Verification of Multimodel Ensemble Forecasts of North Atlantic Tropical Cyclones

NICHOLAS M. LEONARDO AND BRIAN A. COLLE
School of Marine and Atmospheric Sciences, Stony Brook University, State University of New York, Stony Brook, New York

(Manuscript received 1 May 2017, in final form 19 August 2017)

ABSTRACT

North Atlantic tropical cyclone (TC) forecasts from four ensemble prediction systems (EPSs) are verified using the National Hurricane Center’s (NHC) best tracks for the 2008–15 seasons. The 1–5-day forecasts are evaluated for the 21-member National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS), the 23-member UKMO ensemble (UKMET), and the 51-member European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, as well as a combination of these ensembles [Multimodel Global (MMG)]. Several deterministic models are also evaluated, such as the Global Forecast System (GFSdet), Hurricane Weather Research and Forecasting Model (HWRF), the deterministic ECMWF model (ECdet), and the Geophysical Fluid Dynamical Laboratory model (GFDL). The ECdet track errors are the smallest on average at all lead times, but are not significantly different from the GEFS and ECMWF ensemble means. All models have a slow bias (90–240 km) in the along-track direction by 120 h, while there is little bias in the cross-track direction. Much of this slow bias is attributed to TCs undergoing extratropical transition (ET). All EPSs are underdispersed in the along-track direction, while the ECMWF is slightly overdispersed in the cross-track direction. The MMG and ECMWF track forecasts have more probabilistic skill than the ECdet and comparable skill to the NHC climatology-based cone forecast. TC intensity errors for the HWRF and GFDL are lower than the coarser models within the first 24 h, but are comparable to the ECdet at longer lead times. The ECMWF and MMG have comparable or better probabilistic intensity forecasts than the ECdet, while the GEFS’s weak bias limits its skill.

1. Introduction

a. Background

Landfalling tropical cyclones (TCs; see the appendix for a list of key terms and acronyms used in this study) cause a wide range of costly impacts, including widespread damaging winds, coastal storm surge, and inland flooding from heavy precipitation. For example, Hurricane Katrina (2005) caused about 1200 fatalities and over $108 billion (U.S. dollars) in property losses (Blake et al. 2011). Seven years later, Hurricane Sandy (2012) impacted the metropolitan regions of New York and New Jersey, damaging or destroying at least 650,000 homes, leaving approximately 8.5 million people without power for weeks, and causing more than $50 billion in damage (Blake et al. 2013). Accurate forecasts of TCs several days prior to landfall are essential for determining emergency preparations, evacuations zones, and relief supplies. Responses to an approaching TC are often commenced well before the NHC officially issues a tropical storm or hurricane watch or warning (48 and 36 h prior to the expected arrival of tropical storm or hurricane force winds, respectively).

NHC’s track forecasts of North Atlantic TCs have significantly improved over the decades. The official forecast track errors at 48-h lead time steadily decreased ~3% yr\(^{-1}\) between 1990 and 2008. The 72-h track errors throughout 2000 and 2008 were comparable to the 48-h track errors throughout 1990 and 1999 (Rappaport et al. 2009).

Much of the improvement in TC forecasts corresponds to improvements in dynamical models. In 1997, NHC and the Hurricane Research Division began “synoptic surveillance” missions, which use observations from the Gulfstream IV-SP jet to improve the initialization of the near-TC environment in forecasts for storms threatening the United States. Aberson (2010) verified the GFS and GFDL for the 1997–2006 seasons and noted a 10%–15% improvement in the
average track errors of GFS forecasts incorporating synoptic surveillance data. The GFDL track errors did not improve significantly and were on average more than 100 km larger than the GFS at 96 h. However, the GFS track improvement was only statistically significant for lead times less than 72 h, possibly suggesting a predictability limit posed by upstream features not sampled by the surveillance.

Intensity forecasts of TCs have improved much less compared to track forecasts, with dynamical models showing only comparable skill to statistical techniques that use climatology and persistence (Rappaport et al. 2009). Numerous studies show that the interplay of scales and inadequate model resolution contribute to the slower improvement in hurricane intensity forecasts (e.g., Gall et al. 2013). Some evidence suggests that recent upgrades in high-resolution regional models, such as HWRF, have significantly reduced their average intensity errors (Goldenberg et al. 2015) and that the rate of improvement of models at 48–96-h lead times may be outpacing that of the NHC forecasts (DeMaria et al. 2014). However TCs that are initially weak and/or undergoing rapid intensification continue to prove difficult to forecast consistently from case to case (Kaplan et al. 2010).

Studies have also verified the TC forecasts of ensemble systems, such as the Global Ensemble Forecast System (GEFS). Most operational ensembles have limited resolution, so much of the emphasis has been on track validation. Buckingham et al. (2010) verified the GEFS for all TCs of the 2006–08 North Atlantic and western North Pacific seasons. They noted that the track errors of the North Atlantic and North Pacific basins linearly increased at rates of ~90 and ~110 km day$^{-1}$, respectively, during the first 144 h. The average tracks of North Atlantic TCs exhibited a significant slow bias that grew with lead time. Much of this slow bias in the North Atlantic was attributed to TCs undergoing extratropical transition (ET). The ensemble dispersion was shown to be adequate in the North Atlantic and underdispersive in the North Pacific.

Hamill et al. (2011) verified their own 30-km ensemble, in which they ran the Global Forecast System (GFS) and the Flow-Following Finite-Volume Icosahedral Model (FIM) initialized with the ensemble Kalman filter (EnKF). They simulated all North Atlantic and North Pacific TCs of the 2009 season. Their verification also compared four operational EPSs: the GEFS and those of the European Centre for Medium-Range Weather Forecasts (ECMWF), Met Office (UKMO), and Canadian Meteorological Centre (CMC). The GFS-EnKF and FIM-EnKF mean track errors were lower than the GEFS, UKMO, and CMC results and comparable to those of the ECMWF. The GFS-EnKF and ECMWF both showed the most consistency between the magnitudes of spread and error.

Nixon (2012) verified and compared the ECMWF, GEFS, and UKMO ensembles for the 2008–11 North Atlantic seasons. He found that while the ECMWF had the lowest ensemble mean track errors for all TCs across the entire basin, the GEFS had the lowest errors over the U.S. East Coast region. The models all showed positive biases in the along-track direction, particularly for TCs over the Gulf of Mexico region. However, the sample variances over most of the individual regions were too large for the biases to be statistically significant. The ECMWF was the most reliable ensemble in that its spread most frequently captured the observed TC between 24 and 120 h. The other two EPSs were somewhat underdispersed. The combination of the three ensembles showed even better spread, though most of this was contributed by the ECMWF.

Some studies have shown the benefits of combining ensemble systems and models. Goerss (2000) verified the track forecasts from several deterministic global and regional models for the 1995 and 1996 North Atlantic hurricane seasons and found that their consensus forecast produced 72-h track errors that were 16%–23% lower than the best individual model. This was also shown to be true for western North Pacific TCs, with successive improvements in the skill of the consensus forecast when adding more models (Goerss et al. 2004). The average track of several deterministic models is now often used as a tool to aid and evaluate forecasters (Cangialosi and Franklin 2014).

There is some evidence that EPS mean forecasts may also show deterministic skill in the long run. For the 2005–07 North Atlantic seasons, the deterministic GFS track errors were on average lower than the GEFS mean at 24–72 h; however, the GEFS mean outperformed the GFS for lead times longer than 96 h (Rappaport et al. 2009).

b. Motivation

Although there has been some model evaluation for TCs, limitations in previous ensemble validation of TCs leave unanswered questions. Nixon (2012) verified the combination of three EPSs for a 5-yr period, but these thesis results were not more formally published and reevaluated for subsequent years. The sample sizes for most other studies are limited to ranges from only one to three seasons. This small range in time prevents an analysis of any trends in the EPS errors over time. Furthermore, most of these studies analyzed the EPSs individually. It has not been shown in the formal literature whether TC forecasts benefit from the combination...
of EPSs in the same manner as the deterministic model consensus used by Goerss (2000).

There is also a need for a more comprehensive verification of the dispersion and probabilistic skill of EPS forecast tracks, extending beyond a simple comparison of the ensemble mean and spread values. While metrics such as probabilities within spread and ellipse reliability are commonly used, they are within a relative-to-mean Lagrangian framework that assumes a specific distribution in the clustering of ensemble member positions around the mean position. Furthermore, these metrics provide limited information regarding how well the EPSs forecast the likelihood that TCs will impact specific regions, such as major metropolitan areas.

Attempts to compare the probabilistic skill of the EPSs with the cone forecasts from operational centers have also been limited. The widths of the NHC cones are seasonally fixed, given by the 87th percentile of the NHC track errors over the previous 5 yr (http://www.nhc.noaa.gov/aboutcone.shtml). This climatology-based representation of uncertainty does not account for the case-dependent sensitivity of track to specific external and internal processes. Dupont et al. (2011) verified forecasts of Indian Ocean TCs and compared the ECMWF against cone probability forecasts that they derived from the climatology of track errors of the Regional Specialized Meteorological Centre of La Réunion in the Indian Ocean (Météo-France). While the ECMWF had more skill than the cone during the first 72 h, the sample size in this study was limited to the 2007–09 seasons.

Specifically, this study will address the following questions:

1) What is the accuracy of the operational EPSs and deterministic model track and intensity forecasts for North Atlantic TCs from day 1 to 5?
2) Have the 72–120-h forecasts improved throughout the previous several years?
3) Are there any biases in the model track and intensity forecasts? How much of the slow bias in track can be attributed to ET events?
4) Do the ensemble track and intensity forecasts show better probabilistic skill compared to the best deterministic model? How does the probabilistic skill for track compare against the NHC climatology-based cone forecasts?
5) What are the merits of combining multiple EPSs?

Section 2 describes the datasets and methods used throughout this study. Section 3 presents the deterministic scores of multiple models and EPSs, including their biases and trends. Section 4 presents the probabilistic verification of the global ensembles and how their skill compares to that of the most deterministically skillful model and the NHC forecast. Section 5 concludes with a summary of the results.

2. Data and methods

a. Datasets

The main goal of this study is to verify operational model forecasts for all North Atlantic TCs from 2008 to 2015. Much of this study's focus is on evaluating three commonly used global EPSs: the ECMWF (51 members), the Met Office ensemble (UKMET; 23 members), and the NCEP GEFS (21 members). The combination of these ensembles is called the Multimodel Global (MMG; 95 members) ensemble. The cyclone tracks from these EPSs and the deterministic ECMWF (ECMWFdet) are archived in near–real time by the THORPEX Interactive Grand Global Ensemble (TIGGE) database, available online through NCAR (http://rda.ucar.edu/datasets/ds330.3/). No pregenesis forecasts (e.g., in which the TC had yet to attain tropical depression status) are available in this archive. Furthermore, the ECMWF ensemble prior to 2014 only has the forecast times in 12-h intervals up to 120 h. Thus, the verification in this study is limited to postgenesis stages of named TCs at lead times up to 120 h.

The tracks from several deterministic models are analyzed for comparison: the operational UKMET (UKMdet) and GFS (GFSdet), the Canada Global Environmental Multiscale Model (CMC), the Navy Operational Global Prediction System (NGPS), the NWS/Hurricane Weather Research Model (HWRF), the NWS/Geophysical Fluid Dynamics Laboratory Model (GFDL), and the official NHC forecast (OFCL). The HWRF and GFDL are regional models in that their domains cover a limited area. Global models, such as the GFSdet, have coarser grids that encompass the entire globe and provide initial and lateral boundary conditions (ICs and BCs) for the higher-resolution regional models. Note that beyond the first 48 h, the OFCL is incremented every 24 h. All models are verified by the NHC’s best-track dataset: the best estimate of the observed TC's position and intensity based on the synthesis of various observations (Cangialosi and Franklin 2014). The tracker data of these models come from various agencies and is aggregated and archived along with the best-track data by NHC. This archive can be accessed online (ftp://ftp.nhc.noaa.gov/atcf/archive/).

Tables 1 and 2 provide the summary characteristics of these EPSs and models. With no set standard for cyclone tracking, the tracking algorithms used by these models (not shown) and their respective operational agencies...
 vary, but generally they use lower-tropospheric vorticity maxima and/or mean sea level pressure (SLP) minima within a given search radius to define the TC center (Van der Grijn and Grazzini 2003).

b. Model forecast data homogeneity

The forecast lead time during which each model or ensemble member first detects a TC can vary. Similarly, the time when a model’s tracking algorithm can no longer track a TC can differ. In this study, all model comparisons are homogenous in that the statistics of each model are drawn from the same set of forecasts available from all models. Homogeneity is maintained for the ensemble verification in the following manner: for each of the ensembles, a forecast hour is verified only if at least eight of their members have cyclone track data available. This number was empirically determined by Buckingham et al. (2010) to be adequate for producing smooth realistic mean track forecasts. Hence, in the creation of the MMG composed of three EPSs, a forecast hour is only verified if each EPS has at least eight members with a TC.

The number of homogenous forecasts with respect to all EPSs and models is shown as a function of lead time in Table 3. The 2011 and 2012 seasons were the most active, while 2009, 2013, and 2015 were significantly less active. The year 2013 was especially inactive in terms of long-lived TCs, with only two hurricanes. Many of the models failed to maintain the TCs by 72 h in various forecasts (not shown). Thus, for the analysis of model trends, homogeneity is maintained for forecasts common to only the OFCL, HWRF, GFDL, and the EPSs, and inactive years are combined (e.g., 2008 with 2009 and 2013–15). The result is that each time period averaged includes at least 30 forecasts.

c. Track and intensity validation

This study focuses on the validation of the TC intensity and tracks in several deterministic models and EPSs. The ensemble mean position at a given lead time is only verified if each EPS has at least eight members with a TC.

<table>
<thead>
<tr>
<th>Model</th>
<th>Horizontal resolution (km)</th>
<th>No. of vertical levels</th>
<th>DA initial conditions</th>
<th>Convective scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13 (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25 (2010–14)</td>
<td>70 (&gt;2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 (&gt;2014)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>42 (2010–14)</td>
<td>42 (2009–14)</td>
<td>Naval Research Laboratory Atmospheric Variational Data Assimilation System with the accelerated representer (NAVADAS-AR), 4DVAR (&gt;2009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31 (&gt;2014)</td>
<td>60 (&gt;2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GFDL</td>
<td>55/18/12 (&gt;2014)</td>
<td>42</td>
<td>Bogus vortex, GFS ICs/BCs</td>
<td>Simplified Arakawa–Schubert</td>
</tr>
<tr>
<td></td>
<td>55/18/6 (&gt;2014)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>18/6/2 (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of homogenous forecasts with respect to all EPSs and models is shown as a function of lead time in Table 3. The 2011 and 2012 seasons were the most active, while 2009, 2013, and 2015 were significantly less active. The year 2013 was especially inactive in terms of long-lived TCs, with only two hurricanes. Many of the models failed to maintain the TCs by 72 h in various forecasts (not shown). Thus, for the analysis of model trends, homogeneity is maintained for forecasts common to only the OFCL, HWRF, GFDL, and the EPSs, and inactive years are combined (e.g., 2008 with 2009 and 2013–15). The result is that each time period averaged includes at least 30 forecasts.

c. Track and intensity validation

This study focuses on the validation of the TC intensity and tracks in several deterministic models and EPSs. The ensemble mean position at a given lead time

<table>
<thead>
<tr>
<th>EPS name</th>
<th>Horizontal resolution (km)</th>
<th>No. of vertical levels</th>
<th>Perturbation/consensus model methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35 (2015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UKMET (23)</td>
<td>60 (&lt;2015)</td>
<td>70</td>
<td>44-member EnKF</td>
</tr>
</tbody>
</table>
is the average of all member TC positions. The ensemble mean track error is then given by the great-circle distance from this point to the position of the best track. Other distance calculations are performed in the same manner.

TC intensity is quantified by either the maximum sustained surface wind or the minimum central SLP. The intensity errors are then given by subtracting the best-track intensity from the model TC intensity. Both the raw and absolute averages of these errors are analyzed. The raw errors give the sign of a bias while the mean absolute errors (MAEs) give the magnitude. The UKMET ensemble and UKMdet do not have intensity archived prior to 2014 and are thereby excluded from the verification. The CMC stopped including model TC wind estimates after 2011 and is excluded from the wind intensity verification. The OFCL only forecasts the maximum wind speed and is excluded from the SLP intensity verification.

As demonstrated by Froude et al. (2007) for extratropical cyclone tracks, the total track error (TTE) can be decomposed into along-track (ATE) and cross-track (CTE) errors relative to the bearing of the best track (Fig. 1). By this convention, a positive (negative) ATE corresponds to a forecast TC that is too fast (slow) relative to the observed TC. Similarly, a positive (negative) CTE corresponds to a forecast TC that is too far to the right (left) of the observed TC.

NHC annotates the status of the observed TCs in their best-track dataset, including when the TC becomes extratropical. The contribution of ET events to average along-track biases is assessed by repeating the verification after removing all forecasts in which the best-track TC is labeled as extratropical.

Table 3. Number of TCs (events) verified at each lead time and in each year, with all forecasts available from all models.

<table>
<thead>
<tr>
<th>Lead time (h)</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>14 (45)</td>
<td>8 (47)</td>
<td>18 (81)</td>
<td>17 (149)</td>
<td>19 (185)</td>
<td>8 (30)</td>
<td>7 (39)</td>
<td>10 (48)</td>
</tr>
<tr>
<td>24</td>
<td>13 (42)</td>
<td>6 (40)</td>
<td>16 (70)</td>
<td>16 (125)</td>
<td>17 (160)</td>
<td>6 (20)</td>
<td>6 (37)</td>
<td>9 (41)</td>
</tr>
<tr>
<td>48</td>
<td>9 (32)</td>
<td>5 (33)</td>
<td>13 (57)</td>
<td>11 (100)</td>
<td>15 (125)</td>
<td>3 (7)</td>
<td>6 (32)</td>
<td>7 (26)</td>
</tr>
<tr>
<td>72</td>
<td>9 (24)</td>
<td>4 (24)</td>
<td>10 (45)</td>
<td>8 (74)</td>
<td>12 (101)</td>
<td>2 (3)</td>
<td>5 (23)</td>
<td>6 (27)</td>
</tr>
<tr>
<td>96</td>
<td>6 (17)</td>
<td>3 (17)</td>
<td>10 (35)</td>
<td>7 (53)</td>
<td>8 (73)</td>
<td>1 (1)</td>
<td>4 (18)</td>
<td>4 (16)</td>
</tr>
<tr>
<td>120</td>
<td>5 (8)</td>
<td>2 (10)</td>
<td>6 (27)</td>
<td>5 (30)</td>
<td>7 (53)</td>
<td>1 (1)</td>
<td>3 (10)</td>
<td>3 (8)</td>
</tr>
</tbody>
</table>

d. Analysis of ensemble dispersion and probabilistic skill

A straightforward way of evaluating the ensemble dispersion of forecast tracks is through the use of probabilities within spread (PWS). This metric quantifies the likelihood that the observed TC will fall within the ensemble spread. First, the spread $\sigma$ is calculated from the $M$ members of each forecast at a given lead time, using the distances of each $i$th member from the ensemble mean position $s_i$:

$$\sigma^2 = \frac{1}{M} \sum_{i=1}^{M} s_i^2.$$  \hspace{1cm} (1)

For each of the $N$ ensemble forecasts, an integer multiple ($k = 1, 2, 3, \ldots$) of $\sigma$ is compared against the distance of the observed TC from the mean position $\overline{TTE}$:

$$\text{PWS} = \frac{1}{N} \sum_{n=1}^{N} \left\{ \begin{array}{ll} 0: & \text{TTE} > k(\sigma)_n \\ 1: & \text{TTE} \leq k(\sigma)_n \end{array} \right\}.$$  \hspace{1cm} (2)

An EPS that has a high (low) PWS implies that the member tracks tend to (not) have enough spread to...
reflect the track uncertainty. Assuming that the members are sampled from a Gaussian distribution with a standard deviation of $\sigma$, then $1\sigma$ and $2\sigma$ should correspond to PWS values of 0.67 and 0.95, respectively. That is, the $1\sigma$ and $2\sigma$ spreads should contain the best-track position 67% and 95% of the time, respectively.

Ensemble track dispersion can also be diagnosed using rank histograms (Hamill 2001). Rank histograms show the frequency with which an observation falls between each pair of members in an ordered set of ensemble values. In this case, the values are the 72–120-h member ATEs and CTEs. The observation has a value of zero and is ranked among eight sorted values drawn randomly from an EPS without replacement. For the MMG ensemble, eight members are drawn randomly from each of the three EPSs for a total of 24 values. A flat rank histogram implies adequate dispersion, with the observation having an equal chance of falling anywhere within the ensemble. Unlike PWS, this method accounts for the anisotropy of ensemble spread and does not assume a Gaussian distribution.

The probabilistic skill of the EPSs relative to certain models is assessed with the Brier skill score (BSS). Given $N$ forecasted probabilities $f$ of an event paired with the condition $o$ of whether that event was observed ($o = 1$) or not observed ($o = 0$) in the best track, the Brier score $b$ is calculated as follows (Wilks 2011):

$$ b = \frac{1}{N} \sum_{n=1}^{N} [f(n) - o(n)]^2. \quad (3) $$

The BSS is then given by

$$ \text{BSS} = 1 - \frac{b}{b_{\text{ref}}}, \quad (4) $$

where $b_{\text{ref}}$ is the Brier score of some reference model. A BSS of one indicates that the EPS gives a perfect forecast relative to the reference, while a BSS of zero (or less than zero) indicates that the EPS provides no improvement over (or is less skillful than) the reference.

The Brier score can be decomposed into three components: reliability, resolution, and uncertainty (Murphy 1973). These components can be shown in reliability diagrams, which are derived in the following manner. The probability forecasts for an event are binned into $I$ discreet intervals of probability $y_i$, which in this case are chosen to be every 20% ($\pm 10\%$) from 0% to 100%. The percentage of forecasts for $y_i$ of which the event is observed is the conditional observed frequency $o_i$. Reliability diagrams plot $o_i$ against $y_i$. The reliability component evaluates the statistical accuracy of a forecast, quantifying how much the forecast deviates from a 1:1 relationship between $o_i$ and $y_i$ (e.g., for all forecasts of a $y$-percent chance of an event, the event occurs in $y$ percent of those forecasts). Resolution addresses the ability to distinguish between different regimes from climatology and is given by the difference between $\bar{o}_i$ and the unconditional observed frequency of the event $\bar{o} = (1/N)\sum_{n=1}^{N} o(n)$. The uncertainty term $\sigma(1 - \sigma)$ is a measure of the variability of the system and is not influenced by the forecast.

In this study, the BSS is used to evaluate the day 3–5 forecast strike probabilities relative to fixed geographical points. The strike probability from an ensemble forecast is given by the fraction of members within a threshold distance from a grid point. In the example shown in Fig. 2a, two out of five members are within the 200-km circle from the grid point, such that the strike probability for this one event is 40% [e.g.,
while the verifying observed position is outside of the circle \( o(n) = 0 \). The strike probabilities are calculated at each of the 24 different points within a 5° grid along the U.S. Atlantic coast (Fig. 2b). Note that certain points on this grid are located close to metropolitan areas, such as New York, New York; Miami, Florida; and New Orleans, Louisiana. The BSS calculation is repeated using different threshold distances between 300 and 600 km from the points. The lower limit was chosen such that an adequate number of events (at least 100) have an observed TC within that distance from a point. The upper limit was a compromise intended to reduce the overlap of threshold radii between multiple grid points. Forecasts for a point in which the verifying best-track position is more than 2500 km away are excluded to minimize the number of cases in which all models correctly predict a 0% probability of a hit (e.g., so \( \sigma \) is not too small). This distance was chosen in that it is more than 600 km larger than the largest TTE forecast from any of the EPSs.

As a reference in the BSS, the strike probabilities from NHC’s climatology-based cone approach are quantified by fitting concentric circles onto the OFCL forecast track. The radius of each circle of \( p \) probability is given by the \( p \)th percentile of all OFCL TTEs during the previous five seasons (Fig. 3a). The percentiles are extracted every 10% up to 100%, such that there are 10 concentric rings that each contain 10% of the full probability distribution. The OFCL forecast probability is then calculated by summing the fractions of each ring’s area overlapping with the threshold circle, as demonstrated in Fig. 3b. It is important to note that the operational EPSs take several hours to complete after the valid time they are initialized, while the OFCL forecast is issued no more than 3 h after the initial valid time. To account for the lateness of the EPSs, each OFCL 72-h (96 h) forecast is compared with the 84-h (108 h) forecast of the EPS run initialized 12 h earlier. The EPS member tracks are interpolated, such that the 12-h positions are shifted to the current 0-h best-track position, and the subsequent positions are displaced by the same vector.

A BSS analysis is also used to verify the 72–120-h probabilistic intensity forecasts. In this case, the event centers on whether the wind speed of a TC will exceed a specific threshold. The BSS is repeated for thresholds ranging from 18 \( \text{m s}^{-1} \) (35 kt) to 44 \( \text{m s}^{-1} \) (85 kt). The lower limit is approximately minimal tropical storm intensity. The upper limit is close to minimal category 2 hurricane intensity and was chosen because the GEFS never forecasted a probability over 60% of winds exceeding 46 \( \text{m s}^{-1} \).

3. Multimodel deterministic validation

The model TTEs calculated for all years range between 350 and 680 km by 120 h (Fig. 4a). Overall, TTE increases linearly during the 5-day forecast period, with the error increase ranging from 75 to 125 km day\(^{-1} \). The NGPS has larger errors than the other models, especially between 48 and 96 h. The higher-resolution regional models (e.g., GFDL and HWRF) show comparable track errors to the CMC and the UKMET after 72 h, while the other models (besides NGPS) are clustered with TTEs \( \sim 400–450 \text{ km by 120 h} \). The UKdet, ECdet, and GFSdet each show slightly lower TTEs than the means of their coarser ensemble counterparts, though the overlapping error bars suggest that the
difference is not statistically significant. The ECdet has the smallest errors by 120 h, followed closely by the MMG mean and the ECMWF mean.

For most models, a large fraction of the TTE is from a slow bias in the along-track component, which exceeds 90 km in magnitude by 120 h (Fig. 4b). The ECMWF ensemble mean has one of the worst slow biases, ~37% larger than the ECdet. Comparatively, there is greater variability in the cross-track biases among models, most of which are not significantly different from 0 (Fig. 4c). The GEFS and GFSdet have a right-of-track bias (~75 km) that is only significant at 120 h.

The percentage of forecasts for which each model has the smallest TTE relative to all other models was also calculated (Fig. 5a). The EPSs are again outperformed by their deterministic versions, with the exception of the UKMET. The ECdet does the best beyond 72 h, followed by the GFSdet and CMC, both of which have comparable percentages by 120 h. The OFCL has the highest percentage of best initial positions. Interestingly, the HWRF least often has the relative best forecast at 120 h, despite its finer grid spacing and coupling with an ocean model. Meanwhile, the percentages of forecasts from each model that produced the largest TTE (Fig. 5b) indicate that the EPSs are less often the worst compared to their respective deterministic counterparts. The MMG in particular is never the worst at any lead time, even relative to the OFCL. The GEFS, ECdet, ECMWF, and OFCL are each only the worst for less than 5% of all forecasts beyond 72 h.

The TC intensity verification shows that the regional models and the OFCL have the lowest MAEs in minimum central SLP and maximum surface wind speeds during the first 24 h (Fig. 6). The global models have significantly larger MAEs consistent with weak biases in both intensity parameters (Figs. 6c,d). However, the ECdet errors do not grow as fast, such that its MAEs in intensity are comparable to the GFDL and HWRF results between 36 and 48 h. Both the ECdet and GFSdet outperform their respective ensembles, the GFSdet in particular with respect to wind speed errors (Figs. 6b,d). By 96 h, the intensity errors of the regional models are comparable to those of most of the global models and are associated with growing strong biases. The ECdet intensity biases become comparable to zero by 120 h.

The percentage best and worst forecasts from each model were calculated using MAEs in SLP and wind speed (Fig. 7). For SLP (Figs. 7a,c), the HWRF most often has the best initial intensity, followed by the GFDL. These two regional models each have the lowest SLP errors for more than 25% of all forecasts within the first 48 h, but the ECdet has a comparable share of best forecasts for longer lead times (Fig. 7a). The number of poorest SLP forecasts (Fig. 7c) from the regional models increases more than 20% by 120 h. As with the track errors in Fig. 6b, the NGPS has the worst MAEs in SLP for more than 25% of all forecasts within 96 h and the MMG is never the worst model. The MAEs in wind speed (Figs. 7b,d) have similar results, though the OFCL (in place of the CMC) has the best forecasts over 20% of the time within the first 72 h.

Figure 8 shows the annual averages of 72- and 120-h TTEs, ATEs, and CTEs for the regional models, the OFCL, and the ensemble means of each EPS. Inactive years were combined to increase the sample size of each interval of time. The average error of each season between 2013 and 2015 was heavily weighted by one of three TCs: Humberto (2013), Cristobal (2014), and
Joaquin (2015). For example, the EPS mean TTEs for two successive 120-h forecasts of Cristobal exceeded 900 km. Excluding these two forecasts from the sample reduces the annual average TTE of 2014 by up to ~130 km (~31%). Similarly, the 2013 average 120-h ATEs of most models showed a fast bias, but excluding two poor forecasts of Humberto reduces the average bias by up to ~81 km (~54%).

All models and EPSs show an overall decrease in 72-h TTEs between 2010 and 2012 (Fig. 8a); the net improvement ranged between ~45 and 130 km (between ~15% and 36%). As implied by the error bars of 2010

Fig. 5. Percentage of forecasts for which each deterministic model, ensemble mean, and OFCL forecast had (a) the smallest and (b) largest TTEs relative to all other models and ensembles.

Fig. 6. Average 2008–15 (a) SLP (hPa) and (b) wind speed (m s⁻¹) MAEs as a function of lead time for the deterministic models, ensemble means, and OFCL forecast. (c),(d) As in (a) and (b), but for the average biases.
and 2012, the difference is statistically significant for all models except the GFDL and ECMWF mean. The 120-h TTEs have large year-to-year variability (Fig. 8b), such that there is no significant difference between any two years common among the models.

The negative 72-h along-track biases of all models (Fig. 8c) improve by \(-40\)–\(-110\) km (\(-36\%\)–\(-84\%\)) overall throughout the entire sample period, albeit having a minimum in magnitude in 2011. However, the difference between the 2013–15 period and 2010 is only significant for the UKMET, ECMWF, and MMG results. The 120-h ATEs of most models also show an overall improvement of \(\sim 30\)–\(\sim 280\) km, but the change is only significant for the UKMET. Meanwhile, the biases in the cross-track direction (Figs. 8e,f) do not appear to undergo a period of significant improvement common among the models. While the 120-h CTEs (Fig. 8f) switch from slightly positive to slightly negative throughout the sample period, the difference from zero is not consistently significant from year to year.

The annual MAEs in SLP and wind speed (Fig. 9) were also calculated for the same set of forecasts in Fig. 8. Between 2010 and 2012, the 72-h SLP errors (Fig. 9a) of the EPSs significantly improve by 5–11 hPa (\(31\%\)–\(53\%\)). The regional model errors fluctuate significantly during the sample period. The 120-h SLP errors (Fig. 9b) of the EPSs also significantly improve by at least 8 hPa between 2010 and 2012. Similarly, the 72- and 120-h wind speed errors (Figs. 9c,d) of the EPSs significantly improve by more than 5 m s\(^{-1}\) (\(\sim 10\) kt) over the same period. The GEFS mean initially has the largest wind speed errors in 2008 and 2009, but steadily improves at a rate of \(\sim 2\) m s\(^{-1}\) yr\(^{-1}\) up to 2012. Comparatively, the OFCL wind speed errors are consistently lower than the other models and change very little throughout the entire sample period.

Comparing Fig. 9 with Figs. 8c and 8d, there appears to be some consistency between the overall improvements in the EPS track errors and intensity errors, both decreasing between 2010 and 2012. As shown in Table 2, the GEFS and ECMWF both received resolution upgrades in 2010. Afterward, the GEFS received upgrades to its data assimilation and perturbation methods in 2012, while the ECMWF further increased its model resolution in 2013. The gaps in major upgrades make it difficult to associate the steady decrease in track and
intensity errors between 2010 and 2012 to specific improvements in model resolution or initialization.

A significant portion of the slow along-track bias is related to ET events. Comparing the same models in Fig. 8, the sample-averaged ATE is recalculated after omitting all forecast hours when the observed TC was classified by NHC as extratropical (Fig. 10). Comparing Figs. 10a and 10b, all models show a 25%–40% reduction in the slow bias after excluding these events. The ECMWF shows the most significant improvement, with the mean 120-h ATE increasing from $220$ to $120$ km. However, the number of ET events varies from year to year between 0 and 28 (0%–40% of each year’s total number of forecasts). These small sample sizes limit any statistical analysis from determining whether the overall decrease in the slow bias with time (e.g., Figs. 8c,d) corresponds to improved track forecasts of ET events or changes in the frequency of ET events.

4. Probabilistic verification and dispersion

This section evaluates the probabilistic skill and dispersion of ensemble tracks with various methods, starting with PWS. Except for the MMG, the PWS of each ensemble is significantly underdispersed at 0 h (Fig. 11), implying that their initialized TCs are inaccurate relative to the best-track position. The track diversities of the UKMET and ECMWF quickly adjust within the first 12 h such that their PWSs sharply increase toward the $1\sigma$ and $2\sigma$ values. However, the UKMET then decreases at 24–72 h, corresponding to
the ensemble mean track error tending to grow faster than the spread. Meanwhile, the ECMWF steadily increases, remaining slightly underdispersed with respect to 1σ and almost perfectly dispersed with respect to 2σ. The GEFS increases before leveling off by ~24 h and becoming similar to the UKMET after ~48 h. Beyond the first 24 h, the GEFS and UKMET are significantly less dispersed than the ECMWF for both σ values. The MMG is the most consistently dispersed with respect to 1σ, though it shows modest signs of overdispersion between 12 and 48 h. The MMG is somewhat overdispersed with respect to 1σ, implying that the addition

**FIG. 9.** Annual averages of 72-h MAEs in (a) minimum SLP (hPa) and (b) maximum wind speed (m s⁻¹). (c),(d) As in (a) and (b), but for 120-h forecasts.

**FIG. 10.** Average 2008–15 ATEs when (a) including and (b) excluding forecasts in which the observed TC was extratropical.
of UKMET and GEFS member outliers produces too much spread.

The 72–120-h EPS dispersion with respect to the observed TC’s motion is demonstrated in rank histograms (Fig. 12). The observations in the MMG and ECMWF histograms are somewhat clustered toward the middle ranks for the cross-track direction, implying slight overdispersion. Conversely, the GEFS and UKMET are underdispersed in the cross-track direction, with the observations most frequently among the leftmost and rightmost ranks. In the along-track direction, all of the EPSs have underdispersed histograms. The observed TC is most frequently placed among the highest (fastest) ranks, leading all ensemble members in the along-track direction for ~30% of all forecasts and trailing all members for only 5%–10% of all forecasts. The histograms for the along-track direction are thus consistent with the significant slow bias of each EPS.

Since the ECdet track forecasts have the most deterministic skill at 72–120 h (Figs. 4a and 5a), the ECdet was first used as a reference model in the BSS calculation, evaluating the forecasted strike probabilities at the fixed grid points in Fig. 2b. Note that the probability forecasted by a deterministic model can only be zero or one, whereas the ensembles can have probabilities that range anywhere from zero to one. The BSSs indicate that all of the EPSs have positive skill for threshold distances less than 400 km (Fig. 13a), implying that their strike probabilities outperform the ECdet for smaller targets. The MMG and ECMWF are the most skillful, with scores that are at least 0.25 in value and significantly different from zero for all distances. The other two EPSs have smaller slopes that are below the 1:1 line, tending to overforecast strike probabilities between 60% and 100%. A TC is only observed within 300 km from a point in ~62%—71% of the forecasts in which the UKMET and GEFS predict a 100% probability of a hit (Fig. 13b). This overprediction of large strike probabilities becomes less severe for larger threshold distances mainly as a result of a greater number of observed TCs falling within the larger targets (e.g., fewer false alarms from the UKMET and GEFS). The MMG and ECMWF models have good resolution in that they are able to distinguish between regimes differing from climatology (e.g., the horizontal dashed line). The uncertainty term is only a function of threshold distance, increasing from 0.03 to 0.12 with increasing distance. The differences in the BSSs among EPSs largely correspond to how much the resolution term cancels out the uncertainty term.

Using the OFCL cone as a reference to evaluate strike probabilities, the ECMWF and MMG again have skill scores that are positive and greater than 0.06 at all distances, but the difference from zero is not significant (Fig. 14a). The GEFS’s skill is negative and less than ~0.08, but is comparable to the reference, while the UKMET has significantly negative skill that is less than ~0.22 for all distances. From the reliability diagrams (Figs. 14b–d), it appears that the MMG is closest to the 1:1 slope, while the other EPSs again overforecast the number of hits, especially for smaller threshold distances. Conversely, the OFCL cone appears to underforecast hit probabilities of 60%–80%. The forecasts in which the OFCL cone predicts a 60% chance of a TC within 600 km are verified by the observations 72% of
the time (Fig. 14d). This corresponds to the OFCL cone having poorer reliability relative to the MMG. However, the larger deviation of the OFCL from the observed climatological frequency corresponds to a better resolution term, such that the Brier scores of the OFCL and MMG are comparable. The OFCL never forecasts probabilities ~100% in Figs. 14b and 14c because the 80th percentile of TTEs is >500 km (e.g., Fig. 3a). Even if the center of the OFCL cone is directly over a grid point, more than 20% of the climatology-based cone
area is more than 500 km from that point, such that the forecasted strike probability is less than 80%.

As demonstrated in Figs. 6 and 7, the ECdet is one of the best dynamical models in terms of 72–120-h deterministic intensity forecasts and is thus chosen as the reference for evaluating the EPS probability forecasts for intensity. Figure 15a shows the BSS evaluating the 72–120-h forecast probabilities of TCs reaching different wind speeds. The GEFS is significantly outperformed by the ECdet, showing especially poor skill for weak tropical storms (e.g., winds less than 25 m s$^{-1}$). The ECMWF and MMG outperform the ECdet for TCs near the lower and upper wind thresholds and are only comparable to the ECdet for intermediate thresholds. In the reliability diagrams (Figs. 15b–d), all of the models have unconditional “dry biases” in that their forecasted probabilities are consistently too small relative to all conditional observed frequencies. The GEFS in particular has the worst dry bias for forecasting winds exceeding 18 m s$^{-1}$ (Fig. 15b). The dry biases are less apparent for forecasts of weak hurricanes (Fig. 15d), but the GEFS still has poor resolution in that it overforecasts large probability events and underforecasts small probability events. However, the reliability of the ECdet (not shown) degrades faster than the GEFS for forecasting larger wind speeds, such that the BSS approaches zero. Therefore, the GEFS and ECdet have comparable difficulty in forecasting that a TC will reach or maintain category 2 status at 72–120 h.

5. Summary and conclusions

The track forecasts of the ECMWF, GEFS, and UKMET ensembles, as well as their combination (MMG), are verified and compared against several deterministic models for the 2008–15 North Atlantic tropical cyclone (TC) seasons. On average, the ECdet model shows the most deterministic accuracy in terms of mean total track errors (TTEs), though the differences from the ECMWF, GEFS, and MMG ensemble means are not statistically significant. The ECdet also most often produces the best-track forecast relative to all other models. However, the MMG ensemble mean never produces the worst track forecast at 120 h.
Previous studies have shown that the ECMWF model has the best deterministic scores based on domain-averaged errors in upper-level geopotential height (Bourke et al. 2004; Buizza et al. 2005) and wind and rainfall patterns (Kerns and Chen 2014). In relation to TC forecasts, the ECMWF’s extensive assimilation of satellite data using four-dimensional variational data assimilation (4DVAR) has been suggested to aid its superior typhoon track forecasts over the GEFS’s 3DVAR data assimilation (Weissmann et al. 2011). The importance of flow-dependent data assimilation coupled with high resolution was demonstrated by Munsell and Zhang (2014), in which their Weather Research and Forecasting (WRF) Model run with EnKF data assimilation and 3-km resolution produced track forecasts of Hurricane Sandy (2012) that were comparable to or better than those of the ECMWF. On the other hand, the success of the ECMWF track forecast for Sandy has also been attributed to the model convective parameterization (Bassill 2014). The role of model physics compared to initial conditions or model resolution in the ECMWF TC forecasts is beyond the scope of this paper.

All models have a significant slow bias in the along-track direction at 120 h. TCs undergoing extratropical transition (ET) contribute up to 40% to the magnitude of the slow bias. Many factors can limit the predictability of ET events in numerical models (Jones et al. 2003; Harr et al. 2008). The progression of the midlatitude system and the nature of its interaction with the TC may be coupled with the size and intensity of the TC. The systematic slow biases in the models may be from the upper-level flow erroneously decoupling from the lower-level flow (Carr and Elsberry 2000; Payne et al. 2007).

The 72- and 120-h TTEs of many models have improved, particularly between 2010 and 2012. However, the role of model upgrades during this time period is not conclusive. The limited number of ET cases per any one year limits an analysis from determining if the change in the slow bias over time corresponds to changes in ET events.

The intensity forecasts were also verified in terms of minimum central pressure and maximum wind speed. The higher-resolution regional models and the OFCL
forecast have the lowest MAEs in intensity within the first 24 h of lead time, while the global models have significantly large weak biases within 72 h. However, the intensity errors of the models become more comparable at lead times > 72 h. The benefit of an accurately initialized TC and the convection-resolving grids of the HWRF and GFDL may be offset by the track errors becoming large. That is, much of the intensity errors by 72 h may be from the model displacing the TC in a significantly more or less favorable environment than observed (e.g., DeMaria 2010; Emanuel and Zhang 2016). Consistent with this concept, the intensity errors also seem to improve with the track errors between 2010 and 2012.

EPS track dispersion is also evaluated in this study with different metrics. The PWSs of the EPSs suggest that the GEFS and UKMET tracks are underdispersed at all lead times, while the ECMWF and MMG show adequate dispersion, albeit the MMG may be somewhat overdispersed at times. Rank histograms reveal that much of this underdispersion at 72–120 h is in the along-track direction and corresponds to the slow bias. The histograms also suggest that the ECMWF and MMG are slightly overdispersed in the cross-track direction. Overall, with regard to ensemble dispersion, there appear to be some benefits to combining the EPSs into the MMG, though much of this appears to be contributed by the ECMWF.

Using BSSs to evaluate the EPS’s strike probabilities at fixed grid points along the U.S. North Atlantic coast, the ECMWF and MMG show superior probabilistic skill relative to the ECdet. The GEFS and UKMET only show comparable skill to the ECdet for larger threshold distances from the points, as a result of the ensembles overforecasting high strike probabilities. In comparison to the climatology-based OFCL cone forecast, the ECMWF and MMG show comparable skill. The relatively large full distribution of OFCL TTEs causes the corresponding cone to underforecast medium/high strike probabilities for smaller targets. From the reliability diagrams, the OFCL cone has poorer reliability but comparable or better resolution than the EPSs.

The BSS verification of 72–120-h forecast probabilities of wind intensity reveals that the ECMWF and
MMG have comparable or better skill than the ECdet intensity forecast. The GEFS has significantly poor skill when forecasting weak tropical storms, for which the ensemble has an unconditional underforecasting bias. The poor skill of the GEFS is consistent with its significant weak biases. However, the GEFS becomes more comparable to the ECdet when forecasting larger wind speed events, for which both models struggle similarly.

Future work will aim at better understanding large track error cases, such as Joaquin (2015), for which even a multimodel ensemble can have mean errors well above climatology. Individual case studies have attributed these poor track forecasts to uncertainties in TC structure (Brennan et al. 2015; Colby 2015), feedbacks of the TC diabatic outflow to the environment (Sun et al. 2015), and the initialization of synoptic-scale steering systems (Brennan and Majumdar 2011; Munsell et al. 2015). Ensemble-based tools will be used to analyze the contributions of these factors for different cases. The role of ETs will also be further assessed to determine if the development of the slow bias is preceded by a bias in the transition’s timing.

Acknowledgments. The authors thank the two anonymous reviewers for their thoughtful comments to help improve the manuscript. We also thank the National Science Foundation for financial support of this research under Award CMMI-1331269. We thank Dr. Michael Brennan of NHC for his input and for providing us with the link to the archive of the best tracks and forecast tracks of the deterministic models. We also thank NHC for maintaining this archive and the THORPEX TIGGE team for maintaining the archive of ensemble forecast tracks.

APPENDIX

Key Terms and Acronyms Used in This Study

- ATE: Along-track error
- BSS: Brier skill score
- CTE: Cross-track error
- EPS: Ensemble prediction system
- ET: Extratropical transition
- MAE: Mean absolute error
- MMG: Multimodel Global
- NHC: National Hurricane Center
- SLP: Sea level pressure
- PWS: Probability within spread
- TC: Tropical cyclone
- TTE: Total track error

REFERENCES


