Seasonal forecasts of North Atlantic tropical cyclone activity in the North American Multi-Model Ensemble

Seasonal Forecasts of North Atlantic Tropical Cyclone Activity in the North American Multi-Model Ensemble

Julia V. Manganello[1], Benjamin A. Cash[1],Kevin I. Hodges[2], James L. Kinter III[1,3]

[1] Center for Ocean-Land-Atmosphere Studies (COLA), George Mason University (GMU), Fairfax, Virginia, USA

[2] Dept. of Meteorology, University of Reading, Reading, UK

[3] George Mason University (GMU), Fairfax, Virginia, USA

For Submission to the NMME Special Issue in Climate Dynamics

Corresponding Author:
Julia V. Manganello
Center for Ocean-Land-Atmosphere Studies
113 Research Hall, Mail Stop 2B3
George Mason University
4400 University Drive
Fairfax, VA 22030 USA
jvisneva@gmu.edu
Phone:703-993-5716
Fax: 703-993-5700

Abstract

The North American Multi-Model Ensemble (NMME)-Phase II models are evaluated in terms of their retrospective seasonal forecast skill of the North Atlantic (NA) tropical cyclone (TC) activity, with a focus on TC frequency. The TC identification and tracking algorithm is modified to accommodate model data at daily resolution. It is also applied to three reanalysis products at the spatial and temporal resolution of the NMME-Phase II ensemble to allow for a more objective estimation of forecast skill. When used with the reanalysis data, the TC tracking generates realistic climatological distributions of the NA TC formation and tracks, and represents the interannual variability of the NA TC frequency quite well.

Forecasts with the multi-model ensemble (MME) when initialized in April and later tend to have skill in predicting the NA seasonal TC counts and TC days. At longer leads, the skill is low or marginal, although one of the models produces skillful forecasts when initialized as early as January and February. At short lead times, while demonstrating the highest skill levels the MME also tends to significantly outperform the individual models and attain skill comparable to the reanalysis. In addition, the short-lead MME forecasts are quite reliable. It is found that the overall MME forecast skill is limited by poor representation of the low-frequency variability in the predicted NA TC frequency, and large fluctuations in skill on decadal time scales. Addressing these deficiencies is thought to increase the value of the NMME ensemble in providing operational guidance.

1. Introduction

Recognizing high socioeconomic significance of tropical cyclone (TC) prediction, dynamical seasonal forecasts of TC activity have been pursued since the early 2000s using low-resolution climate models (see reviews by Camargo et al. 2007; Camargo and Wing 2016). These efforts have been gaining ground in recent years with the improvements in the prediction systems including the increase of horizontal and vertical resolutions of the component models (Molteni et al. 2011; Vecchi et al. 2014; Camp et al. 2015; Manganello et al. 2016) and wider use of ensemble forecasting and multi-model ensemble approach (MME; Vitart 2006; Vitart et al. 2007). One such system is the North American Multi-Model Ensemble (NMME) experimental multiagency seasonal forecasting system (Kirtman et al. 2014), which is currently delivering real-time seasonal-to-interannual predictions used for operational guidance. In the second stage of this project (NMME-Phase II), improvements to the modeling and data assimilation systems have been introduced, the size of forecast ensembles has increased, and more complete and higher temporal frequency data has become available. In light of these developments, it has become possible to evaluate the skill of dynamical seasonal forecasts of TC activity by the individual NMME models and the corresponding MME to determine whether these forecasts are skillful enough to be used in operational hurricane outlooks.

In this paper, we examine the performance of the NMME-Phase II retrospective forecasts of the North Atlantic (NA) seasonal mean TC activity where predicted storms are identified directly in the model data using a feature-tracking algorithm.

Due to data limitations and relatively coarse horizontal resolution of the NMME models (see Sections 2a and b), our analysis is largely limited to TC frequency, and we briefly examine TC days[1] and regional TC activity as represented by track density (see Vecchi et al. 2014; Manganello et al. 2016). For verification purposes, we use three different reanalysis products in addition to the postseason best track data, such as IBTrACS (see Section 2c). This is done to isolate the influence of model resolution and the TC identification approach on the verification results. In addition to assessing the overall level of skill, our goal is to identify aspects of the simulations that could lead to potential improvements in the TC forecast skill and translate into further developments of the NMME models.

Section 2 presents the NMME-Phase II models and hindcast datasets, and introduces the observational and reanalysis data used to assess the skill of TC hindcasts. It also describes the methodology of identifying and tracking the TCs in the model data and reanalysis. Assessment of the seasonal forecast skill of the NA TC activity, its dependence on the month of initialization and low-frequency variability are presented in Section 3, along with a brief description of the climatology of TC formation and tracks. Discussion of the results and concluding remarks are included in Section 4.

[1] "TC days" is defined as a lifetime of all TCs accumulated over a season, measured in days.

2. Data and Methods

a. NMME-Phase II models and data

The NMME-Phase II ensemble consists of coupled prediction systems from North American modeling centers and the Canadian Meteorological Centre (CMC). Table 1 contains information about the NMME-Phase II models and hindcast datasets used in this study[2]. The NMME System Phase II hindcasat data is available for download from the Earth System Grid at the National Center for Atmospheric Research (NCAR) (https://www.earthsystemgrid.org/search.html?Project=NMME).

Atmospheric horizontal resolution of the models in Table 1 is relatively coarse (between about 1 and 2 degrees), which is common to most present-day operational seasonal prediction systems. (The output resolution is 1°x1° grid for all models.) Daily frequency is the highest temporal output resolution for the majority of the NMME-Phase II models. This rather coarse horizontal and temporal resolution of the data puts additional constraints on the choices of objective criteria used for TC identification, which is further elaborated below. A roughly 30-year period is considered long enough to evaluate the skill of long-range predictions. The hindcast start times include all 12 calendar months, which in addition to a large number of lead times allows for an assessment of long-lead (forecasts initialized as early as January) and short-lead (initialization as late as August) predictions.

b. Tracking of tropical cyclones

[2] At the time of this writing, daily dynamical fields for a common 1982-2012 hindcast period were available for download only for a subset of the NMME-Phase II models, which are listed in Table 1.

Identification and tracking of TCs in coarse- (horizontal) resolution models has been done since the early 1980s, and a variety of methods exist to minimize the effect of resolution on detection criteria (e.g., Walsh et al. 2007; Strachan et al. 2013). On the other hand, to resolve the TC trajectory, including its pre- and post-TC stages, a sufficiently high temporal resolution is generally required with the 6-hourly output frequency preferred for direct comparison with the best track data. Tracking with daily data is not usually done, except in Smith et al. (2010) where TCs are identified as minima in daily sea level pressure as they are tracked, which reduces the number of possible matches but only captures the most intense part of the lifecycle. In their study, the analysis is also restricted to the region between 0° and 25°N. Recently, Vitart (2016) has successfully adjusted the tracking scheme used at the European Centre for Medium-Range Weather Forecasts (ECMWF) to evaluate the skill of sub-seasonal TC predictions using daily data.

In this study, the initial TC identification and tracking is based on the objective feature-tracking methodology of Hodges (1995, 1999) and is tuned to work with daily data, as opposed to 6-hourly data. The detection algorithm identifies vortices as maxima in the 850-hPa relative vorticity field (in the Northern Hemisphere) spectrally truncated at T42 with an intensity threshold of $1x10^{-5}$ s^{-1} and lifetimes greater than 2 days (2 time steps). This tracking method allows TC tracks to be captured in the deep tropics quite well but may underrepresent the extra-tropical extensions of the tracks (see also Section 3a).

To separate predicted TCs from other synoptic-scale features, a set of TC identification criteria needs to be applied to the raw tracks generated above. This should include (1) a structural requirement of a warm core, (2) an intensity threshold, along with (3) the formation region and (4) duration requirements. Due to the coarseness of the spatial and temporal resolutions of the NMME-Phase II models and limited availability of the surface wind data, we decided to base our TC identification criteria solely on multi-level relative vorticity (at 850-hPa, 500-hPa and 200-hPa levels common to all models in Table 1). To derive detection thresholds in this case, simulated TC counts need to be calibrated against observations. In this respect, our approach is similar to the method of Strachan et al. (2013).

We have tested seven sets of TC identification criteria using May-November[3] (MJJASON) reanalyses and model data (forecasts initialized in April). We varied the number of levels used to define the vertical structure, assessed the sensitivity to the presence of vorticity center at each level and monotonic reduction of vorticity with height, and varied the minimum number of days when structural conditions need to be satisfied (see Supplementary Material for more detail). In all cases, a warm core condition remained the same, cyclogenesis was restricted to 0°-20°N over land and 0°-30°N over oceans, and 850-hPa vorticity at output resolution was used to calibrate seasonal TC counts. For each reanalysis and NMME model, we have chosen a set of TC identification criteria that maximizes their MJJASON TC frequency correlation skill. These criteria are therefore not the same for all the datasets, although the sensitivities are not large and are further discussed in the Supplementary Material. While this is not a general practice, we believe that the

[3] The MJJASON period encompasses most of the TC season in the NA basin.

above approach allows to better gauge the skill of each individual reanalysis and model. These dataset-specific criteria do not change for the rest of the analysis, including the skill assessment of long- and short-range predictions.

c. Observational and reanalysis data

For comparison with observations, we use data from the International Best Track Archive for Climate Stewardship (IBTrACS, version v03r07; Knapp et al. 2010; available online at https://www.ncdc.noaa.gov/ibtracs/). IBTrACS makes available for public use a global dataset of post season analysis of TC position and intensity (also know as "best track") by merging storm information from multiple centers into one product. The observed tracks are further processed here by retaining systems with lifetimes greater than 2 days, of tropical storm strength for at least 1 day and with first identification occurring between 0°-20°N over land and 0°-30°N over oceans, to be more in line with the model and reanalysis tracks (see Section 2b). We also use sea surface temperature (SST) data from the National Oceanic and Atmospheric Administration (NOAA) Optimum Interpolation SST version 2 data set (OISSTv2; Reynolds et al. 2002).

Since our choice of TC identification criteria (Section 2b) does not imply a close match with the observational ones, it is prudent to use reanalysis data for more direct verification of model results. In reanalyses, historical observations are objectively ingested into the models with a goal to produce a consistent estimate of the state of the climate. As such, reanalyses have an advantage of models by providing a more comprehensive dataset. They are constrained by the observations but limited by the raw input data and its quality, the resolution of the models used, and the capabilities of the data assimilation system. Overall, applying the same tracking methodology to the reanalysis and model data of the same spatial and temporal resolution would allow a more objective estimation of the model skill.

We have used the following three reanalysis datasets: the National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR; Saha et al. 2010); the Interim ECMWF Re-Analysis (ERA-I; Dee et al. 2011); and the National Aeronautics and Space Administration (NASA) Modern Era Retrospective-Analysis for Research and Applications (MERRA; Rienecker et al. 2011). The spatial resolution of all reanalysis data was downgraded to the 1°x1° grid of the NMME-Phase II model data. The temporal resolution was converted to daily, and the period of 1982-2014 was used for analysis.

3. Results

a. Climatologies of TC formation and tracks

Prior to evaluating the skill of TC frequency forecasts, we verify whether the TC identification and tracking approach chosen here generates realistic distributions of genesis locations and tracks. Figs. 1 and 2 show NA genesis and track densities, respectively, for the IBTrACS, reanalyses and the NMME-Phase II retrospective seasonal forecasts. Reanalysis products reproduce main features of the genesis pattern quite well, with varying levels of success depending on the specific cyclogenesis center (Figs. 1a-d). CFSR is most accurate in representing the Main Development Region (MDR; 10°-25°N, 80°-20°W), whereas in ERA-I and MERRA, activity in this area is largely concentrated near the west coast of Africa. (Origin of some tracks over West Africa is likely related to their tropical easterly wave precursors being captured by the tracking algorithm (see also Manganello et al. 2012). For the same reason, the bulk of the MDR genesis is shifted further to the east compared to observations.) The Gulf of Mexico center is underrepresented in all reanalysis products, whereas the western Atlantic center is quite realistic across the board. The Caribbean genesis is shifted southeast and is somewhat overactive in ERA-I. This shift has been noted earlier and linked to the coarse spatial resolution of the models (Manganello et al. 2012, 2016). The associated track density is overall well reproduced (Figs. 2a-d), except in the extra-tropics which is likely a consequence of tracking using daily data (see Section 2b).

Predicted genesis and track densities on the whole are less realistic compared to observations and reanalyses, where formation regions are strongly concentrated in space (Figs. 1e-h), and track density is overpredicted and too zonal in the tropics and quite weak further north (Figs. 2e-h). However, the MDR genesis is rather active in all the hindcasts, and other centers are well defined, except for the Gulf of Mexico and the western Atlantic centers being absent in the CanCM3 forecasts. In addition, the Gulf of Mexico center, where present, is more realistic than in the reanalysis. On the other hand, the Caribbean genesis is too strong, and the associated tracks are largely confined to the northern tip of South America. To summarize, the tracking algorithm is capable of generating climatologies of the NA TC formation and tracks with many realistic features, particularly when applied to reanalysis products.

b. April forecasts of the North Atlantic seasonal mean TC activity

1). TC frequency

Fig. 3 shows the interannual variability of the observed and reanalyses-based NA TC frequency, which is another demonstration of the utility of the TC tracking method in estimating seasonal mean TC activity using daily data. The reanalysis datasets reproduce the interannual variability quite well, with major peaks of 1995 and 2005 to the most part realistically represented. The correlation coefficients between the reanalyses and the observed time series are also quite high ranging from 0.67 to 0.81 (see Table 2). The reanalyses do differ considerably in terms of their skill in representing multidecadal changes characterized by low activity in the 1980s and early 1990s and high activity in the latter part of the record (e.g., Goldenberg et al. 2001). ERA-I is the most successful in capturing this trend, whereas CFSR displays no trend (see Fig. 3).

Retrospective correlation skill varies markedly among the NMME-Phase II models (see Table 2 for MJJASON forecasts initialized in April). It is quite high for CCSM4 and CanCM4 and is in fact similar to the skill of experimental high-atmospheric-resolution coupled prediction systems in Project *Minerva* (Manganello et al. 2016), whereas it is close to zero for GEOS-5 and CanCM3. As a consequence, correlation of the MME mean[4] is significant but rather modest and does not exceed

[4] The MME mean is defined as the average over all the hindcasts, with all ensemble members of each model having equal weight.

the skill of all models in the ensemble. The root-mean-square error[5] (RMSE), which a measure of forecast accuracy, is fairly large, although the differences are not major when the MME mean is compared to reanalyses (Table 3). RMSE for the detrended time series is smaller across the board suggesting that low-frequency variability is not well reproduced in the forecasts (see below). For short-range predictions, the overall skill improves, and the advantages of the MME approach become more evident (see Secion 3d).

A natural question arises whether the individual NMME-Phase II models are indeed more or less skillful than their MME mean, and whether these models including the MME display skill that is significantly different from the skill based on the reanalyses data. The correlation coefficient is not considered a very good measure to compare skill, as the presence of noise may lead to large differences in this quantity. It is found that the squared error is a more appropriate metric (DelSole and Tippett, 2014), and we choose the Wilcoxon signed-rank test for the forecast skill comparison since it is not sensitive to the type of distribution (ibid.). We find that at the 95% confidence level, the differences in skill among the four NMME models and their MME mean are insignificant, except that the skill of GEOS-5 and CanCM3 is significantly lower that the skill of CanCM4. We also find that all NMME models and the MME mean are as skillful as CFSR and ERA-I but less skillful than MERRA. (The skill of CanCM3 is also significantly lower compared to ERA-I). It

[5] Forecasts are calibrated (without cross-validation) where each ensemble member is multiplied by a constant factor so that the predicted ensemble-mean and observed climatologies become equal.

is worth emphasizing that the above skill comparison is based on the MJJASON season (forecasts initialized in April).

Ensemble forecasts have an additional advantage of being able to quantify uncertainty based on the probabilistic approach. One such measure is statistical reliability, which can be expressed as a ratio of the ensemble spread and the RMSE (SPRvERR). In a perfectly reliable ensemble forecast, forecast probabilities match the observed frequencies, and the SPRvERR is equal to one. Individual NMME and the MME mean April forecasts are found to be underdispersed (or overconfident; Table 4). Detrending the time series enhances reliability quite a bit which indicates that poor low-frequency variability of the predicted NA TC frequency is indeed a distinct source of forecast error. These results are similar to our findings in Project *Minerva* (Manganello et al. 2016).

To further illustrate the above results, Fig.4 shows seasonal mean TC frequency predicted by the CCSM4 and CanCM4 models along with their ensemble information compared with observations. Both models capture year-to-year fluctuations quite well, particularly in the 1990s and early 2000s where only several seasons fall outside the 10th-90th percentile range (1992, 1997, and 2005 for CCSM4; and 1992, 1995, 1997 and 2005 for CanCM4). Neither of the models reproduces the secular trend, and the hindcast skill appears to be inferior in the 1980s and 2010s, which is further discussed below.

2) TC days and TC track density

Seasonally accumulated lifetime of all TCs in the basin, or “TC days” (see definition in Section 1), exhibits retrospective correlation skill behavior quite comparable to TC frequency (Table 5). The forecasts that are skillful in predicting TC frequency are to the most part also skillful in predicting TC days. For MJJASON forecasts initialized in April the correlation of the MME mean TC days is not high but significant (0.46), and increases to 0.59 at shorter leads (July and August initializations). It is curious that reanalyses reproduce variability of TC days seemingly better than TC frequency (using current tracking), where correlation for TC days doesn't drop below 0.76 (Table 5).

One of the current challenges of seasonal TC forecasting is to provide regional information, such as local TC occurrence or probability of landfall, which is more relevant for decision-making (e.g., Vecchi et al. 2014; Camp et al. 2015; Manganello et al. 2016; Murakami et al. 2016). Here we examine whether MME forecasts of the NA TC activity have retrospective skill on sub-basin scales using track density as a metric and Spearman rank correlation as a measure of performance (see Manganello et al. 2016 for more detail). We compare this skill to the rank correlation between the seasonal mean observed and reanalyses-derived track densities. All three reanalysis products are quite successful at reproducing interannual variability of regional TC activity over most of the NA domain (Figs. 5a-c). The regions with significant correlations common to all products are the MDR, the Caribbean Sea, the Gulf of Mexico and central subtropical North Atlantic. These regions also tend to show the highest correlation values. The results do not seem to be particularly sensitive to whether the extended MJJASON season or the peak ASON season is examined (Figs. 5e-g). In comparison, for the longer-lead MME forecasts initialized in April the regions with significant skill are rather sparse and limited to some parts of the MDR and the westernmost margins of the Caribbean Sea and the Gulf of Mexico (Fig. 5d). The absence of any skill north of about 30°N is likely related to strong underprediction of climatological tracks at these latitudes in the NMME models (see Section 3a). At shorter leads (MME forecasts initialized in July), the region with significant skill markedly increases and now covers the western part of the MDR and the whole Caribbean Sea (Fig. 5h). Fairly high retrospective forecast skill in the vicinity of Caribbean islands suggests that predictions of TC landfall frequency in this region may also be skillful. Overall, the skill of regional TC activity forecasts in the NMME is rather modest compared to other coupled prediction systems that employ atmospheric models with much higher horizontal resolution (see Vecchi et al. 2014; Manganello et al. 2016; Murakami et al. 2016).

c. Low-frequency variability in prediction skill

The NMME-Phase II ensemble exhibits variability in the retrospective forecast skill of the NA TC frequency (Fig. 6). Compared to the reanalyses, which maintain relatively constant skill throughout the hindcast period, the MME mean displays markedly lower skill in the 1980s and early 1990s, and also late 2000s and 2010s (Fig. 6a). During these two periods, the model skill deviates from the reanalyses. In contrast, it is quite comparable to the reanalyses in the late 1990s and early 2000s. Since the NA TC season peaks in August-October, forecasts initialized in June could be considered short-lead forecasts of the full hurricane season. We find that at shorter leads (Fig. 6b), forecast skill becomes more in line with the reanalyses in the latter part of the record. This tendency is also present in forecasts initialized in May (not shown).

Loss of skill in the 1980s is not unique to the NMME-Phase II models. Similar behavior was also found in all *Minerva* hindcasts (Manganello et al. 2016) where it was linked to more deficient initialization of ocean fields. It is also feasible that predictability of the NA TC activity can fluctuate from one decade to another. The influence of certain climatic factors that serve as predictors of the NA TC activity may depend on the underlying climate conditions (Fink et al. 2010; Caron et al. 2015). Current seasonal prediction systems are perhaps able to reproduce some of the relationships but not others or do not time them correctly, which may contribute to the drop in skill.

While a detailed analysis of these influences is beyond the scope of the current paper, as a first step we examine here the relationship between the NMME forecasts of TC frequency and several well established predictors of the NA TC genesis, and compare results to observations and reanalyses. The selected climate indices are: 1) SST averaged over the MDR; 2) relative SST index[6], and 3) the Niño-3.4 index[7] (see, e.g., Villarini et al. 2010; Vecchi et al. 2011; Caron et al. 2015 and the extensive lists of references in these papers). Both observations and reanalyses suggest a stronger relationship between the MDR SSTs and the NA TC frequency in the late 1990s and early 2000s compared to the earlier and latter parts of the record where correlations become marginally significant (Fig. 7a). The correlation with the

[6] Relative SST index is defined as the difference between MDR SST and global tropical-mean SST (e.g., Zhao et al. 2010).
[7] Niño-3.4 index is defined as SST averaged over 5°S-5°N, 120°-170°W.

relative SST index is higher and more constant throughout the time period (Fig. 7b), as is the negative connection with the El Niño and the Southern Oscillation (ENSO) except perhaps in 2000s where reanalyses data suggest a weakening of this relationship (Fig. 7c). The NMME models and their MME mean tend to display rather different behavior. During the earlier and latter parts of the hindcast period, TC frequency forecasts appear to be much stronger driven by variations in the predicted MDR SSTs and the relative SST index compared to the middle part of the record, opposite to what observations and reanalyses demonstrate (Figs. 7a and b). It is curious that the late 1990s and early 2000s when the MME correlations with the MDR SSTs and the relative SST index are most realistic coincide with the period of the highest MME TC frequency forecast skill (Fig. 6a). On the other hand, the rest of the hindcast period when these correlations are too high and markedly outside the range of the observed/reanalyses values is also when the forecast skill is at the lowest levels as described above and shown in Fig. 6a. In addition, the retrospective forecast skill of the MDR and relative SST indices is generally quite high except in the 1980s and early 1990s when forecasts of the relative SST index are not skillful (see Fig. S1 in the Supplementary Material). This could further limit the quality of the TC frequency predictions during this time period. In contrast, the influence of ENSO appears to be captured quite well by the MME forecasts, except possibly in the 1980s and late 2000s when it appears to be somewhat stronger (Fig. 7c); the hindcast skill of the Niño-3.4 index is the highest among the indices examined and also fairly constant throughout the record (Fig. S1).

d. Long- and short-lead forecasts

The NA TC hindcast skill as a function of the initialization month is shown in Fig. 8, along with the results for the reanalyses and measures of "null skill". At longer lead times (earlier than April), the MME mean shows marginal skill when initialized in February relative to the IBTrACS trailing 5-yr average, which is a skill metric recommended by the World Meteorological Organization (WMO 2008; Fig. 8a). In this reference forecast, the interannual variability is smoothed out but the interdecadal variability is preserved to some extent. The best performing forecasts at long leads are produced by CanCM4 and are skillful for January and February initializations. It is notable that for most models and the MME mean the skill curves in Fig. 8a display substantial variability from month to month. This "noisiness" is largely due to low-frequency variability being forecasted at varying levels of skill depending on the initialization month. (Compare also with Fig. 8b that shows similar metrics computed for the detrended time series and displaying a more consistent increase in skill with lead time.) Relative to persistence, or the previous season's TC count, the detrended MME mean shows no long-lead skill except perhaps when initialized in March. All detrended long-lead CanCM4 forecasts show skill albeit marginal.

When the hurricane season is approached (March and June initializations) the skill drops somewhat (Figs. 8a and b). At short lead times (July and August), it rebounds and displays the highest levels overall (see also Table 2). It is notable that all detrended MME mean forecasts initialized in April and later are consistently skillful relative to persistence (Fig. 8b). The short-lead MME mean correlation skill

(RMSE) also shows the highest (lowest) value among all the models (detrended only; see Tables 2 and 3). In addition, it becomes comparable to the skill of the reanalyses. For instance, RMSEs of forecasts initialized in July are lower than for CFSR and ERA-I (detrended only in the latter case; Table 3). The short-lead MME mean forecasts are also quite reliable, although somewhat over-dispersed when detrended (Table 4). It is curious that among the forecasts initialized in June through August the best performing model is CanCM3, whereas it is one of the worst performing at longer leads. If April forecasts were chosen as a benchmark and the MME are based on two models with skill (CCSM4 and CanCM4), the resultant correlation at short leads is markedly lower compared to the MME based on *all* available models (not shown). This is one of the advantages of the multi-model ensemble approach that is not always obvious.

The skill of the MME mean relative to the individual NMME-Phase II models and the reanalyses is further assessed using the difference between the squared error as a skill metric and testing the significance by applying the Wilcoxon signed-rank test (see Fig. 9; DelSole and Tippett, 2014). In the vast majority of cases, the MME mean outperforms the individual model with differences being statistically significant at short lead times (June and July initializations). Relative to the reanalyses, the MME mean shows larger error most of the time (except at short leads with respect to CFSR), although it is significant primarily at long leads and when compared to MERRA only. It is also notable that at most lead times, the reliability is improved slightly for the MME mean and to a larger extent when the time series are detrended (not shown).

4. Summary and conclusions

In this study, the NMME-Phase II models are interrogated in terms of the retrospective seasonal forecast skill of the NA TC frequency. The TCs are identified explicitly in the model data by means of an objective feature-tracking methodology. Due to the synoptic nature of these storms, daily resolution (the highest available for the ensemble) is generally considered coarse for TC tracking. As part of this work, we have adjusted the TC identification and tracking algorithm to work with daily data and also applied it to three reanalysis products (CFSR, ERA-I and MERRA) that were coarsened to have the same spatial and temporal resolution of the NMME-Phase II ensemble. The latter step provides additional verification data (apart from best track data) where the effects of resolution and the TC identification approach have been isolated which allows for a more objective estimation of forecast skill.

The TC tracking method used here, when applied to reanalysis data, produces realistic climatological distributions of the NA TC formation and tracks. Low track density in the extra-tropics is a common deficiency, which is a result of tracking using daily data. The tracking is also quite skillful in reproducing the interannual variability of the TC frequency relative to the IBTrACS with correlations ranging between 0.67 and 0.81 depending on the reanalysis product. These values are quite comparable to the estimates obtained in Strachan et al. (2013) and Roberts et al. (2015) where both studies utilized six-hourly data.

Long-lead (March and earlier) retrospective seasonal forecasts of the NA TC frequency with the MME based on the available NMME-Phase II models are found to have low or marginal skill, although one of the models (CanCM4) produces skillful forecasts when initialized as early as in January and February. At shorter leads (April and later), the MME mean forecasts are largely skillful with the best performance for July and August initializations. Skill metrics evaluated for the detrended time series display a more systematic increase in skill with shorter lead time, and all detrended MME mean forecasts initialized in April and later are consistently skillful. At short lead times (June through August), the MME mean also tends to significantly outperform the individual models and attain skill comparable to the reanalysis. The short-lead MME mean forecasts are also quite reliable, while being under-dispersed at longer leads.

We have identified several deficiencies in the simulations that likely limit the NMME-Phase II seasonal hindcast skill of the NA TC frequency.

1. None of the models or the MME mean independent of the initialization month can realistically represent low-frequency variability characterized by low activity in the 1980s and early 1990s and higher activity thereafter. The skill metrics computed for the detrended time series show higher scores in the vast majority of cases. This suggests that poor multi-year variability in the forecasts may indeed be a source of forecast error. This problem is not trivial and is characteristic of other prediction systems like *Minerva* (Manganello et al. 2016) and several reanalysis products, e.g., MERRA and CFSR. It could be related, for instance, to poor skill in reproducing downward trends in upper tropospheric temperature (Emanuel et al. 2013; Vecchi et al. 2013), inadequate representation of the effects of aerosols and ozone (Evan et al.

2009, 2011; Emanuel et al. 2013), possibly deficiencies in simulating tropical heating and atmospheric teleconnections (Manganello et al. 2016), and the sensitivity to the identification of weak and short-lived TCs in the model and reanalysis data.

2. We have shown that the MME mean forecasts exhibit a large drop in skill in the 1980s and early 1990s and also late 2000s and 2010s (mostly at longer leads). It is curious that during the rest of the period (late 1990s and early 2000s), the MME mean skill is quite comparable to the reanalyses, which maintain relatively constant skill throughout the hindcast time period. Early in the record, forecast errors could be partly related to deficiencies in the model initialization. Although the problem as a whole may be more complex and indicate that certain physical relationships that underline predictability of the NA TC activity may not be consistently reproduced or properly timed.

Addressing the above issues, while not an easy task, could lead to marked improvements in the seasonal forecast skill and increase the value of the NMME ensemble in providing operational guidance.

Acknowledgements. Funding of COLA for this study is provided by grants from NSF (AGS-1338427), NOAA (NA09OAR4310058 and NA14OAR4310160), NASA (NNX14AM19G), and the ONR grant (N00014-15-1-2745). We acknowledge NOAA MAPP, NSF, NASA, and the DOE that support the NMME-Phase II system, and we thank the climate modeling groups (Environment Canada, NASA, NCAR, NOAA/GFDL, NOAA/NCEP, and University of Miami) for producing and making available their model output. NOAA/NCEP, NOAA/CTB, and NOAA/CPO jointly provided coordinating support and led development of the NMME-Phase II system. We also gratefully acknowledge computing resources on the Yellowstone supercomputer provided by the National Center for Atmospheric Research.

References

Camargo, S. J., A. G. Barnston, P. J. Klotzbach, and C. W. Landsea, 2007: Seasonal tropical cyclone forecasts. *World Meteorological Organization Bulletin*, **56**, 297-309.

Camargo, S. J., and A. A. Wing, 2016: Tropical cyclones in climate models. *WIREs Climate Change*, **7**, 211-237.

Camp, J., M. Roberts, C. MacLachlan, E. Wallace, L. Hermanson, A. Brookshaw, A. Arribas, A. A. and Scaife, 2015: Seasonal forecasting of tropical storms using the Met Office GloSea5 seasonal forecast system. *Q.J.R. Meteorol. Soc.*, doi: 10.1002/qj.2516.

Caron, L.-P., M. Boudreault, and C. L. Bruyere, 2015: Changes in large-scale controls of Atlantic tropical cyclone activity with the phases of the Atlantic multidecadal oscillation. *Clim. Dyn.*, **44**, 1801-1821.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q.J.R. Meteorol. Soc.*, **137**, 553–597. doi: 10.1002/qj.828

DelSole, T., and M. K. Tippett, 2014: Comparing Forecast Skill, *Mon. Weather Rev.*, ***142*(12)**, 4658–4678.

Emanuel, K. A., S. Solomon, D. Folini, S. Davis, and C. Cagnazzo, 2013: Influence of tropical tropopause layer cooling on Atlantic Hurricane activity. *J. Clim.*, **26**, 2288–2301.

Evan, A. T., G. R. Foltz, D. Zhang, and D. J. Vimont, 2011: Influence of African dust on ocean-atmosphere variability in the tropical Atlantic. *Nature Geoscience*, **4**, 762-

765.

Evan, A. T., D. J. Vimont, A. K. Heidinger, J. P. Kossin, ad R. Bennartz, 2009: The Role of Aerosols in the Evolution of Tropical North Atlantic Ocean Temperature Anomalies. *Science*, **324**, 778-781.

Fink, A. H., J. M. Schrage, and S. Kotthaus, 2010: On the potential causes of the nonstationary correlations between West African Precipitation and Atlantic Hurricane Activity. *J. Climate*, **23**, 5437-5456.

Goldenberg, S. B., C. W. Landsea, A. M. Mestas-Nu.ez, and W. M. Gray, 2001: The recent increase in Atlantic hurricane activity: Causes and implications. *Science*, **293**, 474–479.

Hodges, K. I., 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.,* **123,** 3458–3465.

Hodges, K. I., 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.,* **127,** 1362–1373.

Kirtman, B.P., and Coauthors, 2014: The North American multimodel ensemble: phase 1 seasonal-to-interannual prediction; phase-2 toward developing intraseasoanl prediction. *Bull. Am. Meteorol. Soc.*, **95**, 585–601.

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). *Bull. Amer. Meteor. Soc.*, **91**, 363–376.

Manganello, J. V., K. I. Hodges, B. A. Cash, J. L. Kinter III, E. L. Altshuler, M. J. Fennessy, F. Vitart, F. Molteni, and P. Towers, 2016: Seasonal Forecasts of Tropical Cyclone

Activity in a High Atmospheric Resolution Coupled Prediction System, *J. Climate*, **29**, 1179-1200.

Manganello, J. V., K. I. Hodges, J. L. Kinter III, B. A. Cash, L. Marx, T. Jung, D. Achuthavarier, J. M. Adams, E. L. Altshuler, B. Huang, E. K. Jin, C. Stan, P. Towers and N. Wedi, 2012: Tropical Cyclone Climatology in a 10-km Global Atmospheric GCM: Toward Weather-Resolving Climate Modeling. *J. Climate*, **25**, 3867-3893.

Merryfield, W. J., and Coauthors, 2013: The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart, 2011: The new ECMWF seasonal forecast system (System 4). *ECMWF Technical Memorandum*, No. **656**, pp. 49.

Murakami, H., G. A. Vecchi, G. Villarini, T. L. Delworth, R. Gudgel, S. Underwood, X. Yang, W. Zhang, and S.-J. Lin, 2016: Seasonal forecasts of major hurricanes and landfalling tropical cyclones using a high-resolution GFDL coupled climate model. J. Climate, 29, 7977-7989.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609-1625.

Rienecker, M. M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. J. Climate, 24, 3624–3648.

Roberts, M. J., P. L. Vidale, M. S. Mizielinski, M.-E. Demory, R. Schiemann, J. Strachan,

K. Hodges, R. Bell, and J. Camp, 2015: Tropical Cyclones in the UPSCALE Ensemble of High-Resolution Global Climate Models. *J. Climate*, **28**, 574-596.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Am. Meteorol. Soc.*, **91(8)**, 1015–1057.

Smith, D. M., R. Eade, N. J. Dunstone, D. Fereday, J. M. Murphy, H. Pohlmann, and A. A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. *Nature Geoscience,* **3**, 846–849.

Strachan, J., Vidale, P. L., Hodges, K., Roberts, M., and Demory, M.-E., 2013: Investigating global tropical cyclone activity with a hierarchy of AGCMs: the role of model resolution. *J. Climate*, **26**, 133–152.

Vecchi, G. A., T. Delworth, R. Gudgel, S. Kapnick, A. Rosati, A. T. Wittenberg, F. Zeng, W. Anderson, V. Balaji, K. Dixon, L. Jia, H.-S. Kim, L. Krishnamurthy, R. Msadek, W. F. Stern, S. D. Underwood, G. Villarini, X. Yang, and S. Zhang, 2014: On the Seasonal Forecasting of Regional Tropical Cyclone Activity. *J. Climate*, **27**, 7994–8016.

Vecchi, G. A., S. Fueglistaler, I. M. Held, T. R. Knutson, and M. Zhao, 2013: Impacts of atmospheric temperature trends on tropical cyclone activity. *J. Clim*., **26**, 3877–3891.

Vecchi, G. A., M. Zhao, H. Wang, G. Villarini, A. Rosati, A. Kumar, I. M. Held, and R. Gudgel, 2011: Statistical-dynamical predictions of seasonal North Atlantic hurricane activity. *Mon. Wea. Rev.*, **139**, 1070-1082.

Vernieres, G., C. Keppenne, M.M. Rienecker, J. Jacob, and R. Kovach, 2012: The GEOS-ODAS, description and evaluation. NASA Technical Report Series on Global

Modeling and Data Assimilation, NASA/TM–2012–104606, Vol. 30.

Villarini, G., G. A. Vecchi, and J. A. Smith, 2010: Modeling the dependence of tropical storm counts in the North Atlantic basin on climate indices. *Mon. Wea. Rev.*, **138**, 2681-2705.

Vitart, F., 2006: Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Q. J. R. Meteorol. Soc.*, **132**, 647-666.

Vitart, F. 2016: Tropical cyclogenesis in the S2S Database. *S2S News Letter*, **No. 3**, 3-6.

Vitart, F., M. R. Huddleston, M. Déqué, D. Peake, T. N. Palmer, T. N. Stockdale, M. K. Davey, S. Ineson, and A. Weisheimer, 2007: Dynamically-based seasonal forecasts of Atlantic tropical storm activity issued in June by EUROSIP. *Geophys. Res. Lett.*, **34**, L16815, doi:10.1029/2007GL030740.

Walsh, K. J. E., M. Fiorino, C. W. Landsea, K. L. McInnes, 2007: Objectively Determined Resolution-Dependent Threshold Criteria for the Detection of Tropical Cyclones in Climate Models and Reanalyses. *J. Climate*, **20**, 2307–2314.

WMO, 2008: Report from expert meeting to evaluate skill of tropical cyclone seasonal forecasts. World Meteorological Organization. Tech. Doc. 1455, Geneva, Switzerland. 27 pp.

Zhao, M., I. M. Held, and G. A. Vecchi, 2010: Retrospective forecasts of the hurricane season using a global atmospheric model assuming persistence of SST anomalies. *Mon. Wea. Rev*., **138**, 3858-3868.

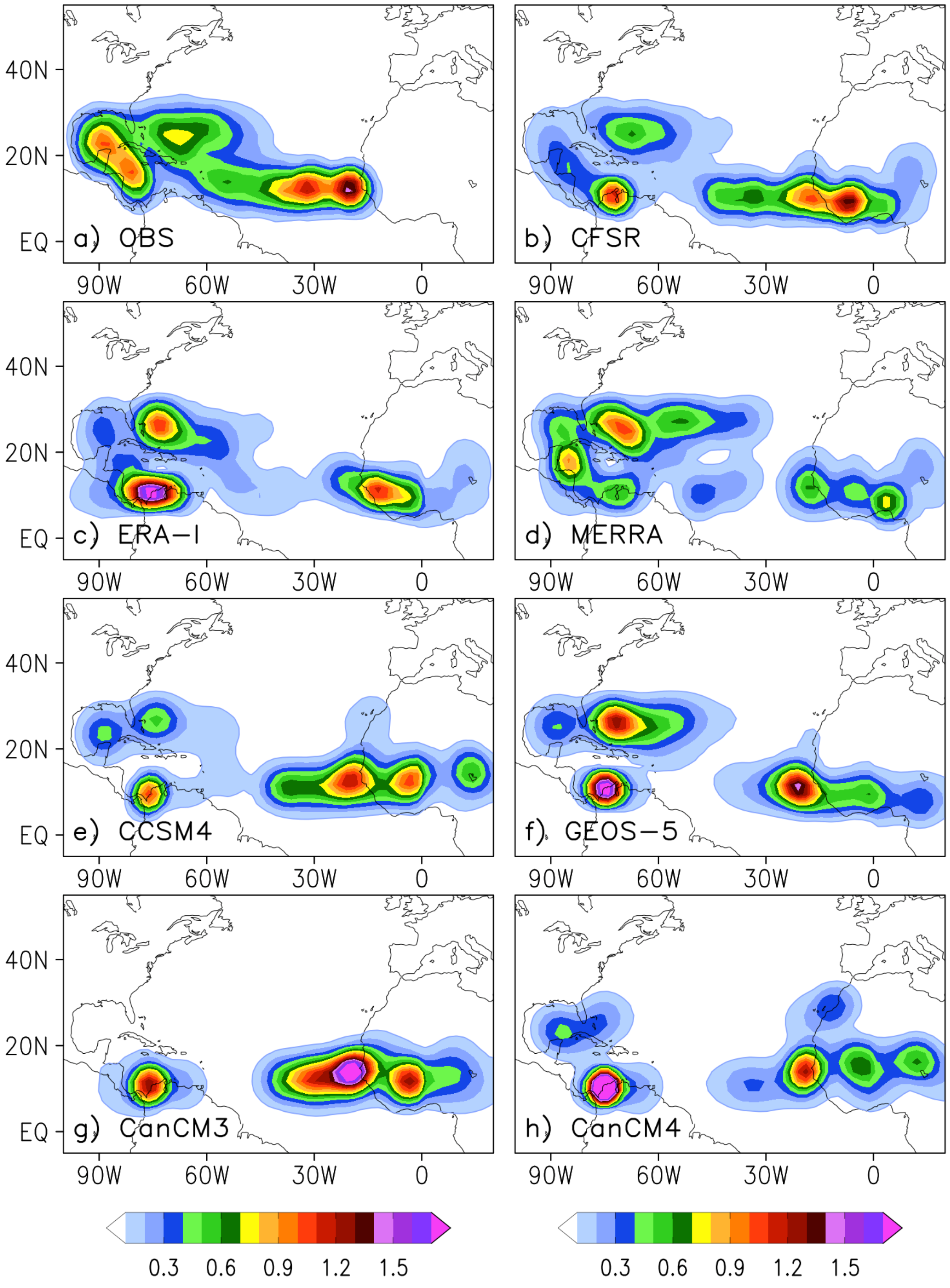

Figure 1: NA genesis densities for the MJJASON season as number density per season per unit area equivalent to a 5° spherical cap for (a) IBTrACS (OBS), (b) CFSR, (c) ERA-I, and (d) MERRA reanalyses based on 1982-2014, and (e) CCSM4, (f) GEOS-5, (g) CanCM3, and (h) CanCM4 seasonal hindcasts (all ensemble members) based on the time periods listed in Table 1.

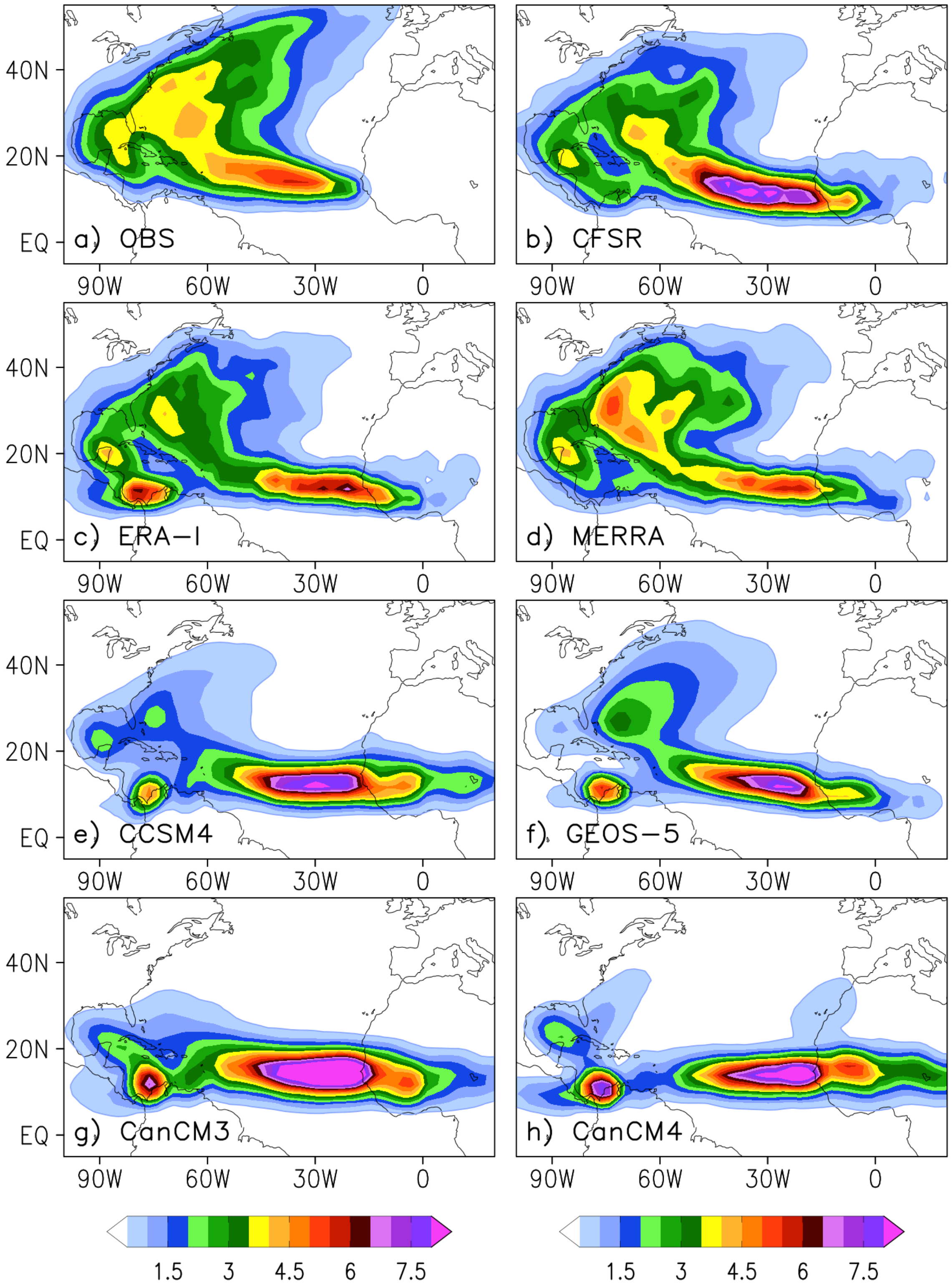

Figure 2: As in Fig. 1, but for the track density.

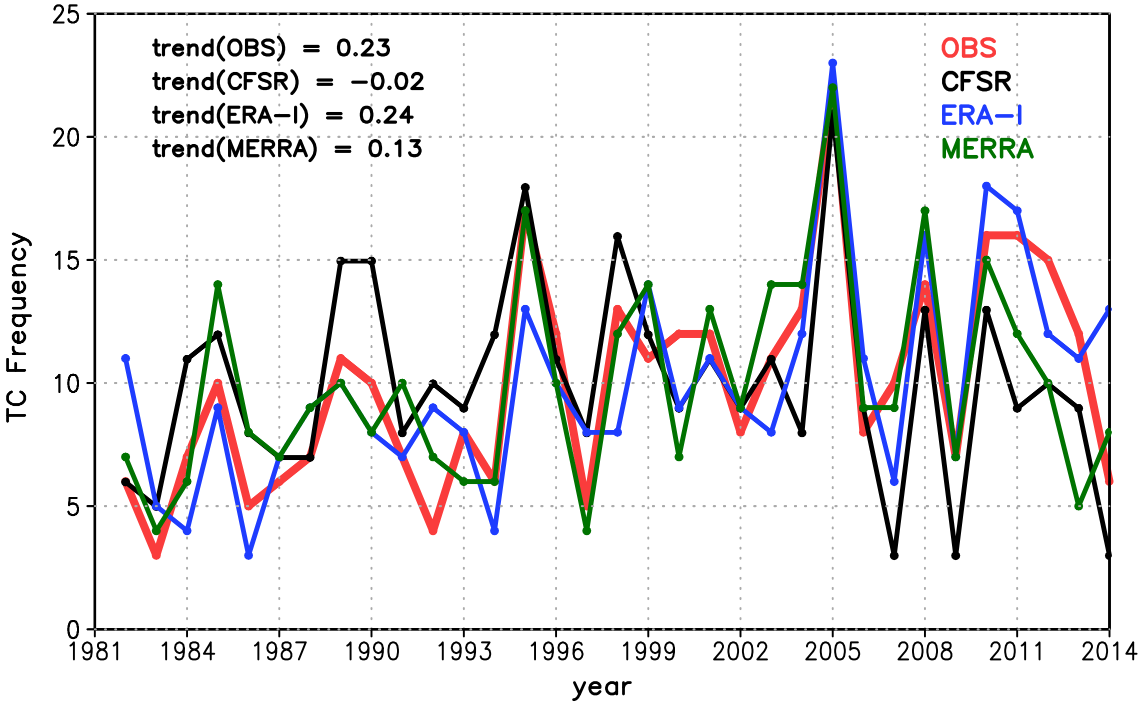

Figure 3: Time series of the NA MJJASON TC frequency based on the IBTrACS (OBS) data (red), and the CFSR (black), ERA-I (blue) and MERRA (green) reanalysis data sets. Linear trends for each time series are shown in the upper-left corner, units are counts per season per year.

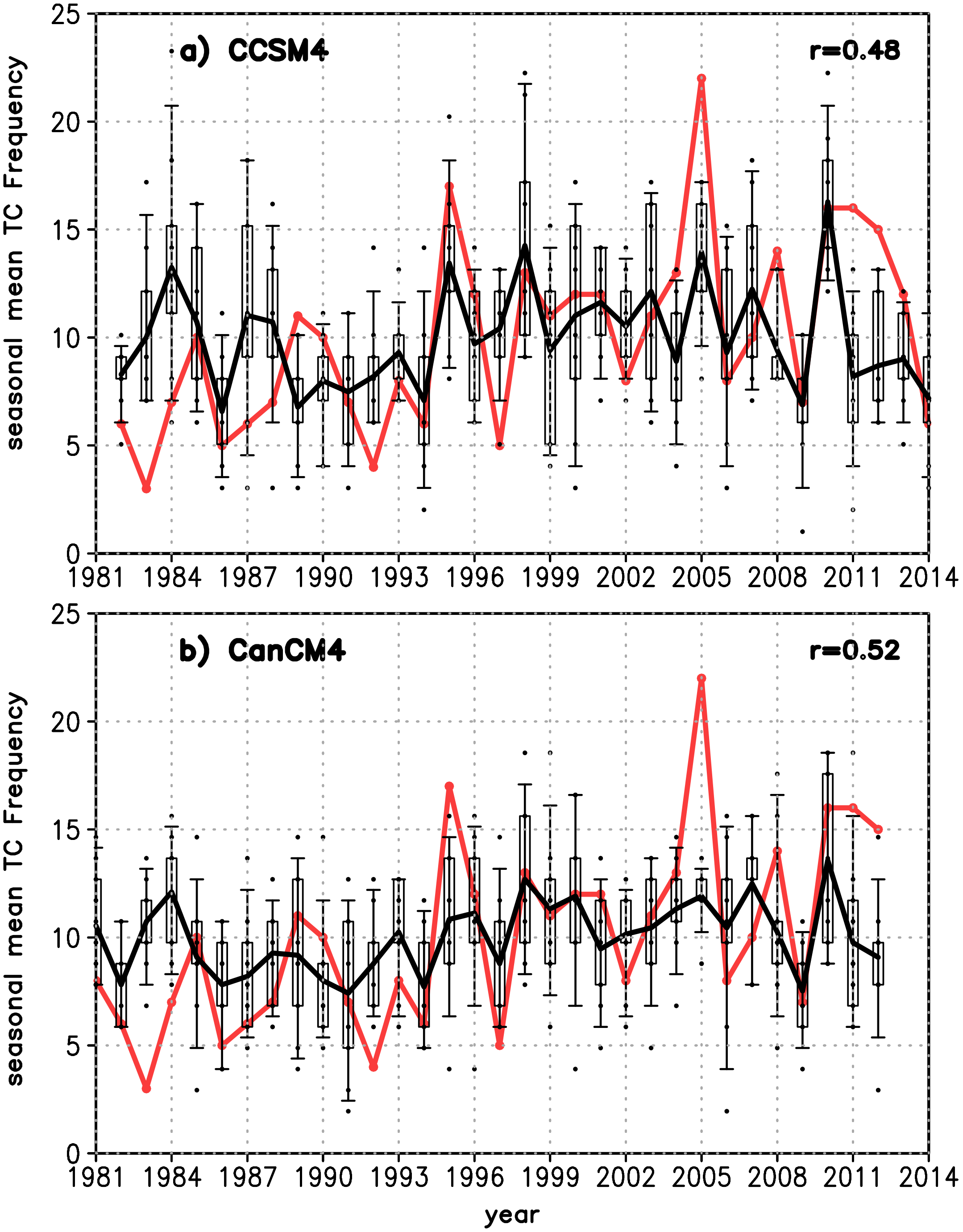

Figure 4: Retrospective forecasts (initialized in April) of the NA MJJASON TC frequency for the (a) CCSM4 and (b) CanCM4 NMME-Phase II models. Red and black lines show the observed time series and the ensemble-mean forecasts, respectively. Black dots mark predictions from the individual ensemble members. Box-and-whisker plots denote the 25th-75th and 10th-90th percentile ranges.

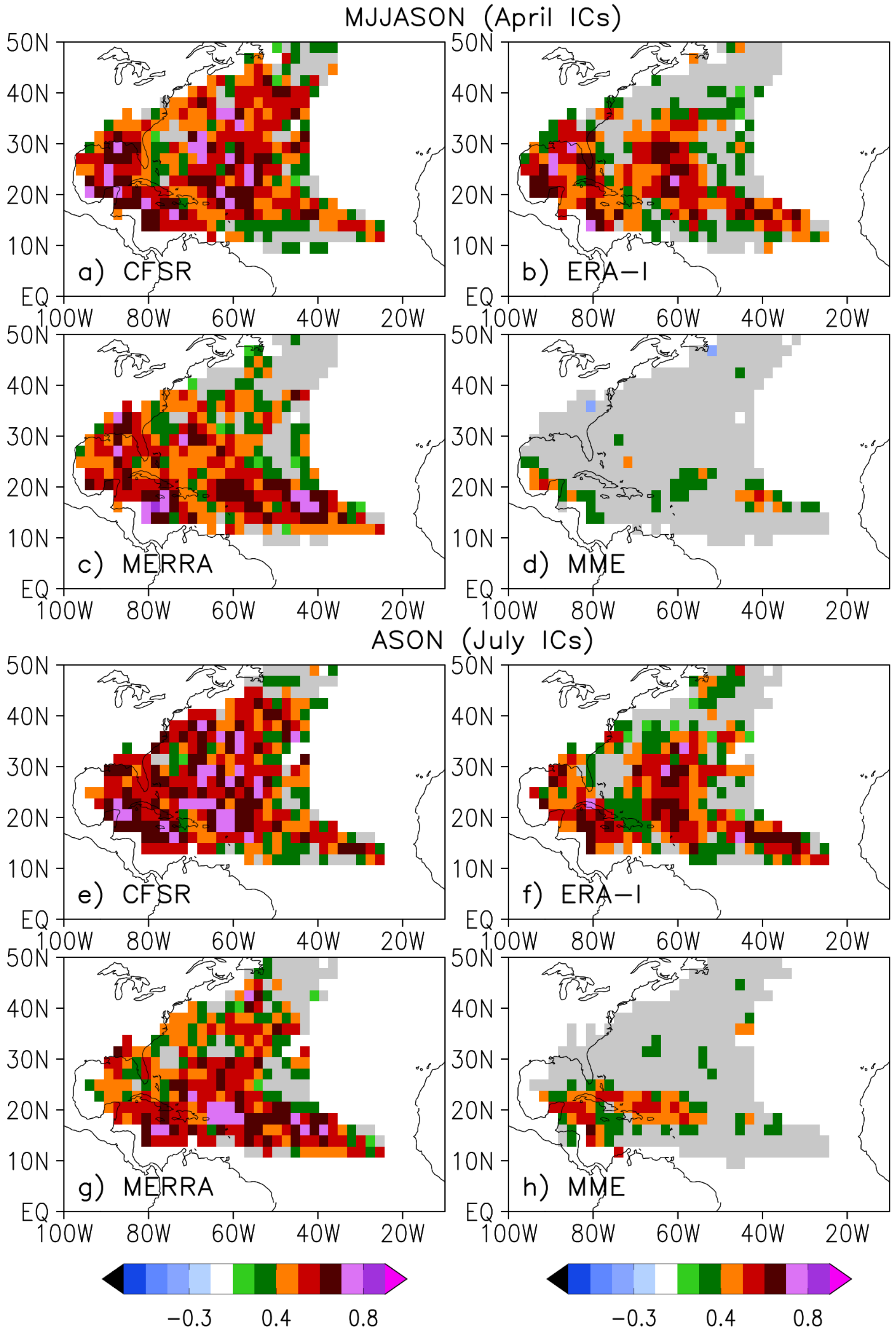

Figure 5: Rank correlation between the MJJASON observed (IBTrACS) and reanalysis-derived TC track densities for 1982-2014 using (a) CFSR, (b) ERA-I, and (c) MERRA. TC track density is defined as number density per season per unit area equivalent to a 5° spherical cap. (E)-(g) are the same as (a)-(c) but for the ASON season. (D) and (h) show retrospective rank correlation of the observed vs. MME predicted TC track density for MJJASON (April ICs) and ASON (July ICs) of 1982-2012, respectively. Values statistically significant at a two-sided p=0.1 level are shown by color shading. Grey shading marks the regions where the observed track density above zero for at least 25% of the years.

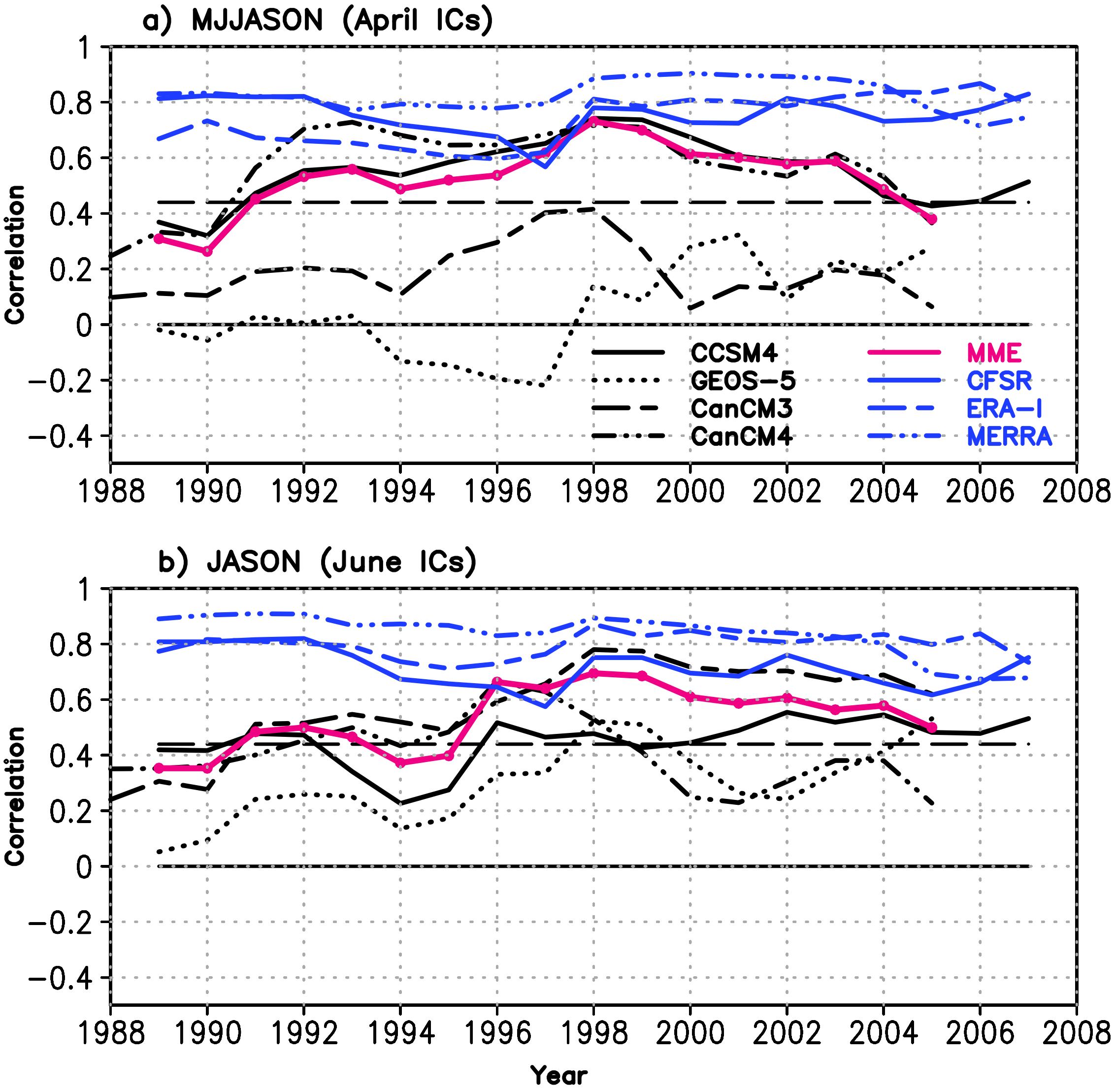

Figure 6: Sliding 15-year correlation of the predicted (ensemble mean) and reanalysis NA TC frequency with the observed (IBTrACS) for the (a) May-November season (forecasts initialized in April), and (b) July-November season (forecasts initialized in June). NMME-Phase II model results are shown in black and solid line for CCSM4, dotted for GEOS-5, long-dash-short-dash for CanCM3, and dot-dot-dash for CanCM4. Results for the MME mean are shown in magenta, and blue for the reanalyses (solid line for CFSR, long-dash-short-dash for ERA-I and dot-dot-dash for MERRA). Horizontal dashed line signifies statistically significant correlation. Horizontal axis marks the central year in the 15-year window.

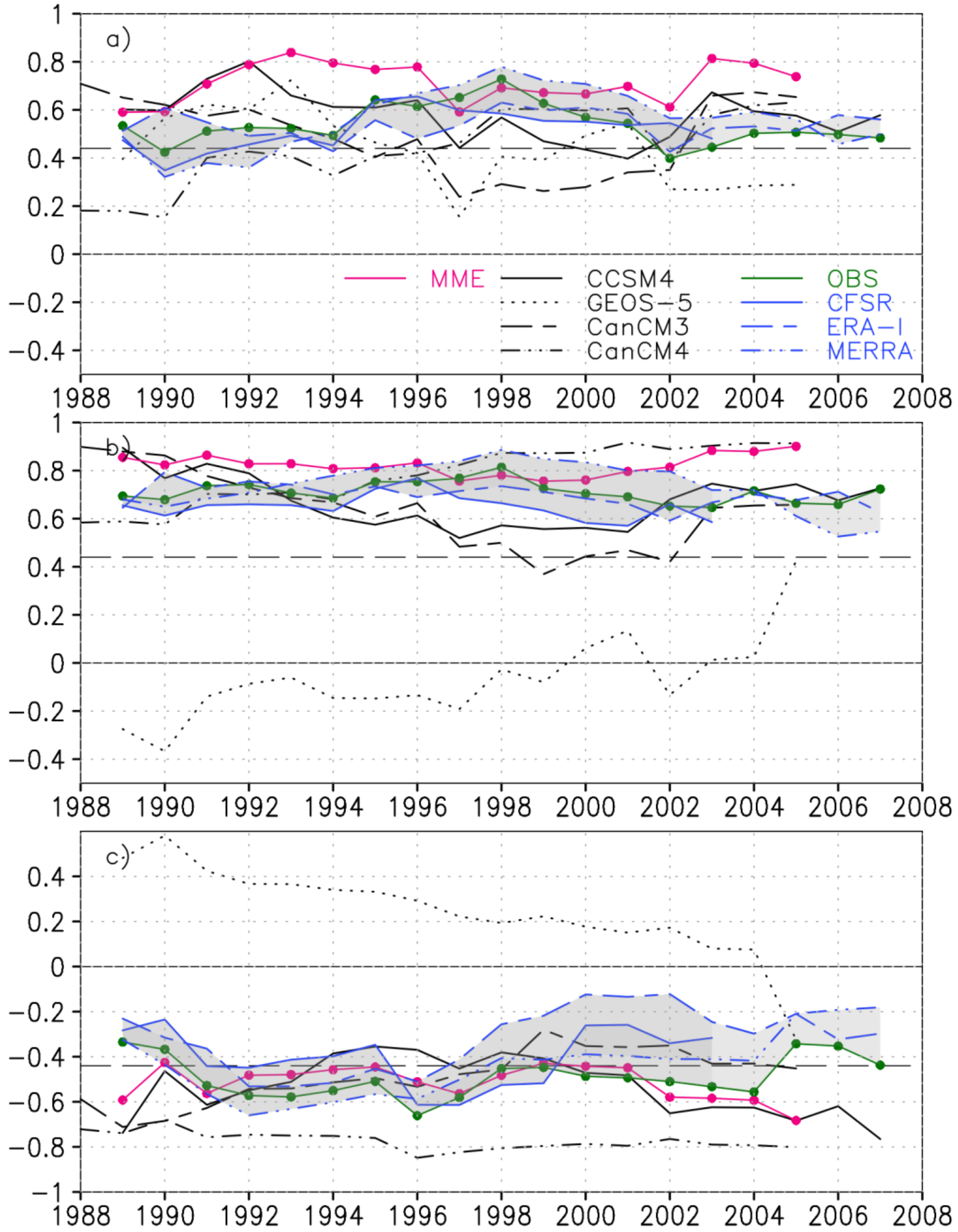

Figure 7: Sliding 15-year correlation of the MJJASON NA TC frequency with the ASO mean (a) MDR SST index; (b) relative SST index; and (c) Niño-3.4 index (see definitions in the text) for observations (IBTrACS vs. OISSTv2), reanalysis and ensemble mean forecasts (initialized in April). NMME-Phase II model results are shown in black and solid line for CCSM4, dotted for GEOS-5, long-dash-short-dash for CanCM3, and dot-dot-dash for CanCM4. Results for the MME mean are shown in magenta, green for observations, and blue for the reanalyses (solid line for CFSR, long-dash-short-dash for ERA-I and dot-dot-dash for MERRA). Grey shading denotes the range of observed/reanalysis values. Horizontal dashed line signifies statistically significant correlation. Horizontal axis marks the central year in the 15-year window.

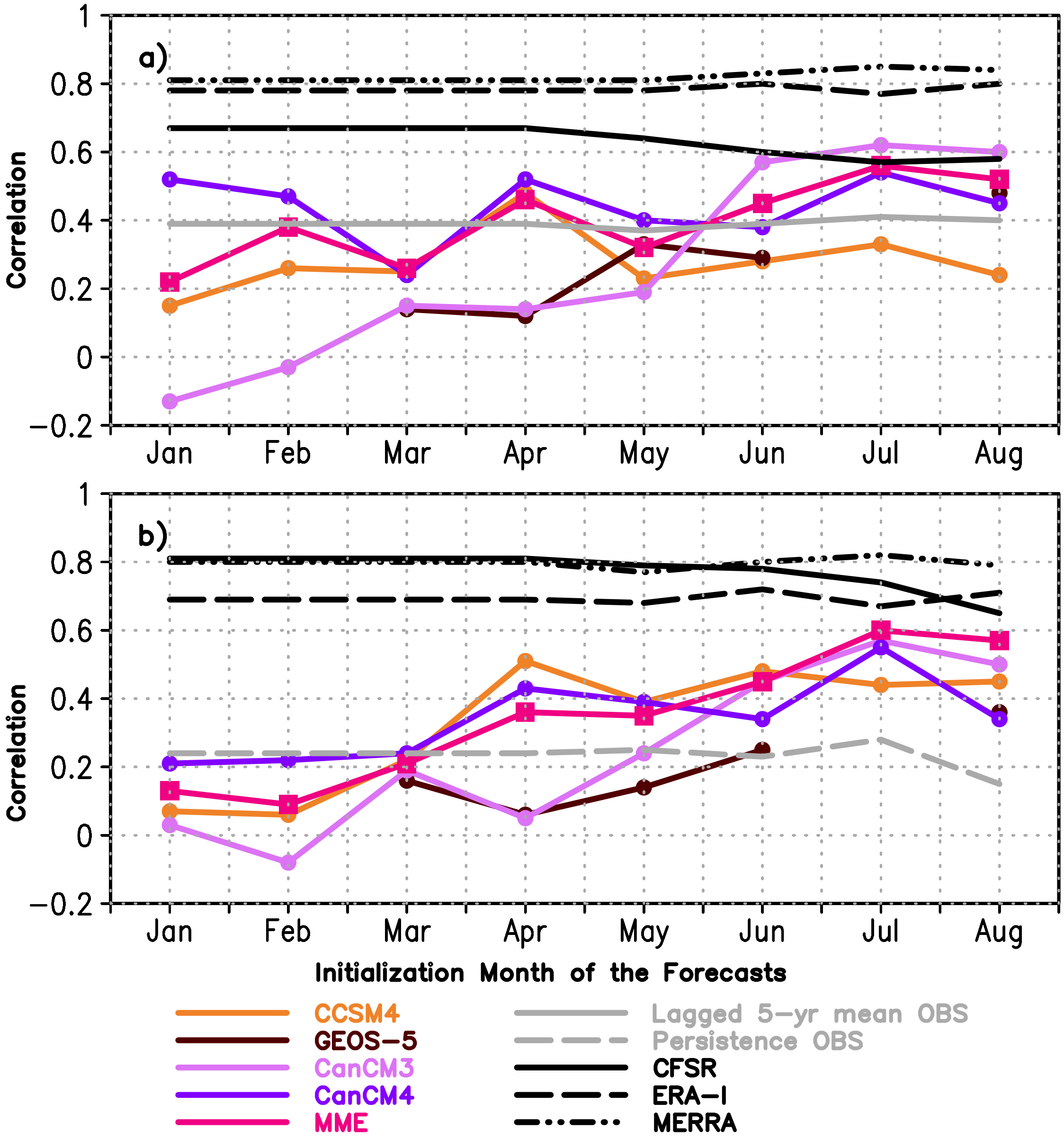

Figure 8: Correlation skill of the seasonal mean NA TC frequency for the NMME-Phase II models, the MME mean and the reanalyses as a function of forecast lead time, shown for the (a) full time series, and the (b) detrended time series. The solid colored lines display the skill of the CCSM4 (orange), GEOS-5 (brown), CanCM3 (lilac), CanCM4 (violet), and the MME mean (magenta). The black lines show the skill of CFSR (solid), ERA-I (long-dash), and MERRA (dot-dot-dash). Results shown are for the May-November average for forecasts initialized in January through April; June-November, July-November, August-November and September-November means when initialized in May, June, July and August, respectively. For the full time series, the skill is compared to a reference forecast comprising of the lagged 5-yr average of the observed TC frequency (solid gray; WMO 2008), and to persistence, or the previous season's observed TC frequency, (long-dash grey) for the detrended cases.

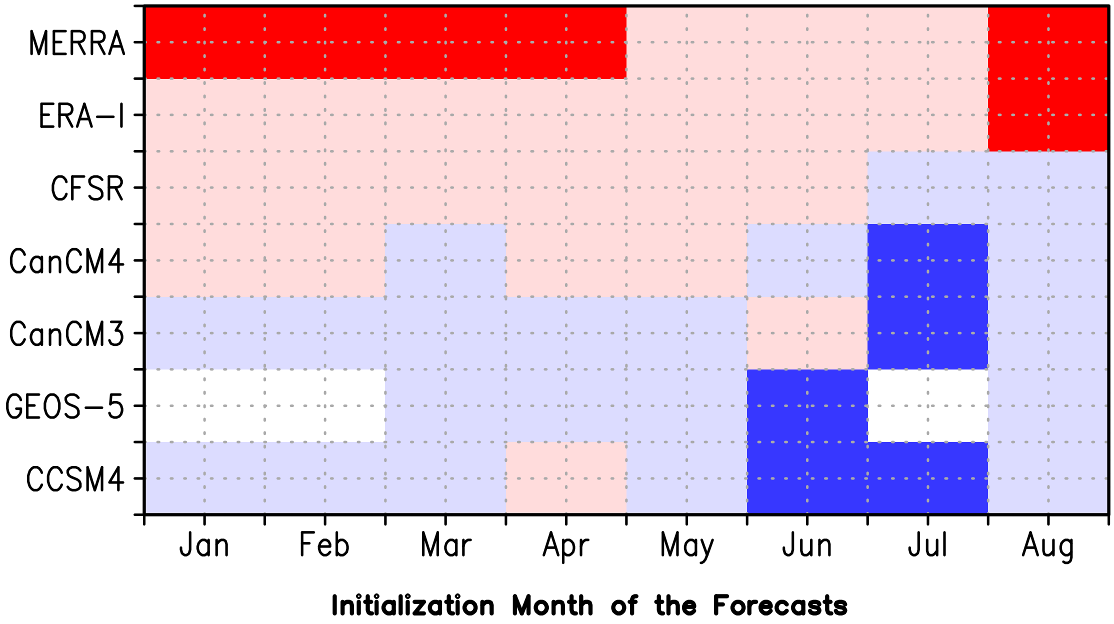

Figure 9: Difference between the squared error of the MME mean hindcasts and the squared error of the NMME-Phase II model or reanalysis indicated on the vertical axis, as a function of forecast lead time. Light blue (light red) color indicates that the MME mean squared error is smaller (larger) than the respective model/reanalysis. Dark blue (dark red) color indicates that the squared error of the MME mean is significantly smaller (larger) than the comparison model/reanalysis at the 95% confidence level using Wilcoxon signed-rank test. White blanks indicate that there are no results due to incompleteness/unavailability of the model data.

Table 1. NMME-Phase II models and forecasts.

Model Name	Modeling Center	Reference	Hindcast Period	Ensemble Size	Lead Times (months)	Atmospheric Model Resolution
CCSM4	University of Miami-Rosenstiel School for Marine and Atmospheric Science (UM-RSMAS)	Kirtman et al. (in prep.)	1982-2014	10	0-11	0.9x1.25 deg. L26
GEOS-5	National Aeronautics and Space Administration (NASA)	Verniers et al. (2012)	1982-2012	10	0-8	1x1.25 deg. L72
CanCM3	Canadian Centre for Climate Modeling and Analysis (CCCMA)	Merryfield et al. (2013)	1981-2012	10	0-11	T63L31
CanCM4	Canadian Centre for Climate Modeling and Analysis (CCCMA)	Merryfield et al. (2013)	1981-2012	10	0-11	T63L35

Table 2. Linear correlation of the predicted (ensemble mean) and reanalysis NA TC frequency with the observed (IBTrACS) for 1982-2014 for the reanalyses data sets, and the time periods listed in Table 1 for the forecasts. Results are shown for May-November (MJJASON), August-November (ASON) and September-November (SON) seasons with forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is based on four or three models listed depending on data availability, as indicated. Values in parentheses show correlation coefficients computed for the detrended time series. Boldface marks values that are statistically significant at the 95% confidence level.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	**0.48** (**0.51**)	0.12 (0.06)	0.14 (0.05)	**0.52** (**0.43**)	**0.46** (**0.36**)	**0.67** (**0.81**)	**0.78** (**0.69**)	**0.81** (**0.80**)
ASON (July ICs)	**0.33** (**0.44**)	-*	**0.62** (**0.57**)	**0.54** (**0.55**)	**0.56** (**0.60**)	**0.57** (**0.74**)	**0.77** (**0.67**)	**0.85** (**0.82**)
SON (August ICs)	0.24 (**0.45**)	**0.48** (**0.36**)	**0.60** (**0.50**)	**0.45** (**0.34**)	**0.52** (**0.57**)	**0.58** (**0.65**)	**0.80** (**0.71**)	**0.84** (**0.79**)

-* incomplete data

Table 3. RMSE between the calibrated ensemble-mean forecasts and the observations (IBTrACS) of the NA TC frequency based on the time periods listed in Table 1, and between the reanalyses and observed NA TC frequency for 1982-2014. Results are shown for May-November (MJJASON), August-November (ASON) and September-November (SON) seasons with forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is based on four or three models listed depending on data availability, as indicated. Values in parentheses show RMSE for the detrended time series.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	3.73 (3.15)	4.32 (3.54)	4.27 (3.58)	3.66 (3.06)	3.87 (3.18)	3.37 (2.37)	2.81 (2.80)	2.57 (2.40)
ASON (July ICs)	3.73 (3.05)	-*	2.89 (2.39)	3.09 (2.44)	3.09 (2.28)	3.34 (2.46)	2.44 (2.43)	1.95 (1.84)
SON (August ICs)	2.93 (2.25)	2.61 (2.23)	2.32 (2.09)	2.59 (2.30)	2.56 (2.02)	2.42 (2.01)	1.79 (1.78)	1.57 (1.54)

-* incomplete data

Table 4. The SPRvERR for the calibrated predicted NA TC frequency based on the time periods listed in Table 1. Results are shown for May-November (MJJASON), August-November (ASON) and September-November (SON) seasons with forecasts initialized in April, July and August, respectively. Multi-model ensemble mean (MME) is based on four or three models listed depending on data availability, as indicated. Values in parentheses show SPRvERR for the detrended time series.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME
MJJASON (April ICs)	0.79 (0.91)	0.59 (0.70)	0.60 (0.69)	0.74 (0.86)	0.74 (0.88)
ASON (July ICs)	0.74 (0.88)	-*	0.93 (1.07)	0.93 (1.11)	1.00 (1.31)
SON (August ICs)	0.75 (0.93)	0.77 (0.88)	0.96 (1.04)	0.90 (0.99)	0.97 (1.20)

-* incomplete data

Table 5. As in Table 2 but for TC days. Only values for the full time series are shown.

Season (ICs)	CCSM4	GEOS-5	CanCM3	CanCM4	MME	CFSR	ERA-I	MERRA
MJJASON (April ICs)	**0.39**	0.21	0.29	**0.57**	**0.46**	**0.85**	**0.82**	**0.82**
ASON (July ICs)	**0.37**	-*	**0.67**	**0.55**	**0.59**	**0.80**	**0.82**	**0.83**
SON (August ICs)	**0.37**	**0.54**	**0.66**	**0.38**	**0.59**	**0.76**	**0.80**	**0.79**

-* incomplete data