is $\overline{p}_0 = 0.00015$ while $\overline{o}_0 = 0.028$. With the number of ensemble members being ⩽50, this means that at least 99% of such forecasts have no members predicting the event (hence the term "complete misses"). Because n_0 is so large (94% of forecasts fall into bin b_0), the contribution to $\overline{E}_p(\alpha_1)$ in Equation (25) from forecasts in bin b_0 (the contribution with $k = 0$ in the last term on the first line, which corresponds to no action taken but the event occurred) is large. This leads to the negative relative value. As α increases (larger j), this last term tends towards

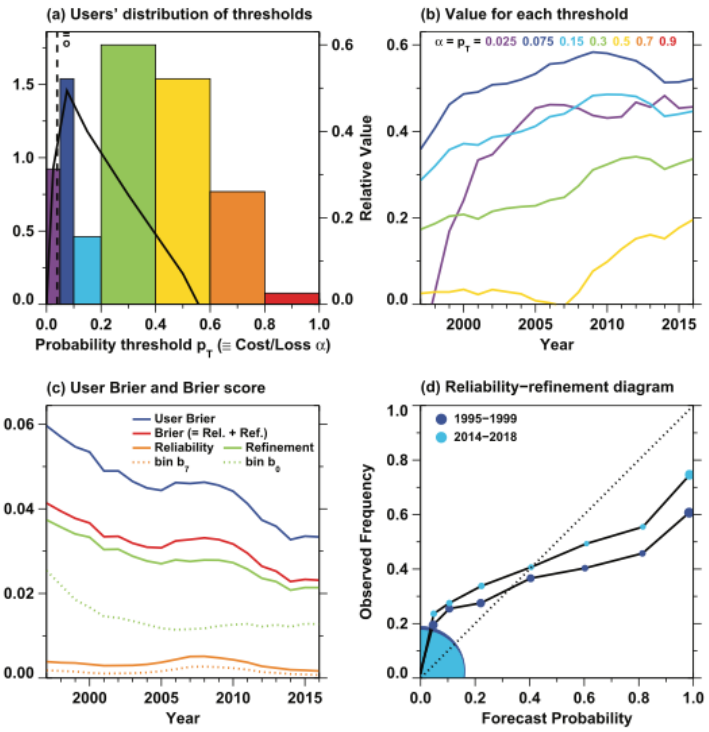


FIGURE 3 As Figure 2, but for the camping scenario. In (c) the Reliability contribution from the highest probability bin is shown (orange dashed) since this is largest

$\overline{o}$ (regardless of the forecast quality) and so the effect of complete misses diminishes. By the later 2014–2018 period, the frequency of complete misses has been halved, and this explains the sharp rise in $V(\alpha_1)$ from ≈0 to ≈0.45: a result which should be very welcome to users who do prefer to avoid danger at low probabilities.

Although the users' distribution of threshold probabilities may not perfectly represent the feelings that the users *should* have in a potentially dangerous situation, we nevertheless compute the UBS. Figure 3c shows that the UBS (blue curve) is much larger than the BS (red curve). Again, this is partly because the distribution of users' cost/loss ratios for this scenario places less weight on high values of α (Figure 3a) than does the uniform distribution, and so does not benefit from the small relative expense near $\alpha = 1$. For this scenario, however, the mean of the users' distribution of cost/loss ratios is $\tilde{\alpha} = 0.40 \neq 1/2$ and interannual variations of the denominator in Equation (26) range within $[0.401, 0.407]$. Since these values are less than 1/2, they will tend to increase the UBS relative to the BS. Since the denominator is fairly constant, the asymptotic limit for UBS propriety has probably been reached. Notice that the UBS improves more rapidly than the BS over the 24-year period. This is because participants place more weight than the uniform distribution on the intermediate values of α (Figure 3a), where the larger relative expenses are decreasing more rapidly.

The BS for this weather event also suffers from complete misses. This issue largely enters the BS via the Refinement term for bin b_0: $k = 0$ in the third line of Equation (32), as seen in Figure 3c; green dotted curve. In the early years, 62% of the BS is associated with the bin b_0 refinement contribution alone. 70% of the subsequent improvement in the BS is associated with improvements in this refinement term too. Over the 24-year period, there is an increase in the fraction of forecasts falling into bin b_0. While this might partly reflect a reduction in the base-rate $\overline{o}$ over this period (not investigated here), it is seen (Figure 3d) that $\overline{o}_0$ actually reduces: with the large light blue circle for (2014–2018), near forecast probability 0, being lower (and closer to the diagonal) than is the large dark blue circle for (1995–1999). The difference might not look much, but the size of the bin makes this important. This improvement is also seen in the bin b_0 reliability contribution (Figure 3c; orange dotted curve), but it is much smaller because it is essentially quadratic in $\overline{o}_0$ while the refinement contribution is essentially linear in $\overline{o}_0$. Notice that Figure 3d shows a general improvement

TABLE 8 Users' replies when asked to identify the key factors that influenced their decision for the Camping scenario

p_T	Participants' comments on key factors (paraphrased)	*P,R,S,T* interpretation
90%	I chose what for me would be the most likely camping scenario, which is a long way from home so that a return home is not an option. In this case I was fairly prepared to ride out the risk of a summer storm.	High pain due to need to travel a long distance. No discussion of Regret.
70%	I don't really go camping. If I'm already there, I may as well stay as long as possible. A case of making it an adventure with the family pulling together to stop the tent being blown away.	Low Regret due to an appetite for adventure (and lack of experience?)
50%	A 50% chance of strong winds might only be a 5% chance of the tent blowing down, and even smaller chance of serious harm. With a low probability, I'd feel responsible for taking away my family's fun. However, as a parent, I wouldn't want to put very young kids at risk of flying branches. If it had been a very high probability and I hadn't done anything I'd feel responsible.	High Pain associated with spoiling the family's holiday. High Regret associated with a small probability of serious harm.
30%	I have experienced this before and it is better to leave before the tent does!	We can perhaps trust this user to have a good appreciation of the potential Pain involved.
8%	I am not a keen camper. I would struggle to deal with elderly or young people if things were collapsing around us. I thought that 5% means 'highly improbable'.	Little pain due to dislike of camping. High regret. The forecast has its maximum relative value at this probability.
3%	I can only stand camping if I am sure that it is sunny.	Very low pain due to dislike of camping. Regret not mentioned.

TABLE 9 *Vox pop* for the camping scenario. Respondents' chosen actions $\mathbb{I}_p$ are shown for a given forecast probability p for strong winds.

	Respondent										
p	1	2	3	4	5	6	7	8	9	10	11
20%	1	0	0	0	1	0	0	0	0	0	0
40%	1	1	0	0	1	0	0	—	0	1	1
80%	1	1	1	1	1	0	1	1	0	1	1
90%	1	1	1	1	1	1	1	1	1	1	1

Note: $\mathbb{I}_p = 0$ means stay, $\mathbb{I}_p = 1$ means leave, and "—" means no response.

in reliability; the improvement in the highest probability bin b_7 is particularly important. The drop in intermediate bin sizes (circle areas) indicates a general increase is sharpness.

It is clear that low-probability events can cause problems for forecasters. A simple calibration of all 0% probabilities to 2.8% for a life-threatening event without any dynamical reasoning might provoke the 'crying wolf' effect. Clever fitting of distributions could do better. However, it is clearly desirable to improve the reliability of the raw ensemble forecast as this can represent flow-dependent aspects. Such improvement appears to have happened over this 24-year period. Future improvements might benefit from a closer monitoring of the contributions to forecast scores, such as the terms represented in the first line of Equation (25) and the central line of Equation (32).

Knowledge of key user priorities might also inform forecast configuration. For example, if users have primarily low cost/loss ratios, where actions are taken at low probabilities, then this might shift the emphasis towards a better sampling of the predicted state space, and thus a prioritisation of computing resources on an increased ensemble size (Palmer, 2002; Leutbecher, 2019).

7 | DISCUSSION AND CONCLUSIONS

In his investigation into the elicitation of personal probabilities Savage (1971) suggested that some commodity traders do know their costs and losses, but can be reluctant to reveal them to experimenters in case this information is used against them. Other difficulties including the nonlinearity of the decision maker's utility

of money, risks of bankruptcy, changes in their preferences within a sequence of similar proposals, and situations where the user might be (irrationally) "risk-averse" (Roebber and Bosart, 1996) also cause complications for experimenters. Although detailed discussions with selected users have led to estimates of their individual cost/loss ratios (Kolb and Rapp, 1962; Murphy, 1977; Roebber and Bosart, 1996; Thornes and Stephenson, 2001), these do not give the complete distribution for the user community as a whole. Faced with an incomplete knowledge of the user's (or users') costs and losses, Murphy (1966) proposed that forecasters estimate expected utility based on a uniform distribution of the cost/loss ratios, while Murphy (1969) suggested the more flexibly defined Beta distribution. Roebber and Bosart (1996) highlighted considerable sensitivity of the overall value of forecasts to the prescribed shape of the Beta distribution. Palmer (2002) hypothesised that users might self-select based on the relative value of the forecasts to each of them individually, so that the users' distribution of cost/loss ratios might be strongly peaked around $\alpha = \bar{o}$ where relative value is maximised. The "Diagonal Score" of Ben Bouallègue *et al.* (2018) takes this to the limit by employing a Dirac delta function at $\alpha = \bar{o}$. While it could be argued that users would still use a forecast even if its relative value was somewhat less than the maximum, the "Diagonal Score" may represent an improved user-oriented summary of forecast performance compared to the CRPS.

Here, from our *Live Science* event, and for scenarios which are more concerned with subjective feelings than hard cash, we have had some success in eliciting personal probabilities, and have been able to estimate the users' distributions of cost/loss ratios directly from their intended actions. (Morss *et al.*, 2010 describe a similar study of the American public.)

In order to engage with the users, we chose scenarios and associated weather events which are of direct concern to users. Rather than scoring Northern Hemisphere temperatures at 850 hPa, for example, we scored probabilistic forecasts of a bivariate 'user-oriented variable' (UOV) involving rainfall and temperatures at a point – which could represent the location of the user's towel on the beach! We also scored an extreme near-surface wind event, which is above the 96th percentile for point observations. Recently the Predictability, Dynamics and Ensemble Forecasting (PDEF) working group of the World Meteorological Organization has proposed more use of such UOVs as a means of engaging better with users and allowing forecasters to learn more about how their forecasts are used and calibrated. It will be important to use proper scores to evaluate forecast performance for such UOVs.

If a user decides on the 'Bayes Action' (the action which minimises their expected expense based on the forecast probability), then their expense represents a proper score of the forecast (Dawid, 2007; Gneiting and Raftery, 2007). In this way, the aims of users and forecasters are seen to be aligned. Scores of utility can be expected to complement rather than conflict greatly with proper scores of more traditional flow variables when it comes to deciding on forecast system developments. Here we have introduced a 'User Brier Score', which is (asymptotically) proper. This is better tailored to the users' distribution of cost/loss ratios. By focusing less on high cost/loss ratios (in the scenarios discussed), it indicates larger errors than does the Brier Score. Such a score could be effective in user-oriented comparisons of forecast systems. It could also be integrated over event thresholds in a similar way to the Continuous Rank Probability Score.

The first scenario we investigated involved a trip to the beach with the potential for warm, dry weather. The key factors cited as important in the decision-making process can be characterised as the 'pain' of a bad day at the beach and the 'thrill' of a good day at the beach. Some participants appeared to wait for slightly higher probabilities than might be indicated by their expressed feelings but, from some responses, it was clear that this can be due to the event definition not perfectly matching their criteria for going to the beach: this argues for an ability to tailor event definitions to individual users. With this reasoning we conclude that, for this scenario, the users are attempting to make their Bayes Action and that the distribution of threshold probabilities represents reasonably well their various feelings of 'pain' and 'thrill'. A potential complication here is the possibility that a user might change their preferences during the season – for example if they have already had a few good days at the beach, the thrill of another day might be diminished. Perhaps a bigger concern is the representativeness of the *Live Science* audience for the population as a whole. Nevertheless, the distribution we have obtained represents a proof-of-concept and a first estimate which may be of use to forecasters wishing to tailor their advice to users. In this case, for example, the most popular threshold probability is 0.7, and so a forecaster might recommend a beach trip if the forecast probability for warm dry weather exceeds 0.7. For this user-relevant weather event, it was encouraging to see the forecasters' goal being realised: improvements to refinement while maintaining or improving reliability.

The second scenario involved a risk of destructive winds during a camping holiday. The key factors cited as important in the decision-making process can be characterised as the 'pain' of ending a holiday early and the 'regret' of experiencing major loss or injury. In this

scenario, the emphasis different users placed on these feelings appeared to relate well to their threshold probabilities. However, some inexperienced campers did not appear to appreciate the risks they were taking. Morss *et al.* (2010) indicate a 'risk-seeking' bias in their respondents which could also be the case here. Since the feedback during the *Live Science* event was only at the end, we did at least avoid the concept of 'reinforcement learning' (Millner, 2009): "I didn't get injured by that low probability extreme event last time, so I'll make the same decision again!" For such scenarios, it is possible that more direction from forecasters and authorities might help the user make a better decision. It is not only campers that had difficulties since a clear issue of "complete misses" (zero forecast probability, but non-negligible outcome frequency) was identified in the forecasts. This is evident in the relative value of forecasts for low cost/loss ratios and in a decomposition of the Brier Score – particularly the refinement aspect at low forecast probabilities. Monitoring of such aspects would seem useful for informing better user-oriented decisions in system development. Model and assimilation improvements seem to have helped reduce the frequency of complete misses over the period 1995–2018 – something which should be of great value to many users who wish to avoid danger even at low probabilities. Further such improvements and forecast calibration (post-processing) might help. Possibly the users' interest in low probability–high impact events might be a strong argument for an increase in ensemble size.

We conclude with a few more general questions raised by this study

- How best do scientists communicate risk/probability when, ultimately, the audience's decisions from the scientific information they receive are put through their own personal 'filter' of subjective experiences/emotions?
- Moreover, should presenters impose value judgements in their messaging, with the intention of helping their audience's decision-making? In the context of this study, we have observed that the statement: "A good day for the beach", may mean one thing for some but a different thing for others.
- Is there a strong case that technology can now allow audiences to choose their own focused thresholds of sensitivity to various weather parameters, a basis from which notifications can be personalised?
- We acknowledge that our sample of scientists, weather enthusiasts and shoppers in leafy Buckinghamshire will not have fully sampled society in general. Is there a greater role for social science to help understand different audience's aspirations? This may then help forecasters steer resources towards optimising forecast configuration.

ACKNOWLEDGEMENTS

The authors would like to thank Tilmann Gneiting, Zied Ben Bouallègue and David Lavers for helpful discussions about this work. We thank Liz Bentley, Marcia Spencer, Sylvia Knight and Richard Parsons from the Royal Meteorological Society for assisting in the hosting of the *Live Science* event. Thanks also to Saleh Abdalla, Seonaid Anderson, James Cosgrove, Richard Forbes, Anna Ghelli, Peter Gibbs, Brian Golding, Sue Gray, Tim Hewson, Ken Mylne, Philip Newton, Christel Prudhomme, Florence Rabier, Nigel Roberts, Alex Sherred, and all the *Live Science* participants. Finally, we would like to thank the reviewers of this study for their diligence and insightful comments which helped improve this paper.

ORCID

M. J. Rodwell https://orcid.org/0000-0001-5986-5218

REFERENCES

Ångström, A. (1922) On the effectivity of weather warnings. *Nordisk Statistisk Tidskrift*, 1, 394–408

Ben Bouallègue, Z., Haiden, T. and Richardson, D.S. (2018) The diagonal score: definition, properties, and interpretations. *Quarterly Journal of the Royal Meteorological Society*, 144, 1463–1473

Ben Bouallègue, Z., Magnusson, L., Haiden, T. and Richardson, D.S. (2019) Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, 145, 1741–1755

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3

Chng, L., Gurvitch, R. (2018) Using Plickers as an Assessment Tool in Health and Physical Education Settings. *Journal of Physical Education, Recreation & Dance*, 89, 19–25. http://dx.doi.org/10.1080/07303084.2017.1404510.

Dawid, A.P. (2007) The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59, 77–93

Degroot, M.H. and Fienberg, S.E. (1983) The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32, 12–22

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016) Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 505–562. https://doi.org/10.1111/rssb.12154

Fundel, V.J., Fleischhut, N., Herzog, S.M., Göber, M. and Hagedorn, R. (2019) Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quarterly Journal of the Royal Meteorological Society*, 145, 210–231

Gandin, L.S. and Murphy, A.H. (1992) Equitable skill scores for categorical forecasts. *Monthly Weather Review*, 120, 361–370

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29, 411–422

Granger, C.W.J. and Pesaran, M.H. (2000) Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19, 537–560

Haiden, T., Janousek, M., Bidlot, J.R., Buizza, R., Ferranti, L., Prates, F. and Vitart, F. (2018). Evaluation of ECMWF forecasts, including the 2018 upgrade. Technical Memorandum 831, ECMWF, Reading, UK.

Kolb, L.L. and Rapp, R.R. (1962) The utility of weather forecasts to the raisin industry. *Journal of Applied Meteorology*, 1, 8–12

Leutbecher, M. (2019) Ensemble size: how suboptimal is less than infinity?. *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 107–128. https://doi.org/10.1002/qj.3387

Liljas, E. and Murphy, A.H. (1994) Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bulletin of the American Meteorological Society*, 75, 1227–1236

Millner, A. (2009) What is the true value of forecasts?. *Weather Climate and Society*, 1, 22–37

Morss, R.E., Lazo, J.K. and Demuth, J.L. (2010) Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication. *Meteorological Applications*, 17, 149–162

Murphy, A.H. (1966) A note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation. *Journal of Applied Meteorology*, 5, 534–537

Murphy, A.H. (1969) Measures of the utility of probabilistic predictions in cost–loss ratio decision situations in which knowledge of the cost–loss ratios is incomplete. *Journal of Applied Meteorology*, 8, 863–873

Murphy, A.H. (1973) Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12, 215–223

Murphy, A.H. (1977) The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Monthly Weather Review*, 105, 803–816

Murphy, A.H. and Epstein, E.S. (1967) A note on probability forecasts and "hedging". *Journal of Applied Meteorology*, 6, 1002–1004

Palmer, T.N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128, 747–774

Peirce, C.S. (1884) The numerical measure of the success of predictions. *Science*, 4, 453–454

Richardson, D.S. (2000) Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126, 649–667

Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127, 2473–2489

Rodwell, M.J. (2011) On Peirce's motivation for equitability in forecast verification. *Monthly Weather Review*, 139, 3667–3669

Rodwell, M.J., Richardson, D.S., Parsons, D.B. and Wernli, H. (2018) Flow-dependent reliability: a path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, 99, 1015–1026

Roebber, P.J. and Bosart, L.F. (1996) The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather and Forecasting*, 11, 544–559

Savage, L.J. (1971) Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783–801

Schervish, M.J. (1989) A general method for comparing probability assessors. *Annals of Statistics*, 17, 1856–1879

Stephenson, D.B., Coelho, C.A.S. and Jolliffe, I.T. (2008) Two extra components in the Brier Score decomposition. *Weather and Forecasting*, 23, 752–757

Thompson, J.C. (1952) On the operational deficiencies in categorical weather forecasts. *Bulletin of the American Meteorological Society*, 33, 223–226

Thornes, J.E. and Stephenson, D.B. (2001) How to judge the quality and value of weather forecast products. *Meteorological Applications*, 8, 307–314

Winkler, R.L. and Murphy, A.H. (1968) "Good" probability assessors. *Journal of Applied Meteorology*, 7, 751–758

Youden, W.J. (1950) Index for rating diagnostic tests. *Cancer*, 3, 32–35

How to cite this article: Rodwell MJ, Hammond J, Thornton S, Richardson DS. User decisions, and how these could guide developments in probabilistic forecasting. *QJR Meteorol Soc*. 2020;146:3266–3284. https://doi.org/10.1002/qj.3845

APPENDIX: PROPERTIES OF EXPECTED RELATIVE VALUE

Here we derive some useful properties of the expected relative value $\mathbb{E}(V)$ for reliable forecast systems when $0 < \alpha < 1$ and $0 < \mathbb{E}(o) < 1$ and the user is assumed to take their Bayes Action under all forecast probabilities $(p, \mathbb{E}(o), o)$. Note that we think of V as a "collective skill score" (Murphy, 1973) so that

$$\mathbb{E}(V) \equiv \frac{\mathbb{E}(E_{\mathbb{E}(o)} - E_p)}{\mathbb{E}(E_{\mathbb{E}(o)} - E_o)}. \tag{A1}$$

To obtain the expected relative value, we average over a continuous distribution of forecast probabilities $g(p)$. We will consider the two possibilities for $\mathbb{E}(E_{\mathbb{E}(o)})$ in Equation (14). Firstly, for $\alpha \geqslant \mathbb{E}(o)$, we have $\mathbb{E}(E_{\mathbb{E}(o)}) = \mathbb{E}(o)$. Together with $\mathbb{E}(E_p)$ from Equation (5) and $\mathbb{E}(E_o)$

A6. Published Article: On the ROC Area of Ensemble Forecasts for Rare Events

This appendix contains the published version of the following paper (summarised in Section 6.5.2):

Bouallègue, Z. Ben and Richardson, D.S. (2022) 'On the ROC Area of Ensemble Forecasts for Rare Events', *Weather and Forecasting*, 37(5), pp. 787–796. Available at: https://doi.org/10.1175/WAF-D-21-0195.1.

On the ROC Area of Ensemble Forecasts for Rare Events

ZIED BEN BOUALLÈGUE[a] AND DAVID S. RICHARDSON[a,b]

[a] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*
[b] *Department of Geography and Environmental Science, University of Reading, Reading, United Kingdom*

(Manuscript received 2 December 2021, in final form 22 February 2022)

ABSTRACT: The relative operating characteristic (ROC) curve is a popular diagnostic tool in forecast verification, with the area under the ROC curve (AUC) used as a verification metric measuring the discrimination ability of a forecast. Along with calibration, discrimination is deemed as a fundamental probabilistic forecast attribute. In particular, in ensemble forecast verification, AUC provides a basis for the comparison of potential predictive skill of competing forecasts. While this approach is straightforward when dealing with forecasts of common events (e.g., probability of precipitation), the AUC interpretation can turn out to be oversimplistic or misleading when focusing on rare events (e.g., precipitation exceeding some warning criterion). How should we interpret AUC of ensemble forecasts when focusing on rare events? How can changes in the way probability forecasts are derived from the ensemble forecast affect AUC results? How can we detect a genuine improvement in terms of predictive skill? Based on verification experiments, a critical eye is cast on the AUC interpretation to answer these questions. As well as the traditional trapezoidal approximation and the well-known binormal fitting model, we discuss a new approach that embraces the concept of imprecise probabilities and relies on the subdivision of the lowest ensemble probability category.

KEYWORDS: Statistical techniques; Ensembles; Forecast verification/skill

1. Introduction

Over the past decades, the popularity of the relative operating characteristic (ROC) curve has steadily increased with applications in numerous fields (Gneiting and Vogel 2018). In meteorology, verification of weather forecasts based on signal detection theory has been in usage since the seminal works of Mason (1982) and Harvey et al. (1992), and recommended as a standard verification tool by the Word Meteorological Organization in Stanski et al. (1989). In the framework of probabilistic forecast verification, the area under the ROC curve (AUC) is often used as a summary measure of forecast discrimination. Discrimination is the ability to distinguish between event and nonevent and, along with calibration, it is one of the key attributes of a probabilistic forecast (Murphy 1991). While calibration deals with the meaning of probabilities (its estimation is an attempt to measure whether taking a forecast at face value is an optimal strategy), discrimination appraises the existence of a signal in the forecast when an event materializes and its absence in the opposite situation.

Practically, the ROC plots the hit rate (HR) versus the false alarm rate (FAR) of an event for incremental decision thresholds. Examples are provided in Fig. 1 for probability forecasts derived from the 50-member ensemble run at the European Centre for Medium-Range Weather Forecasts (see more details about the data in section 2). Corresponding probability fields for 15 February 2021 over the British Isles, are shown in Fig. 2. The targeted events correspond to precipitation exceeding the following thresholds: 1, 20, and 50 mm (24 h)$^{-1}$. A ROC curve is defined by the line joining successive ROC points, where each point corresponds to results for increasing decision thresholds, from the top right to the bottom left corners of the plot. When the decision variable is the number of members exceeding the event threshold (interpreted as a raw probability forecast), the issued forecast can take values in $[0, 1/M, 2/M, \ldots, 1]$ for an ensemble of size M. As a consequence, the resulting ROC curve is based on (up to) $M + 1$ points.

The area under the straight lines formed by connecting the $M + 1$ points [including the (0, 0) and the (1, 1) points] of the ROC plot correspond to the AUC with the so-called trapezoidal approximation (T-AUC). This nomenclature comes from the fact that the area is estimated considering straight lines between two consecutive points of the plot and so as a sum of trapeziums. Interestingly, T-AUC is equivalent to the result of a two alternative forced choice (2AFC) test for dichotomous events (Mason and Weigel 2009). The 2AFC test consists in checking whether, for two different observations in a verification sample, one event and one nonevent, the forecast associated with the former is larger than the forecast associated with the later (provided that the decision variable is oriented so that large implies more likely). If we denote x_1 a forecast when the event occurs and x_0 a forecast issued when the event does not materialize, the scoring function associated with the test is

$$S(x_1, x_0) = \begin{cases} 0 & \text{if } x_1 < x_0 \\ 1 & \text{if } x_1 > x_0 \\ 0.5 & \text{if } x_1 = x_0 \end{cases}. \quad (1)$$

The result of the test is the average score over all event/nonevent pairs in the verification sample. In Eq. (1), the test returns a value of 0.5 when the forecasts are indistinguishable, that is, in our examples, when two probability forecasts are

Denotes content that is immediately available upon publication as open access.

Corresponding author: Z. Ben Bouallègue, zied.benbouallegue@ecmwf.int

DOI: 10.1175/WAF-D-21-0195.1

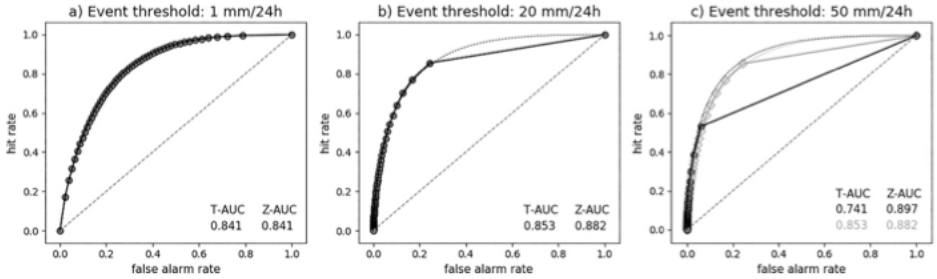


FIG. 1. Examples of ROC curves for common and rare events. ROC curve and corresponding AUCs for precipitation forecasts with event thresholds: (a) 1, (b) 20, and (c) 50 mm (24 h)$^{-1}$. In gray, the results obtained when using the probability of exceeding 20 mm (24 h)$^{-1}$ for predicting the occurrence of precipitation exceeding 50 mm (24 h)$^{-1}$.

identical. An average score of 0.5 is also the expected mean result for a random forecast.

For rare events, there is "a tendency for the points on the ROC to cluster toward the lower left corner of the graph" as noted by Casati et al. (2008) and illustrated in Fig. 1. When computing T-AUC, a straight line is drawn between the last meaningful point on the ROC curve and the top-right corner to close the ROC curve, giving the impression that part of the

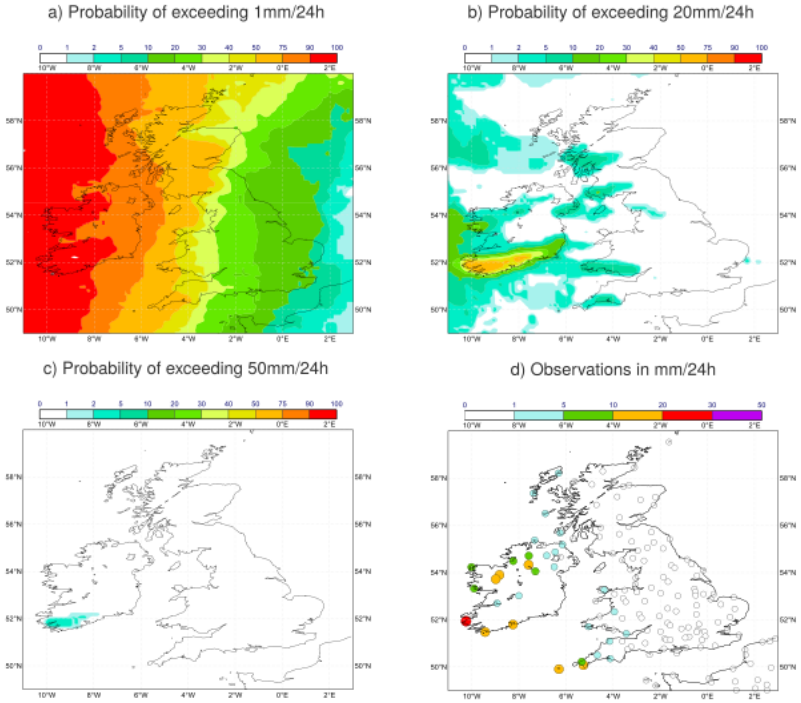


FIG. 2. Storm Karim and probabilistic precipitation forecasts valid on 15 Feb 2021 over the British Isles. Probability of precipitation derived from a 50-member ensemble, 5 days in the lead time. Probability of exceeding (a) 1, (b) 20, and (c) 50 mm (24 h)$^{-1}$, and (d) corresponding observations at synoptic stations.

curve is missing. How much of the curve is "missing" depends on the lowest category,[1] defined here by the ensemble size,[2] and the base rate of the event. As a rule of thumb, half of the ROC curve (from the apex to the right corner) is missing when assessing the performance of an ensemble of size M focusing on an event with a base rate $\alpha = 1/M$. This rule is valid when the probability forecast is close to calibration, that is when the probability can be taken at face value for decision making. The rule is derived from the relationship between an optimal decision threshold, the slope of the line tangent to the ROC curve, and the event base rate [see e.g., Eq. (25) in Ben Bouallègue et al. 2015].

To draw a "full" ROC curve, one can apply the so-called binormal model (Harvey et al. 1992; Wilson 2000; Atger 2004).[3] The fitting of the ROC curve with the binormal model is based on the assumption that HR and FAR are integrations of normal distributions, a signal and a noise distribution, respectively. A closed-form for the computation of the AUC exists [see Eqs. (2) and (3) in Harvey et al. 1992]. The fitting of the HR and FAR requires a Z-transformation based on the unit normal distribution. For this reason, the resulting AUC is denoted here Z-AUC. When applied to ensemble-derived probability forecasts for rare events, this approach consists effectively in an extrapolation to a hypothetical continuous decision variable based on the limited set of decision thresholds materially assessable. Because such a decision variable may not be achievable in practice, Z-AUC is sometime considered as a measure of the potential discrimination that could be achieved, for example "for an unlimited ensemble size" (Bowler et al. 2006).

T-AUC and Z-AUC summary metrics can provide very different comparative results. The statistics reported in Fig. 1c are striking in that respect. In gray, we report the verification results obtained when using the probability of exceeding 20 mm (24 h)$^{-1}$ to predict the occurrence of precipitation exceeding 50 mm (24 h)$^{-1}$. On the one hand, Z-AUC is (slightly) smaller than for the original forecast (in black): the interpretation is that the probability forecast for 50 mm (24 h)$^{-1}$ is potentially more informative than the probability forecast for 20 mm (24 h)$^{-1}$ when we focus on the higher event threshold. On the other hand, T-AUC statistics point toward a larger predictive skill of the low event-threshold probability forecast practically users may benefit more from using the lower threshold unless additional postprocessing is carried out to realize the potential additional benefit from using the higher threshold implied by the Z-AUC. As illustrated in Fig. 2, in many cases no ensemble member exceeds the high event threshold. The small proportion of distinguishable forecasts explains the poor results of the original forecast in terms of T-AUC: the discrimination ability of two forecasts with the same value is equivalent to the discrimination ability of two random forecasts [see Eq. (1)].

When verifying ensemble forecasts focusing on rare events, the AUC users face a dilemma: should they use T-AUC that relies on a clustering of points in the bottom left corner of the ROC plot, or should they use Z-AUC that extrapolates the results to compute scores based on a "full" ROC curve? The user's preference depends on the scientific question at hand, and in particular on whether the practical usefulness or the intrinsic information content of the ensemble forecast is the key aspect to be assessed. AUC assesses the discrimination ability of a decision variable, so special attention should be paid on how this decision variable is derived from the ensemble forecast. A decision variable defined as the number of ensemble members exceeding a threshold can appear appropriate for common events but may be less useful when forecasting rare events as illustrated in Fig. 1.

Aiming at bridging the gap between T-AUC and Z-AUC results, we propose using a new decision variable that encompasses more information from the ensemble than simply the number of members exceeding the event threshold for the computation of T-AUC. Our approach is inspired by a suggestion in Casati et al. (2008): "one solution to this problem [the clustering of the ROC points] is to subdivide the lowest-valued forecast probability bins. The verification sample can usually support subdividing the lower-valued probability bins when fitting the ROC for low base-rates." In other words, in the situation where no ensemble member exceeds the event threshold, the probability of occurrence should (more than ever) be interpreted as an imprecise probability (IP), a probability over an interval. A refinement of the forecast probability on that interval is possible using additional information from the ensemble itself such that different levels of near 0 chances of event occurrence can be distinguished. In the following, we show how to draw additional information from the ensemble about low chances of occurrence when using the ensemble mean (EM). EM is called "secondary" decision variable in this process and AUC estimated with this approach is referred to as IP/EM-AUC.

Having in mind the key question "How to interpret AUC results of ensemble forecasts when focusing on rare events?", we design a series of verification experiments in order to analyze T-AUC and Z-AUC in context. The verification experiments are chosen to show:

1) the loss of predictability with forecast lead time;
2) the impact of a postprocessing step, which accounts for subgrid variability;
3) the impact of increasing the forecast probability categories;
4) how to isolate the ensemble size effect with the help of a parametric model; and
5) the impact of subdividing the lowest category with the help of an ensemble summary statistic.

The verification experiments and derived results are described and presented in section 2 before drawing recommendations in section 3.

[1] More generally the conditional probability of the event given the lowest value of the discrimination variable.
[2] The ensemble members are equally probable in our example.
[3] More recently, Gneiting and Vogel (2018) introduce a flexible two-parameter beta family for fitting empirical ROC curves. This approach is not investigated or discussed further in our study.

2. Verification experiments

a. Verification setup

1) DATASET

Forecasts of daily precipitation are used in the following verification experiments, but similar qualitative results can be obtained with other accumulation periods or weather variables. The probability forecasts are derived from the ensemble prediction system run operationally at the ECMWF. The interpretation of the 50-member ensemble in terms of probability follows a simple (but common) approach. It consists in counting the number of members exceeding a threshold. Observation measurements at surface synoptic observation stations over the globe are compared with forecasts at the nearest grid point over a verification period running from 1 September 2019 to 31 August 2020.

Probability derived from ensemble forecasts are interpreted as imprecise probabilities as we are in a situation where the source of probabilistic information is incomplete and imperfect (Bradley 2019). For example, when no member exceeds the event thresholds of interest, the derived probability belongs to a probability interval close to 0. A ranking of the probability forecasts in that category can, however, be expressed with the help of an ensemble summary statistic such as the ensemble mean or an ensemble quantile. For illustration purposes, the ensemble mean is chosen as a secondary decision variable which is used to refine the ensemble interpretation for the lowest probability interval. Other choices can be valid as well and further research is encouraged in order to determine if an optimal summary statistic as a secondary decision variable exists in such a context.

Statistical postprocessing of precipitation forecasts is also envisaged here and tested applying a parametric approach, that is relying on a predefined type of probability distribution. In the following, censored shifted gamma distributions are used to describe appropriately precipitation forecast distributions (a detailed description of the statistical method can be found in Ben Bouallègue et al. 2020). Postprocessing aims here at correcting for the scale mismatch between forecasts (as model outputs on a grid) and observations (as point measurements at stations). We follow a so-called perturbed ensemble approach, which consists in adding uncertainty to the forecast in order to represent the larger uncertainty at a finer scale. Practically, random perturbations drawn from a parametric distribution are added to each ensemble member.

Other ensemble postprocessing techniques and their impact on the ROC are not investigated here. Techniques such as the neighborhood method or the use of lagged ensemble lead to an increase of the effective ensemble size at much lower computational cost than running additional ensemble members (Ben Bouallègue et al. 2013). The availability of more members allows in any case a finer probability discretization. In general terms, the impact of discretization is discussed below along with experiment III.

2) VERIFICATION METHODOLOGY

In our experiments, the central step of the verification process consists in populating contingency tables. The 2 × 2 tables are the raw material for generating:

- ROC curves (Mason 1982),
- performance diagrams plotting hit rate versus success ratio (Roebber 2009), and
- potential economic value (PEV) plotted as a function of the user's cost–loss ratio (Richardson 2000, 2011).

Contingency tables are populated for incremental decision thresholds. In the traditional ensemble verification case, the decision variable is the number of members exceeding the event threshold [e.g., 50 mm (24 h)$^{-1}$] and the number of decision thresholds is $M + 1$ with M the ensemble size. In our experiment dealing with imprecise probabilities, the "zero category" is subdivided in additional subcategories using the ensemble mean EM as a secondary decision variable. We are considering EM as a continuous decision variable rather than focusing on the binary forecast EM being greater than the event threshold. Dealing here with daily precipitation, the following (secondary) decision thresholds are used in this context: [0.1, 0.2, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 10, 15] in millimeters per 24 h [mm (24 h)$^{-1}$]. The basic principle governing this choice of thresholds is to span decision thresholds over the range of values that the secondary decision variable can take.

So in practice, the following is performed:

1) choose an event threshold;
2) count the number of members exceeding this event threshold (this decision variable is called raw probability forecast);
3) if no members exceed the event threshold, compute the ensemble mean EM and count the number k out of K cases for which EM is greater than each of the K secondary decision thresholds;
4) adjust the raw decision variable by considering a probability of $k/[M(K + 1)]$ (rather than 0) for that forecast; and
5) derive a contingency table for all distinct probability values in the forecast sample.

The trapezoidal approximation applied to the corresponding set of (HR, FAR) pairs correspond to IP/EM-AUC. This acronym reflects that the ensemble mean is used as a secondary decision variable.

Scores are aggregated over different domains, but mainly results for the globe are shown. Event thresholds are defined in absolute terms with a focus on an exceeding event threshold of 50 mm (24 h)$^{-1}$. In the global verification sample, this event has a base rate of around 1%, but with more occurrence in certain regions of the world than others. Hamill and Juras (2006) recommend the use of thresholds expressed in relative terms (quantile of a climatology) in order to avoid over-interpretation of the AUC results by mixing the forecast ability to distinguish between wet and dry regions and genuine predictive skill. Results for precipitation exceeding the 99th percentile of the local climatology is also discussed as a final example.

b. Verification experiments results

Each of the following figures (Figs. 3–7) comprises three plots: a performance diagram, a ROC plot, and a potential economic value plot (with a log scale on the x axis). AUC estimates are indicated for both the trapezoidal approach

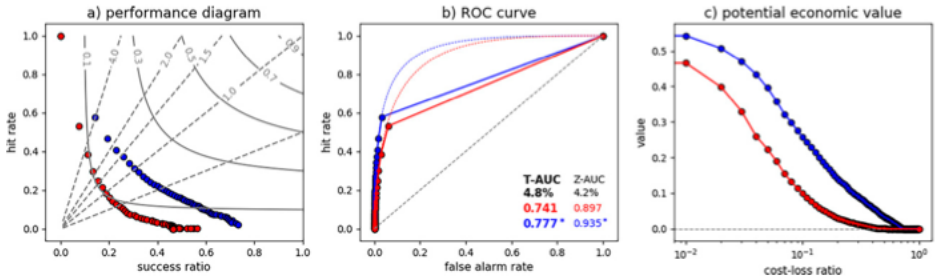


FIG. 3. Results of experiment I, the benchmark experiment comparing forecast performance at day 1 (blue) and at day 5 (red). On the performance diagram, dashed and solid lines refer to the forecast bias and critical success index, respectively.

(T-AUC) and the binormal method (Z-AUC). For each figure, two different sets of forecasts are compared. An asterisk indicates which of the two sets of forecasts (the red or the blue) has the best score in terms of T-AUC and Z-AUC. The relative superiority of the best forecast is indicated in percent. For all plots (except for experiment IV in Fig. 6), the results in red correspond to the raw ensemble-derived probability forecasts at day 5.

1) EXPERIMENT I: THE IMPACT OF THE FORECAST LEAD TIME

The first experiment compares the performance of forecasts at two different lead times. This experiment illustrates how different verification tools are impacted by a genuine change in forecast predictive skill. Figure 3 illustrates the visual impact one can expect to see in such a situation where a difference in verification results can be explained only by a difference in predictive skill.

In Fig. 3a, the performance diagram shows the blue points (day 1 results) lie closer to the top right corner than the red ones (day 5 results). In Fig. 3b, on the ROC plot, the blue points are distinct from the red ones and are closer to the top left corner. In Fig. 3c, on the PEV plot, the blue curve lies above the red one, except for cost–loss ratios close to 1 for which both short and medium range forecasts have no value. This first experiment provides typical results expected from an increase in forecast predictive skill and, as such, serves as a benchmark for the following experiments. In terms of summary metrics, the relative skill improvement between day 5 and day 1 is of the same order of magnitude for T-AUC and Z-AUC, with 4.8% and 4.2% measured improvement, respectively.

2) EXPERIMENT II: THE IMPACT OF A POSTPROCESSING STEP

In this experiment, ensemble postprocessing is applied in order to account for the scale mismatch between forecasts and observations. The method can be applied in a verification context to account for observation uncertainty, but also in a postprocessing context to provide a forecast valid for all points within a model grid box. The postprocessed forecast is derived by adding a random perturbation independently to each member. The random perturbation is draw from a distribution described by the value of the forecast member (i.e., the predicted gridscale precipitation) and with fixed parameters for all forecasts (as in Ben Bouallègue et al. 2020). In other words, a constant piece of information is added in the process: the expected subgrid scale uncertainty expressed as a function of the grid-scale precipitation value. The main impact on the forecast is a significant increase of the ensemble spread with, in particular, larger distributional tails of the postprocessed ensemble distribution compared with the original one while the ensemble size remains unchanged.

In Fig. 4 and on, both blue and red curves are valid for forecasts at day 5. All three plots in Fig. 4 show that the blue (postprocessed ensemble results) and red points (original ensemble results) are overlapping with the exception of a couple of points in each plot. In the case of the performance diagram and ROC curve, blue and red points belong to the same underlying curves. In the PEV plot, we see larger value for users with a cost–loss ratio smaller than or equal to 2%. In terms of Z-AUC, the results are identical before and after postprocessing up to the second decimal indicating that the underlying forecast information content is not altered by the process. The information has increased in the sense that information about the subgrid uncertainty has been added to the forecast,[4] but this bit of information is not different from one forecast to another. The improvement by postprocessing as measured by T-AUC is 4.5%, close to the improvement measured between day 5 to day 1 discrimination ability (see Fig. 3b). While the latter is attributed to an improved predictability at shorter lead times, the former, the T-AUC improvement with postprocessing, is attributed to a change in the frequency of forecast events: postprocessed ensemble members exceed the event threshold more often due to the larger spread (as in the example in Fig. 2c).

3) EXPERIMENT III: THE IMPACT OF DISCRETIZATION

The postprocessing technique used in experiment II is based on a parametric approach. A random perturbation drawn from a parametric distribution is added to each member. The random draw is made from a distribution for which the parameters are known. Now, in experiment III, rather than a single perturbation

[4] With a positive impact on the forecast calibration (not shown).

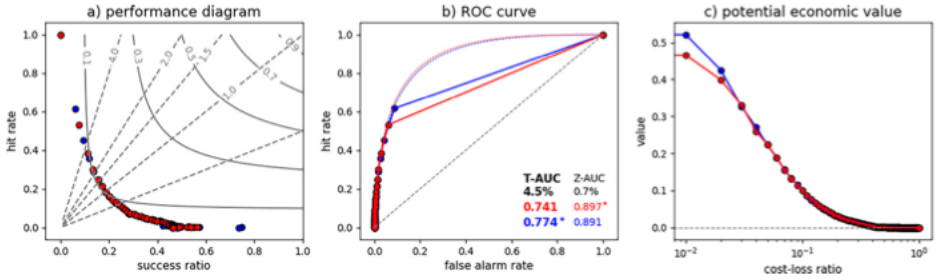


FIG. 4. Results of experiment II illustrating the impact of postprocessing. Results when accounting for representativeness (blue) compared with results for the corresponding raw day 5 forecasts (red).

for each member, we consider two random draws for each of the 50 raw ensemble members. So, the resulting postprocessed ensemble has a size of 100. Let us recall that the ensemble perturbations are based on random draws from a distribution whose form depends only on the forecast value itself. The model for the precipitation subgrid variability is a new but constant piece of information added to the forecast. Drawing additional random numbers from a parametric distribution better captures the form of the distribution but does not change the distribution itself or the quality of the underlying model.

Figure 5 is similar to Fig. 4. The major difference between the two figures is the fact of a single point on each plot such as on the ROC plots with one point on the blue ROC curve closer to the top right corner when increasing the forecast discretization. In terms of T-AUC, the change in the number of probability thresholds from 51 to 101 leads to a jump from 4.5% to 7.9% of improvement with respect to the original forecast. The larger the number of members describing the underlying forecast distribution, the better the decision variable defined as the number of members exceeding a threshold. In terms of Z-AUC, the doubling of the number of categories as part of the postprocessing has no impact on performance. The underlying forecast distribution is the same, no additional information is provided that improves the forecast discrimination ability. With this experiment, we see that the choice regarding the categorization of the probability forecast considerably influences the T-AUC results but not the Z-AUC ones.

4) EXPERIMENT IV: ISOLATING THE ENSEMBLE SIZE EFFECT

Here we again exploit the parametric nature of the postprocessing approach we have followed. The goal here is to assess the ensemble-size effect on the forecast discrimination independently from the forecast discretization effect. This experiment is designed to compare probability forecasts with the same discretization (the same decision thresholds) but derived from raw ensembles with different sizes. So, we distinguish raw ensemble size (the source of the forecast information) and the postprocessed ensemble size which is an arbitrary choice in our setting. In this experiment, we compare raw ensembles of size 10 and 50, both postprocessed in order to get 50 postprocessed "members" (drawing 5 and 1 random perturbations for each member, respectively).

Figure 6 shows the positive impact of increasing the raw ensemble size from 10 to 50 members. Derived probabilities from 50 postprocessed forecasts (in both cases) exhibit better performance on the three plots. In particular, as expected, users with smaller cost–loss ratios) benefit most of the ensemble size

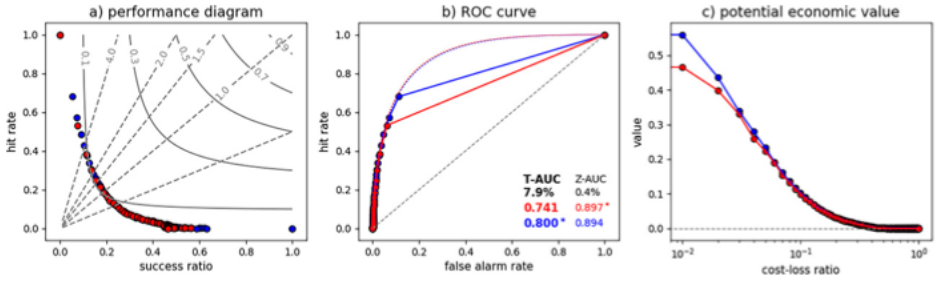


FIG. 5. Results of experiment III illustrating the impact of discretization. Results for 50 (red) and 100 (blue) postprocessed members from a parametric model. Many points are indistinguishable because they are very close one to another.

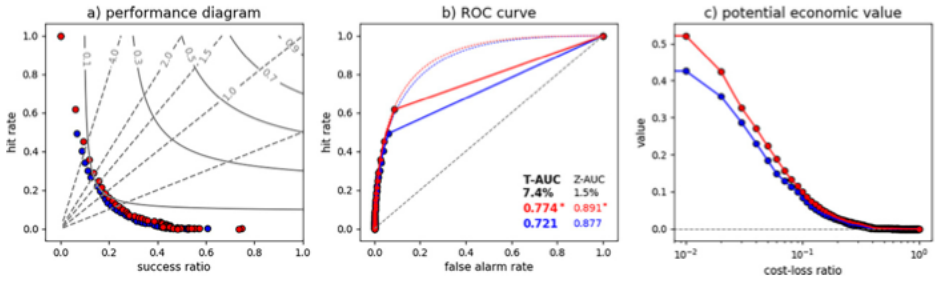


FIG. 6. Results of experiment IV illustrating the expected impact of the ensemble size on verification results. Results for an ensemble of size 10 (red) and 50 (blue) postprocessed to 50 members each. Note that the red points on this figure are the same as the blue points in Fig. 4.

increase. The discrimination ability as measured by T-AUC and Z-AUC shows an increase of 7.4% and 1.5%, respectively. These values can be compared with a 7.3% theoretical improvement of a proper score,[5] the continuous ranked probability score, which encompasses both discrimination and calibration contributions.

5) EXPERIMENT V: SUBDIVIDING THE LOWEST CATEGORY

In many cases, none of the ensemble members forecast a rare event, i.e., precipitation exceeding 50 mm (24 h)$^{-1}$. The raw probability forecast is 0, but it is interpreted as an imprecise probability on an interval close to 0. Using a secondary decision variable, it is possible to distinguish between lower and higher chances within the pool of "0 probability" forecasts. In this experiment, we consider a forecaster that has access to the ensemble mean in addition to the probability forecast itself. In the situation where a forecast probability of 0 is issued, the forecaster infers that the chance of occurrence of an event is higher when the ensemble mean is higher. Quantitatively, within an interval close to 0, a larger probability is assigned to a forecast with a larger ensemble mean (see methodology in section 2). The verification results obtained with this ensemble interpretation are shown in Fig. 7 and compared with results obtained when using only the ensemble mean as a decision variable in Fig. 8. We can recall that we are here considering the ensemble mean EM as a continuous variable, not drawing information from a binary forecast derived as EM being greater than the event threshold.

Figures 7a and 7b show the continuity between the results based on the original data (red points) and the ones when subdividing the lowest probability category with the help of the ensemble mean (blue points). The so-called IP/EM-AUC corresponds to T-AUC computed using the blue points. In Fig. 7c, PEV results diverge only for cost–loss ratio values smaller than 2%. These results share similarities with the ones obtained with postprocessing in Fig. 4. When using the ensemble mean as a secondary decision variable, we obtain T-AUC and Z-AUC results that are (almost) identical, as a "full" ROC curve is now available based on the enhanced interpretation of the ensemble forecast. Interestingly, T-AUC with the lowest category subdivision is very close to the Z-AUC estimate for the original data. In other words, similar results can be obtained by extrapolation using Z-AUC or with T-AUC applied to an appropriate decision variable for rare events. On the plot in Figs. 7 and 8, the blue points follow the red line derived with the binormal approximation, in one case they are above and in the other below. The results with the two approaches are consistent but subject to different level of uncertainty as discussed in section 2c.

The ensemble interpretation in two steps with the help of a secondary decision variable refines the imprecise probabilities close to 0%. The ensemble mean forecast serves as a secondary decision variable in our example. This approach is different than using the ensemble mean as a unique decision variable as illustrated in Fig. 8. Results here are produced using the 99% quantile of the local climatology as an event threshold. This choice allows to better highlight the major benefit of integrating the ensemble distribution into the decision variable rather than using the ensemble mean only. Indeed, not only the ROC plot but also the performance diagram and potential economic value plot show the overall poor results of the ensemble mean (in cyan) compared with the probability forecasts (in red and blue).

Focusing on the T-AUC results in Fig. 8b, we see that the ensemble mean EM appears as a better decision variable than the original ensemble interpretation as measured with the trapezoidal approximation. However, EM as a decision variable does not seem to perform better anywhere on the performance diagram (see Fig. 8a) or to benefit any user except possibly with very low cost–loss ratios (see Fig. 8c), this result illustrates how misleading conclusions can be drawn when comparing T-AUC results from different "sources."

The ensemble size has an impact on both the raw probability and the ensemble mean estimates. The impact of the ensemble size on the discrimination ability as estimated when combining information from both ensemble aspects is not explored here.

[5] Derived from Eq. (9) in Leutbecher (2019).

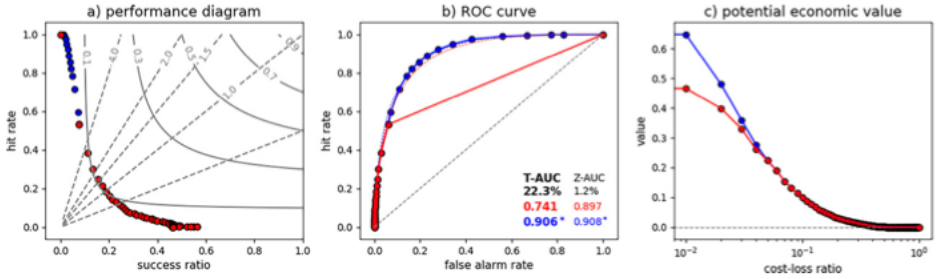


FIG. 7. Results of experiment V showing the impact of subdividing the lowest category with the help of the ensemble mean (in blue) with respect to the original forecast results (in red).

c. Impact of the event base rate

As a final investigation, we analyze AUC estimations and corresponding uncertainty as a function of the event base rate. We consider three event thresholds: 20, 40, and 50 mm $(24\ \text{h})^{-1}$, and build verification datasets for four different geographical domains (global, Northern Hemisphere, Southern Hemisphere, Europe) and four different seasons (autumn 2019, winter 2019, spring 2020, and summer 2020). The event base rate is different for each domain and threshold combination. The score uncertainty associated with each AUC estimation is derived as the inter-quantile range (5%–95%) of the score empirical bootstrapping distribution. Results are shown in Fig. 9 for T-AUC, Z-AUC, and IP/EM-AUC, the trapezoidal approximation when interpreting the ensemble in terms of imprecise probabilities and using the ensemble mean as secondary decision variable.

In Fig. 9a, we observe a slight increase of the discrimination ability as a function of the rarity of the event, with the Z-AUC and IP/EM-AUC approaches displaying consistent estimates with respect to each other. However, a drop in discrimination is measured with T-AUC for event base rates smaller than 3%. When applied to ensemble probability forecasts, T-AUC appears base rate dependent. The drop is the result of the application of the trapezoidal approximation to raw probabilities derived from a non-infinite size ensemble: the AUC computation is confined to a smaller part of the full ROC curve as the event's base rate decreases. The drop in discrimination is also an indicator of when the ensemble interpretation into a decision variable by simply counting the number of members exceeding a threshold is no longer appropriate. The distinction between low probability of occurrence would require more ensemble members, or some sort of statistical postprocessing, or simply to categorize low probability forecasts based on an additional ensemble summary statistic as for example the ensemble mean.

In Fig. 9b, larger score uncertainty is seen for more extreme events. The increase in uncertainty is visible for all AUC methods, but T-AUC and Z-AUC estimates are more impacted than IP/EM-AUC ones. The binormal model exhibits the largest level of uncertainty for rare events because the original points are too close to get good estimates of the slope in the Z-transformed space, but the trapezoidal approximation is also subject to large variations. In practice, for event base rates larger than 99%, the level of uncertainty appears too large to draw any useful conclusion with T-AUC or Z-AUC applied to raw probability forecast while the results for the enhanced probability using the ensemble mean as secondary decision variable appears more robust and reliable.

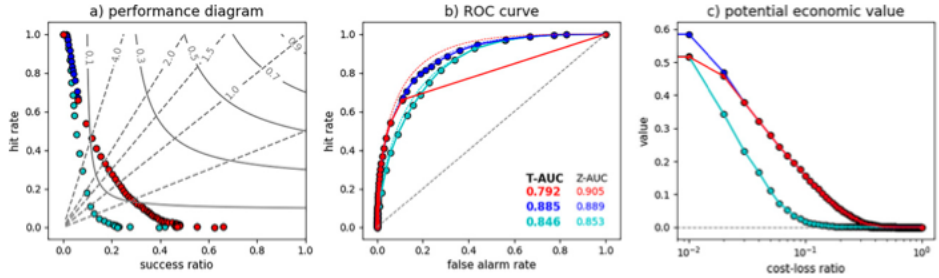


FIG. 8. Results of experiment V as in Fig. 7, but using a threshold defined as the 99% of the local climatology. In addition to the results for raw probability (red) and when subdividing the lowest probability category with the ensemble mean (blue), we show results when using only the ensemble mean as a decision variable (cyan).

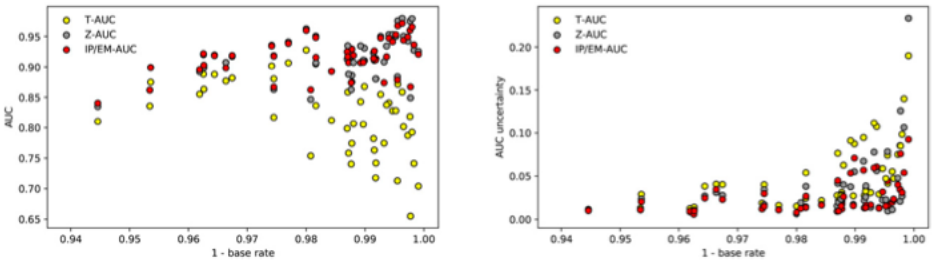


FIG. 9. (left) Area under the ROC curve and (right) corresponding uncertainty estimated with the trapezoidal approximation (T-AUC, in yellow), a binormal model (Z-AUC, in gray), and interpreting the ensemble in terms of imprecise probabilities (IP/EM-AUC, in red) plotted as a function of the event base rate. Each dot corresponds to the result for one domain, one season, and one threshold (see text).

3. Recommendations

T-AUC and Z-AUC provide complementary information about the discrimination ability of a forecasting system. It is important to understand the differences between them and to use each appropriately depending on the question under investigation. Z-AUC measures the inherent discrimination ability of a forecasting system, while T-AUC measures how well this is achieved by a given implementation. Computing IP/EM-AUC shows a path to practically building a bridge between the two approaches.

Awareness of AUC properties is key in situations which combine 1) the assessment of probability forecasts derived from an ensemble and 2) a focus on rare events. Differences between T-AUC and Z-AUC are largest in such situations and it is especially important to interpret the results carefully. The recommendations below are targeted to this specific type of situation. In other cases (when focusing on more frequent events or when assessing probability forecasts derived from a parametric distribution for example), the different approaches for computing AUC converge to identical results as illustrated for instance in Fig. 1a. T-AUC is not the only verification metrics whose interpretation can be altered when focusing on rare events. Forecast verification of rare events has gathered increasing attention over the last decade and the interested reader could refer to Ben Bouallègue et al. (2019) and references within.

T-AUC, the AUC estimation with the trapezoidal approximation, is the traditional way to measure forecast discrimination. As illustrated, differences in T-AUC can be attributed to multiple sources, for example the level of forecast discretization, the presence of ensemble forecast biases, or the way probabilities are derived from the ensemble forecast. So, when using the empirical ROC and summarizing performance using the T-AUC, it is important to consider the following:

- Comparing T-AUC results should be done carefully. For example, a positive difference in T-AUC should be scrutinized before being interpreted as an improvement in intrinsic discrimination ability (or as an improvement of the forecast performance in a broader sense).
- T-AUC is a measure of the performance of the forecasts using a particular discretization of a chosen decision variable. Therefore, both the discretization and the decision variable should be clearly described.
- The maximum available discretization should be used (e.g., each member rather than fixed percentage bins), to ensure the ROC is as complete as possible.
- A comparison of the practical benefits of two competing forecast configurations should follow the above approach, describing the decision variable and discretization used for each configuration. However, this approach should not be used to draw conclusion about the underlying discrimination ability of the different systems.
- A fair comparison for the underlying discrimination ability of different systems would rely on using the same decision variable and the same discretization. A sanity check for no significant differences in the underlying forecast bias is also important.
- Comparing competing forecasts with T-AUC can lead to misleading conclusions if the above is not accounted for.
- To evaluate the impact of changing discretization T-AUC should be used (Z-AUC will not be sensitive to this).

Z-AUC, the AUC estimate with a binormal model, is an alternative to the traditional trapezoidal approach. Z-AUC is a measure of "potential" discrimination ability of a system, in the sense that the extrapolation of the performance with the curve fitting has no practical meaning in terms of forecast usefulness. The binormal model is based on the assumption that HR and FAR follow the characteristics of normality distributed parameters. This assumption is not tested in our experiments, but our examples show good fits between model and data. As a summary, we state the following conclusions:

- Results with Z-AUC are not sensitive to forecast discretization and to simple ensemble postprocessing, in contrast to the T-AUC results.
- Z-AUC should be used to compare the potential discrimination ability of different forecasting systems. It does not indicate how this can be realized, but it does show which system has the better underlying performance.
- Z-AUC is useful to compare for example ensemble forecasts with different number of members it gives a better indication of the skill that could be achieved if sufficient discretization is available.

- Z-AUC cannot distinguish whether a chosen discretization is sufficient or can be improved; instead it shows what would be achieved if a sufficient discretization was available.
- Computing both T-AUC and Z-AUC allows useful comparison of potential and actual discrimination ability.
- If T-AUC and Z-AUC are similar, the ensemble interpretation allows a discretization of the derived decision variable which is sufficient and there is not much to be gained from postprocessing to generate a better interpretation or a finer discretization.
- If T-AUC is lower than Z-AUC there is the potential to improve the forecast performance by a better ensemble interpretation. This will be most beneficial to users with low C/L and most likely to happen for rare events (especially where low probability categories are not well resolved in the forecast).

We have demonstrated two methods to improve the ensemble interpretation and thus increase the forecast discretization in such situations. Postprocessing to account for subgrid-scale variability introduces a continuous distribution and allows arbitrary fine discretization. The choice of discretization should be sufficient to generate the full ROC (as well as plotting the ROC, a simple way to check this is to compare the T-AUC and Z-AUC).

IP/EM-AUC refers to the AUC estimated with a new approach which involves subdividing the lowest probability category by ranking the forecasts in this category according to an ensemble summary statistic, here the ensemble mean. We showed that this approach can provide sufficient discretization to generate a full ROC curve based directly on the ensemble forecasts. The score estimations using this method have been shown to be more robust than with the binormal model. They also represent the real skill of the forecast system since users can act on each of the discretization categories, rather than the potential skill that is shown by the extrapolation using a statistical model. Using this method, we have shown that while Z-AUC is strictly only a measure of potential discrimination skill, it may actually be straightforward to achieve in practice.

Other implications of this interpretation in the context of ensemble forecast verification and postprocessing are topics for future research.

Acknowledgments. The authors are grateful to two anonymous reviewers for their valuable comments.

Data availability statement. Ensemble forecasts used in this study are publicly available. For more info, please visit the website https://www.ecmwf.int/en/forecasts/accessing-forecasts. SYNOP observations used in this study cannot be shared with third parties.

REFERENCES

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **130**, 627–646, https://doi.org/10.1256/qj.03.23.

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteor. Z.*, **22**, 49–59, https://doi.org/10.1127/0941-2948/2013/0374.

——, P. Pinson, and P. Friederichs, 2015: Quantile forecast discrimination ability and value. *Quart. J. Roy. Meteor. Soc.*, **141**, 3415–3424, https://doi.org/10.1002/qj.2624.

——, L. Magnusson, T. Haiden, and D. S. Richardson, 2019: Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quart. J. Roy. Meteor. Soc.*, **145**, 1741–1755, https://doi.org/10.1002/qj.3523.

——, T. Haiden, N. Weber, T. Hamill, and D. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148**, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1.

Bowler, N. E., C. E. Pierce, and A. W. Seed, 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quart. J. Roy. Meteor. Soc.*, **132**, 2127–2155, https://doi.org/10.1256/qj.04.100.

Bradley, S., 2019: Imprecise probabilities. *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, E. N. Zalta, Ed., Stanford University, https://plato.stanford.edu/archives/spr2019/entries/imprecise-probabilities/.

Casati, B., and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.*, **15**, 3–18, https://doi.org/10.1002/met.52.

Gneiting, T., and P. Vogel, 2018: Receiver operating characteristic (ROC) curves. arXiv, 1809.04808, https://arxiv.org/abs/1809.04808.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Harvey, L. O., J. K. Hammond, C. Lusk, and E. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883, https://doi.org/10.1175/1520-0493(1992)120<0863:TAOSDT>2.0.CO;2.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145** (Suppl. 1), 107–128, https://doi.org/10.1002/qj.3387.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349, https://doi.org/10.1175/2008MWR2553.1.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, https://doi.org/10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

——, 2011: Economic value and skill. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 167–184.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Stanski, H., L. Wilson, and W. Burrows, 1989: Survey of common verification methods in meteorology. 2nd ed. Research Rep. MSRB 89-5, WWW Tech. Rep. 8, WMO/TD-358, World Meteorological Organization, 114 pp., http://www.cawcr.gov.au/projects/verification/Stanski_et_al/Stanski_et_al.html.

Wilson, L. J., 2000: Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system." *Wea. Forecasting*, **15**, 361–364, https://doi.org/10.1175/1520-0434(2000)015<0361:COPPOP>2.0.CO;2.