On the Ability of Global Ensemble Prediction Systems to Predict Tropical Cyclone Track Probabilities

SHARANYA J. MAJUMDAR AND PETER M. FINOCCHIO

University of Miami, Miami, Florida

(Manuscript received 1 July 2009, in final form 19 November 2009)

ABSTRACT

The ability of ensemble prediction systems to predict the probability that a tropical cyclone will fall within a certain area is evaluated. Ensemble forecasts of up to 5 days issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Met Office (UKMET) were evaluated for the 2008 Atlantic and western North Pacific seasons. In the Atlantic, the ECMWF ensemble mean was comparable in skill to a consensus of deterministic models. Dynamic “probability circles” that contained 67% of the ECMWF ensemble captured the best track in 67% of all cases for 24–84-h forecasts, and were slightly underdispersive beyond 96 h. In contrast, the Goerss predicted consensus error (GPCE) was overdispersive. The addition of the UKMET ensemble yielded improvements in the short range and degradations for longer-range forecasts. The ECMWF ensemble performed similarly when the size was reduced from 50 to 20. On average, it produced a lower measure of independence between its members than an ensemble comprising different deterministic models. The 67% circles normally captured the best track during straight-line motion, but less so for sharply turning tracks. In contrast to the Atlantic, the ECMWF ensemble (and GPCE) was unable to capture sufficient verifications within the 67% probability circles in the western North Pacific, in part because of a less skillful ensemble mean (and consensus). Though further evaluations are necessary, the results demonstrate the potential for ensemble prediction systems to enhance probabilistic forecasts, and for The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) to be embraced by the operational and research communities.

1. Introduction

The most fundamental metric in tropical cyclone (TC) prediction is a storm’s location, or track. Improved short-range (1–3 day) track forecasts lead to fewer unnecessary warnings and evacuations. More accurate track forecasts at longer lead times (>3 days) will allow more time for emergency management to prepare and mobilize resources for disaster recovery. Additionally, a better track forecast will yield improved forecasts of other metrics such as wind speed, storm surge, and precipitation.

This paper focuses on uncertainty prediction, which is required given that deterministic forecasts yield limited information due to the inherent uncertainty in TC forecasts. Operational centers such as the National Oceanic and Atmospheric Administration’s (NOAA) National Hurricane Center (NHC) commonly compute probabilities that are centered on their official forecasts, and based on error distributions over recent years (Rappaport et al. 2009). These probabilities are presently static, in that there is no forecaster input, and as such the probabilities do not vary from case to case and thereby ignore the predictability of the flow. But given that the uncertainty in a forecast is dependent on the atmospheric conditions of that forecast, dynamic predictions of uncertainty are expected to be superior. For example, in a high-predictability scenario such as a TC propagating westward for days along the southern periphery of a subtropical ridge, the uncertainty (and therefore warning area) may be considerably smaller than a low-predictability scenario such as a TC that recurves close to the coastline. Prediction of forecast uncertainty is one of the primary components of the Hurricane Forecast Improvement Project (HFIP), established by NOAA in 2007 (information available online at http://www.nrc.noaa.gov/plans_docs/HFIP_Plan_073108.pdf).

Ensemble forecasts have been widely proposed as a method of producing quantitative predictions of uncertainty. Until recently, ensembles of global models have
primarily been applied to TC forecasting to determine a consensus forecast of up to 5 days, in which an average of several different models has shown demonstrable improvement over selecting the best-performing single model (Rappaport et al. 2009). Another method of consensus prediction is via a weighting of the component model predictions based on past history (Krishnamurti et al. 2000; Williford et al. 2003; Weber 2003). It is generally accepted that increasing the number of forecasts and the variety of models in the ensemble will yield an improved consensus track forecast, via a reduction in the variance of the consensus track errors (Goerss 2000; Goerss et al. 2004; Sampson et al. 2006). The variance changes depending on the effective degrees of freedom of the models included in the consensus forecast. The relationship between the variance and the errors of a consensus of dynamical models has been explored by Goerss (2000) and Elsberry and Carr (2000). Further efforts to produce quantitative predictions of track forecast uncertainty include those of Weber (2003, 2005), who extended his statistical track prediction system by adding a Gaussian distribution of strike probabilities to a statistically weighted track. These results were installed on the Automated Tropical Cyclone Forecast (ATCF) at the Joint Typhoon Warning Center (JTWC; see Sampson and Schrader 2000). Goerss (2007) developed a regression-based technique to predict the error of consensus forecasts with multiple models, based on a range of predictors including model spread, initial and forecast TC tracks, intensity, and motion speed. Using this technique, state-dependent estimates of the track error were created via what is now known as the Goerss predicted consensus error (GPCE). The GPCE product, and the multimodel consensus forecast upon which it is centered, have been operational on the ATCF at NHC since 2005 and at the JTWC since 2004. The GPCE technique is currently being incorporated into new wind probability products (expanding upon DeMaria et al. 2009) and is also being extended to along- and cross-track errors to produce elliptical probability areas (J. Goerss 2009, personal communication).

Ensembles based on multiple integrations of a single global model (commonly known as ensemble prediction systems or EPSs) have been operational since the early 1990s. TC strike probabilities from the European Centre for Medium-Range Weather Forecasts (ECMWF) EPS are provided on their Web site, and decisions on mission planning around typhoons during The Observing System Research and Predictability Experiment (THORPEX) Pacific Asian Regional Campaign (T-PARC) were based in part on the uncertainty exhibited in EPS track forecasts from multiple centers. Operational planning of Atlantic synoptic surveillance missions utilizes the variance of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) ensemble in the synoptic environment of the TC (Aberson 2010). However, to the authors’ knowledge, a formal evaluation of operational global EPS for TC track uncertainty prediction has not been published prior to the study presented in this paper.

To prepare its “cone of uncertainty,” the NHC uses the 67th percentile of its cumulative track forecast errors over the previous 5 yr to define the radius of a circle centered on its official forecast track for each time. The cone is then constructed by a smooth curve that traverses the radii of the circles from 0 to 5 days (Franklin 2009). We emphasize again that the size of the circles is fixed for each hurricane season and, therefore, does not vary from forecast to forecast. In this paper, we evaluate the ability of the ECMWF and the Met Office (UKMET) EPS, and a combination of the two, to produce dynamic forecasts of track probabilities out to 5 days that are comparable in principle to the circles that are used to construct the cone of uncertainty. Unlike NHC or GPCE, our EPS-based probabilistic forecasts do not contain any dependence on statistics from previous seasons. Our primary motivation is to evaluate whether global model ensembles are ready for use in the probabilistic prediction of the track, via simple metrics and its comparison versus GPCE, which represents the present state of the art. A satisfactory evaluation would raise the potential for the introduction of EPS into products such as surface wind speed probabilities, predicted consensus errors (Goerss 2007), along- and cross-track probabilities, and probabilistic predictions of threshold warning areas.

At the Sixth World Meteorological Organization (WMO) International Workshop on Tropical Cyclones (IWTC-VI), it was recommended that all TC-related numerical weather prediction products be made available to all operational and research users in real time (IWTC-VI recommendations are documented online: http://severe.worldweather.org/iwtc/). In 2008, various TC forecast parameters from EPS were disseminated daily by multiple operational centers and uploaded onto the THORPEX Interactive Grand Global Ensemble (TIGGE) database in a unified format, known as Cyclone XML or CXML. For the first time, this database has offered an opportunity to the community to access operational data from multiple centers without the overhead of signing agreements and writing extensive software to read data in multiple formats. This paper is organized as follows: the data and methods are described in section 2, followed by illustrations of the EPS forecasts and probability circles in section 3. The evaluation of the ensemble means, EPS probability circles, and GPCE is presented in section 4, followed by concluding remarks in section 5.
2. Data and methods

a. Ensemble prediction systems (EPSs)

The TIGGE database, established in 2008, provides ensemble data from multiple operational centers in near-real time, in a standardized, easy-to-read CXML format (the TIGGE TC database is available online: http://www.bom.gov.au/bmrc/projects/THORPEX/CXML/index.html). The format allows for a large number of storm parameters, including position, pressure, maximum wind speed, and location of maximum winds. In 2008, the most consistently available and reliable ensemble data in TIGGE were from ECMWF and UKMET.

The ECMWF EPS (Buizza et al. 2007) comprises 50 perturbed members initialized at 0000 and 1200 UTC every day, at T1399 (~50 km) resolution out to day 10 (and at T1255 resolution from days 10 to 15). All integrations have 62 levels. Model errors due to physical processes and subgrid-scale effects are represented in the ensemble by stochastic perturbations of tendencies of physical processes. Initial condition perturbations are constructed using singular vectors (SVs), which represent the fastest linear growing error structures over a 48-h period. In addition to the 50 routine hemispheric SVs that use a dry total energy norm (Buizza and Palmer 1995), 5 extra SVs with a diabatic version of the tangent-linear model and moist total energy norm are computed in up to six subspaces enclosing tropical cyclones, at T42 resolution and with 42 vertical levels (Barkmeijer et al. 2001; Puri et al. 2001).

The UKMET ensemble (known as the global version of MOGREPS) comprises 24 members integrated at 0000 and 1200 UTC each day, at ~90 km resolution with 38 levels. The ensemble is integrated out 15 days. The local ensemble transform Kalman filter (LETKF) is used to initialize the ensemble, with model uncertainty perturbations prescribed via a stochastic kinetic energy backscatter scheme (Bowler et al. 2009). A combined 74-member ECMWF + UKMET ensemble is also employed in this paper, in order to quantify the effects of using a larger ensemble with multiple models.

The NCEP Global Ensemble Forecast System comprises 20 members integrated 4 times per day at T126 L28 (~90 km) resolution, out to 16 days. Unlike ECMWF and UKMET, a tropical cyclone relocation procedure is used. An ensemble transform method is employed to create initial perturbations (Wei et al. 2008). While not included in 2008, stochastic model perturbations will be created in the operational ensemble in the near future (together with a resolution upgrade to T190 L28). In this study, the NCEP ensemble tracks were lost more often than those in ECMWF and UKMET and therefore the NCEP ensemble is only used here for illustrative purposes in section 3.

All cases that were available at 0000 and 1200 UTC in the TIGGE database for the 2008 season were included. We do not claim statistical independence between the cases, given that a separation of at least 30 h between forecasts of the same storm has been recommended in order to perform statistical significance tests (DeMaria et al. 1992). A few TC cases were omitted due to a lack of data.

b. Mean and probability circles

In this paper, we assume that the ensemble mean represents the most likely forecast scenario, and the center of the distribution. Counterexamples such as a bimodal distribution can easily be identified. For each forecast time, the ensemble mean is simply computed as the average of the respective track forecasts in the particular ensemble (ECMWF, UKMET, or ECMWF+UKMET). The mean is only used if at least two-thirds of each ensemble exists at that time, consistent with the procedure used at NOAA to compute the ensemble mean (T. Marchok 2008, personal communication).

In addition to assuming that the mean represents the center of the distribution of all feasible track forecasts, we also assume that the distribution of track probabilities is independent of direction (isotropic) about the ensemble mean, and therefore along-track and cross-track errors in any direction are equally likely. This latter assumption is consistent with that implicit in NHC’s cone of uncertainty, although NHC constructs its cone about the official NHC forecast and not an ensemble mean. We decided against constructing probabilities centered on the NHC or a consensus model forecast point, given that the distribution of ensemble members would not be centered about those points (or may even lie entirely outside the forecast point), thereby rendering any probability distribution about the point highly anisotropic.

We define the radius of an X% “probability circle” about the ensemble mean at a particular time as the radius that encloses X% of the ensemble forecasts valid at that time. If the probability circle is accurate, one would expect the best track to fall within the probability circle in X% of all realizations. The NHC uses a series of 67% probability circles at each forecast time to construct its cone of uncertainty, based on 5 yr of their forecast errors for that time interval. In addition to computing 67% the probabilities, we also explore circles that encapsulate 33%, 50%, and 100% of the ensemble forecasts. An ensemble is said to be “underdispersive” if the probability circle is too narrow, while the ensemble is “over-dispersive” if the circle is too broad. However, it is not always clear whether an ensemble is underdispersive, given that the ensemble mean may be in large error, which leads to the best track being shifted outside the
circle. Finally, we emphasize here that we will strictly be evaluating whether the best track lies within the *probability circle appropriate for that particular discrete forecast time*, as opposed to whether the track lies within a cone that is an integral of the circles over time and fills in the area connecting the tangents to the time series of circles.

c. Deterministic forecasts

The competitiveness of the ensemble mean forecasts is assessed by evaluating their skill against operational models used by forecasters at NHC. Global models include the ECMWF and UKMET deterministic models, plus the NCEP GFS and the Navy Operational Global Atmospheric Prediction System (NOGAPS) model. The two regional models, the Geophysical Fluid Dynamics Laboratory (GFDL) model and the Hurricane Weather Research and Forecasting (HWRF) model, run at NCEP using GFS boundary conditions, are also included. All of the deterministic models are initialized at the same time as the ensembles, and are referred to as “late cycle” forecasts by NHC. Note that NHC uses a time-interpolated version of these forecasts, instead of the raw forecast products evaluated in this study. The ensemble mean forecasts are also compared against the climatology and persistence (CLIPER) model, which serves as a “no skill” benchmark for forecasts of length up to 5 days (Aberson 1998; Aberson and Sampson 2003).

d. Consensus forecasts and GPCE

One of the consensus track forecasting products used by NHC in 2008 is TVCN, which represents the average of at least two of the following seven deterministic models: GFS, ECMWF, GFDL, the GFDL model nested within NOGAPS (GFDN), HWRF, NOGAPS, and UKMET. Unlike the late-cycle deterministic forecasts outlined in section 2c above, the products that form TVCN are interpolated from forecasts initialized at the previous assimilation time and are, therefore, “early cycle” forecasts. The corresponding consensus forecast aid used by the Joint Typhoon Warning Center for the western North Pacific basin is CONW, which comprises at least two of the following interpolated models: GFS, ECMWF, NOGAPS, UKMET, GFDN, the Australian regional model, the Japan Meteorological Agency (JMA) Global Spectral Model, and a locally run barotropic model (Weber 2001; Sampson et al. 2006). As with TVCN, CONW is an early cycle prediction.

GPCE is based on several predictors that are available prior to the forecast: consensus model spread, initial and forecast TC intensities, initial and forecast TC positions, motion speed of the TC, and the number of available members in the consensus model. Using this pool of predictors, a stepwise linear regression technique is combined with additive values (that vary with time) to provide GPCE radii that contained the consensus forecast position approximately two-thirds of the time, based on recent historical forecasts. The GPCE circles with these radii are centered on the TVCN and CONW forecasts for the Atlantic and western North Pacific basins, respectively, and are therefore early cycle predictions. For further details on the methodology of GPCE, the reader is referred to Goerres (2007).

e. Verification

The verifications are conducted using the postseason best tracks produced by the NHC for the Atlantic Basin, and the JMA for the western North Pacific basin. The ensembles are evaluated for TCs that verified at tropical storm strength (34 kt) or higher over the 2008 Atlantic and western North Pacific seasons. The track error is defined as the great circle distance between the forecast and best-track position of the tropical cyclone center. Cases in which the cyclone was officially verified as being a depression, extratropical, or as a remnant low by the operational center are omitted. A few cases in which the tropical cyclone was only a weak/small tropical storm were not captured by many ensemble members and, thus, are omitted. A summary of the sample is given in Table 1 (Atlantic) (see also Table 3 for the western North Pacific). The full set of cases is evaluated using the ECMWF ensemble. A reduced homogeneous sample of cases is used when comparing the ECMWF, UKMET, and combined ECMWF + UKMET ensembles.

3. Example: Hurricane Ike

The concepts of ensemble track forecast scatter, ensemble means, and circles of uncertainty are illustrated using two challenging forecasts during the life cycle of Hurricane Ike (2008). Ike was a long-lived tropical cyclone that developed from an African easterly wave on 1 September 2008 and had subsequently intensified rapidly into a major hurricane with maximum sustained winds of 115 kt (category 4 on the Saffir–Simpson scale) by 0000 UTC 5 September 2008. At this time, Ike was moving westward under the influence of the mid- to upper-tropospheric Atlantic subtropical ridge. The primary challenge in the track forecast was to determine whether the ridge would weaken, taking Ike toward the west-northwest and resulting in the eventual recurvature or landfall on the eastern seaboard of the United States, or if the ridge would build further, resulting in a south-westward track toward Cuba. The scatter of tracks from the ECMWF, UKMET, and NCEP ensembles illustrates this broad range of possibilities in the 3–5-day forecast.
In reality, Ike took the track toward Cuba, due to the ridge persisting over the western Atlantic. A minority of ECMWF and NCEP ensemble members captured this track, indicating that this was a low-probability track according to the ensembles, with most ensemble members signifying that the ridge would weaken considerably. The ensemble mean forecasts are accordingly all situated to the north of the actual track, with no obvious forecast gain over the corresponding deterministic forecasts, of which the GFDL model possessed the highest skill (Fig. 1d). The NCEP ensemble produced a superior ensemble mean track to those of ECMWF and UKMET. However, at least half of the NCEP ensemble members were dissipated within 3 days, rendering the ensemble of limited use even for an intense hurricane.

The second example illustrates another difficult track forecast for Ike initialized 4 days later, on 0000 UTC 9 September 2008 (Figs. 1e–h). At this time, Ike was traveling toward the west-northwest, with an inevitable landfall on the coast of the Gulf of Mexico within 3–5 days. The forecast scenario was governed by two contrasting synoptic patterns: a weak subtropical ridge in the northeastern Gulf of Mexico, possibly gaining strength and inducing a more westward motion of Ike, and a shortwave trough situated over the Rocky Mountains that was predicted to move southward and induce a right turn. Most of the deterministic models initialized at 0000 UTC 9 September 2008, including the deterministic UKMET (UKX), downplayed this trough interaction and predicted a westward track and landfall (Fig. 1h). And while the deterministic ECMWF (EMX) and HWRF indicated a turn toward the north, the landfall location was still west of the verification. Strikingly, the ECMWF ensemble mean was on the best-track trajectory out to 5 days, although its 4–5-day forecast was lagging by about 12 h. And contrary to the corresponding deterministic forecast, the UKMET ensemble included several members that exhibited a pronounced northward component, with its resulting mean being the outlier to the right of the best track (Fig. 1f). Similarly, the NCEP ensemble (Fig. 1g) was shifted to the right of its deterministic counterpart. The reasoning for these significant differences between deterministic and ensemble forecasts from the same operational center requires a detailed investigation, which is beyond the scope of this paper. In conclusion, we speculate that the global model EPS had the potential to add

Table 1. List of Atlantic tropical storm (TS) and hurricane (H) cases for which UKMET and ECMWF ensemble data exist, for each tropical cyclone. The last 4 columns give the ratio of the number of cases in which the best track is contained within the ECMWF 67% probability circles (E) and GPCE circles (G) to the total number of cases.

<table>
<thead>
<tr>
<th>No.</th>
<th>Storm name</th>
<th>Dates</th>
<th>UKMET</th>
<th>ECMWF</th>
<th>48 h</th>
<th>72 h</th>
<th>96 h</th>
<th>120 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>03L</td>
<td>TS Cristobal</td>
<td>19–23 Jul</td>
<td>8</td>
<td>8</td>
<td>E 2/3</td>
<td>G 3/3</td>
<td>E 1/4</td>
<td>G 4/4</td>
</tr>
<tr>
<td>04L</td>
<td>H Dolly</td>
<td>21–24 Jul</td>
<td>8</td>
<td>8</td>
<td>E 1/4</td>
<td>G 4/4</td>
<td>E 1/2</td>
<td></td>
</tr>
<tr>
<td>05L</td>
<td>TS Edouard</td>
<td>4–5 Aug</td>
<td>4</td>
<td>4</td>
<td>E 12/12</td>
<td>G 12/12</td>
<td>E 10/10</td>
<td>G 10/10</td>
</tr>
<tr>
<td>06L</td>
<td>TS Fay</td>
<td>16–23 Aug</td>
<td>16</td>
<td>16</td>
<td>E 7/11</td>
<td>G 10/11</td>
<td>E 4/9</td>
<td>G 7/7</td>
</tr>
<tr>
<td>08L</td>
<td>H Hanna</td>
<td>28 Aug–7 Sep</td>
<td>20</td>
<td>20</td>
<td>E 18/22</td>
<td>G 21/22</td>
<td>E 15/20</td>
<td>G 18/20</td>
</tr>
<tr>
<td>09L</td>
<td>H Ike</td>
<td>1–14 Sep</td>
<td>26</td>
<td>26</td>
<td>E 3/3</td>
<td>G 0/3</td>
<td>E 1/1</td>
<td></td>
</tr>
<tr>
<td>10L</td>
<td>TS Josephine</td>
<td>2–5 Sep</td>
<td>7</td>
<td>7</td>
<td>E 0/3</td>
<td>G 0/3</td>
<td>E 0/1</td>
<td></td>
</tr>
<tr>
<td>11L</td>
<td>H Kyle</td>
<td>26–29 Sep</td>
<td>5</td>
<td>7</td>
<td>E 2/3</td>
<td>G 3/3</td>
<td>E 1/1</td>
<td></td>
</tr>
<tr>
<td>12L</td>
<td>TS Laura</td>
<td>30 Sep–1 Oct</td>
<td>3</td>
<td>3</td>
<td>G 2/5</td>
<td>G 0/3</td>
<td>G 0/1</td>
<td></td>
</tr>
<tr>
<td>14L</td>
<td>TS Nana</td>
<td>13 Oct</td>
<td>1</td>
<td>1</td>
<td>G 2/3</td>
<td>G 1/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15L</td>
<td>H Omar</td>
<td>14–18 Oct</td>
<td>2</td>
<td>9</td>
<td>E 0/3</td>
<td>G 2/3</td>
<td>E 0/1</td>
<td>G 1/1</td>
</tr>
<tr>
<td>17L</td>
<td>H Paloma</td>
<td>6–9 Nov</td>
<td>7</td>
<td>7</td>
<td>G 2/3</td>
<td>G 1/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>153</td>
<td>165</td>
<td>E 68%</td>
<td>G 82%</td>
<td>E 64%</td>
<td>G 82%</td>
</tr>
</tbody>
</table>
FIG. 1. The 5-day ensemble forecasts for Hurricane Ike, initialized (left) at 0000 UTC 5 Sep and (right) at 0000 UTC 9 Sep 2008. (a)–(c) and (e)–(g) Ensemble track forecasts for ECMWF, UKMET, and NCEP, respectively. The length of each forecast in days is depicted by the integers 1–5. (d),(h) Boldface type indicates the respective ensemble mean forecasts. Lightface type shows the deterministic model forecasts initialized at the same time. The NHC best track is given by the line connecting the hurricane symbols, plotted at 0000 UTC on each day.
significant benefit to the landfall forecast of Ike at this particular time.

It is also evident from Fig. 1 that the distribution of the ensemble forecasts can be highly anisotropic. For example, the ECMWF ensemble predicted a wider distribution of cross-track uncertainty than along-track uncertainty at 3–5 days (Figs. 1a and 1e). In this paper, we assume that the most probable distribution of track forecasts is isotropic, consistent with the 2008 version of GPCE and the NHC cone of uncertainty. The construction of anisotropic ensemble-based estimates of uncertainty is left to future work.

Next, the isotropic estimates of uncertainty based on the ECMWF ensemble prediction system, GPCE, and NHC (OFCL) are compared in Fig. 2. For the first forecast of Ike initialized at 0000 UTC 5 September 2008, the 67% circle of uncertainty based purely on the ECMWF ensemble is situated to the north of the best track out to 3 days, suggesting that the probability of the best track verifying would be at most one-third (Fig. 2a). The best track lies within the probability circle at 4 and 5 days. In contrast, the GPCE and NHC circles are significantly broader than their ECMWF counterparts, capturing the best track on most occasions for 2–5-day forecasts (Figs. 2b and 2c). For the second forecast of Ike initialized at 0000 UTC 9 September 2008, the ECMWF ensemble provides a 67% circle of uncertainty centered on an accurate ensemble mean track (Fig. 2d). The radii of the ECMWF circles are larger than those in Fig. 2a, due to a higher spread of tracks in Fig. 1e compared against Fig. 1a. The GPCE and NHC circles are again significantly broader than those based on the ECMWF ensemble, and are shifted toward the west in accordance with the TVCN and the NHC’s official track forecast (Figs. 2e and 2f). The GPCE circles in Fig. 2e are slightly broader than those in Fig. 2b, reflecting a larger spread in the deterministic models used in TVCN for the later forecast. By definition, the radii of the NHC circles are identical in each case (Figs. 2c and 2f).

4. Evaluation

The qualitative presentation in section 3 illustrates the potential for probabilistic track forecasts using dynamic ensemble-based circles. Given that the probability circles are constructed about the ensemble mean track, it is of paramount importance for the ensemble mean to have skill. In this section, the mean of the ECMWF ensemble is first evaluated for the 2008 Atlantic season, followed by the equivalent evaluation with the UKMET ensemble added, and a subsequent evaluation of the probability circles in comparison to those provided using GPCE. Finally, the same evaluation is performed in the western North Pacific basin for the ECMWF ensemble and GPCE.

a. Ensemble mean

1) ECMWF ENSEMBLE

The ECMWF ensemble is first chosen since it possesses the largest sample size at longer forecast times, due to the highest number of available cases and a relative lack of premature dissipation of the TC compared with other global model EPSs. The first stage is to evaluate whether the ensemble mean is competitive with the corresponding deterministic and consensus forecasts. These forecasts, initialized at 0000 and 1200 UTC are evaluated here, for a homogeneous sample of cases in which at least 33 ECMWF ensemble members and all of the global and regional deterministic forecasts exist. The number of cases at the initial time is 159, dropping to 56 cases at 5 days. It is evident from Fig. 3 that the ensemble mean (EEMN) performed better in 2008 than all of the deterministic models except for the high-resolution ECMWF (EMX). Furthermore, the EEMN forecast was of comparable skill to the TVCN forecast, although it is important to note that the TVCN forecast comprised interpolated deterministic model integrations that were initialized 6–12 h earlier than were those in the ECMWF ensemble. All of the models were far superior to CLIPER over 5 days (not shown).

Further insights into the track forecast errors are gained via the cumulative distribution functions (CDFs) of the errors. For each model, the number of forecasts in each bin with an interval of 50-km error (0–50, 50–100 km, etc.) is counted, and the CDF is then deduced as a percentage of all cases. For example, for 3-day forecasts, the 80th percentile of errors in the ECMWF ensemble mean is around 250 km, while those of the other global and regional models (except EMX), and the early cycle TVCN, are at least 300 km (Fig. 4a). Similar results are found for 2–5-day forecasts. Additionally, the ECMWF ensemble mean possesses a smaller number of high-error (>400 km) 5-day forecasts than TVCN and all the late-cycle models (including EMX; see Fig. 4b).

In summary, the ECMWF ensemble mean was found to be comparable in skill to TVCN, and superior to all global models except its deterministic counterpart. It was also superior to the regional models, GFDL and HWRF, in 2008. Given that the ECMWF ensemble mean exhibited an acceptable level of skill, we suggest that it is appropriate to construct probabilistic forecasts centered on the mean values.

2) COMBINED ECMWF AND UKMET ENSEMBLE

An ensemble from a single model may contain biases inherent to that model, which precludes a realistic
distribution across the entire space of possibilities from being spanned. It is therefore appealing to use more than one global EPS in TC track prediction, as in Fig. 5, where the ECMWF tracks for Hurricane Gustav lie predominantly to the right of the best track, the UKMET tracks are mostly to the left of the best track, and the NCEP members take a pronounced westward turn away from the best track. The best track would therefore lie on...
the edges of the respective probability density functions (pdfs) for each individual EPS, while it would be closer to the center of the pdf produced from a combined ensemble and, therefore, be interpreted as a more likely solution.

The UKMET ensemble is added to the ECMWF ensemble, producing a combined ensemble of 74 members. The number of cases is reduced from 165 to 153 at the initial time to produce a homogeneous sample. At later times, the UKMET ensemble members dissipate the TC more readily than ECMWF, thereby providing sixty 5-day forecasts for ECMWF by itself, but only thirty-one 5-day forecasts for ECMWF + UKMET (Fig. 6a). For this reduced sample, EEMN outperforms the UKMET ensemble mean (UEMN), with an average gain in lead time of 12–18 h over UEMN for forecasts longer than 2 days (Fig. 6b). UEMN is still competitive with the deterministic models presented in Fig. 3. The track errors in the mean of the combined ensemble (GEMN) are moderately smaller on average than those of the individual ensembles for forecasts out to 1 day, and are comparable but not superior to the ECMWF ensemble mean. The ECMWF ensemble mean over the full sample of 165 cases possesses a larger average error than that for the homogeneous sample, due to the fact that 7 of the extra 12 cases are for Hurricane Omar, which produced some of the largest track errors of the 2008 season. Finally, the CDF for a homogeneous sample of seventy-two 3-day forecasts demonstrates that the addition of the UKMET ensemble to the ECMWF ensemble helped produce more ensemble mean forecasts of very low error (Fig. 6c). For example, the 20th percentile error of the combined ensemble mean was 35 km, while the 20th percentile error of the ECMWF ensemble mean was over 60 km. However, the addition of the UKMET ensemble did serve to produce more large errors in the combined ensemble (80th percentile = 295 km) than if the ECMWF ensemble were solely used (80th percentile = 240 km). It is worth noting that this 80th percentile value for the combined ensemble mean is lower than that found for each of the deterministic models (>325 km) in Fig. 4b. Similar results were found for 4- and 5-day forecasts (not shown).

b. Probability circles

1) ECMWF AND UKMET ENSEMBLES

Attention is now turned to the dynamic probability circles as constructed in Figs. 2a and 2d. For the homogeneous sample, it is particularly encouraging that the best track lies within the ECMWF 67% probability circle close to 67% of the time, for forecasts of 1–3½ days, while being slightly underdispersive (60%–62%) for 4–5-day forecasts (Fig. 7c). Unlike for ECMWF, the best track is situated within the UKMET 67% circles too infrequently (40%–60%). However, it cannot be concluded that the UKMET ensemble is underdispersive. Instead, part of the reasoning for the poor probabilistic
prediction may be attributed to inferior forecasts of the ensemble and therefore its mean (UEMN) about which the probability circles are centered, thereby leading to the best track falling outside the circles more often than desired. The 67% probability circles constructed using the combined ECMWF + UKMET ensemble encapsulated the best track between 0% and 15% less than the desired 67% of all cases. Probability circles of different sizes were also evaluated. For the 33% circle, the combined ensemble was slightly overdispersive at most times, while both individual ensembles were closer to the 33% mark (Fig. 7a). For the 50% circles, the combined ensemble was superior, with the UKMET not capturing the best track often enough, and the ECMWF varying between slightly overdispersive (1.5–3 days) and then slightly underdispersive (3.5–5 days) (Fig. 7b). Finally, the circle whose radius is equal to that of the outermost ensemble member is expected to contain the best track $100 \times N/(N + 1)$% of the time, for an $N$-member ensemble (i.e., approximately 98% for the ECMWF ensemble). For the ECMWF and combined ensembles, this mark is approached for >1 day forecasts, suggesting that only on
a few occasions does the best track fall outside the entire ensemble (Fig. 7d). In contrast, the best track falls outside the entire UKMET ensemble around 20% of all cases at short lead times, and up to 40% of all cases for 5-day forecasts. As a side note, the ECMWF ensemble is consistently underdispersive at 12 h, as is also evident in the tiny circles in Figs. 2a and 2d. This is most likely due to the small initial perturbations constructed using moist singular vectors, which are conditioned to grow rapidly over the first 48 h of the forecast. In contrast, the UKMET ensemble likely has a larger spread of tracks at the earliest times.

A breakdown of the number of times the best track falls within the ECMWF 67% probability circle is listed for each Atlantic basin storm in the last four columns of Table 1 (numbers preceded by an E). The statistics are dominated by five storms (Bertha, Fay, Gustav, Hanna, and Ike), each of which exhibits distinct trends in the ensemble’s ability to capture the best track. For Bertha, the majority of the ECMWF ensemble members systematically failed to capture the turn toward the northwest 4 days into the storm’s existence, instead predicting a more westward path (Fig. 8a). (This case also serves to illustrate that the size of the circles does not necessarily increase with forecast time.) The subsequent recurvature and unusual southward turn of Bertha between 30° and 35°N were also not captured by most ensemble members. This is in contrast to the GPCE circles, which captured the best track in nearly all 4–5-day forecasts (Table 1, numbers preceded by a G). The same was true for the equally challenging sharp southward dip and loop of Hanna (Fig. 8c). The ECMWF ensemble also barely captured the best track for 3–5-day forecasts of Hurricane Gustav, due to several erroneous westward tracks toward the Yucatan Peninsula in the early stages, and slightly too slow forward motion later on in the Gulf of Mexico and erroneously low uncertainty predicted by the ensemble (Fig. 5). In contrast, for Tropical Storm Fay, the best track verified within the circles for nearly every case. For example, as Fay was approaching Florida, the 5-day ECMWF ensemble produced an accurate mean cross-track position and a circle spanning a broad range of uncertainty, including Fay’s abrupt westward turn across northern Florida (Fig. 8b). For the latter stages of Hurricane Ike in the Gulf of Mexico, the best track also fell within the ECMWF probability circles (Fig. 2d). Unsurprisingly, the cases in which the best track commonly fell outside the circles often coincided with a significant change in the track of the storm. For near-straight-line motion, the ensemble was better able to capture the best track, except in cases of large along-track error.

A summary of the ability of the respective ensembles’ X% probability circles to capture the best track X% of
FIG. 6. (a) Number of cases for the ECMWF ensemble only in the Atlantic, a homogeneous sample of ECMWF and UKMET ensemble data in Atlantic, and the ECMWF ensemble in the western North Pacific. (b) Average error of the ensemble mean track forecasts in the Atlantic, for the homogeneous sample (153 cases) and for the full set of ECMWF data (165 cases). (c) CDF of the ensemble mean 3-day track forecast errors for the same ensembles as in (b).
The time is presented in Table 2, for three forecast time ranges: 12–36, 48–84, and 96–120 h. The classification of “over,” “well,” and “under,” respectively, refers to whether the best track resided inside the circle too often, about the right frequency (within 5% of X% for the majority of the time range), or too rarely. The main result is that the ECMWF ensemble was able to predict the probabilities well between 48 and 84 h, and it was only slightly underdispersive at 96–120 h. In contrast, UKMET underpredicted the higher probabilities at all times. The combined ensemble performed similarly to ECMWF, except for the 12–36-h category where the addition of the UKMET ensemble to ECMWF increased the spread in the tracks, leading to a more appropriate probability circle than ECMWF produced alone.

The average radii of the 67% probability circles are illustrated in Fig. 9. For the ECMWF and combined ensembles, the radii increase nearly linearly with time, up to 350 km for 5-day forecasts (Fig. 9a). In contrast, the average radii are considerably larger for the UKMET ensemble out to 4 days, thereby dispelling the notion that UKMET may be underdispersive (i.e., that the circles are too narrow). Given that part of the overarching goal of advancing hurricane prediction is to reduce the uncertainty associated with warnings and evacuations, it is prudent to compare the sizes of the 67% circles predicted using the ensembles with those produced by an operational center. To provide a fair comparison with the 5-yr average employed by NHC, the ensemble forecasts that may be available at the time that NHC prepares a forecast are considered. A 12-h time lag is chosen here, such that a 60-h-old ensemble forecast is compared against the NHC radii for 48 h (Fig. 9b). Using this estimation, the radius of the 67% circle for ECMWF and the combined ensemble is very similar to that of NHC up to 3 days. Beyond 3 days, the average ensemble-based radii are smaller than those of NHC by 50 km, although this statement lacks substance since the ensembles are slightly underdispersive at these forecast times.

The correlation between the dynamic uncertainty explained by the 67% probability circles and the corresponding ensemble mean forecast error is now explored.
A low predicted uncertainty (small radius) implies that it should be very unlikely to produce a high forecast error. A high predicted uncertainty (large radius) is expected to exist in cases of high forecast error. However, a high predicted uncertainty does not preclude the existence of cases of low forecast error. In Fig. 9c, the predicted 4-day forecast uncertainty based on the ECMWF ensemble is well correlated with the 4-day forecast error of the ECMWF ensemble mean over all 165 cases, as indicated by a Pearson product moment correlation coefficient of $r = 0.54$. There are no cases of low predicted uncertainty and high forecast error (>400 km). On the other hand, those cases of high forecast error (such as the outlier from the only 96-h forecast of Hurricane Omar, with error 1084 km) did have a relatively high predicted uncertainty. Overall, the $r$ values were above 0.3 for the full sample of 165 ECMWF cases for 1–5-day forecasts (Fig. 8d). However, the sensitivity of this result to the inclusion of the cases that were not in the homogeneous sample is evident, with the corresponding $r$ values dropping below 0.3 for the sample of 153 ECMWF cases for which UKMET data also existed. In contrast to the ECMWF, the UKMET ensemble uncertainty explained by the 67% radii was not correlated to the forecast.
errors of its ensemble mean, with values of $r$ around zero for all forecast times. In summary, it is encouraging that the dynamic uncertainty based on the ECMWF ensemble may possess value in discriminating a priori between cases of high and low forecast uncertainty.

2) RANGE OF ECMWF AND GPCE RADII

The characteristics of the 67% probability circles derived from the best-performing ensemble, the ECMWF, are now compared against GPCE over the full 165-case sample. The relative size ranges of their respective probability circles over this sample are illustrated in Fig. 10. It is evident that the median and mean radii of the GPCE circles are considerably larger than those constructed using ECMWF, consistent with the illustrations in Fig. 2, which serve as a typical example. The range of ECMWF radius values over all cases is usually broader than that of GPCE, between the extrema and also between the 25th and 75th percentiles. The fixed NHC radii are close to the GPCE mean and median values, well within the 25th–75th percentiles for all forecast times. In contrast, the NHC radii are larger than the 75th percentiles of radii constructed using the ECMWF ensemble, for all forecast times, although it is worth recalling that the NHC radii are based on historical early cycle forecasts whose errors were on average larger than those in 2008. It is also evident that a few very large 67% probability circles are constructed in cases of high uncertainty in the ECMWF ensemble, for forecasts of 3 days or longer. One example is the outlying 750-km radius for a 4-day forecast of Hurricane Omar in Fig. 9c.

3) INDEPENDENCE OF ENSEMBLE MEMBERS

A measure of the independence of the track forecast errors of the individual ensemble members is achieved by first computing the “effective degrees of freedom,” which is the square of the ratio of the average track error of the ensemble members to the track error of the ensemble mean (Goerss 2000; Sampson et al. 2006). This value, averaged over the season sample, is considerably less than the number of members in the ensemble. The value for ECMWF is smallest (1.7) for 12-h forecasts, and it steadily increases for longer-range forecasts, lying in the 2.2–2.4 range for 2–5-day forecasts (Fig. 11). In contrast, the effective degrees of freedom for UKMET are largest (nearly three) for 12-h forecasts, and decreases for longer-range forecasts, down to less than two for 5-day forecasts. The contrasting values are likely due to the differences in the size and structure of the initial ensemble perturbations between ECMWF and UKMET, and their different ensemble means and perturbation growth or decay at later times. The effective degrees of freedom are also computed for an ensemble of six late-cycle deterministic models: AVNO, EMX, GFDL, HWRF, NGX, and UKX. The value is found to be smaller than those of either the ECMWF or UKMET ensembles, between 1.6–1.8 for all forecast times. To provide a quantitative measure of the independence of the track forecast errors for individual ensemble members, the effective degrees of freedom are then divided by the number of members. For the 50-member ECMWF ensemble, this value ranges between 0.044 and 0.048 for 2–5-day forecasts, while the corresponding values for the six-member multimodel ensemble range between 0.27 and 0.30, about 6 times greater. Therefore, the forecasts from the multimodel ensemble are considerably more independent than are the members of the ECMWF EPS (and UKMET EPS).

A related question concerning the number of ensemble members required to produce effective predictions of the ensemble mean tracks and 67% probabilities is also investigated. Only subsamples of the ECMWF ensemble are considered here. First, the average error of the ECMWF ensemble mean over the sample of 165 forecasts is found to remain almost the same for all forecast times, even when the number of ensemble members is reduced from 50 to 20 (Fig. 12a). If the ensemble size is reduced further to 10, the average errors of the ensemble mean become larger, although by less than 10%. Similarly, the percentage of all cases in which the best track falls within the 67% cone is generally similar for ECMWF ensembles of size between 20 and 50 (Fig. 12b). The ensemble becomes more underdispersive as the number of members is reduced to 10. It can therefore be concluded

---

**Table 2. Summary of the ability of the ensemble-based X% probability circles to capture X% of the best tracks over the 2008 Atlantic cases.** “Over” implies that the best track falls within the probability circle more than $X + 5\%$ of the time, and “under” implies that the best track falls within the probability circle less than $X - 5\%$ of the time. “Well” indicates that the probability circles are adequately dispersive on average, with the best track existing within the circles between $X - 5\%$ and $X + 5\%$ of all cases.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>ECMWF</th>
<th>UKMET</th>
<th>ECMWF + UKMET</th>
</tr>
</thead>
<tbody>
<tr>
<td>12–36 h</td>
<td>33%, under</td>
<td>33%, well</td>
<td>33%, well</td>
</tr>
<tr>
<td></td>
<td>50%, under</td>
<td>50%, well</td>
<td>50%, under</td>
</tr>
<tr>
<td></td>
<td>67%, under</td>
<td>67%, well</td>
<td>67%, under</td>
</tr>
<tr>
<td></td>
<td>98%, under</td>
<td>98%, well</td>
<td>98%, well</td>
</tr>
<tr>
<td>48–84 h</td>
<td>33%, well</td>
<td>33%, well</td>
<td>33%, under</td>
</tr>
<tr>
<td></td>
<td>50%, well</td>
<td>50%, under</td>
<td>50%, under</td>
</tr>
<tr>
<td></td>
<td>67%, under</td>
<td>67%, under</td>
<td>67%, under</td>
</tr>
<tr>
<td></td>
<td>96%, under</td>
<td>96%, under</td>
<td>96%, under</td>
</tr>
<tr>
<td>96–120 h</td>
<td>33%, over</td>
<td>33%, well</td>
<td>33%, well</td>
</tr>
<tr>
<td></td>
<td>50%, well</td>
<td>50%, well</td>
<td>50%, under</td>
</tr>
<tr>
<td></td>
<td>67%, well</td>
<td>67%, well</td>
<td>67%, under</td>
</tr>
<tr>
<td></td>
<td>99%, well</td>
<td>99%, well</td>
<td>99%, well</td>
</tr>
</tbody>
</table>
that, for the Atlantic basin season in 2008, a 20-member ECMWF ensemble would likely have been equally as effective, on average, as the full ensemble at producing forecasts of the mean track and associated 67% probability circles. However, we emphasize again that the sample size is small, particularly at later times, and is dominated by just five storms.

4) ATLANTIC AND WESTERN NORTH PACIFIC: ECMWF AND GPCE

The same evaluations were performed for TCs in the western North Pacific basin during 2008, for the ECMWF ensemble, which had proven to be the most reliable in the Atlantic, and for the early cycle deterministic model consensus and GPCE. First, the average track errors of the ECMWF ensemble mean were higher for the western North Pacific than for the Atlantic in 2008, for forecasts of 3 days or more (Fig. 13a). In the western North Pacific, the errors for CONW were strikingly higher than those produced by the ECMWF ensemble mean. This discrepancy is in contrast to the Atlantic basin, in which the errors in the ECMWF ensemble mean and TVCN were comparable. The forecast skill, defined by NHC (Franklin 2009) as $s_f(\%) = 100(e_b - e_f)/e_b$, where $e_b$ is the error in the baseline (CLIPER here) and $e_f$ is the error in the forecast (ensemble mean or consensus), is illustrated for both basins in Fig. 13b. The equally skillful ECMWF ensemble mean and TVCN forecasts (over 60% for 2–5-day forecasts) in the Atlantic basin allow for a fair comparison between the dispersiveness of the ECMWF and GPCE probability circles. As had been shown earlier (Figs. 7c and 12b), the ECMWF ensemble was found to be mostly appropriately dispersive with the best track falling within the 67% probability circles 60%–68% of the time. In contrast, the GPCE circles were found to be overdispersive, containing the best track between 78% and 90% of the time over all Atlantic basin cases (Fig. 13c). In the western North Pacific, the results are less straightforward to interpret, since the ECMWF ensemble mean exhibited significantly less skill than in the Atlantic,
FIG. 10. Range of values of radii for (a) ECMWF 67% probability circles and (b) GPCE circles for the Atlantic basin, over a homogeneous sample of 165 cases. The ends of the sticks refer to the lowest and highest radii in the sample. The 25th, 50th (median), and 75th percentiles are also illustrated, together with the average radius for each forecast time and the corresponding fixed NHC value.
and the CONW forecast exhibited nearly no skill for forecasts beyond 4 days. Given these relatively poor forecasts about which the probability circles were centered, it is not surprising that the best tracks lie outside the ECMWF and GPCE circles too often, with the circles capturing the best track only 40%–60% of the time for 2–5-day forecasts. The average size of the GPCE circles in both basins was considerably larger than the 67% circles constructed using the ECMWF ensemble (Fig. 13d). The average width of the ECMWF (GPCE) circles in the western North Pacific was similar to that for the Atlantic for forecasts of up to 4 (3) days, with the circles being larger in the western North Pacific for forecasts beyond those ranges.

The numbers of cases for which the best track resides within the 67% ECMWF circles are listed for each TC in the final four columns of Table 3. As for the Atlantic storms, there are distinct groups of cases for which the best track resides inside or outside the circle. For straight-line-moving storms over the open ocean such as 03W Rammasun and 13W Nuri, the best track lies within the circles for most forecasts. In contrast, 07W Fengshen was a notoriously difficult forecast using any suite of models, with the models consistently insisting that the typhoon would recurve instead of making its actual landfall over China. Storm 08W Kalmaegi was a similar case in which the models, including most of the ECMWF ensemble, exhibited tracks toward the right of the best track. For the longest-lived storm, 13W Sinlaku, the ECMWF ensemble produced a realistic scatter of track positions, but the storm stalled over Taiwan for 2 days in a weak steering flow between two subtropical ridges, which led to very high along-track errors. Additionally, the timing of recurvature and subsequent rapid motion in the mid-latitude westerly flow was a challenge for the ensemble, with the ensemble producing a mean track too far to the north and with narrow 67% circles. For typhoons 18W Hagupit and 19W Jangmi, the along-track errors were sometimes large. Similar conclusions were drawn for GPCE, which was centered on the CONW forecasts that exhibited low skill in 2008.

5. Conclusions

The ability of global model ensemble prediction systems (EPSs) to provide added benefit to operational forecasts of tropical cyclone track was evaluated for the 2008 Atlantic and western North Pacific seasons, using newly available Cyclone XML (CXML) data in the TIGGE database. In addition to the skill of the ensemble mean forecast, this paper represents a first investigation into the characteristics of dynamic EPS-based “probability circles,”
FIG. 12. (a) Average forecast error of the ensemble mean over 165 Atlantic basin cases, for ECMWF ensembles of different sizes. (b) Percentage of cases in which the best track falls within the 67% probability circle, constructed using the same ECMWF ensembles.
including the extent to which they were able to capture the best track an appropriate number of times. These circles were also compared against the Goerss predicted consensus error (GPCE) that is used at the NHC and JTWC.

For the Atlantic basin, the ECMWF ensemble mean was found to be of similar skill to the TVCN consensus, and it possessed considerably lower errors than all the global and regional models used at NHC except for the deterministic ECMWF model, which was outstanding in 2008. The ensemble mean produced fewer forecasts of high error than any of the deterministic models. For the probabilistic forecasts, the circles were centered on the ensemble mean. The most significant and encouraging result of this investigation is that the ECMWF ensemble was able to produce accurate 67% probability circles for 1–3½-day forecasts, and it was only weakly underdispersive for 4–5-day forecasts. For forecasts of 2 days or more, the ECMWF ensemble generally produced appropriately dispersive probability circles at the 33%, 50%, and 98% levels. In contrast, the GPCE was overdispersive for the 2008 Atlantic season, with the best track falling within the GPCE circles in between 78% and 90% of the cases. This was partly due to the high skill of the TVCN forecasts in 2008. This result is consistent with previous years, in which the GPCE was overdispersive for seasons in which the consensus errors were lower than average and underdispersive for the season in which the consensus errors were greater than average (J. Goerss 2009, personal communication).

Given that the ECMWF ensemble produced admirably accurate mean and probabilistic forecasts over the 2008 season, there was little room for improvement by adding the UKMET ensemble. Nevertheless, the combined ECMWF + UKMET ensemble produced more mean forecasts of small error than either individual ensemble. And the combined ensemble also yielded more accurate probabilistic 12–36-h forecasts. The UKMET ensemble produced an ensemble mean that was inferior to ECMWF (~50 km average difference for 2–4-day
forecasts), although it was competitive with most deterministic and regional models. For probabilities higher than 50%, the best track was situated outside the UKMET probability circles more often than was desired. The ECMWF and UKMET ensembles were found to possess more effective degrees of freedom than a smaller ensemble comprising six late-cycle deterministic models. However, the six individual members of the multimodel ensemble possessed greater independence than either EPS system.

The range of the ECMWF circle radii was generally broader than that of GPCE. A linear regression analysis on the ensemble mean forecast errors versus the forecast circle radii yielded an increasing slope, with a correlation coefficient between 0.3 and 0.6 for 1–5-day forecasts. Examining individual cases, it was found that the best track fell outside the probability circles particularly for recurving tropical cyclones, which is to be expected since their predictability is accordingly lower (Aberson and Sampson 2003), implying that the truth is more likely to occur on the edges of the distributions than in high-predictability cases.

The advantages and disadvantages of the EPS and GPCE approaches are evident from this study. The EPS approach is appealing in that it aims to provide probabilities (and distributions) based purely on the flow and its inherent uncertainty governed by the spread in the model. In practice, the EPS may not capture the best track often enough (as in the western North Pacific). Furthermore, the position and size of the EPS probability circles may fluctuate markedly from one initial time to the next, leading to decreased confidence in the average and probabilistic forecast. Since the GPCE approach includes a dependence on the past several seasons, in addition to other parameters, it may yield statistically more reliable results over a longer period with fewer fluctuations in circle size than with the EPS. The authors are of the opinion that a hybrid of the two approaches may yield superior results, maintaining run-to-run consistency while exploiting probabilistic information from a large number of numerical forecasts.

To conclude, based on the 2008 Atlantic data, the ECMWF ensemble holds promise for probabilistic track forecasting, as well as for use in predictions of consensus error and dynamic wind speed probabilities (Goerss 2007; DeMaria et al. 2009), and new metrics such as the probability of gale force winds. However, we must caution that the sample size is limited, and many additional cases are required to corroborate the main conclusions with statistical significance testing. Investigations are under way into the characteristics of EPS perturbations from different operational centers (Yamaguchi and Majumdar 2010), and advancements of the perturbation methods are ongoing at all centers. Given that different ensembles can produce distinct probability distributions, the addition of further global EPSs from the TIGGE database, such as NOGAPS, NCEP, and JMA.

### Table 3. List of western North Pacific cases for which ECMWF ensemble data exist, for each tropical cyclone. The last 4 columns give the ratio of the number of cases in which the best track is contained within the ECMWF 67% probability circles to the total number of cases.

<table>
<thead>
<tr>
<th>No.</th>
<th>Storm name</th>
<th>Dates</th>
<th>No. of ECMWF cases</th>
<th>48 h</th>
<th>72 h</th>
<th>96 h</th>
<th>120 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>03W</td>
<td>TY Rammusun</td>
<td>7–12 May</td>
<td>11</td>
<td>6/7</td>
<td>5/5</td>
<td>3/3</td>
<td>1/1</td>
</tr>
<tr>
<td>04W</td>
<td>TS Matmo</td>
<td>15–17 May</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05W</td>
<td>TY Halong</td>
<td>15–20 May</td>
<td>10</td>
<td>4/6</td>
<td>2/4</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>06W</td>
<td>STY Nakri</td>
<td>27 May–3 Jun</td>
<td>14</td>
<td>6/10</td>
<td>5/8</td>
<td>4/6</td>
<td>4/4</td>
</tr>
<tr>
<td>07W</td>
<td>TY Fengshen</td>
<td>19–25 Jun</td>
<td>14</td>
<td>2/10</td>
<td>2/7</td>
<td>1/6</td>
<td>1/4</td>
</tr>
<tr>
<td>08W</td>
<td>TY Kalmaegi</td>
<td>14–18 Jul</td>
<td>9</td>
<td>1/5</td>
<td>1/3</td>
<td>0/1</td>
<td></td>
</tr>
<tr>
<td>09W</td>
<td>TY Fung-Wong</td>
<td>26–29 Jul</td>
<td>7</td>
<td>3/3</td>
<td>1/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10W</td>
<td>TS Kammuri</td>
<td>5–7 Aug</td>
<td>5</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12W</td>
<td>TS Vongfong</td>
<td>15–17 Aug</td>
<td>5</td>
<td>1/2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13W</td>
<td>TY Nuri</td>
<td>17–22 Aug</td>
<td>11</td>
<td>7/7</td>
<td>5/5</td>
<td>2/2</td>
<td></td>
</tr>
<tr>
<td>15W</td>
<td>STY Sinlaku</td>
<td>10–20 Sep</td>
<td>22</td>
<td>11/18</td>
<td>7/16</td>
<td>7/13</td>
<td>5/12</td>
</tr>
<tr>
<td>18W</td>
<td>TY Hagupit</td>
<td>19–24 Sep</td>
<td>12</td>
<td>5/8</td>
<td>3/6</td>
<td>2/4</td>
<td>0/2</td>
</tr>
<tr>
<td>19W</td>
<td>STY Jangmi</td>
<td>24–30 Sep</td>
<td>14</td>
<td>6/10</td>
<td>2/8</td>
<td>3/6</td>
<td>1/4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>138</td>
<td>53/87</td>
<td>33/63</td>
<td>23/43</td>
<td>12/27</td>
</tr>
<tr>
<td>Percentage</td>
<td></td>
<td></td>
<td>61</td>
<td>52</td>
<td>53</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>
plus ensembles constructed in research, may be expected to improve the quality of the ensemble mean and probabilistic forecasts. Other near-term improvements include the extension of these forecasts beyond 5 days, and the introduction of probabilities in along-track and cross-track directions. A combination of the EPS and GPCE may provide improved results by exploiting the strengths of the two respective methods.

Acknowledgments. The authors gratefully acknowledge James Goerss for his thorough review and recommendations, and Sim Aberson, Tim Marchok, and multiple staff at NHC for constructive discussions. Two anonymous reviewers provided assistance in correcting several statements. We also thank Matthew Niznik for decoding the CXML data, Piers Buchanan for providing UKMET data, Buck Sampson for providing the western North Pacific CLIPER forecast data, and the THORPEX TIGGE team for creating and maintaining the database. The authors acknowledge funding from ONR Grant N000140810250, Frank Marks and the NOAA Hurricane Forecast Improvement Project, and two undergraduate research fellowships from RSMAS and the University of Miami.

REFERENCES


