A global evaluation of multi-model ensemble tropical cyclone track probability forecasts

Received: 18 July 2019 | Revised: 21 November 2019 | Accepted: 24 November 2019 | Published on: 1 January 2020

DOI: 10.1002/qj.3712

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

A global evaluation of multi-model ensemble tropical cyclone track probability forecasts

Helen A. Titley[1,2] | Rebecca L. Bowyer[1] | Hannah L. Cloke[2,3,4,5]

[1]Weather Science, Met Office, Exeter, 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

[5]Centre of Natural Hazards and Disaster Science, CNDS, Uppsala, Sweden

Correspondence
Helen A. Titley, Met Office, FitzRoy Road, Exeter, EX1 3PB, UK.
Email: helen.titley@metoffice.gov.uk

Funding information
Newton-Bhabha Fund, Met Office WCSSP India Programme

Abstract

At the Met Office, dynamic ensemble forecasts from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G), the European Centre for Medium-Range Weather Forecasts Ensemble (ECMWF ENS) and National Centers for Environmental Prediction Global Ensemble Forecast System (NCEP GEFS) global ensemble forecast models are post-processed to identify and track tropical cyclones. The ensemble members from each model are also combined into a 108-member multi-model ensemble. Track probability forecasts are produced for named tropical cyclones showing the probability of a location being within 120 km of a named tropical cyclone at any point in the next 7 days, and also broken down into each 24-hour forecast period. This study presents the verification of these named-storm track probabilities over a two-year period across all global tropical cyclone basins, and compares the results from basin to basin. The combined multi-model ensemble is found to increase the skill and value of the track probability forecasts over the best-performing individual ensemble (ECMWF ENS), for both overall 7-day track probability forecasts and 24-hour track probabilities. Basin-based and storm-based verification illustrates that the best performing individual ensemble can change from basin to basin and from storm to storm, but that the multi-model ensemble adds skill in every basin, and is also able to match the best performing individual ensemble in terms of overall probabilistic forecast skill in several high-profile case-studies. This study helps to illustrate the potential value and skill to be gained if operational tropical cyclone forecasting can continue to migrate away from a deterministic-focused forecasting environment to one where the probabilistic situation-based uncertainty information provided by the dynamic multi-model ensembles can be incorporated into operational forecasts and warnings.

KEYWORDS

ensembles, probabilistic forecasting, tropical cyclones, uncertainty, verification

1 | INTRODUCTION

In dynamic ensemble forecasting, instead of making a single forecast of the most likely weather conditions, a forecast model is run multiple times to produce an ensemble of forecasts. These ensemble forecasts take into account uncertainty in the initial conditions and imperfections in the model formulation, and aim to give an indication of the range of possible future states of the atmosphere. For over 25 years, these dynamical ensemble model forecasts have been routinely produced by several global numerical weather prediction modelling centres, including the European Centre for Medium-Range Weather Forecasts Ensemble (ECMWF ENS: Palmer, 2019), the National Centers for Environmental Prediction Global Ensemble Forecast System (NCEP GEFS: Toth and Kalnay, 1997) and the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G: Bowler *et al.*, 2008).

These ensemble forecast models have an important role to play in tropical cyclone forecasting, through their ability to highlight the situation-dependent uncertainty and provide probabilistic forecast information to help inform decision makers. Consequently, many global modelling centres produce tropical cyclone track forecasts from their ensemble forecast models, develop ensemble tropical cyclone forecast products, verify these forecasts, and share the forecast track data via the TIGGE (The International Grand Global Ensemble) cyclone exchange programme (Swinbank *et al.*, 2016). Several studies have shown the benefits of dynamic ensemble forecasting for both existing tropical cyclones (e.g. Dupont *et al.*, 2011; Yamaguchi *et al.*, 2012; Du *et al.*, 2016; Leonardo and Colle, 2017; Zhang and Yu, 2017), and for providing probabilistic information about tropical cyclone genesis (e.g. Yamaguchi *et al.*, 2015; Yamaguchi and Koide, 2017). However, a recent survey among operational tropical cyclone forecasters (Titley *et al.*, 2019) showed that although ensemble forecasts are used and valued by almost all forecasters, the deterministic-focussed forecasting environment has often limited the extent to which the full probabilistic information provided by ensembles has been pulled through into operational tropical cyclone forecasts and warnings.

Objectively identifying the forecast track of each ensemble member is essential both for post-event model evaluation and for the generation of forecast guidance products in real time. Various tracking techniques are used by operational centres around the world (Vitart and Stockdale, 2001; Van der Grijn, 2002; Tallapragada *et al.*, 2013) and in research (Hodges, 1995). At the Met Office, a bivariate (850 hPa relative vorticity and mean sea-level pressure) tropical cyclone tracker known as MOTCTracker (Heming, 2017) is run in real time on the Met Office MOGREPS-G ensemble, and the tracks are made available to Regional Specialised Met Centres (RSMCs) and via the research TIGGE cyclone CXML archive.

Although ensemble forecasts are a good way of assessing forecast uncertainty, they are limited to the uncertainty captured by a specific modelling system, and there is a tendency for single-model ensembles to be under-spread, with the observations too often falling outside of the range of solutions. A multi-model ensemble approach, where dynamic ensemble systems from multiple centres are combined together into a grand multi-centre ensemble (thereafter called multi-model ensemble), can help address this shortcoming and provide a more complete representation of the uncertainty in the model structure, also potentially reducing the errors. The rationale behind multi-model ensemble forecasting was summarised by Hagedorn *et al.* (2005) who stated that "the key to the success of the multi-model concept lies in combining independent and skilful models, each with its own strengths and weaknesses". Several studies have demonstrated that probabilistic forecast skill and reliability can be improved through the use of multi-model ensembles, including Park *et al.* (2008), Johnson and Swinbank (2009), Hagedorn *et al.* (2012), Hamill (2012) and Matsueda and Nakazawa (2015).

The application of a grand combined multi-model ensemble approach in tropical cyclone forecasting is a natural extension of the "consensus" forecasting approach that has been a valuable cornerstone of tropical cyclone track and intensity forecasting for many years, where traditionally three or more deterministic forecasts have been combined or averaged into a "consensus" forecast and also used to give a prediction of likely forecast error (e.g. Goerss, 2000; 2007; Sampson *et al.*, 2008; Goerss and Sampson, 2014; Yamaguchi *et al.*, 2017). The process of combining together several ensemble forecast models into a multi-model ensemble combines the strengths of the consensus and ensemble approaches, by pulling through the full probabilistic forecast information from several dynamic ensembles and model formulations into multi-model ensemble probability forecasts.

At the Met Office, in order to produce real-time multi-model ensemble tropical cyclone products, MOTC-Tracker is also run in real time on the direct input data from the ECMWF ENS and NCEP GEFS ensembles. The ensemble forecast tracks from the ensemble members from these models are combined with the Met Office MOGREPS-G ensemble tracks, to create a 108-member multi-model ensemble. The three corresponding deterministic models (the Met Office Unified Model, the ECMWF IFS, and the NCEP GFS model) are also tracked. A range of products, including track and intensity forecasts for both named and developing tropical cyclones, are produced and used by forecasters in the Met Office Global Guidance Unit. An example of the track and track

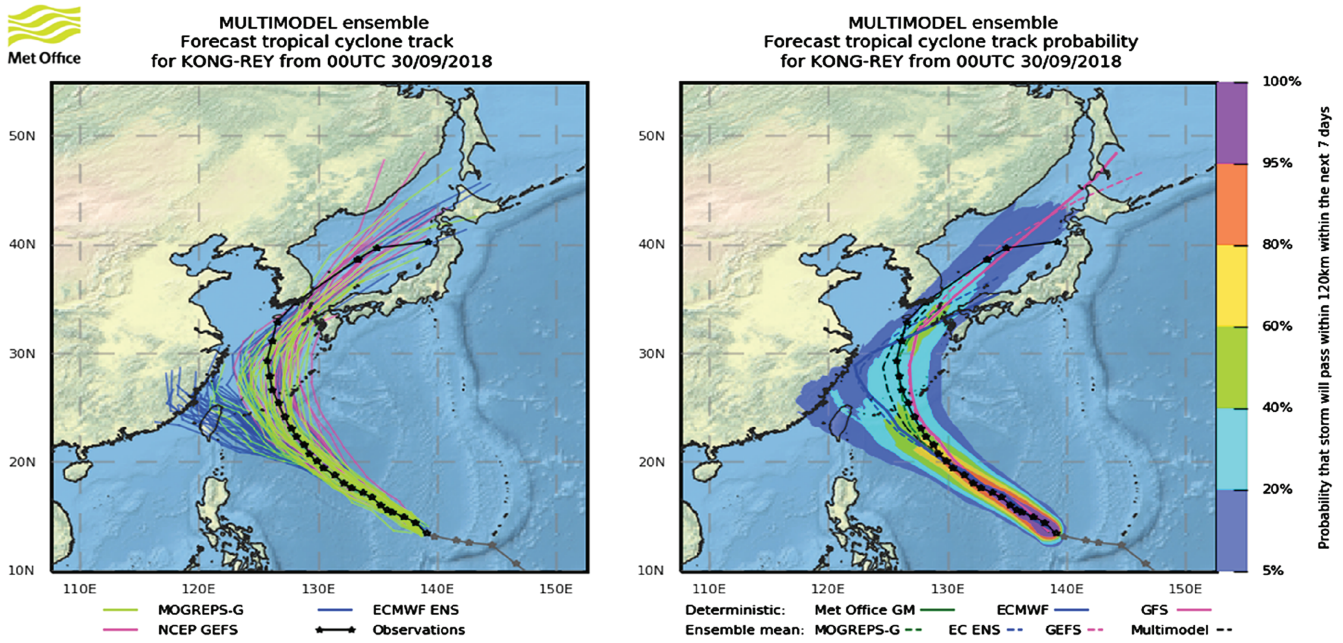

FIGURE 1 Multi-model ensemble forecasts for Typhoon Kong-rey from 0000 UTC 30 September 2018: Tracks coloured according to model (left), and multi-model ensemble track probability with deterministic (solid) and ensemble mean (dashed) tracks (right)

probability products produced for named tropical cyclones from the multi-model ensemble is shown in Figure 1 for Typhoon Kong-rey. The products are also distributed to several operational tropical cyclone forecasting centres, including the RSMCs in Miami and La Réunion. The tracks and track probabilities from each of the ensembles and the multi-model ensemble are also displayed in the Met Office Global Hazard Map (Robbins and Titley, 2018), where they can be viewed alongside the associated wind and rain hazard information, and overlain on vulnerability and exposure fields.

To fully assess the skill and value of the multi-model ensemble tropical cyclone forecasts produced at the Met Office, a framework to produce objective verification of named tropical cyclone track probability forecasts has been developed, the results of which are presented in this article. Previous studies evaluating named tropical cyclone track probability forecasts using multi-model ensembles have focussed on only one or two Northern Hemisphere basins, e.g. Majumdar and Finocchio, 2010 (North Atlantic and Northwest Pacific); Yamaguchi *et al.*, 2012 (Northwest Pacific); and Leonardo and Colle, 2017 (North Atlantic). In this article the verification results are analysed for all named tropical cyclones in all tropical cyclone basins around the globe during 2017 and 2018, and then split into each basin in order to compare the absolute and relative performance of the ensembles across basins. Storm-based verification results have also been calculated for some high-profile tropical cyclones in order to compare the results from storm to storm.

The key questions addressed by this article are:

- How do forecast performance and characteristics vary between the global ensembles, and with lead time?
- Is there benefit in terms of probabilistic forecast skill, reliability and value, from combining the three global ensembles into a multi-model ensemble?
- Does using the full probability forecast information via the multi-model ensemble add skill compared to a "consensus" forecast of the parent deterministic models?
- When comparing basin to basin, does forecast performance vary, both overall compared to a reference consensus forecast, and relatively between the different ensembles?
- When comparing storm to storm, how does forecast performance vary, and what does it tell us about the benefit of multi-model ensembles?

The verification framework and methodology are described in Section 2. Section 3 presents the results, split into three sections: (a) 7-day ("overall") track probability forecast results, (b) 24-hour track probability forecast results, and (c) storm-specific verification results for two high-profile North Atlantic hurricanes. Section 4 provides a discussion, focussed around the implications of the results for operational forecasters, decision-makers, and model developers, while Section 5 states some key conclusions and ideas for future work.

2 | VERIFICATION METHODOLOGY

The three ensembles tracked in real time at the Met Office using MOTCTracker (Heming, 2017), and included in the multi-model ensemble products, are MOGREPS-G (24 members until 11 July 2017, then 36 members thereafter), ECMWF ENS (51 members) and NCEP GEFS (21 members). The ensemble forecast tracks from the three ensembles are combined to create a 108-member multi-model ensemble (96 members until 11 July 2017), with each individual member from each ensemble equally weighted when creating the track probability forecasts. The track probability forecasts are calculated for named tropical cyclones from each of the three global ensembles, and the combined multi-model ensemble, both for the next 7 days (with no lead time component, to directly verify the forecast product shown in Figure 1 – henceforth described as the "overall" track probability), and for 24-hour forecast periods out to 7 days (in order to evaluate the change in performance over different lead times). Track probabilities are defined as the probability of the named tropical cyclone passing within 120 km of each grid point in the given time period using the commonly used "strike probability" definition first defined in Van der Grijn *et al.* (2004). In this article the term track probability is preferred, because a tropical cyclone is a large weather system that can also "strike" and lead to impacts further away than 120 km from the centre of the storm track, for example from storm surges, flooding, and extreme winds including tornadoes. Thus the term "strike probability" could be misleading to the public as it could falsely imply that those areas outside of 120 km from the centre of the cyclone will not be impacted. Tropical cyclone genesis (storms that are forecast to form during the forecast) is not included in the verification in this article. Observed track data are collected routinely at the Met Office from Global Telecommunications System (GTS) bulletins from the RSMCs, tropical cyclone warning centres, and the Joint Typhoon Warning Center (used if no track is available from the RSMC), to accumulate observed tropical cyclone positions for all named storms to verify the forecasts. The mean track or "consensus" of the three deterministic model tracks is calculated and used as the reference forecast to fully assess the forecast skill.

Since 2015 the verification has been produced biannually, at the end of the Northern and Southern Hemisphere seasons, for tropical cyclones in all global basins over the previous 12 months, in order to evaluate the most recent configurations of each ensemble. This study verifies a 2-year period (January 2017 to December 2018), in order to increase the sample of cases, whilst ensuring the model configurations are still relatively recent.

For each of the named tropical cyclones in the study period, the following steps are carried out in the overall track probability verification process:

1. All observed positions for this storm are read in and the observations are included from the first time period that it has an observed intensity of greater than 34 knots to the last time that it exceeds this value. This ensures that only tropical cyclones that reach tropical storm strength or greater are verified, and that for the included storms, all observations in between these two times are included even if the intensity dips below this threshold in between.
2. A storm-specific verification grid is created for the area covered by the storm track at a resolution of 0.5°.
3. For each of the 0000/1200 UTC forecast run times at which this tropical cyclone was named, a track probability gridded data file is calculated for the observations, the reference consensus forecast, and each forecast model. For the observation data this is a file containing 1's (for grid points which fall within 120 km of the observed track of the storm over the next 7 days), and 0's (those that do not). To create the reference consensus forecast, the mean of the three deterministic forecast tracks is calculated, and a file created containing 1's (for grid points which fall within 120 km of the consensus track of the storm over the next 7 days), and 0's (those that do not). For the ensemble forecast models, each grid point contains a probability value between 0 and 1, calculated by counting how many members have forecast tracks for this storm that lie within 120 km of that point at some point between T+0 and T+168 h, and dividing by the number of ensemble members.
4. For the forecast model data (consensus and ensemble forecasts), the forecasts are only included up to the last time that a matching observation is available. Therefore this verification is focussed on the ability of the models to predict the future track position of the tropical cyclone while it is at tropical storm strength or above, rather than on the ability of the models to weaken and dissipate the storm at the same time as shown in the observations, as in practice the point at which the forecast tracks end is highly dependent on the choice of thresholds for dissipation in the storm tracker.
5. As many forecast runs do not have a full 7 days of matching observed tracks to verify against, there are a considerably smaller overall sample size of tracks verifying at day 7 than at day 1. To prevent the overall track probability verification from being dominated by the verification of shorter lead-time forecasts, which are not so relevant for society given that important decisions such as evacuations generally need to be made on forecasts of two or more days, only those forecast runs

where there are matching observed tracks for T+48 or longer are included.

A similar process is then followed to verify for each 24-hour period from T+0–T+24 through to T+144–T+168. In this case however, the forecasts are verified at the basin-scale, again to replicate the forecast product, which is a basin-based animation of the track probability forecasts for every 24-hour period.

Gridded verification is then carried out comparing the track probability forecast data with the corresponding observed data. A range of probabilistic verification statistics are calculated to assess the skill, reliability and value of the ensemble forecasts. These include the Relative Operating Characteristic (ROC), reliability diagrams, relative economic value and Brier skill score (BSS), as described in Jolliffe and Stephenson (2012), and summarised below:

- The ROC plot assesses the skill of the forecast at discriminating between events and non-events. The points along the curve are the hit rates and false alarm rates for each probability bin. Perfect skill would produce a curve from bottom left to top left to top right, and no skill is indicated by the diagonal line from (0,0) to (1,1).
- Reliability diagrams display how well the predicted probabilities correspond to their observed frequencies. Perfect reliability would be a diagonal line from (0,0) to (1,1), a line above the diagonal indicates under-forecasting and below the diagonal shows over-forecasting, while a line which falls below the diagonal for high probabilities and above it for low probabilities exhibits over-confidence (Wilks, 2011).
- For a given user, their cost-loss ratio is the term given to the ratio of the cost of a preventative measure to the loss averted, and can be used to guide the probability threshold above which to take action. At each probability threshold there will be a 2x2 contingency table containing the number of hits, false alarms, misses and correct rejections. Assuming the user takes action whenever an event is forecast, then a cost (C) can be associated with the hits and false alarms and a loss (L) associated with the misses. The relative economic value can then be calculated for a range of cost-loss ratios and visualised using a relative economic value plot which displays the relative improvement in economic value between the sample climatology and a perfect forecast for all cost-loss ratios (Richardson, 2000). The relative economic value is a useful additional user-focussed measure for comparing forecasts. In practice the user's cost-loss ratio may be difficult to determine and may have to be estimated; however, if one forecast consistently has higher value than another forecast across all cost-loss ratios then it clearly has greater value for any user.
- The BSS assesses the relative skill of the probabilistic forecast over that of a reference forecast in terms of predicting whether an event occurred. A score of 0 indicates no skill when compared to the reference forecast and a score of 1 would be a perfect score. In this article the reference forecasts are the mean or consensus track from the three deterministic forecasts. This is a deliberately challenging reference forecast compared to climatology-based reference forecasts, and the implications of this are discussed in Section 4.

Verification statistics are produced for all tropical cyclones, and then are also split into six global basins to allow for an inter-region comparison, as shown in Figure 2.

3 | RESULTS

3.1 | Overall track probability forecasts for named tropical cyclones

During the verification period of January 2017 to December 2018, 130 named tropical cyclones across the six basins had at least one forecast run where the verification criteria were met and all forecast data were available (24 in NAT, 25 in NEP, 41 in NWP, 7 in NI, 15 in SWI and 18 in AUS, as shown in Figure 2). Overall this two-year period had close to the average number of named tropical cyclones per year, with above-average activity in the North Atlantic basin, and below-average activity in the Southern Hemisphere.

The ROC plot in Figure 3 shows excellent skill for all models; the multi-model ensemble has the largest ROC area (0.985), followed by ECMWF ENS (0.967), MOGREPS-G (0.956) and NCEP GEFS (0.925). All of the models have very low false alarm rates for the majority of forecast probabilities, while the corresponding hit rates vary more significantly across the forecast probabilities. Figure 4 indicates that all models show good reliability, particularly ECMWF ENS and the combined multi-model ensemble. MOGREPS-G and NCEP GEFS are slightly over-confident, with the line falling above the diagonal for low probabilities and below it for higher probabilities. Over-confident reliability diagrams are a common characteristic of ensemble forecasts and are characteristic of an under-spread in the ensemble. So when a high track probability is predicted the ensemble member forecast tracks are sometimes too closely clustered, and the tropical cyclone track probability is less likely than forecast, but when a low probability is forecast the ensemble members are again too closely clustered and so the probability should sometimes be a bit higher.

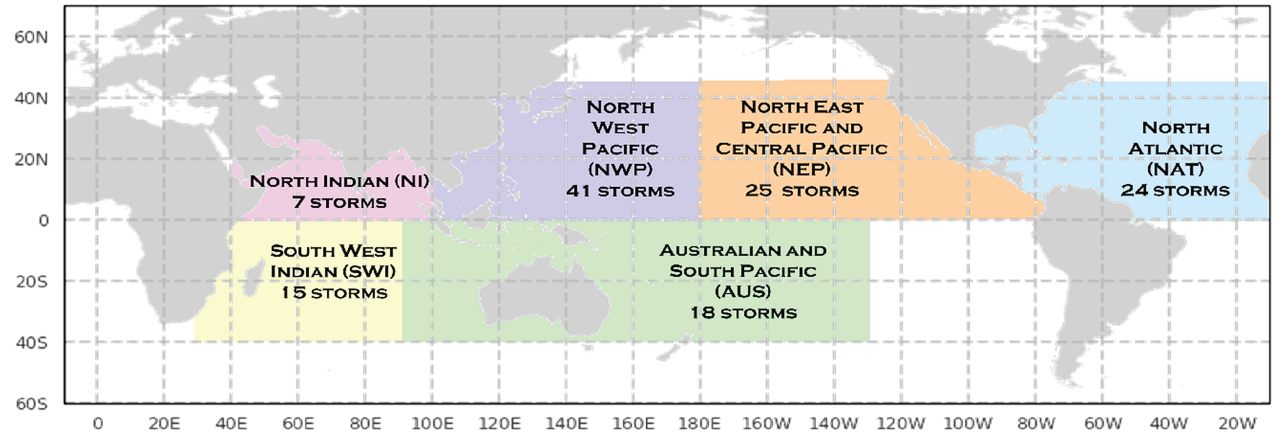

FIGURE 2 The six tropical cyclone basins verified in this study, and the number of named tropical cyclones included in the verification in each basin in the study period (2017 and 2018)

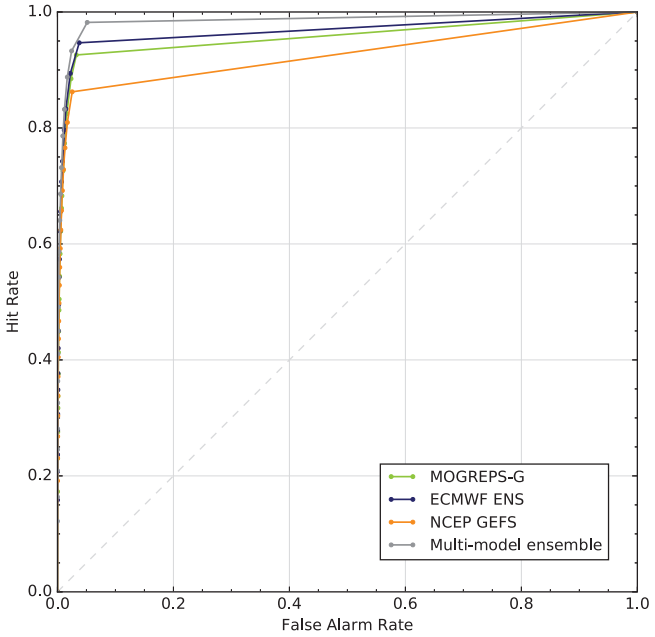

FIGURE 3 ROC plot showing the hit rate and false alarm rate at each probability threshold, for the overall track probability forecasts from the three individual ensembles (MOGREPS-G, ECMWF ENS and NCEP GEFS) and the combined multi-model ensemble

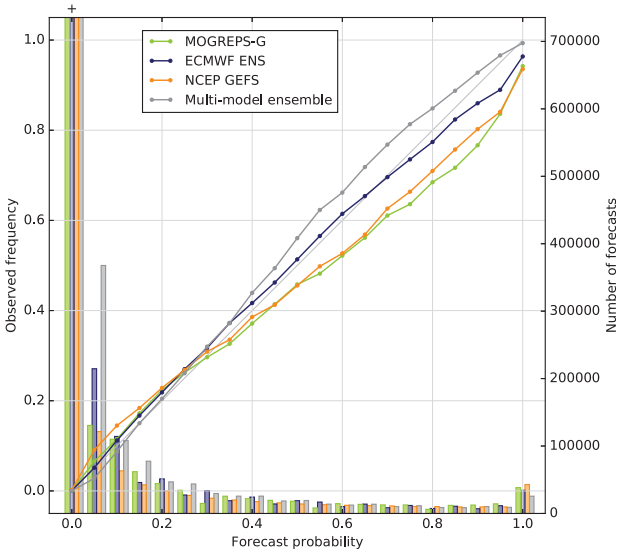

FIGURE 4 Reliability diagram for the overall track probability forecasts for the three individual ensembles (MOGREPS-G, ECMWF ENS and NCEP GEFS) and the combined multi-model ensemble. The number of forecasts (grid points) in each bin is also indicated by bars along the *x*-axis

Figure 5 demonstrates that the combined multi-model ensemble has the greatest relative economic value for all cost-loss ratios, with the multi-model ensemble curve fully encompassing the curves of the three individual models. All models show the greatest relative economic value for low cost-loss ratios (0–0.2) where the loss associated with an event is significantly greater than the cost of acting to mitigate against it when forecast. This is a promising result as although the cost-loss ratios vary for different forecast users, tropical cyclones tend to be associated with low cost-loss ratios due to the severity of their associated hazards and potential impacts.

The BSS, calculated using the consensus of the deterministic models as the reference forecast, is displayed in Figure 6, both for all named tropical cyclones, and split into their relevant basins. Although the absolute and relative performance of the individual ensembles varies from basin to basin, the combined multi-model ensemble has the largest BSS across all basins, showing that the multi-model ensemble adds forecast skill in every basin. The strongest performing individual ensemble is ECMWF ENS in all basins except the Southwest Indian where NCEP GEFS has the largest BSS of the individual ensembles. The relative performance of MOGREPS-G and NCEP GEFS varies between basins.

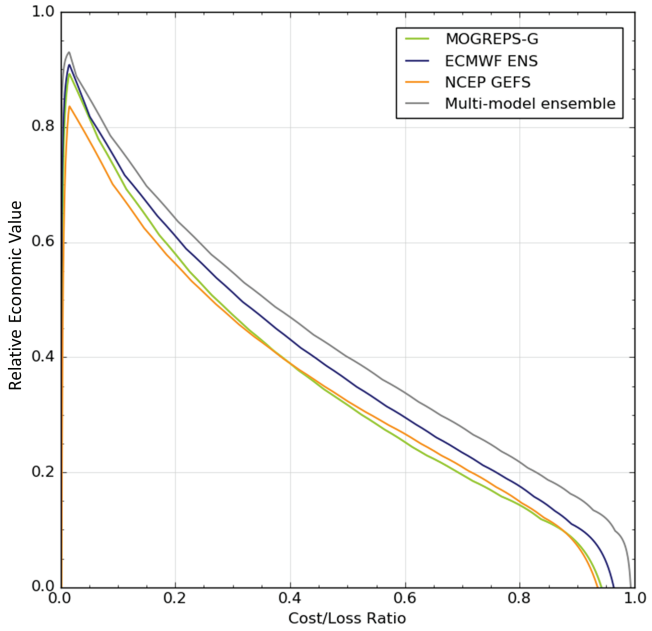

FIGURE 5 Relative economic value plot displaying the relative economic value at each cost-loss ratio for the overall track probability forecasts from the three individual ensembles (MOGREPS-G, ECMWF ENS and NCEP GEFS) and the combined multi-model ensemble

3.2 | Track probability forecasts for named tropical cyclones over each 24-hour forecast period

To illustrate the changes in forecast characteristics and performance against lead time, the reliability diagrams and relative economic value plots shown in Figures 7 and 8 display the 24-hour track probability centred on six lead times from T+24 to T+144. The reliability diagrams (Figure 7) all exhibit some degree of over-confidence in the forecasts, which varies with model and lead time. The lowest probability bin is slightly above the diagonal (as there are some forecasts where the forecasts are confident there will be no tropical cyclone and in fact the tropical cyclone does track that way), but this is not visible by eye due to the very large numbers of correct near-zero probabilities. The other probability bins are below the diagonal, where the forecasts are over-forecasting the tropical cyclone track probability. At T+24 (Figure 7a) there is good reliability for all models, but at the higher probability bins (>0.5), the advantage of the multi-model ensemble is evident compared to the three individual ensembles, which all show a similar degree of over-confidence. As lead time increases, reliability begins to decrease in the NCEP GEFS and MOGREPS-G ensembles, while ECMWF ENS, and especially the multi-model ensemble, maintain very good reliability particularly at probabilities greater than 0.4 at T+72 and T+96 (Figure 7c,d). MOGREPS-G and NCEP GEFS are showing significant over-confidence by T+120 (Figure 7e), indicating that these models are significantly under-dispersive in their five-day track forecasts. At T+120, the multi-model ensemble remains the most reliable at probabilities less than 0.8. This is an important result, as it shows that MOGREPS-G and NCEP GEFS, despite being significantly under-dispersive by this point, are still adding benefit to the ECWMF ENS in the multi-model ensemble at most forecast probabilities. However, at this lead time and in particular at T+144 (Figure 7f)

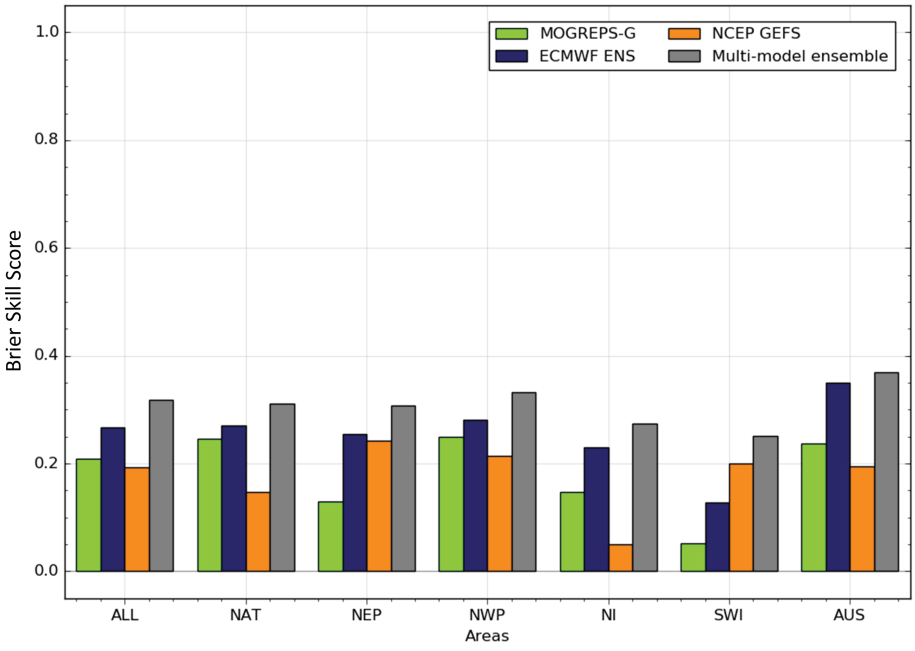

FIGURE 6 BSS for the overall track probability forecasts for all named tropical cyclones (left) and for named tropical cyclones in each of the individual six basins, for each individual ensemble and the multi-model ensemble. The reference forecast used in the skill score calculation is the consensus (mean) track of the three deterministic forecasts

there are far fewer cases in the higher probability bins, leading to more noisy results.

In Figure 8, nearly all cost-loss ratios (0–0.9) have relative economic value at T+24 (Figure 8a). The peak value is at low cost-loss ratios, with approximately 0.9 relative economic value for a cost-loss ratio of 0–0.1. As lead time increases there is a gradual drop in value, particularly in the higher cost-loss ratios, but even by T+144 significant value still remains at 0–0.1 cost-loss ratios (Figure 8f). At all lead times the multi-model ensemble value curve fully encompasses those for the individual ensembles. Out to T+72 (Figure 8c) the value of the multi-model ensemble is significantly greater than any of the individual ensembles, but by T+120 (Figure 8e) the value for the ECMWF ENS is noticeably greater than the other two ensembles, and closer to the value shown by the multi-model ensemble.

The area under the ROC curve (AUC) values are shown against lead time in Figure 9. The AUC decreases with lead time for all models. The multi-model ensemble maintains an AUC of over 0.93 at all lead times, whereas the AUC score for the individual ensemble decreases more significantly with lead time (in particular for NCEP GEFS). The Brier skill score (BSS) results in Figure 10 show that the skill of the 24-hour track probability forecasts varies with lead time in each basin, with the skill compared to the deterministic consensus forecast rising with increasing lead time. This is as expected, and shows that the benefit of probabilistic forecasts over a consensus of deterministic forecasts becomes greater as the forecast lead time increases. When all tropical cyclones are included, irrespective of the basin (the light green line in Figure 10), the multi-model ensemble probabilistic forecasts show positive skill over the multi-model consensus from T+60 onwards. However, this hides big differences between the basins. In the North Atlantic and North East Pacific basins the BSS is positive throughout and rise to a value of 0.3 by T+156. In the Northwest Pacific basin the BSS is positive from T+48 onwards. However, in the North Indian Ocean, the Southwest Indian Ocean and the Australian basins, the BSS stays negative throughout, showing that in these basins the ensembles are not adding skill to the consensus reference forecast for the 24-hour track probability forecasts in this study period. The negative BSS for these basins in their 24-hour track probability forecasts is in contrast to the positive BSS in their overall 7-day track probability forecasts that was shown in Figure 6, and highlights that the 7-day verification is more forgiving of along-track errors. It is also worth noting that these are also the three basins where the sample size of storms is the smallest. A case-by-case assessment of the tropical cyclones in the worst-performing SWI basin reveals that the deterministic errors were relatively low compared to previous seasons, making the consensus reference forecast hard to beat for this sample of storms.

3.3 | Verification for two high-profile tropical cyclones

Figure 11 compares the BSS for the storm-based verification of two high-profile tropical cyclones in the North Atlantic basin: Hurricane Matthew in 2016 and Hurricane Irma in 2017. For each storm a different individual ensemble displays the highest skill (MOGREPS-G for *Matthew* and ECMWF ENS for *Irma*), illustrating that even within the same basin, the strongest performing individual ensemble varies from storm to storm, rather than one ensemble always being the most skilful in a given basin. In both cases, the multi-model ensemble shows comparable forecast skill to the strongest performing model. Figure 12 shows one of the forecasts included in the storm-based verification in each case, illustrating the strong performance of MOGREPS-G for *Matthew* and ECMWF ENS for *Irma*. For Hurricane Matthew, the MOGREPS-G ensemble was the first model to give a strong signal for the storm to track just off the Florida coast with an eventual landfall in South Carolina. For Hurricane Irma, the ECMWF ensemble was the only ensemble to contain the observed track in the ensemble track spread in the early forecast runs. At the time of the forecast it is not known which of the individual ensemble forecast models will have the greatest forecast skill for that particular storm, and so the key result here is that in both cases the multi-model ensemble probability forecasts were able to provide equivalent forecast skill to the best performing individual ensemble forecasts.

4 | DISCUSSION

In Section 1, five key questions were laid out to be addressed by the probabilistic evaluation of tropical cyclone track probability forecasts in this study. This discussion section is organised to answer these questions, and draw out the key implications from these:

- *How do forecast performance and characteristics vary between the global ensembles, and with lead time?*

All ensembles exhibit good reliability and value in the named tropical cyclone track probability forecasts, particularly at low cost-ratios. This shows how the probability forecasts have huge potential to be useful to decision makers and downstream users of tropical cyclone forecasts, who will often have low cost-loss ratios due to high potential losses to property and personal safety, compared

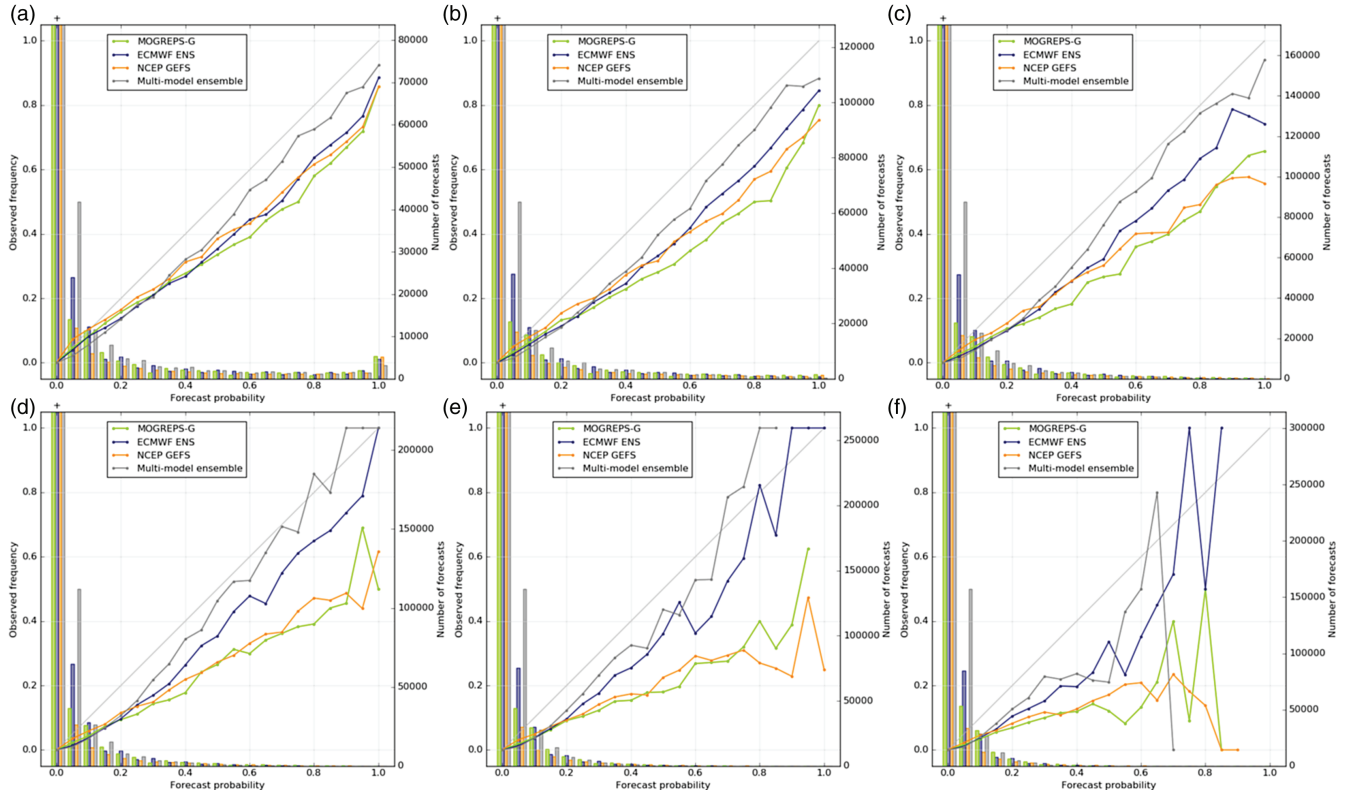

FIGURE 7 Reliability diagrams for 24-hour track probability forecasts for all named storms for each individual ensemble and the multi-model ensemble, for 24-hour periods centred on (a) T+24, (b) T+48, (c) T+72, (d) T+96, (e) T+120 and (f) T+144

to the cost of mitigative efforts such as putting up shutters or evacuating. It highlights the importance of initiatives to increase the pull-through of probabilistic situation-based uncertainty information into operational warnings, such as the collaboration through the WMO HIWeather project that is discussed in Titley *et al.* (2019).

The best performing of the three individual ensembles included in the study, in terms of the verification statistics presented for tropical cyclone track probability forecasts, is the ECMWF ENS, followed by MOGREPS-G and NCEP GEFS, which are both over-confident in their track probability forecasts, indicating the known tendency of these models to be under-dispersive. There are many differences between the three ensemble forecast systems, including differences in data assimilation influencing both the initial conditions of the storm itself and the wider environmental steering flow, the model formulation, and the ensemble perturbation strategies. The relative contribution of perturbations from an ensemble of data assimilations, singular vectors, and stochastic model perturbations to ECMWF ENS track spread was presented in Lang *et al.* (2012). Benefits of the ECMWF ENS perturbation strategy includes the ability to target singular vectors on tropical cyclones, and an enhanced ability to be able to tune perturbations to give improved spread and reliability, compared with the perturbation schemes in the other ensembles.

In the verification of the track probability forecasts for each 24-hour period, all models were shown to be over-confident in their track probability forecasts, but this became much more pronounced at longer lead times for MOGREPS-G and NCEP GEFS, where the higher probabilities were forecast far too often compared to the observed frequency. These results are important to model developers as they show the importance of increasing the ensemble spread in the MOGREPS-G and NCEP GEFS ensembles. In MOGREPS-G, a major upgrade was scheduled to go live in autumn 2019 to try to address this issue. The ensemble perturbation system will be changed from Ensemble Transform Kalman Filter (ETKF) to an ensemble of data assimilations (En-4D-En-Var: Bowler *et al.*, 2017). In the new system, data assimilation is performed for each member, creating increments relative to its own background trajectory. A partial re-centring around the deterministic analysis gives an additional increase in skill and reduces jumpiness. Comparative trials of the new En-4D-En-Var ensemble have shown faster spread growth across many variables including 850 hPa wind speed in the tropics, with a much better match to observed errors. The ensemble trials have also been processed through

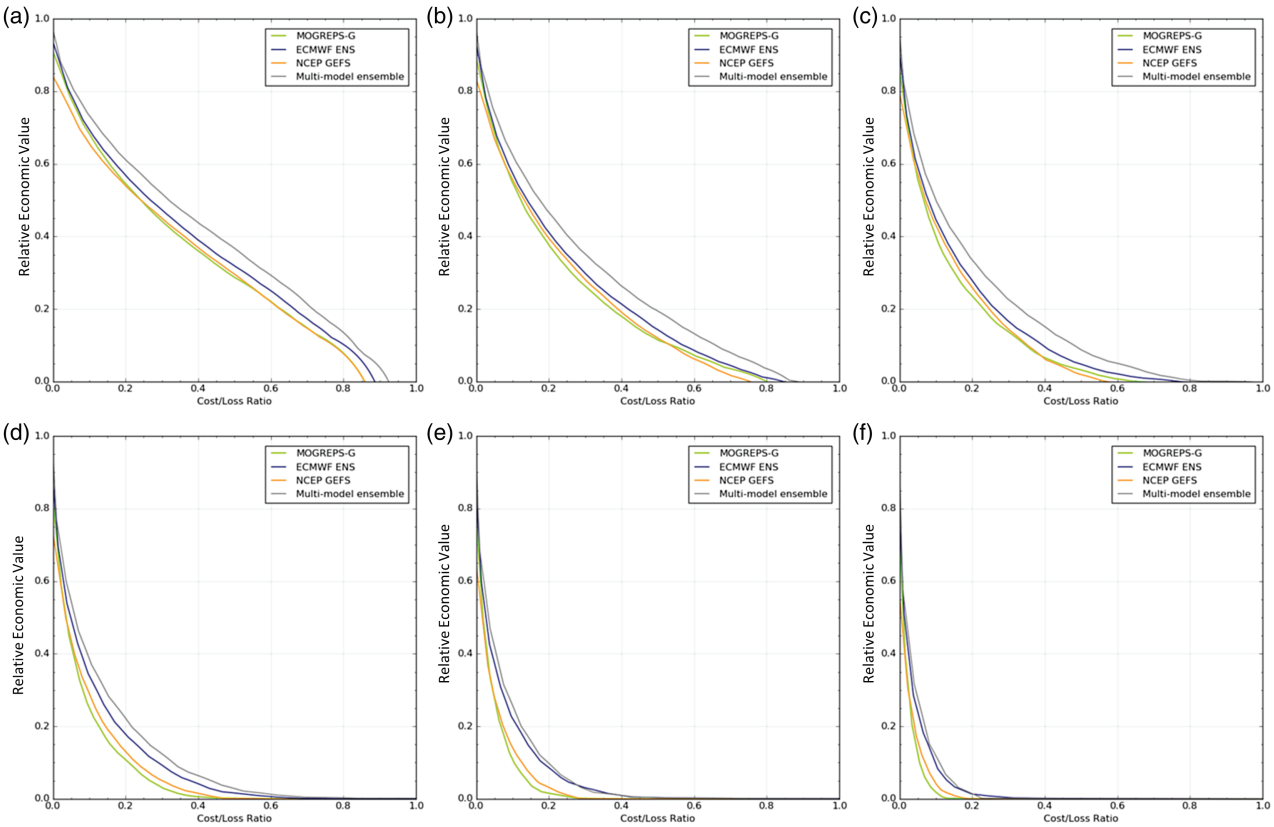

FIGURE 8 Relative economic value curves for 24-hour track probability forecasts for all named storms for each individual ensemble and the multi-model ensemble, for 24-hour periods centred on (a) T+24, (b) T+48, (c) T+72, (d) T+96, (e) T+120 and (f) T+144

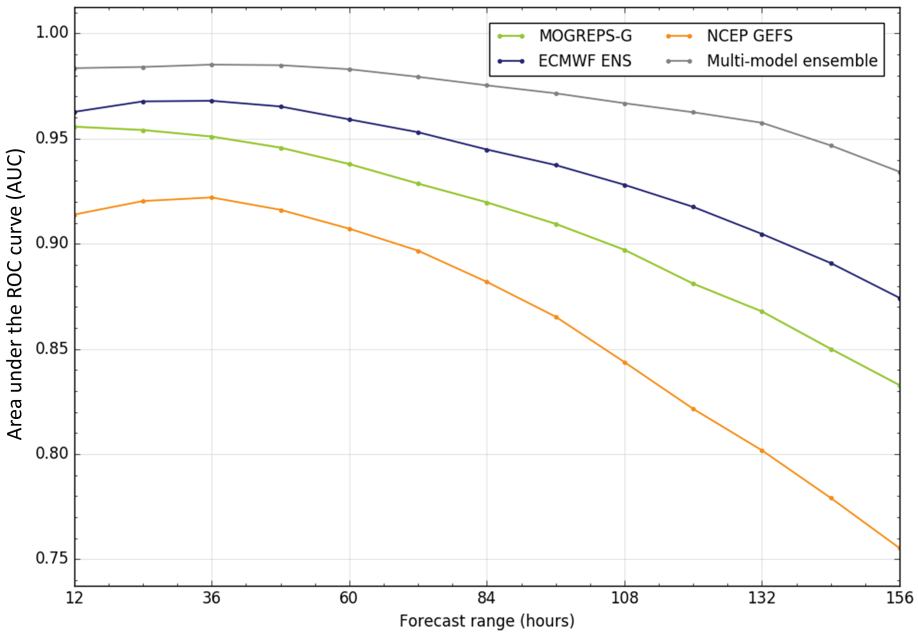

FIGURE 9 The AUC for each individual ensemble and the multi-model ensemble, for the 24-hour track probability forecasts centred on each forecast range

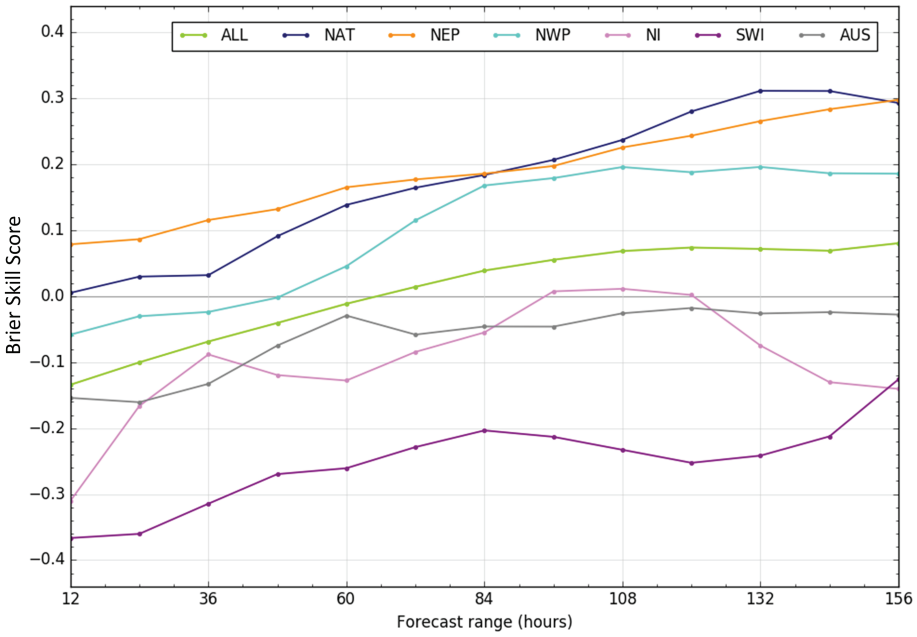

FIGURE 10 BSS for the multi-model ensemble, for all storms (in green) and split into the six tropical cyclone basins, for the 24-hour track probability forecasts centred around each forecast range

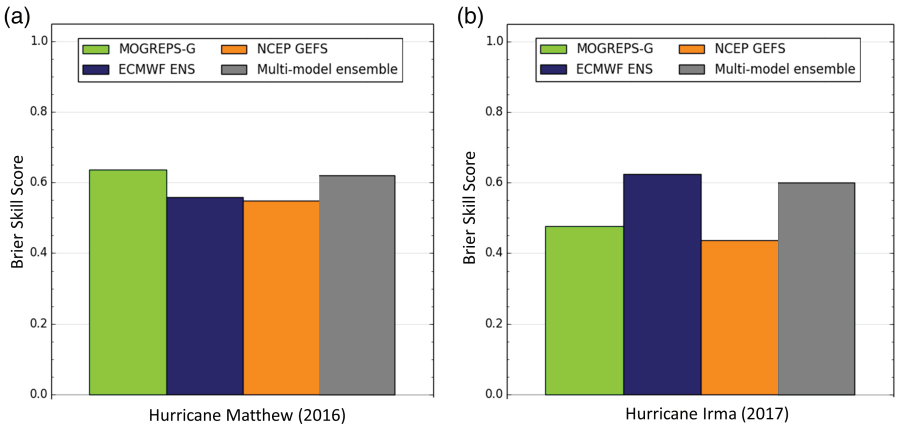

FIGURE 11 Example storm-specific verification for all forecast runs of (a) Hurricane Matthew (2016) and (b) Hurricane Irma (2017): Brier skill score for MOGREPS-G, ECMWF ENS, NCEP GEFS and multi-model ensemble forecasts (reference forecast in this verification is the sample climatology)

the tropical cyclone tracking and post-processing system, and show a significant improvement in track spread at all lead times. Meanwhile at NCEP a new version of the deterministic Global Forecast System (GFS) with a new Finite-Volume Cubed-Sphere Dynamical Core (FV3) went operational in June 2019. The potential of this new model to improve hurricane forecast performance was described by Chen *et al.* (2019), who showed the improved performance in re-runs of the active 2017 Atlantic hurricane season. The FV3 dynamical core will be implemented into the GEFS ensemble in 2020, in an upgrade that will also see improved stochastic physics and an increase in ensemble members. The ECMWF ENS also continues to be upgraded, with IFS Cycle 46r1 implemented in June 2019, including a data assimilation upgrade which improves the initial conditions of the ensemble forecasts. Additional work to improve the tropical cyclone intensity forecasts in the ECMWF ENS is described in Magnusson *et al.* (2019). The track probability verification described in this article will continue to run every 6 months to verify the three ensemble models and their combined multi-model ensemble to investigate the impact of these model upgrades on the skill of tropical cyclone track probability forecasts.

- *Is there benefit in terms of probabilistic forecast skill, reliability and value, from combining the three global ensembles into a multi-model ensemble?*

As established in the introduction, the rationale of multi-model ensemble forecasting lies in combining independent and skilful models, each with its own strengths and weaknesses. The evaluation in this article clearly shows that additional forecast skill and value can be gained from combining the members from the three individual ensembles included in this study into a multi-model ensemble. The three ensembles have different data assimilation strategies, model formulations and ensemble perturbation schemes, that when combined together are

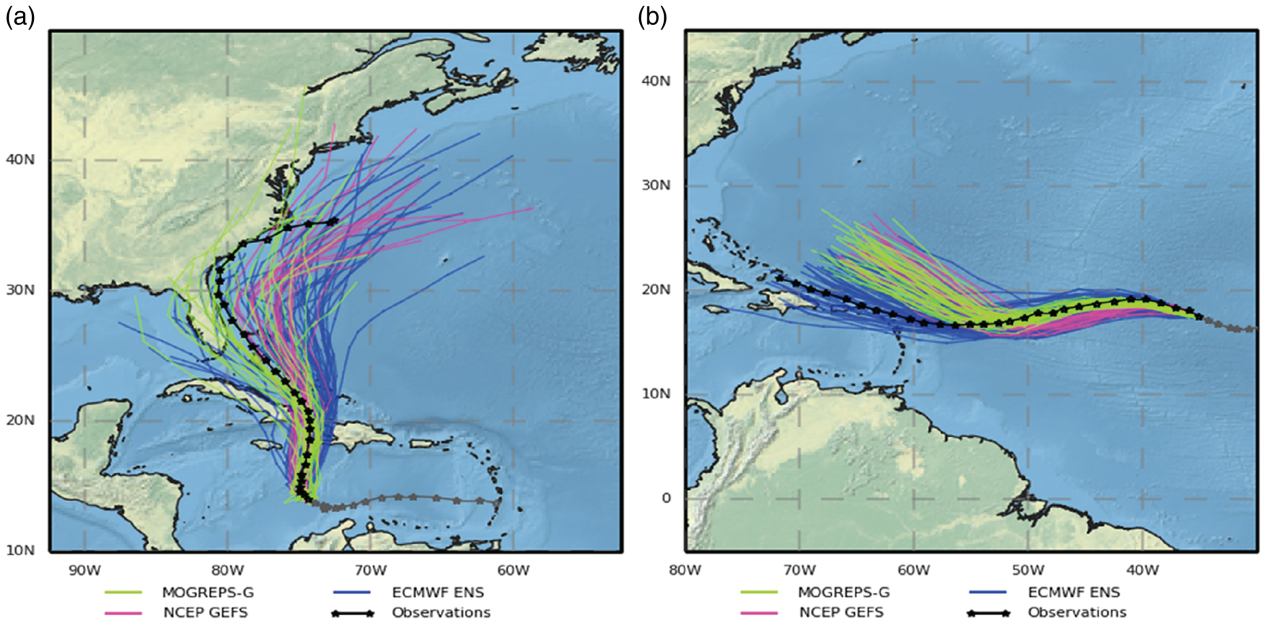

FIGURE 12 Ensemble member tracks from one of the forecast runs included in the verification of each case: (a) 1200 UTC 2 October 2016 forecast for Hurricane Matthew; (b) 0000 UTC 1 September 2017 forecast for Hurricane Irma. Tracks are coloured according to model (green = MOGREPS-G, blue = ECMWF ENS and pink = NCEP GEFS)

shown to collectively provide more realistic estimates of tropical cyclone track probabilities. In the overall track probability verification, the relative economic value curve of the multi-model ensemble fully encompasses that of the individual ensembles, and in the verification for each 24-hour period it is particularly noteworthy that the AUC decreases more significantly with lead time in the individual ensembles compared to the multi-model ensemble. This illustrates that the amount of potential forecast skill to be gained from using a multi-model ensemble to derive your track probabilities increases significantly with forecast lead time. This finding is important for operational tropical cyclone forecasters, forecast centre managers, and numerical weather prediction centres, showing the importance of improving access to multiple ensemble forecast model forecasts in order to allow multi-model ensemble information to be used in the operational tropical cyclone forecasting process.

- *Does using the full probability forecast information via the multi-model ensemble add skill compared to a "consensus" forecast of the parent deterministic models?*

The choice of reference forecast is crucial in the Brier Skill Score (BSS) calculation, as the BSS measures the improvement of the probabilistic forecast relative to a reference forecast. This forecast reference is often calculated from the sample climatology, or, in the case of tropical cyclone forecasts, from a combined climatology and persistence forecast (CLIPER: Knaff *et al.*, 2003). However, in order to be more relevant for the current common forecasting practice of establishing a consensus forecast, and in order to provide a more skilful and challenging reference for the ensemble probabilities to be compared to, a consensus forecast based on the mean track of the three parent deterministic models was created and used as the reference in the BSS calculation. The BSS for the overall 7-day track probability skill scores shows positive skill for all of the ensemble forecast models in all basins when compared to the reference consensus track, with the BSS being highest for the combined multi-model ensemble. This shows the added benefit to be gained in operational tropical cyclone track forecasting if the full probability information provided by the ensembles can be pulled through into operational warnings, as has begun to happen with the incorporation of dynamic uncertainty information in warning products from several centres including RSMC La Réunion and RSMC Tokyo (Titley *et al.*, 2019). When the track probability verification is split into 24-hour periods it is harder to beat the consensus reference, particularly at short lead times, but the BSS increases with forecast range, showing the increasing value of ensemble prediction over consensus forecasting with lead time. Overall there is positive forecast skill from T+60 onwards. The BSS was also calculated for the 24hour track probability verification using the sample climatology (not shown) and showed the reverse pattern, with the highest BSS (0.6) at shorter lead times, slowly decreasing with forecast range. This is to be expected since the sample climatology reference has no forecast capability, whereas the consensus has a very

strong forecast capability at short range which reduces with increasing lead time. This illustrates how important it is to be clear about the implications and rationale of selecting a particular reference forecast.

- *When comparing basin to basin, does forecast performance vary, both overall compared to a reference consensus forecast, and relatively between the different ensembles?*

In the overall track probability forecasts BSS results (Figure 6), the combined multi-model ensemble has the largest BSS of all the models, and varies between 0.25 and 0.4 with each basin, showing that the multi-model ensemble is adding skill in every basin. The strongest performing individual ensemble is ECMWF ENS in all basins except the SWI basin where NCEP GEFS has the highest skill. The relative performance of MOGREPS-G and NCEP GEFS varies between basins, with MOGREPS-G performing well in the NAT, NWP and AUS basins, and NCEP GEFS performing well in the SWI and NEP basins. This result would provide useful guidance to operational tropical cyclone forecasters in these regions, as it provides a breakdown of the current levels of performance of each model in their area of interest, and confirms that all areas would see additional value in computing track probability forecasts from multi-model ensemble forecast data. It is also of interest to model developers, who could investigate why their model performs better relatively for tropical cyclones in one basin over another, leading to potential ideas on how to improve forecast skill. As discussed earlier, each model has different perturbation strategies and data assimilation schemes, with differences in which observations are included (including differences in quality-control systems and in the assimilation of RSMC tropical cyclone observations, known as bogussing, which is not currently carried out at ECMWF). These results show that some aspects of each model system may be better suited to one region over another, and further investigations including the implications of a tropical cyclone being in a data-rich vs. data-poor area may lead to further insights and improvements in the model forecasts.

The inter-basin comparison of BSS using the 24-hour track probability forecasts centred on each lead time (Figure 10) reveals large differences from basin to basin in the skill compared to the consensus reference forecast, with positive skill in the NAT, NWP and NEP, but not in the NI, SWI or AUS basins. This emphasises that the traditional deterministic consensus is hard to beat in regions where the tracks of the named storms in the study period are well forecast by the deterministic models. The lower BSS for the 24-hour track probability forecasts highlights the need to focus on improving the ensemble forecasts for the translation speed of the storm, as the results when splitting the track probability forecasts in to 24-hour periods will be impacted by along-track errors in addition to the cross-track errors, whereas the cross-track errors are the most influential factor in the overall 7-day track probability results. Although forecast users predominantly need to know whether or not they will be impacted by a hurricane rather than when, the timings are also important in forecast preparedness activities, and future verification could investigate the relative along- and cross-track errors in the ensemble vs. the deterministic consensus.

- *When comparing storm to storm, how does forecast performance vary, and what does it tell us about the benefit of multi-model ensembles?*

Although when averaged across all tropical cyclones the ECMWF ENS is clearly currently the most skilful of the individual ensembles, the case-study analysis in Figure 11 showed that in a particular case this will not always hold true. Sometimes a different model is more skilful in a particular storm (as in the case for MOGREPS-G for Hurricane Matthew). At the time of the live forecasts, operational forecasters will not know which individual model is destined to be the most skilful for that storm, and therefore the result showing that the multi-model ensemble skill matches that of the strongest performing individual ensemble is an important result, and clearly illustrates the benefit of combining the ensemble members into a multi-model ensemble when computing track probability forecasts.

5 | CONCLUSIONS AND FUTURE PLANS

This study has shown that combined multi-model ensemble tropical cyclone track probability forecasts, calculated from all members of ECMWF ENS, MOGREPS-G and NCEP GEFS, have increased skill and value over the best-performing individual ensemble. This result is consistent when verifying all global named tropical cyclones together, and when the verification is carried out for each individual basin. The improved skill and value of the multi-model ensemble is found for both the full (up to 7-day) track probability forecasts, and for track probabilities split into 24-hour forecast periods. The verification results from the three individual ensembles show that the track probability forecasts from ECMWF ENS display the best reliability, skill and value. MOGREPS-G and NCEP GEFS become increasingly over-confident and under-dispersive with increasing forecast range, emphasising the importance of ongoing work at both centres to improve the perturbation strategy and increase the spread in the ensemble forecasts. However, even when there is an individual ensemble model that on average performs

better in a particular storm, basin, or overall, there is still skill, reliability and value to be gained from adding additional ensemble models into a combined multi-model ensemble, indicating that they have independent systematic strengths and errors and collectively provide more realistic estimates of tropical cyclone track probabilities than individually.

Storm-based verification illustrated that the best-performing individual ensemble can change from case to case, but that the multi-model ensemble matches the best-performing individual ensemble, which would not be known in advance, in terms of overall probabilistic forecast skill. The mean, or consensus, of the three higher-resolution deterministic forecasts is hard to beat in some basins, but overall the additional probabilistic skill of the ensembles is shown, particularly at longer lead times or when computing the full 7-day track probability forecasts. This study helps to illustrate the potential value and skill to be gained if operational tropical cyclone forecasting can continue to migrate away from a deterministic-focussed forecasting environment to one where the probabilistic situation-based uncertainty information provided by the ensembles can be pulled through into operational forecasts and warnings.

There are many ideas for where this work could be further applied and extended in the future. In this study no weighting is applied to the ensemble members in the multi-model ensemble, with each member from each model given an equal weight. It would be interesting to look at different options for combining together the ensemble members, and also investigate how much of the additional value is from additional members as opposed to the inclusion of members from other models. For example, would the multi-model ensemble still add value to the ECMWF ENS if it were restricted to the same number of members? Additional global ensemble forecast models, including those available in the TIGGE cyclone CXML archive from the Japan Meteorological Agency, the Canadian Meteorological Centre and Météo-France, could be added to the multi-model ensemble, to investigate the optimal combination of ensembles in track probability forecasts. At the Met Office the ensemble tropical cyclone products and verification will continue to be used to evaluate important model upgrades and trials, such as the impact of the forthcoming move from ETKF to Ens-4D-En-Var perturbations. The products and verification are also run on high-resolution convective-permitting ensembles in Southeast Asia. Work is also underway to extend the ensemble tropical cyclone verification capability at the Met Office to incorporate a verification of forming storms (tropical cyclone genesis) and intensity trends. It is also important to look beyond the traditional tropical cyclone track and intensity forecasts and move towards verifying the associated hazards, for example to assess how the uncertainty and predictability of the track translates through to uncertainty and predictability of the precipitation, and the downstream flood hazard. Ongoing collaborations between global numerical weather prediction centres, researchers and operational forecasting centres continue to be essential to ensure that future research and ensemble model developments are of maximum benefit to operational tropical cyclone forecasting.

ACKNOWLEDGEMENTS

The authors would like to thank three reviewers for their helpful comments and suggestions that have led to improvements to the article, Julian Heming for his support with the use of MOTCTracker and for providing the observations for use in the verification, and Ken Mylne, Philip Gill and Rutger Dankers for their support during this work. This work was partially supported by the Met Office Weather and Climate Science for Service Partnership (WCSSP) India Programme as part of the Newton–Bhabha Fund.

ORCID

Helen A. Titley https://orcid.org/0000-0003-1654-9826
Rebecca L. Bowyer https://orcid.org/0000-0001-7584-4425
Hannah L. Cloke https://orcid.org/0000-0002-1472-868X

REFERENCES

Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B. and Beare, S.E. (2008) The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134, 703–722.

Bowler, N.E., Clayton, A.M., Jardak, M., Jermey, P.M., Lorenc, A.C., Wlasak, M.A., Barker, D.M., Inverarity, G.W. and Swinbank, R. (2017) The effect of improved ensemble covariances on hybrid variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703), 785–797.

Chen, J.H., Lin, S.J., Magnusson, L., Bender, M., Chen, X., Zhou, L., Xiang, B., Rees, S., Morin, M. and Harris, L. (2019) Advancements in hurricane prediction with NOAA's next-generation forecast system. *Geophysical Research Letters*, 46, 4495–4501.

Du, Y.G., Qi, L.B. and Cao, X.G. (2016) Selective ensemble-mean technique for tropical cyclone track forecast by using time-lagged ensemble and multi-centre ensemble in the western North Pacific. *Quarterly Journal of the Royal Meteorological Society*, 142, 2452–2462.

Dupont, T., Plu, M., Caroff, P. and Faure, G. (2011) Verification of ensemble-based uncertainty circles around tropical cyclone track forecasts. *Weather and Forecasting*, 26, 664–676.

Goerss, J.S. (2000) Tropical cyclone track forecasts using an ensemble of dynamical models. *Monthly Weather Review*, 128, 1187–1193.

Goerss, J.S. (2007) Prediction of consensus tropical cyclone track forecast error. *Monthly Weather Review*, 135, 1985–1993.

Goerss, J.S. and Sampson, C.R. (2014) Prediction of consensus tropical cyclone intensity forecast error. *Weather and Forecasting*, 29, 750–762.

Hagedorn, R., Buizza, R., Hamill, T.N., Leutbecher, M. and Palmer, T.N. (2012) Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138, 1814–1827.

Hagedorn, R., Doblas-Reyes, F.J. and Palmer, T.N. (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept. *Tellus*, 57A, 219–233.

Hamill, T.M. (2012) Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review*, 140, 2232–2252.

Heming, J.T. (2017) Tropical cyclone tracking and verification techniques for Met Office numerical weather prediction models. *Meteorological Applications*, 24, 1–8.

Hodges, K.I. (1995) Feature tracking on the unit sphere. *Monthly Weather Review*, 123, 3458–3465.

Johnson, C. and Swinbank, R. (2009) Medium-range multimodel ensemble combination and calibration. *Quarterly Journal of the Royal Meteorological Society*, 135, 777–794.

Jolliffe, I.T. and Stephenson, D.B. (2012) *Forecast Verification: A practitioner's guide in atmospheric science*, 2nd edition. Chichester: Wiley and Sons Ltd.

Knaff, J.A., DeMaria, M., Sampson, C.R. and Gross, J.M. (2003) Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Weather and Forecasting*, 18, 80–92.

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. *Quarterly Journal of the Royal Meteorological Society*, 138, 2030–2046.

Leonardo, N.M. and Colle, B.A. (2017) Verification of multimodel ensemble forecasts of North Atlantic tropical cyclones. *Weather and Forecasting*, 32, 2083–2101.

Magnusson, L., Bidlot, J.R., Bonavita, M., Brown, A.R., Browne, P.A., 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. and Malardel, S. (2019) ECMWF activities for improved hurricane forecasts. *Bulletin of the American Meteorological Society*, 100, 445–458.

Majumdar, S.J. and Finocchio, P.M. (2010) On the ability of global ensemble prediction systems to predict tropical cyclone track probabilities. *Weather and Forecasting*, 25, 659–680.

Matsueda, M. and Nakazawa, T. (2015) Early warning products for severe weather events derived from operational medium-range ensemble forecasts. *Meteorological Applications*, 22, 213–222.

Palmer, T.N. (2019) The ECMWF ensemble prediction system: looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 12–24. https://doi.org/10/1002/qj.3383.

Park, Y.Y., Buizza, R. and Leutbecher, M. (2008) TIGGE: preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134, 2029–2050.

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.

Robbins, J.C. and Titley, H.A. (2018) Evaluating high-impact precipitation forecasts from the Met Office Global Hazard Map (GHM) using a global impact database. *Meteorological Applications*, 25, 548–560.

Sampson, C.R., Franklin, J.L., Knaff, J.A. and DeMaria, M. (2008) Experiments with a simple tropical cyclone intensity consensus. *Weather and Forecasting*, 23, 304–312.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T.M., Hewson, T.D., Keller, J.H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H.A., Wilson, L. and Yamaguchi, M. (2016) The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97, 49–67.

Tallapragada, V., Bernardet, L., Gopalakrishnan, S., Kwon, Y., Liu, Q., Marchok, T., Sheinin, D., Tong, M., Trahan, S., Tuleya, R., Yablonsky, R. and Zhang, X. (2013) *Hurricane Weather Research and Forecasting (HWRF) model: 2013 scientific documentation*. HWRF Development Testbed Center, Technical Report 99. Available at: http://www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWRFv3.5a_ScientificDoc.pdf [Accessed 4th December 2019].

Titley, H.A., Yamaguchi, M. and 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, 166–180.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3318.

Van der Grijn, G. (2002) *Tropical Cyclone Forecasting at ECMWF: New products and validation*. ECMWF Technical Memorandum, Vol. 386. Reading, UK: ECMWF, 1–13.

Van der Grijn, G., Paulsen, J.E., Lalaurette, F. and Leutbecher, M. (2004) *Early medium-range forecasts of tropical cyclones*. ECMWF Newsletter, No. 102, Reading, UK, 7–14. Available at: https://www.ecmwf.int/file/28462/download?token=i0qbDD6S.

Vitart, F. and Stockdale, T.N. (2001) Seasonal forecasting of tropical storms using coupled GCM integrations. *Monthly Weather Review*, 129, 2521–2537.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences. International Geophysics Series, 100*, 3rd edition. San Diego, CA: Elsevier, Academic Press.

Yamaguchi, M., Ishida, J., Sato, H. and 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, 2337–2349.

Yamaguchi, M. and Koide, N. (2017) Tropical cyclone genesis guidance using the early stage Dvorak analysis and global ensembles. *Weather and Forecasting*, 32, 2133–2141.

Yamaguchi, M., Nakazawa, T. and Hoshino, S. (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.

Yamaguchi, M., Vitart, F., Lang, S.T., 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. *Weather and Forecasting*, 30, 1695–1709.

Zhang, X.P. and Yu, H. (2017) A probabilistic tropical cyclone track forecast scheme based on the selective consensus of ensemble prediction systems. *Weather and Forecasting*, 32, 2143–2157.

How to cite this article: Titley HA, Bowyer RL, Cloke HL. A global evaluation of multi-model ensemble tropical cyclone track probability forecasts. *Q J R Meteorol Soc*. 2020;146:531–545. https://doi.org/10.1002/qj.3712