2010

Short- and Medium-Range Prediction of Tropical and Transitioning Cyclone Tracks within the NCEP Global Ensemble Forecasting System

Christian E. Buckingham
University of Rhode Island, cbuckingham@gso.uri.edu

Timothy Marchok

Isaac Ginis
University of Rhode Island, iginis@uri.edu

Lewis M. Rothstein
University of Rhode Island, lrothstein@uri.edu

Dail Rowe

Citation/Publisher Attribution

Buckingham, C., T. Marchok, I. Ginis, L. Rothstein, and D. Rowe, 2010: Short- and Medium-Range Prediction of Tropical and Transitioning Cyclone Tracks within the NCEP Global Ensemble Forecasting System. *Wea. Forecasting,* 25, 1736–1754, https://doi.org/10.1175/2010WAF2222398.1
Available at: https://doi.org/10.1175/2010WAF2222398.1

Short- and Medium-Range Prediction of Tropical and Transitioning Cyclone Tracks within the NCEP Global Ensemble Forecasting System

Short- and Medium-Range Prediction of Tropical and Transitioning Cyclone Tracks within the NCEP Global Ensemble Forecasting System

CHRISTIAN BUCKINGHAM

Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island

TIMOTHY MARCHOK

Geophysical Fluid Dynamic Laboratory, National Oceanic and Atmospheric Administration, Princeton, New Jersey

ISAAC GINIS AND LEWIS ROTHSTEIN

Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island

DAIL ROWE

WeatherPredict Consulting, Inc., Narragansett, Rhode Island

(Manuscript received 29 December 2009, in final form 29 July 2010)

ABSTRACT

The NCEP Global Ensemble Forecasting System (GEFS) is examined in its ability to predict tropical cyclone and extratropical transition (ET) positions. Forecast and observed tracks are compared in Atlantic and western North Pacific basins for 2006–08, and the accuracy and consistency of the ensemble are examined out to 8 days. Accuracy is quantified by the average absolute and along- and cross-track errors of the ensemble mean. Consistency is evaluated through the use of dispersion diagrams, missing rate error, and probability within spread. Homogeneous comparisons are made with the NCEP Global Forecasting System (GFS). The average absolute track error of the GEFS mean increases linearly at a rate of 50 n mi day^{-1} [where 1 nautical mile (n mi) = 1.852 km] at early lead times in the Atlantic, increasing to 150 n mi day^{-1} at 144 h (100 n mi day^{-1} when excluding ET tracks). This trend is 60 n mi day^{-1} at early lead times in the western North Pacific, increasing to 150 n mi day^{-1} at longer lead times (130 n mi day^{-1} when excluding ET tracks). At long lead times, forecasts illustrate left- and right-of-track biases in Atlantic and western North Pacific basins, respectively; bias is reduced (increased) in the Atlantic (western North Pacific) when excluding ET tracks. All forecasts were found to lag behind observed cyclones, on average. The GEFS has good dispersion characteristics in the Atlantic and is underdispersive in the western North Pacific. Homogeneous comparisons suggest that the ensemble mean has value relative to the GFS beyond 96 h in the Atlantic and less value in the western North Pacific; a larger sample size is needed before conclusions can be made.

1. Introduction

Numerical modeling of tropical cyclones has improved in recent years, owing to improved observations of the atmosphere, better assimilation methods, improved model physics, and increased model resolution (Rappaport et al. 2009). As numerical weather prediction (NWP) models have improved, scientists have investigated the use of ensemble forecasts for the prediction of a tropical cyclone's path. Zhang and Krishnamurti (1997, 1999) introduced a perturbation technique for tropical cyclone track prediction based on empirical orthogonal functions and applied this technique to a global spectral model with promising results. Goerss (2000) analyzed the mean track of a multimodel ensemble forecast (referred to as a consensus track), and found that the ensemble mean showed 16%–23% improvement in track forecasts over the best member in the ensemble within the first 72 h of the forecast during the 1995–96 Atlantic hurricane seasons. More recently, the GUNA Consensus, a forecast introduced at

Corresponding author address: Christian Buckingham, Graduate School of Oceanography, University of Rhode Island, South Ferry Rd., Narragansett, RI 02882.
E-mail: cbuckingham@gso.uri.edu

DOI: 10.1175/2010WAF2222398.1

the National Oceanic and Atmospheric Administration/National Hurricane Center (NOAA/NHC) consisting of an average of track forecasts from four operational NWP models [Global Forecast System (AVNI), Geophysical Fluid Dynamics Laboratory (GFDI), Met Office (UKMI), and Navy Operational Global Atmospheric Prediction System (NOGAPS) forecasts interpolated ahead 6 h], was shown to have forecasts at 96-h lead time that are 18% more accurate than the best-performing member within the ensemble during the 2004–06 seasons (Rappaport et al. 2009). Another model consensus, the CONU [which is similar to the GUNA Consensus but includes the Navy version of the GFDL Hurricane Model forecast track/intensity (GFNI)], shows similar skill (Goerss 2007). Elsberry and Carr (2000) investigated the use of a selective consensus, in which forecasters remove one or more members from the ensemble at their discretion. Sampson et al. (2007) have found that, in the western North Pacific, forecasters were unable to produce a selective consensus that consistently improved guidance over a nonselective consensus. A final approach to extracting guidance from NWP model forecasts is the so-called superensemble, in which past model performance is used to assign weighting to individual members of an ensemble prior to forming the consensus (Williford et al. 2003).

In addition to multimodel ensemble forecasts, ensemble systems based on the application of perturbations to a control analysis from a global model have gained increasingly wider use in recent years. Such global ensemble prediction systems (EPSs) show promise in medium-range prediction by providing forecasts whose boundaries and dynamics are consistent over longer lead times.

Global EPSs in use today include those at the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction (NCEP), the Meteorological Service of Canada (MSC), and the Japan Meteorological Agency (JMA). See Park et al. (2008) for a more exhaustive list of global ensemble forecasting systems currently in use. Each EPS is unique in its data assimilation system, perturbation method, model physics, and boundary conditions, making comparisons between ensembles difficult. Nevertheless, Buizza et al. (2005) presented a comparison of ECMWF, NCEP, and MSC EPS performance from model data in 2002, and work is currently being conducted in comparing the performance between EPSs as part of The Observing System Research and Predictability Experiment (THORPEX) program (Park et al. 2008).

The present paper examines the performance of the NCEP Global Ensemble Forecasting System (GEFS) in predicting tropical cyclone and extratropical transitioning cyclone tracks. The motivation for including transitioning cyclones in this analysis is that operational forecast centers must make forecasts for all storms currently in their warning areas, including those that may transition within a given forecast period, and accurate prediction of such transitioning cyclones is a significant challenge for forecasters. Given the damage and loss of life that can occur as a result of transitioning storms (Jones et al. 2003), it is desirable to assess the performance of the GEFS in predicting extratropical transitioning cyclone tracks in addition to tropical cyclone tracks.

The NCEP GEFS is a single-model, global ensemble consisting of 21 members and is run 4 times daily (0000, 0600, 1200, and 1800 UTC) out to 384-h (16 day) lead time. The underlying model for the GEFS is the NCEP Global Forecasting System (GFS), a high-resolution (T382L64 for 0–180-h lead time; T190L64 for 180–384-h lead time) spectral atmospheric model run 4 times daily at the Environmental Modeling Center. The GFS analysis is spectrally truncated and interpolated to a lower-resolution analysis (T126L28[1]), which then serves as the control analysis for the ensemble. The analysis field for each member forecast is created by applying a small perturbation to the control analysis. The present perturbation method used by NCEP is referred to as the ensemble-transform bred-vector technique (Wei et al. 2008), and it differs from its predecessor, the bred-vector technique, in that perturbations applied to initial conditions are orthogonal vectors with magnitude and direction, rather than simply positive–negative pairs. This change was implemented in May 2006. The number of perturbation members was also increased at this time, from 11 to 14, and again from 14 to 20 in 2007.

The outline of the paper is as follows. The forecast verification methods used in this study are described in section 2. Results are presented in section 3, followed by a discussion in section 4. The Appendix details the use of the bootstrap method to estimate confidence intervals.

2. Methods

The performance of the NCEP GEFS in predicting tropical cyclone (TC) and extratropical transitioning (ET) tracks is estimated by comparing forecast and observed tracks in the Atlantic and western North Pacific basins. Cyclone intensity is not considered in this study. Forecast verification methods include the computation of the average absolute and along- and cross-track errors, use of dispersion diagrams, and computation of the missing rate error and probability within spread. Before introducing these methods, however, we discuss the set of forecast and observed cyclone tracks used in the study.

[1] As of 1200 UTC 23 February 2010, the horizontal resolution is T190 (McClung 2009).

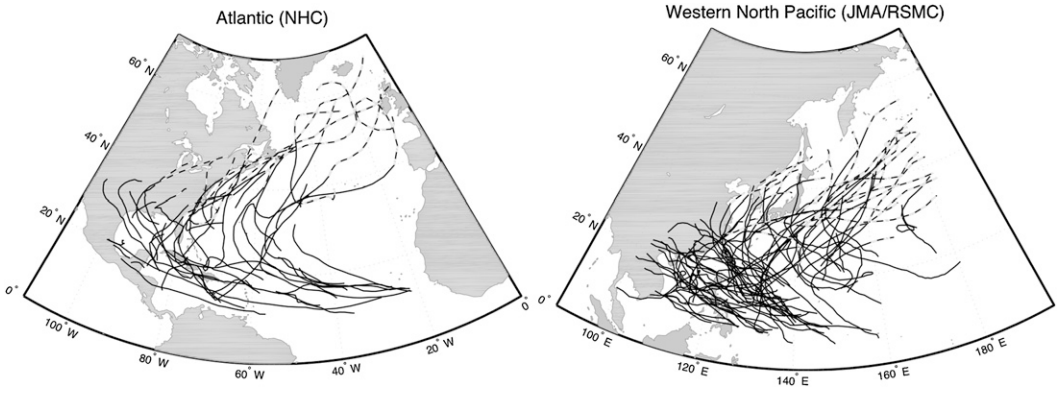

FIG. 1. Observed cyclone tracks within the Atlantic and western North Pacific basins, 2006–08. Cyclone tracks were obtained from NHC and JMA/RSMC best-track records. Solid and dashed lines represent stages when cyclones were classified as TC and ET, respectively.

a. Description of forecast and observed tracks

Forecast tracks were obtained from the GEFS through the use of an automated tracking system (Marchok 2002). Used operationally at NCEP since 1998, the algorithm produces position fixes for several low-level parameters, including relative vorticity at 850 and 700 hPa, minimum sea level pressure, geopotential height at 850 and 700 hPa, and minimum wind speed at 850 and 700 hPa.

To locate a maximum or minimum value for a given variable, the algorithm employs a single-pass Barnes analysis (Barnes 1964) at grid points centered on the observed storm position, as determined by a Regional Specialized Meteorological Center (RSMC). The Barnes analysis provides an array of Gaussian-weighted mean position fixes surrounding the initial-guess position. A position fix is defined as the point at which the Barnes function is maximized or minimized, depending on the parameter being analyzed. After a fix is returned from the first iteration of the analysis, additional iterations are performed. For each iteration, the Barnes analysis grid is centered on the position fix from the previous iteration, and the grid resolution is doubled to obtain a finer estimate. Position fixes for all variables are then averaged together in order to produce a mean position at each lead time. Parameters with position fixes outside a specified distance [usually 150 n mi, where 1 nautical mile (n mi) = 1.852 km] of the guess position for a given forecast hour are excluded from the computation of the mean position. We note that cyclone genesis is not considered in this study. Thus, the forecast tracks are only those corresponding to storms that have been identified and numbered by an RSMC.

Cyclone tracks from the NCEP GFS deterministic forecast are also used to assess the accuracy of the ensemble. GFS tracks are obtained by using the same automated tracking procedure described above. The current operational version of the NCEP tracker produces GFS tracks out to 180-h lead time, thereby limiting comparisons between the two systems. Future versions of the tracker will extend to 384 h.

As mentioned in the introduction, the NCEP GEFS has undergone significant changes in recent years. In addition to those previously mentioned, in May 2006, a modification was made that relocated cyclone vortices closer to observed storm positions. This had the effect of improving initialization. To assess the impacts of increased ensemble size in 2007, the ensemble was randomly sampled so that the ensemble was made up of 15 members for all years. It was found that the differences in the results were negligible between samples with all members present and with 15 members present. Because the differences are negligible, and because subsampling reduces the number of cases (see the discussion regarding the ensemble mean below), the results presented in this paper use the full set of ensemble members.

Observed cyclone positions were obtained from the NHC (information online at http://www.nhc.noaa.gov/pastall.shtml#hurdat) and JMA/RSMC (information online at http://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/besttrack.html). Figure 1 shows the observed cyclone tracks examined in this study, where solid lines correspond to cyclones categorized as TCs and dashed lines represent those categorized as ET. Only cyclones categorized as tropical depressions or stronger are considered in the analysis.

Tables 1 and 2 summarize the number of forecast and observed cyclones available for comparison in the Atlantic and western North Pacific basins, respectively. A number of tropical cyclones of significant duration, and which transitioned into extratropical cyclones, occurred in the Atlantic during the 2006 season, causing

TABLE 1. The number of cyclones and cases available for comparison at each lead time in the Atlantic, 2006–08. Numbers outside parentheses refer to those when considering both TC and ET tracks, while numbers in parentheses refer to those when considering TC tracks alone. The numbers of cyclones and cases were determined by intersecting available forecasts with observations from NHC best-track data. The numbers of cases under homogeneous comparison are less because of limitations of the automated tracking procedure.

Lead time (h)	0	24	48	72	96	120	144	168	192	216	240
No. of cyclones	39 (39)	37 (37)	31 (29)	27 (24)	21 (17)	16 (13)	13 (11)	11 (7)	8 (3)	2 (2)	2 (1)
No. of cases	880 (864)	774 (710)	628 (536)	461 (359)	339 (241)	244 (166)	171 (106)	106 (53)	62 (23)	44 (14)	37 (9)
No. of cyclones (homogenous)	39 (39)	37 (36)	31 (29)	23 (19)	20 (16)	13 (11)	12 (9)	8 (6)	0 (0)	0 (0)	0 (0)
No. of cases (homogenous)	685 (676)	569 (526)	429 (368)	303 (241)	220 (163)	154 (111)	112 (77)	69 (39)	0 (0)	0 (0)	0 (0)

the Atlantic dataset (Table 1) at longer lead times to be dominated by cyclones from 2006. In addition to listing the number of cyclones available for comparison, Tables 1 and 2 list the total number of *cases* for which forecast–observation pairs are available. We note that for a given storm there can be many cases.

b. Definitions

The ensemble mean position is defined as the average of the member forecast locations at a given lead time:

$$\mathbf{x}_{\bar{e}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \tag{1}$$

where $\mathbf{x}_i$ is the position of the ith member of the ensemble, N is the number of ensemble members, and the summation is performed in vector space. All distances have been calculated in a *great circle* sense to avoid inaccuracies when working with projections (Froude et al. 2007). In order for an ensemble mean to exist at a given lead time, we require that at least eight members be present. This number has been determined empirically, noting that a smaller number produces unrealistic, jagged forecasts, while requiring a greater number reduces the total number of cases available for comparison. A less restrictive constraint was used by Froude et al. (2007), who determined that five member tracks provided sufficient statistics for both the NCEP and ECMWF ensemble systems.

The variance of the ensemble is defined as the average of the squared distances of members from the mean:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}s_i^2 = \frac{R_E^2}{N}\sum_{i=1}^{N}[\cos^{-1}(\mathbf{x}_i \cdot \mathbf{x}_{\bar{e}})]^2, \tag{2}$$

where s_i is the distance of the ith member to the ensemble mean and R_E is the mean radius of the earth (3440 n mi). We define the spread σ as the square root of this quantity. We note that Goerss (2000) has defined spread as the average distance of members from the mean, which yields a value slightly smaller than the definition given here. Figure 2 illustrates both the ensemble mean and spread for a GEFS forecast issued at 1200 UTC 10 June 2006. Spread is shown as a circle with radius σ.

c. Measures of accuracy

Forecast accuracy refers to the average correspondence of individual forecasts and the events they predict (Wilks 2006). The average absolute and along- and cross-track errors of the ensemble mean provide us with measures of the ensemble accuracy.

Absolute track error is the distance between observed and forecast cyclone positions. In the case of the ensemble mean, this is expressed as

$$s = R_E \cos^{-1}(\mathbf{x}_{\text{obs}} \cdot \mathbf{x}_{\bar{e}}), \tag{3}$$

where $\mathbf{x}_{\text{obs}}$ is the position of the observed cyclone. Cross-track error is estimated as the minimum distance of a

TABLE 2. As in Table 1, except numbers are shown for the western North Pacific. The numbers of cyclones and cases were determined by intersecting available forecasts with observations from JMA/RSMC best-track data.

Lead time (h)	0	24	48	72	96	120	144	168	192	216	240
No. of cyclones	58 (58)	56 (56)	50 (49)	43 (40)	36 (33)	33 (30)	26 (22)	16 (11)	15 (10)	9 (6)	3 (1)
No. of cases	1099 (1093)	1032 (974)	864 (767)	658 (557)	489 (405)	339 (276)	241 (178)	155 (106)	95 (57)	45 (23)	11 (3)
No. of cyclones (homogenous)	58 (58)	54 (54)	49 (48)	42 (39)	34 (32)	27 (25)	23 (20)	15 (11)	0 (0)	0 (0)	0 (0)
No. of cases (homogenous)	987 (981)	876 (845)	694 (634)	501 (438)	343 (294)	229 (192)	151 (116)	89 (64)	0 (0)	0 (0)	0 (0)

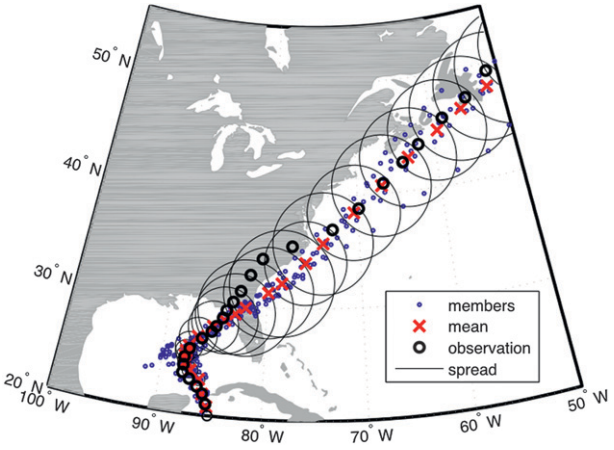

FIG. 2. Illustration of the NCEP GEFS for a forecast issued at 1200 UTC 10 Jun 2006. GEFS members, the GEFS mean, the observations, and the spread are shown.

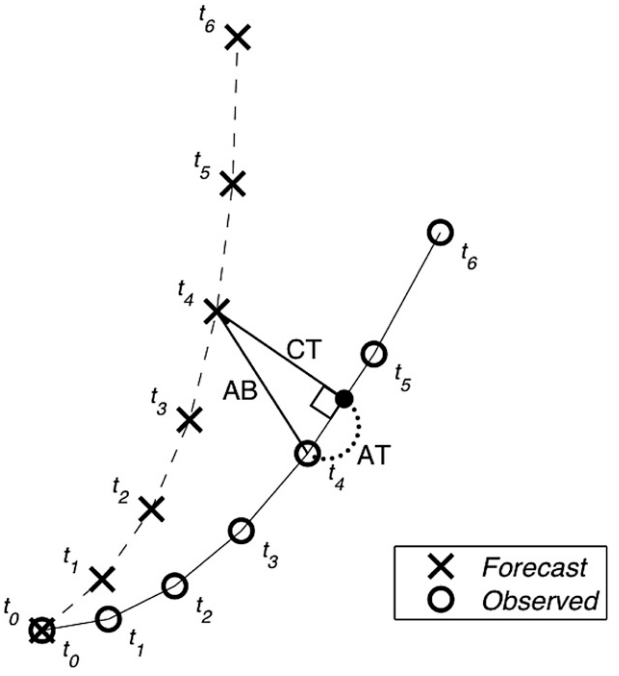

FIG. 3. Graphical illustration of absolute (AB), along- (AT), and cross-track (CT) errors. Absolute track error is computed as the distance between a forecast and an observed cyclone location. Cross-track error is computed as the minimum distance of a forecast to an interpolated observed track. Along-track error is computed as the great circle distance between an observed cyclone position and the intersection of the cross-track line with the interpolated observed track (black dot). The curvature is exaggerated for emphasis.

forecast cyclone position to an interpolated observed track (Fig. 3). Cross-track error is positive (negative) when a cyclone is forecast to the right (left) of the observed track. Interpolation is performed using linear interpolation at a resolution of 1.0 n mi along the length of the observed track. Care is taken to remove forecasts from consideration that lie in front of or beyond the observed track. While this decreases the sample numbers slightly, we believe it is appropriate given that extrapolation of observed tracks can produce large errors. The average cross-track error reveals the left- and right-of-track biases present in an ensemble.

Along-track error is defined as the great circle distance between an observed cyclone and the point of intersection of the cross track with the interpolated observed track (Froude et al. 2007) (Fig. 3). Along-track error is positive (negative) when a forecast lies ahead of (behind) an observed cyclone. The average along-track error reveals the forecast bias in the along-track direction.

In addition to computing the track errors of the ensemble mean, we consider homogeneous comparisons of the GEFS mean, GEFS control, and GFS deterministic forecasts. The motivation for comparing these forecasts is to determine if the GEFS mean is more or less accurate than both the GEFS control and the higher-resolution GFS forecast. Ideally, the GEFS mean should be more accurate at all lead times, but crossover in error between the forecasts may suggest where the value of higher spatial resolution in the GFS deterministic forecast is overcome by having additional members in the GEFS ensemble to account for initial condition errors. One notes that comparisons of average along- and cross-track errors reveal biases in the forecasts relative to each other (not shown). However, since the results are very similar among the forecasts, we display only homogeneous comparisons of average magnitudes of the along- and cross-track errors. Results, therefore, reveal how much along- and cross-track error is present in the forecasts relative to each other.

d. Measures of consistency

The consistency of an ensemble refers to the degree to which observations statistically resemble members of the ensemble (Wilks 2006). Applied to track prediction, one expects distances of observed cyclones to the ensemble mean to resemble distances of ensemble member cyclones to the mean. This serves as the basis for the forecast verification techniques described below.

1) DISPERSION DIAGRAMS

The mean squared error of the ensemble mean is estimated as

$$\mathrm{MSE} = \frac{1}{M}\sum_{m=1}^{M} s^2, \tag{4}$$

and the average ensemble variance is estimated by

$$\mathrm{VAR} = \frac{1}{M}\sum_{m=1}^{M}\sigma^2. \tag{5}$$

Here, s is given by Eq. (3) and M is the total number of cases at a particular lead time. Talagrand et al. (1997) point out that a consistent ensemble has a mean squared error approximately equal to the average variance of the ensemble. For an ensemble of finite size, the ratio of MSE to VAR is expected to be

$$\left(\frac{\mathrm{MSE}}{\mathrm{VAR}}\right)_{\mathrm{exp}} \approx 1 + \frac{2}{n+1} = \frac{n+1}{n-1}, \tag{6}$$

where n is the number of ensemble members (Ziehmann 2000; Eckel and Mass 2005). Rearranging this relationship, one obtains an expression for the average variance of a consistent ensemble given the mean squared error of the ensemble mean:

$$\mathrm{VAR}_{\mathrm{exp}} \approx \mathrm{MSE}\frac{n-1}{n+1}. \tag{7}$$

The reason that the average variance is not exactly equal to the rhs of Eq. (7) is that, in practice, the estimated mean squared error comprises both forecast and observation errors. Observation error refers to the uncertainty in cyclone positions reported by an RSMC. Errors in observed cyclone position arise from several sources, including limited observations, ambiguity associated with assigning single locations to complex atmospheric systems (e.g., vertically sheared cyclones or weak systems containing two cyclonic circulations), and postprocessing of observed cyclone tracks. Observation error in best-track data is not formally quantified (E. Fukada 2010, personal communication; J. Franklin 2010, personal communication). In the present study, we estimate this value by comparing differences in best-track records from JMA/RSMC and the Joint Typhoon Warning Center (JTWC) in the western North Pacific. Comparison of records over the 2006–08 seasons yields a mean squared difference of approximately (30 n mi)^2 = 900 n mi^2. Assuming that differences in cyclone positions reflect uncertainty in the observed position, rather than differences in methods employed at respective operational centers, Eq. (7) can be written as

$$\mathrm{VAR}_{\mathrm{exp}} = \mathrm{MSE}\frac{n-1}{n+1} - \mathrm{VAR}_{\mathrm{obs}}, \tag{8}$$

where we refer to $\mathrm{VAR}_{\mathrm{obs}}$ = 900 n mi^2 as the observation variance.

One difficulty in computing Eq. (8) is that the number of members present in a given forecast may be less than the total number of members of the ensemble. This occurs when the automated tracking algorithm fails to locate cyclones in one or more of the member forecasts. For this reason, the variable n in Eq. (8) is replaced with an effective ensemble number $\langle n\rangle$, where the angle brackets represent averaging over all cases for a given lead time.

Dispersion diagrams illustrate both the average variance of the ensemble and MSE of the ensemble mean as functions of lead time. Since values of VAR and MSE are small at early lead times and large at later lead times, it is useful to plot the square root of these quantities. Thus, dispersion diagrams in the next section illustrate the root MSE, root VAR, and root expected VAR (i.e., the variance expected if the ensemble was consistent).

2) MISSING RATE ERROR

One quantity of interest when examining whether the spread of an ensemble is appropriate is the percentage of observed cyclones falling closer to or farther from the ensemble mean than any one of the ensemble members. Given n equally likely members, one expects this percentage to be $200/(n + 1)$%. The deviation from this value is termed the missing rate error (MRE; Eckel and Mass 2005) and is defined as

$$\mathrm{MRE} = 100\left(\frac{1}{M}\sum_{m=1}^{M}\begin{cases}0: s_{\min} \le s_{\mathrm{obs}} \le s_{\max}\\ 1: \text{otherwise}\end{cases} - \frac{2}{n+1}\right), \tag{9}$$

where $s_{\min}$ and $s_{\max}$ are the minimum and maximum distances of members from the mean, s_{obs} is the distance of the observed cyclone to the mean, and M is the total number of cases at a particular lead time. A positive (negative) value suggests underdispersion (overdispersion) of the ensemble.

3) PROBABILITY WITHIN SPREAD

Another useful metric of ensemble consistency is the probability with spread (PWS). PWS estimates the likelihood of observed cyclones falling within the dispersion of an ensemble, and differs from MRE in that it considers varying distances from the mean. Expressed in terms of integer multiples of spread, PWS is defined as

$$\mathrm{PWS} = \frac{1}{M}\sum_{m=1}^{M}\begin{cases}0: s_{\mathrm{obs}} > k(\sigma)_m\\ 1: s_{\mathrm{obs}} \le k(\sigma)_m\end{cases}, \tag{10}$$

where k is an integer ($k = 1, 2, 3\ldots$), m is an integer, M is the total number of forecasts at a given lead time, s_{obs} is

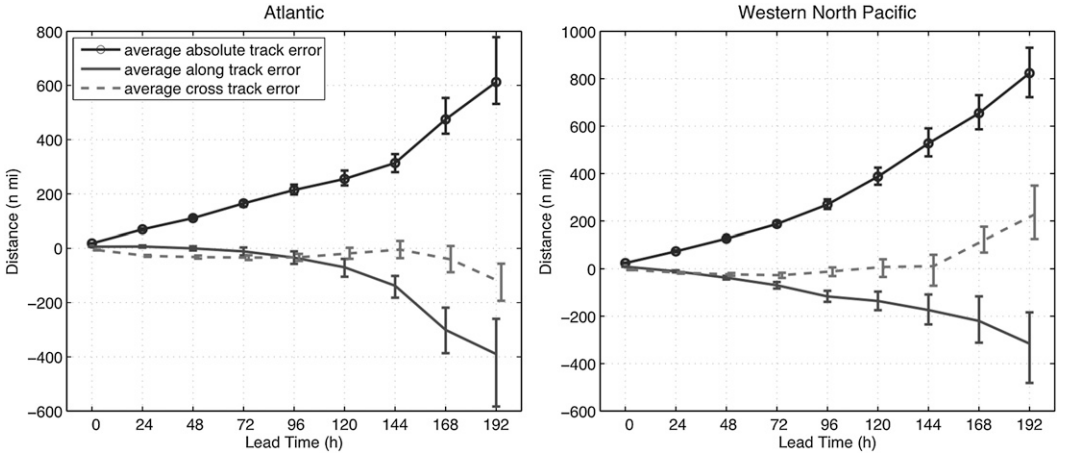

FIG. 4. The average absolute and along- and cross-track errors of the NCEP GEFS mean in the Atlantic and western North Pacific basins, 2006–08. Error bars illustrate 95% confidence intervals on the mean as determined from the bootstrap method. Both TC and ET tracks are included in the analysis. Along-track error is positive (negative) when a forecast lies ahead of (behind) an observed cyclone and cross-track error is positive (negative) when a cyclone is forecast to the right (left) of the observed track.

the distance of the observed cyclone to the ensemble mean, and σ is the spread of the ensemble. If members are sampled from a normal distribution with standard deviation σ, one expects PWS corresponding to 1σ, 2σ, and 3σ to have values near 0.68, 0.95, and 0.99, respectively. These numbers serve as references for results presented in the next section.

e. Confidence intervals

In the following section, confidence intervals are computed in order to bound our estimates of forecast accuracy and ensemble consistency. The method chosen to estimate the confidence intervals is known as the bootstrap technique (Efron 1979; Efron and Tibshirani 1993). The advantage of this method is that it makes no assumptions regarding the distribution of the data. While the samples are correlated, an investigation of the bootstrap technique applied to track data suggests that confidence intervals converge so long as 50 samples are present (see the appendix). In some instances, the data have fewer than 50 cases (see Tables 1 and 2). For this reason, results within this study are limited to lead times of 192 h (8 days) when examining combined TC–ET tracks and 168 h (7 days) when examining TC-only tracks. Confidence intervals are estimated at the 95% level.

3. Results

Utilizing the above methods, the performance of the GEFS is reported upon. Results in this section are presented for the full dataset containing both tropical cyclones and extratropical transitioning cyclones (TC–ET), as well as for the subset containing only tropical cyclones (TC only).

a. Average absolute and along- and cross-track errors

For the full set of TC–ET cases in the Atlantic (Fig. 4), the average absolute track error of the GEFS increases at rates of 50 n mi day^{-1} for lead times of 0–144 h and 150 n mi day^{-1} for lead times 144–192 h. For the subset of TC-only cases (Fig. 5), the rate of error increase is more gradual, with an increase of 45 n mi day^{-1} for lead times of 0–144 h, increasing to 100 n mi day^{-1} at longer lead times. Despite the steady increase in absolute track error, the GEFS remains mainly free of bias in the along-track direction and has only a weakly negative cross-track bias through 96 h for both the TC–ET and TC-only samples (cf. Figs. 4 and 5). After that time, the GEFS develops a negative along-track bias, and the magnitude of that bias at 168 h is significantly stronger in the TC–ET sample than in the TC-only sample. In addition, the GEFS develops a negative cross-track bias at longer lead times (168–192 h), and this bias is likewise stronger at 168 h in the TC–ET sample than in the TC-only sample (cf. Figs. 4 and 5). These results indicate that forecast cyclones fall, on average, behind and to the left of observed cyclones at longer lead times. Furthermore, the differences at later lead times between the TC–ET and TC-only samples suggest that the GEFS may have a slow bias for storms that are recurving into the westerlies.

For the full set of TC–ET cases in the western North Pacific (Fig. 4), the GEFS shows an increase in average

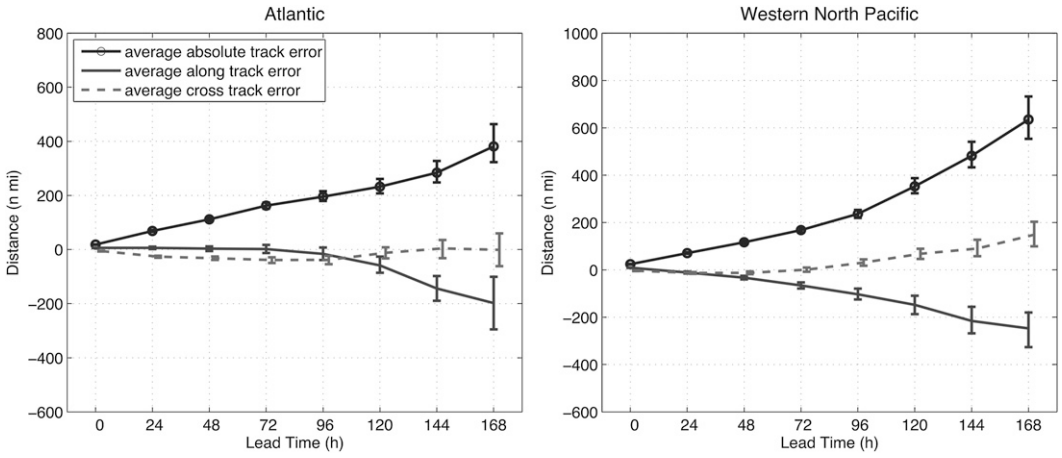

FIG. 5. As in Fig. 4, but only TC tracks are included in the analysis.

absolute track error with increasing lead time that is more gradual than in the Atlantic, increasing from 60 to 150 n mi day^{-1} at 144-h lead time. The rate of absolute track error increase in the western North Pacific is similar between the TC–ET and TC-only samples (cf. Figs. 4 and 5); however, the errors are smaller for the TC-only sample for lead times of 48 h and longer. There is negligible cross-track bias through 96 h for the TC–ET sample (Fig. 4), after which there is a steady positive increase in cross-track error. A negative bias is evident in the along-track direction, beginning at 48 h and increasing steadily through longer lead times. The TC-only sample (Fig. 5) produces results qualitatively similar to those for the TC–ET sample for both the along- and cross-track errors, although the cross-track bias is larger in the TC-only sample for medium-range lead times (96–144 h). These results indicate that forecasts fall, on average, behind and to the right of the observed cyclones at longer lead times. Furthermore, the results indicate an increased cross-track bias in the western North Pacific when considering TC tracks, alone.

b. Comparison with GFS and GEFS control

Homogeneous comparisons of the GEFS mean, GEFS control, and GFS deterministic forecasts for the full set of TC–ET cases (Fig. 6a) show that the GFS has smaller absolute track errors than both the GEFS control and the GEFS mean through 72-h lead time, although this trend is reversed at 96 h, with smaller errors for the GEFS mean relative to the GFS. In the western North Pacific basin (Fig. 6a), the relationship between the lower errors of the GEFS mean relative to the GFS is also evident, but is not established until 120-h lead time. Even then, the relationship is not as pronounced in this basin when compared with that for the Atlantic basin. Analyses in the Atlantic basin using the subset of TC-only cases (Fig. 7a) indicate smaller errors than the TC–ET sample at all lead times beyond 72 h, and they also indicate the same trend of the GEFS mean having smaller errors after 72-h lead time. The results for the TC-only sample in the western North Pacific (Fig. 7a) show the GFS with smaller errors than the GEFS mean through 96 h, and then nearly equal errors between the two models out through 168-h lead time.

A comparison of average magnitudes of along-track errors in the Atlantic for the full set of TC–ET cases (Fig. 6b) reveals smaller errors in the GEFS mean relative to the GFS for lead times beyond 72 h. The trend is the same for the sample of TC-only cases (Fig. 7b), but the magnitude of the along-track errors is smaller for the TC-only cases for all lead times beyond 72 h. In the western North Pacific, similar trends exist but are not established until 120 h (Figs. 6b and 7b). The magnitudes of the along-track errors are larger, on average, for the western North Pacific at lead times beyond 72 h than for the same lead times in the Atlantic. A comparison of the average magnitudes of cross-track error in the Atlantic (Figs. 6c and 7c) reveals larger cross-track errors in the GEFS mean relative to the GFS at almost all lead times for both the TC–ET and TC-only datasets. The cross-track errors are smaller for the TC-only dataset relative to the TC–ET set for lead times beyond 120 h. The results are different for the western North Pacific, where the GEFS mean has smaller cross-track errors than the GFS for all lead times beyond 72 h, both for the TC–ET and TC-only datasets (Figs. 6c and 7c).

c. Dispersion diagrams

The dispersion diagrams in Figs. 8a and 9a offer comparisons of the root MSE and root VAR of the

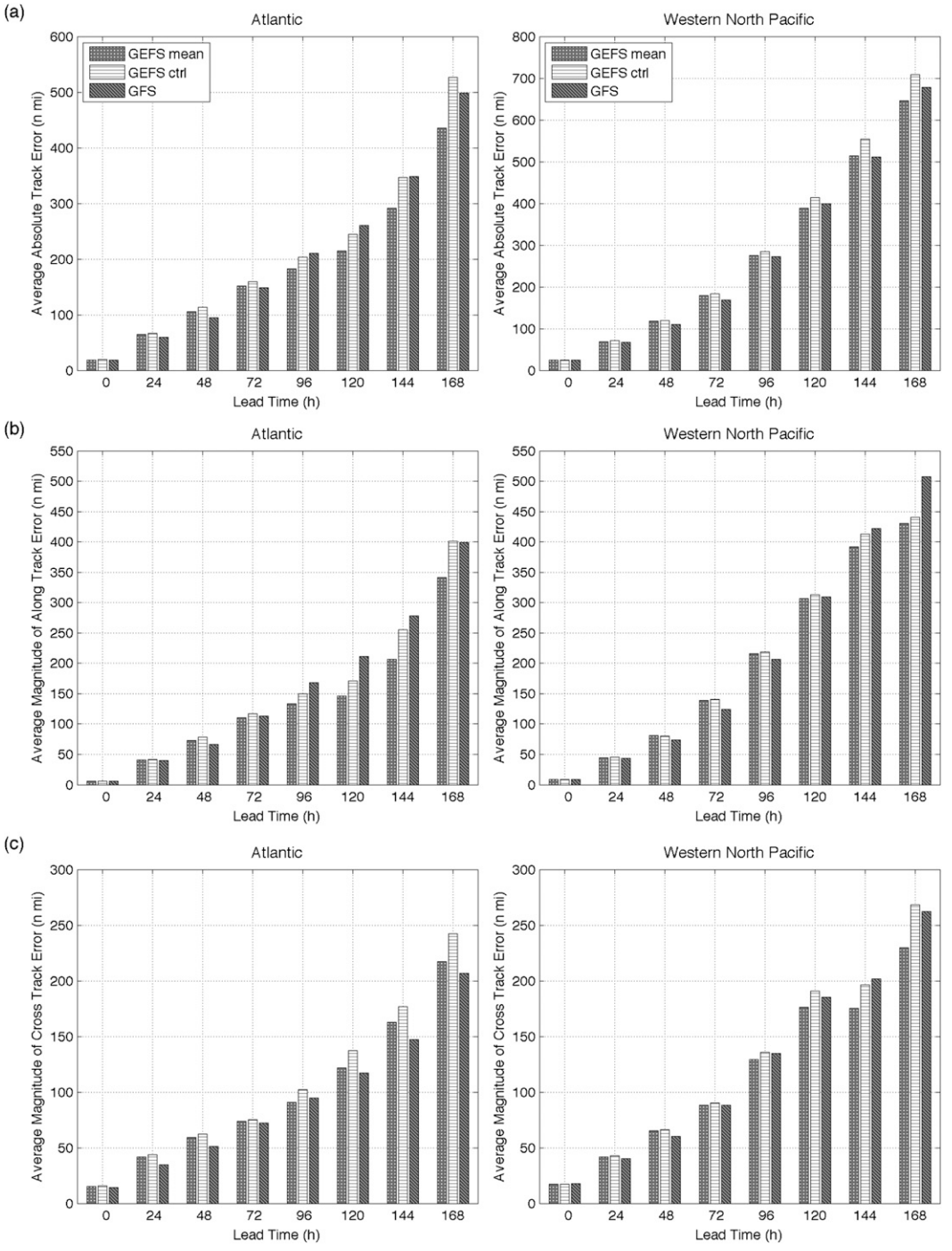

FIG. 6. Homogeneous comparison of the (a) average absolute track error, (b) average magnitude of the along-track error, and (c) average magnitude of cross-track error of the GEFS mean, GEFS control, and GFS deterministic forecasts in the Atlantic and western North Pacific basins, 2006–08. Both TC and ET tracks are included in the analysis.

GEFS mean as functions of lead time. As discussed earlier, the two quantities should be very similar. The expected ensemble variance represents the variance one would expect given error in the GEFS mean and taking into account the effective number of ensemble members and the observation variance [Eq. (8)].

For both the TC–ET and TC-only samples, the variance of the ensemble in the Atlantic basin is nearly appropriate

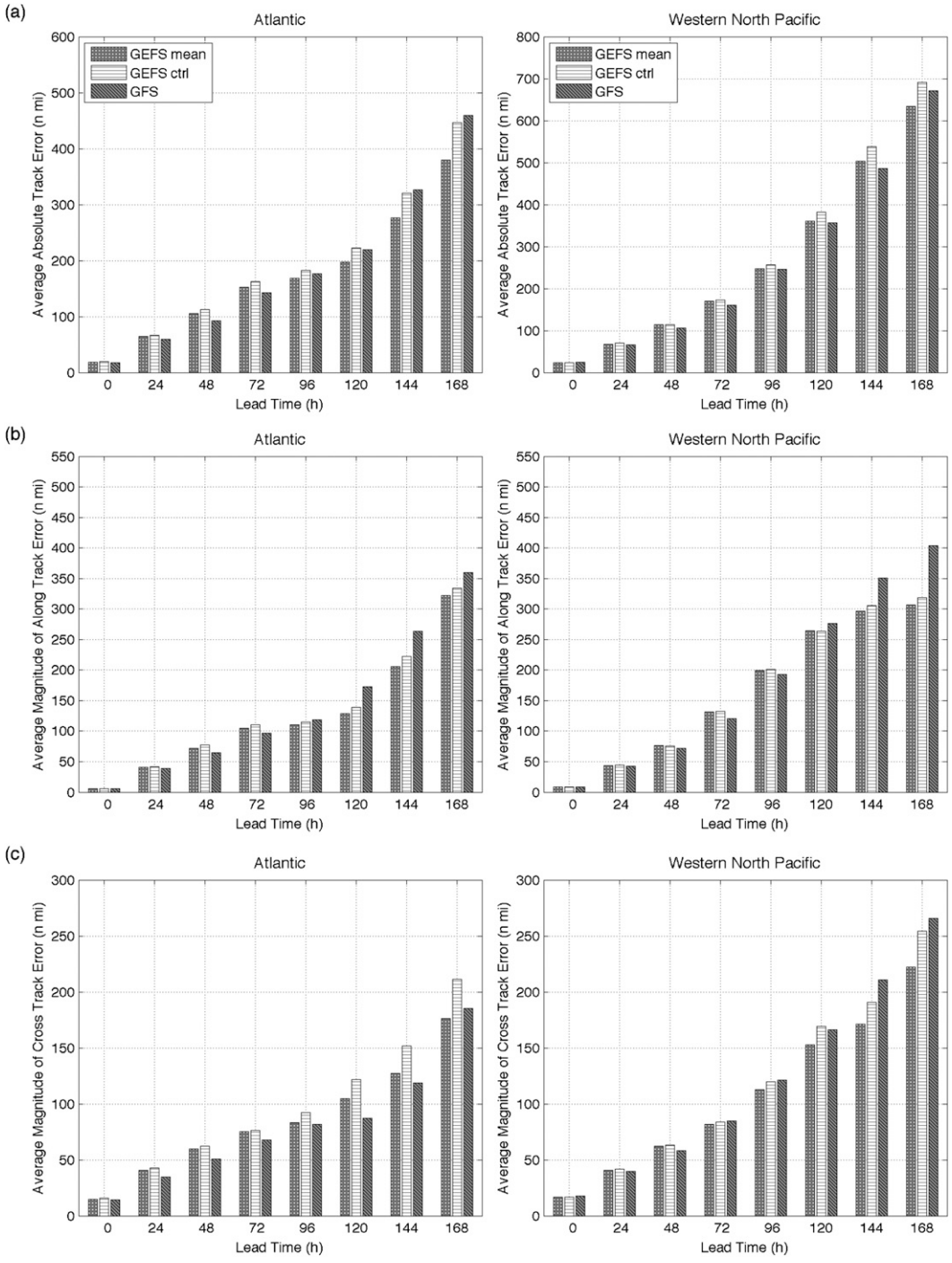

FIG. 7. As in Fig. 6, but only TC tracks are included in the analysis.

(Figs. 8a and 9a); root VAR is within 100 n mi of the expected curve. Taking into account confidence intervals for these quantities, the GEFS appears slightly underdispersive at 24–96 and 168–192 h for the TC–ET sample (Fig. 8a) as well as at 24–120 h for the TC-only sample (Fig. 9a). In contrast, the ensemble in the western North Pacific lacks consistency over all lead times: the square root of the average ensemble variance falls short of the expected value for all lead times, and is less than half its expected value beyond 72 h for both the TC–ET and TC-only samples (Figs. 8a and 9a). The GEFS is inconsistent and very underdispersive in the western North Pacific.

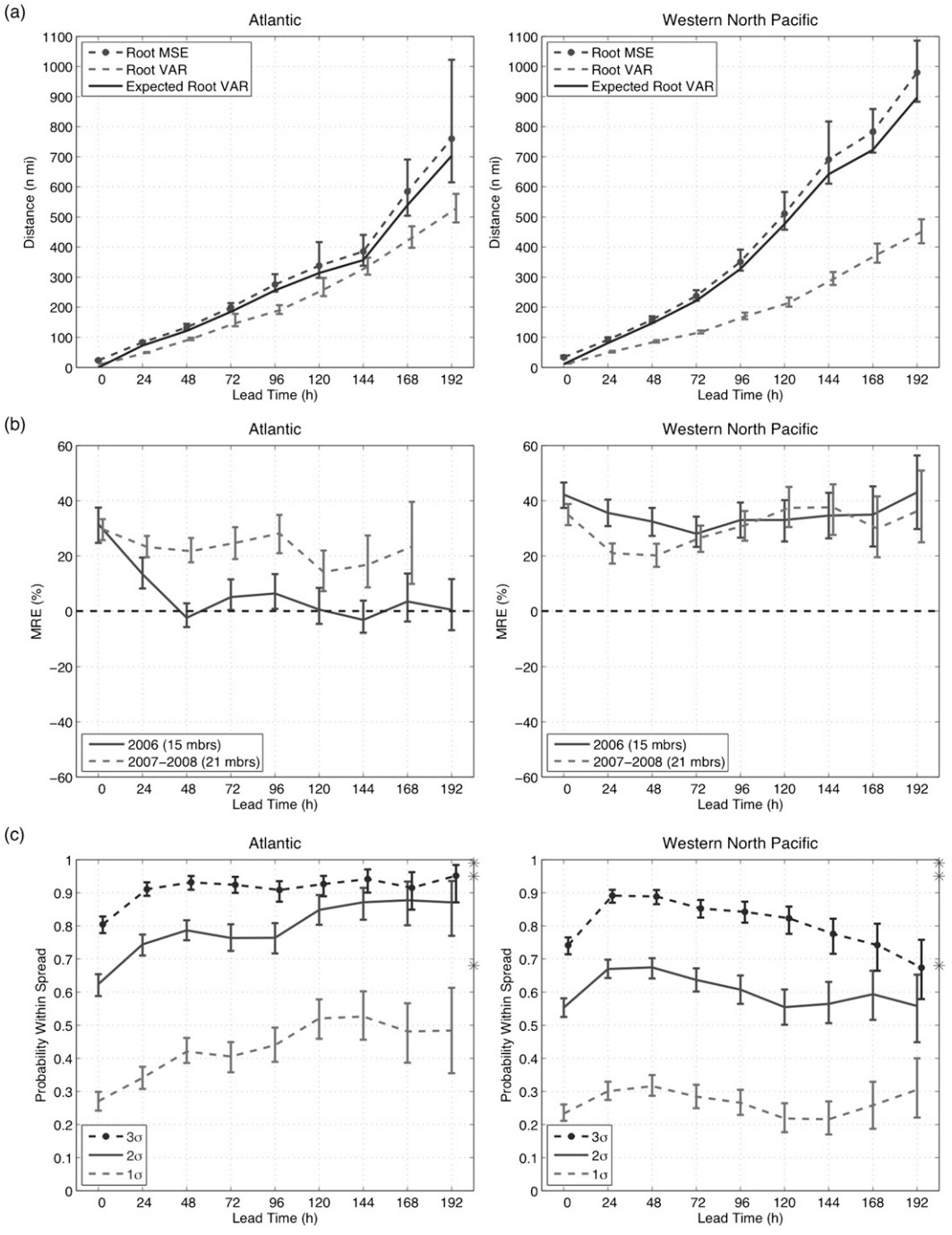

FIG. 8. Measures of ensemble consistency applied to the NCEP GEFS in the Atlantic and western North Pacific basins, 2006–08. Measures include (a) dispersion diagrams, (b) MRE, and (c) PWS. Both TC and ET tracks are included in the analysis. In (c), asterisks denote expected probabilities assuming a normal distribution with standard deviation σ.

d. Missing rate error

Figure 8b illustrates the missing rate error during the 2006 and 2007–08 seasons in the Atlantic and western North Pacific basins. For the 2006 season in the Atlantic, MREs for both the TC–ET and TC-only datasets shows a statistically significant positive value at early lead times (0–24 h), while being close to the expected value of zero for all other lead times (Figs. 8b and 9b). This suggests the ensemble is underdispersive at early lead times, but is

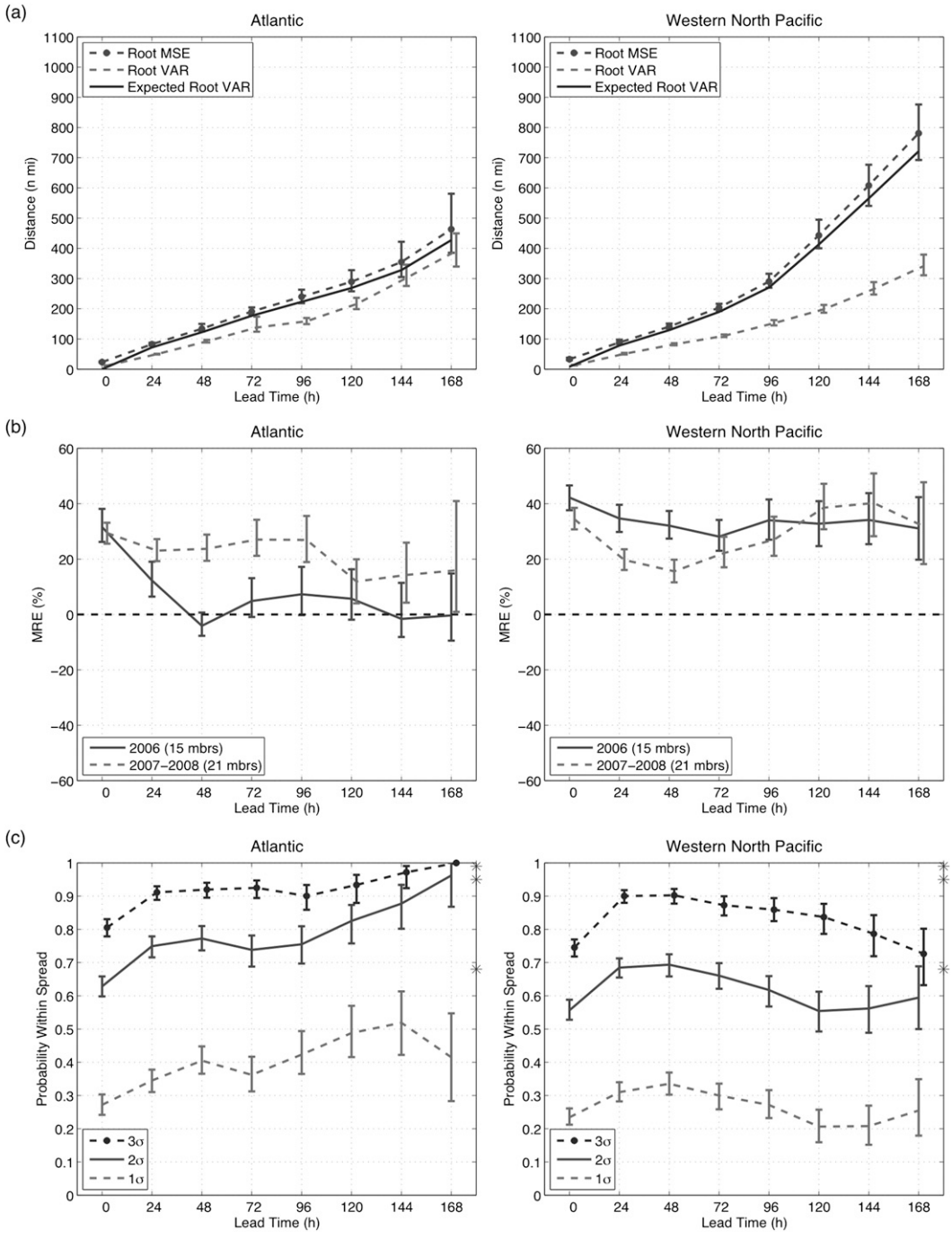

FIG. 9. As in Fig. 8, but only TC tracks are included in the analysis.

potentially consistent at later lead times. MREs in the Atlantic during the 2007–08 seasons show a positive value for all lead times—close to 25%–30% for lead times of 0–96 h and 20% at later lead times for both the TC–ET and TC-only samples (Figs. 8b and 9b). MRE at 192 h in the Atlantic during the 2007–08 seasons is not shown due to the small sample number.

In the western North Pacific, MREs for both TC–ET and TC-only samples indicate positive values for all lead times and both sets of hurricane seasons, with magnitudes ranging between 15% and 50% (Figs. 8b and 9b). This illustrates the inconsistency of the GEFS in the western North Pacific, and agrees with earlier conclusions made from dispersion curves (cf. Figs. 8a and 9a).

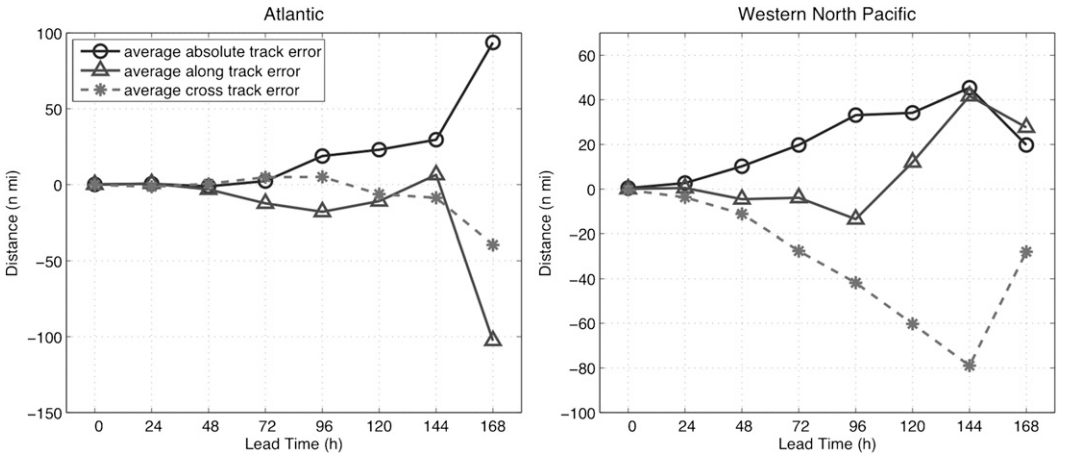

FIG. 10. Differences in the average absolute and along- and cross-track errors of the GEFS mean when including and excluding ET tracks within the analysis. Values are calculated as "with ET minus without ET."

It is not clear why a significant change in MRE exists between the 2006 and 2007–08 seasons in the Atlantic. It is possible that the increase in ensemble membership in 2007 would modify the dispersion characteristics of the ensemble, but one would expect a decrease in MRE to result from an increase in ensemble size (Buizza and Palmer 1998). It is useful to point out that a potential limitation of MRE is that it depends upon the closest and farthest members from the mean, and so can be impacted by outliers. For this reason, we consider another measure of consistency, below.

e. Probability within spread

Plots of the probability within spread for both the TC–ET and TC-only datasets are shown in Figs. 8c and 9c for distances of 1σ, 2σ and 3σ from the ensemble mean position. Asterisks denote expected probabilities, assuming a normally distributed ensemble with standard deviation σ. While the ensemble is not necessarily Gaussian, expected probabilities provide some reference from which to compare results.

In the Atlantic, PWS is less than the expected 0.68 for 1σ for all lead times for both the TC–ET and TC-only samples (Figs. 8c and 9c). In fact, it is lower than 0.50 for 0–96 h. Similarly, probabilities within 2σ and 3σ are less than the expected values of 0.95 and 0.99, respectively, for all lead times. However, it is notable that PWS for 3σ is close to the desired value (0.99) for most lead times. PWS is found to increase with increasing lead time in the Atlantic, indicating that the ensemble is more dispersive at longer lead times. In contrast, PWS in the western North Pacific decreases with increasing lead time for both the TC–ET and TC-only samples (Figs. 8c and 9c), at least for curves corresponding to 2σ and 3σ.

An interesting point is that there is an increase in PWS for 1σ and 2σ at 96-h lead time in the Atlantic (Figs. 8c and 9c), coincident with the decrease in MRE at the same lead time during 2007–08 (Fig. 8b, Atlantic) and in agreement with observations made earlier about the underdispersion of the ensemble at 0–96-h lead time (Fig. 8a, Atlantic).

f. Differences in track error when including and excluding ET tracks

One aspect of GEFS accuracy that can be studied is the error introduced when including extratropical transitioning cyclone tracks. Figure 10 indicates that, in the Atlantic, the inclusion of ET tracks results in an increase in the average absolute track error beginning after 72 h, with a sharp increase at 168-h lead time. This increase in track error is reflected in large negative along- and cross-track errors. One can also see this by examining Figs. 4 and 5 for the Atlantic at 168 h. The inclusion of ET tracks results in an increase of average along-track error from −200 to −250 n mi and average cross-track error from 0 to −40 n mi. This indicates that forecasts of ET cyclones lie farther behind and to the left of observed storms in comparison to tropical cyclones, on average. This error is potentially due to the difficulty the GEFS has both in timing the extratropical transition of cyclones and in forecasting the evolution of the various synoptic features that control the poleward movement of a storm into the westerlies.

In the western North Pacific, the inclusion of ET tracks results in a steady addition of negative cross-track error at lead times between 24 and 144 h. That is, forecast tracks are farther left of the observed tracks than when considering TC tracks alone. This makes sense if observed ET tracks tend to the right more than GEFS forecast tracks.

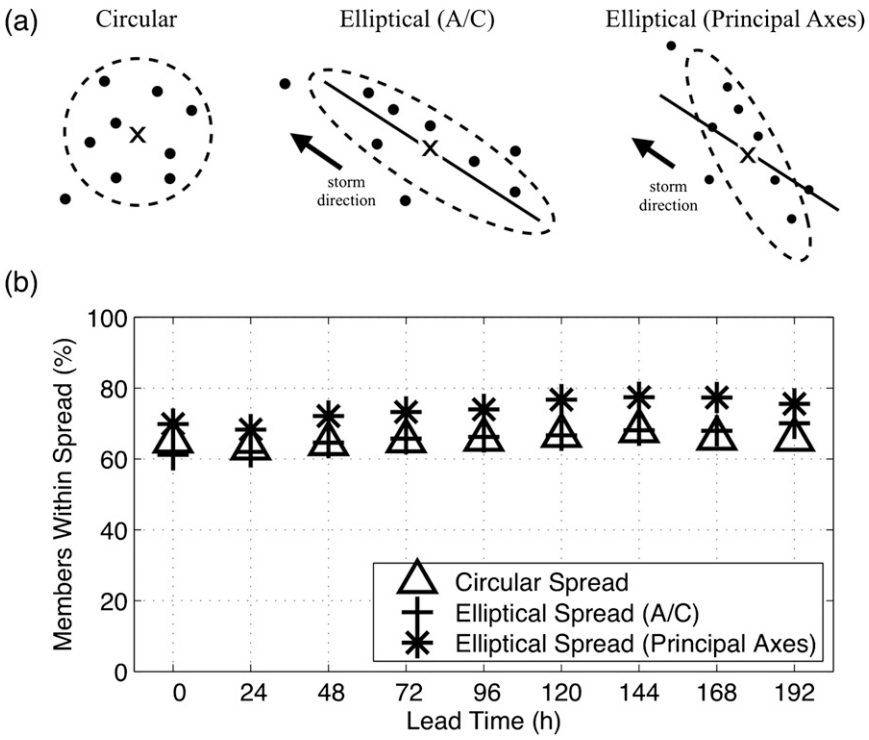

FIG. 11. Three definitions of spread and the percentage of members enclosed by these spreads. (a) Definitions of spread, including circular, elliptical (in the along- and cross-track directions), and elliptical (in the along- and cross-principal axes directions). In determining the percentage of members within the spread, all spreads were scaled to have equal area. Results are shown for the GEFS in the Atlantic, 2006–08, when both TC and ET tracks are considered.

g. Impacts of the definition of spread on the results

Here, we explore how the definition of spread may affect the results presented above. We defined spread as the square root of the average squared distances of members from the mean [cf. Eq. (2)]. Defined in this manner, spread is a scalar measure of the dispersion of the ensemble and is isotropic in its characterization. In the following discussion, we refer to this as a *circular* definition of spread. Another possible definition is the square root of the average squared distances of members from the mean in the along- and cross-track directions. Such a definition is referred to as an *elliptical (A/C)* definition of spread. A third possible definition of spread is the average distances of the members along and across principal axes. This is referred to as an *elliptical (principal axes)* definition of spread. All three definitions are illustrated in Fig. 11a.

To determine the degree to which the above definitions correctly characterize the dispersion of the GEFS, we examined the percent of members enclosed by the three definitions of spread. Figure 11b illustrates this percentage for the Atlantic basin, including both TC and ET tracks during the 2006–08 seasons. (Measures of spread are scaled to have an area equal to that of the circular definition.) Results show that circular and elliptical (A/C) definitions of spread contain the same percentages of members, on average. The elliptical (principal axes) definition of spread best describes the ensemble, containing approximately 5%–10% more of the members. This latter result was expected. The fact that an elliptical (A/C) definition of spread does not describe the ensemble's dispersion any better than the circular definition was not expected. While some attention has been given to spread in the along- and cross-track directions (e.g., Yamaguchi et al. 2009, Fig. 11), the results presented here do not provide support for the use of such a definition in the work pertaining to the GEFS. As for how these findings may affect our results, we believe that the results presented in this study are robust to changes in the definition of spread, given that differences in the percentage of members enclosed by the spread are small and because trends are similar across all lead times. We suggest that it may be worth exploring how the track errors of an ensemble mean relate to its spread in an elliptical (principal axes) coordinate system.

This exercise is considered to be beyond the scope of this study.

4. Summary and discussion

In this section, we summarize the results of our study, discuss the limitations of these results, and recommend areas for further investigation that may lead to reduced track error and improved consistency of the ensemble.

a. Summary

The NCEP GEFS was found to demonstrate a rather linear increase in track error with increasing lead time in the Atlantic. This rate is about 50 n mi day^{-1} but increases abruptly at 144-h lead time to 150 n mi day^{-1} when including extratropical transition (ET) tracks. When considering only tropical cyclone (TC) tracks, the rate of increase of track error is 100 n mi day^{-1} at these longer lead times. In the western North Pacific basin, average absolute track error increases at a rate of 60 n mi day^{-1} at early lead times, and gradually increases to 150 n mi day^{-1} at later lead times when including ET tracks. This value is 130 n mi day^{-1} when considering only TC tracks.

The GEFS ensemble mean was found to display a slight left-of-track bias at early lead times in the Atlantic (approximately 30 n mi at 72 h) while there is little or no bias at these lead times in the western North Pacific basin. This is true when including and excluding ET tracks. However, at longer lead times when including ET tracks, the ensemble is biased to the left of the observed tracks in the Atlantic and to the right of the observed tracks in the western North Pacific. Consideration of TC tracks alone was found to eliminate this bias in the Atlantic, but not in the western North Pacific. All forecast cyclones were found to lie behind observed cyclones at longer lead times in both basins, suggesting that the ensemble is slow to recurve storms into the westerlies.

Homogeneous comparison of the GEFS mean, GEFS control, and GFS deterministic forecasts reveals greater accuracy of the GEFS mean in the Atlantic basin, while there is little increase in accuracy relative to the GFS forecast in the western North Pacific basin. The greater accuracy of the GEFS mean over the GFS in the Atlantic basin takes place at 96 h, and continues for lead times beyond this time.

Ensemble consistency was explored using dispersion diagrams, the missing rate error (MRE), and the probability within spread (PWS). Dispersion diagrams suggest the spread of the ensemble is approximately appropriate in the Atlantic basin, having the square root of the average ensemble variance (root VAR) close to the expected value. There is some evidence of underdispersion at 0–96-h lead time. In the western North Pacific basin, the consistency of the ensemble is poor: root VAR is less than the expected value for all lead times, and nearly one-half the expected value for lead times beyond 96 h. Such a difference in ensemble consistency across the two basins is an important finding of this study. High MRE and low PWS reveal similar aspects of ensemble consistency in the western North Pacific.

b. Limitations

One of the limitations of the results is the small number of cyclones considered in the study. In the Atlantic, results at longer lead times are dominated by cyclones in the 2006 season, which undergo extratropical transition and are, therefore, characteristic of a certain type of cyclone. Results in the western North Pacific basin are derived from a much larger dataset, but even so, cyclones are limited to the 2006–08 seasons. We acknowledge, therefore, that results from this study may not be representative of the ensemble in other years.

c. Potential areas for improving GEFS forecast skill

The analysis of along- and cross-track errors when extratropical tracks are included in the Atlantic suggests that the GEFS has difficulty modeling the transition of cyclones from tropical to extratropical or that it may have difficulty predicting the evolution of large-scale synoptic features that are responsible for steering TCs into the midlatitudes. Left-of-track bias in the Atlantic and a tendency of forecasts to lie behind observed cyclones hint at such a conclusion. In the western North Pacific basin, a slightly different relationship exists. Forecast cyclones are located behind and to the right of observed cyclones at longer lead times.

In a review article summarizing the current level of understanding of transitioning cyclones, Jones et al. (2003) outline difficulties associated with the numerical prediction of extratropical transitions. The migration of tropical cyclones into regions of often drier air, the interaction of cyclones with land and cooler waters, the potential reintensification of cyclones when interacting with midlatitude systems, and asymmetries that develop in the wind field can contribute to significant challenges in modeling transitioning cyclones. The authors further point out that numerical models, often the primary guidance available to operational forecasters, are limited in their ability to resolve the small-scale processes necessary in tropical cyclone prediction, while at the same time accurately depicting midlatitude systems into which these cyclones move. Payne et al. (2007) examined four operational models in use at the JTWC that provide guidance to forecasters predicting TC tracks in the western North Pacific basin. The motivation of the study was to explore

the usefulness of the selective consensus track. One of the four models considered in the study was the GFS. Payne et al. found that large errors in the GFS deterministic forecast of tropical cyclones in the 2005 season resulted mainly from a poor response to the vertical wind shear. They state that, ''95% of large 96- and 120-h track forecast errors were due to an incorrect depiction of the vertical structure of the vortex.'' GFS vortices were consistently weaker than the observed vortices, allowing the environmental vertical wind shear to dominate cyclone translation and, in some cases, dissipation. The underlying model for this type of error is thought to be an erroneous decoupling of upper-level winds from the lower-level winds (Carr and Elsberry 2000; Payne et al. 2007). Payne et al. (2007) note that this type of error was not present in the regional models studied.

In 2005, the GFS deterministic forecast was run at T382L64 within the first 180 h (Campana et al. 2009). Given that the GEFS uses the same model as the GFS, but at lower resolution (T126L28), it follows that problems described by Payne et al. (2007) are present in the GEFS, as well. While not shown, similarities in bias revealed by homogeneous comparisons of average along- and cross-track errors among the GEFS mean, GEFS control, and GFS deterministic forecasts point to this being the case. It may be beneficial to run the GEFS at increased model resolution for several cyclones. One could select two or three tropical cyclones within each basin that undergo extratropical transition, and examine the track error and dispersion characteristics resulting from this change. Also, it would be useful to look at the vertical structure of vortices, to determine if structures are ''resisting'' the vertical wind shear, as described by Payne et al. (2007).

Another trait of the ensemble that should receive attention is the low consistency of the GEFS in the western North Pacific basin. Given the manner in which perturbations are bred from model integrations (Wei et al. 2008), it is likely that errors in the model translate to errors in perturbations applied to the initial conditions. To this end, it may be helpful to compare perturbed analysis fields of the GEFS with those of other EPSs to determine if perturbations are appropriate for this basin. Also, we note that the ensemble is slightly underdispersive at early (0–96 h) lead times in the Atlantic basin, as revealed by dispersion diagrams (Figs. 8a and 9a), lower PWSs (Figs. 8c and 9c, Atlantic), and 2007–08 MRE values (Figs. 8b and 9b, Atlantic). This merits additional attention.

On 23 February 2010, the GEFS was upgraded to have increased horizontal resolution (T190), effectively moving from 105- to 70-km resolution (McClung 2009). The number of vertical levels was unchanged. Also, a stochastic perturbation package was introduced to account for model uncertainty, which is expected to increase the spread of the ensemble. These changes will need to be monitored to assess how they impact the track forecasts and dispersion characteristics of the GEFS. Since the GEFS is one of several global EPS in the THORPEX Interactive Grand Global Ensemble (TIGGE), it will be important to understand the limitations and tendencies of the ensemble in interpreting the performance of the TIGGE. Scientists have conducted preliminary research in this area (e.g., Park et al. 2008), but much work remains. Also, an exciting prospect is the use of multimodel, global ensembles for cyclone track forecasting.

Acknowledgments. We thank Zoltan Toth, Dick Wobus, Yuejian Zhu, Edward Fukada, James Franklin, Buck Sampson, Rich Yablonsky, Chris Sisko, Max Mayfield, Dave Hebert, and Dave Ullman for helpful discussions during the course of this study. Particular thanks are given to three anonymous reviewers who significantly improved the content of the manuscript. This research was supported by a grant from WeatherPredict Consulting, Inc., an affiliate of RenaissanceRe. IG and TM are grateful for additional support provided by NOAA Hurricane Forecast Improvement Project (HFIP).

APPENDIX

Estimating Confidence Intervals Using Bootstrap Methods

a. Background

Efron (1979) introduced a technique to estimate statistical parameters from a set of data when limited numbers of samples are present. The technique, known as the bootstrap method, generates multiple datasets from the available data by selecting random samples with replacement, allowing one to estimate the statistical parameters regardless of the distribution of the underlying data. One proceeds to estimate a statistic from each of these synthesized datasets. If the statistic is the mean, one concludes that the statistic is normally distributed (Rice 1995). In this manner, confidence intervals can be estimated for the average absolute and along- and cross-track errors, as well as other quantities. In practice, however, the bias-corrected and accelerated (BC_a) method provides a more reliable estimate than standard normal theory and has been used in this study. For more information, see Efron and Tibshirani (1993, chapters 13–14).

Unfortunately, the bootstrap method assumes that the samples are statistically independent. In this study,

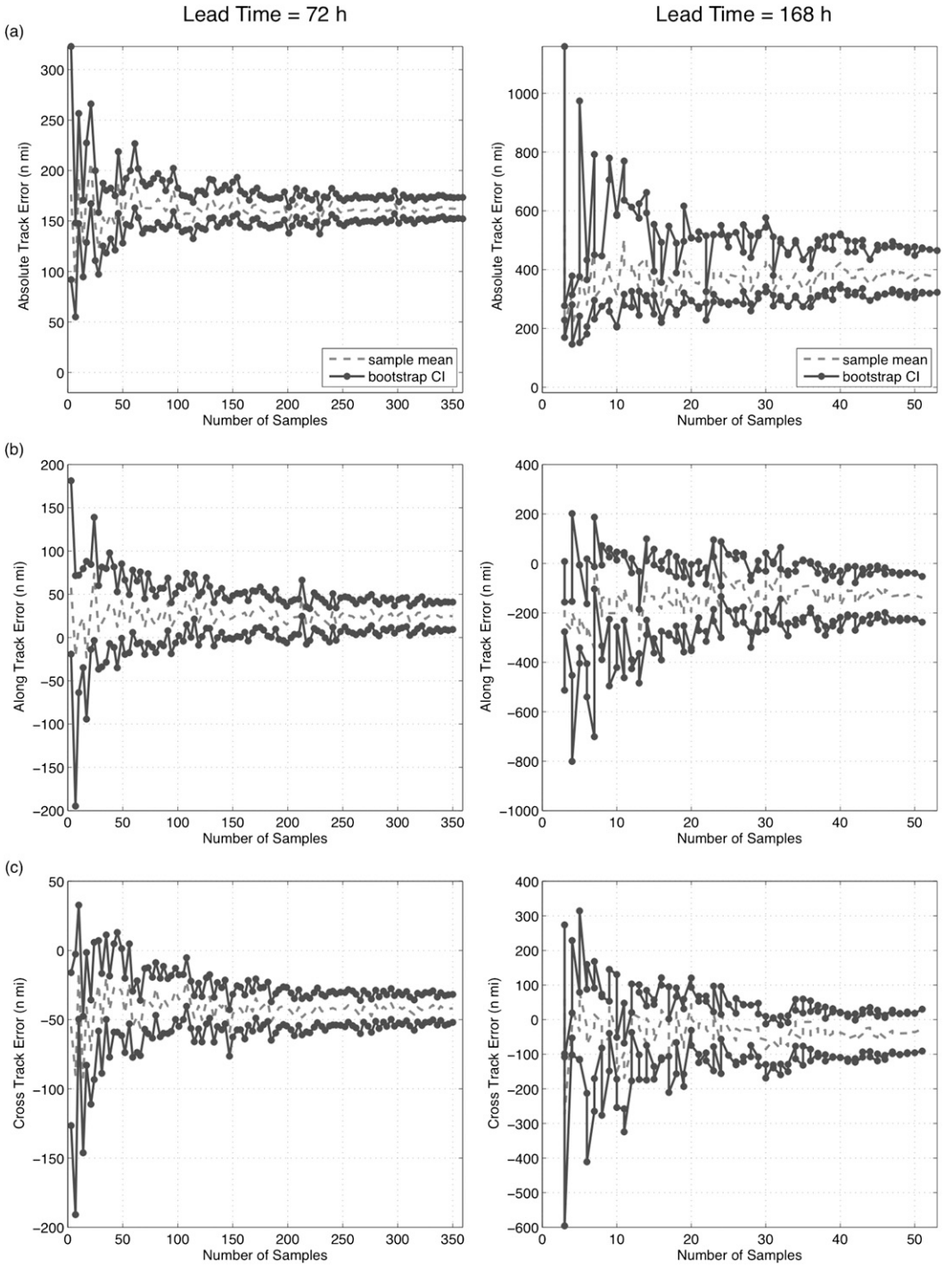

FIG. A1. Illustration of the dependence of confidence intervals on the number of samples used in the bootstrap method. Mean values and confidence intervals for the (a) absolute, (b) along-track, and (c) cross-track errors of the GEFS mean at 72- and 168-h lead times. Only TC tracks are considered in the analysis. Convergence of confidence intervals occurs when 50 or more samples are used.

samples of track error, variance, MRE, and probability within spread are grouped according to lead time. Since forecasts are issued 6 h apart, it is very likely that samples within a given lead time are correlated. A traditional method of handling dependence among samples within a data record is to find the decorrelation time scale using the autocorrelation function and subsample the data, selecting only samples separated by an amount equal to or greater than this time scale (Emery and Thompson 2001). Due to limited sample size, an alternative method has been developed. All samples were used and the sensitivity of the confidence intervals to the number of samples going into the bootstrap estimate was examined. This is illustrated in the next section, and it is suggested that 50 or more cases is a sufficient criteria to obtain good confidence intervals.

b. *The number of samples required for the convergence of confidence intervals*

The confidence intervals of the average absolute and along- and cross-track errors of the NCEP GEFS mean in the Atlantic, excluding extratropical cyclones, were estimated. In Fig. A1, bootstrap estimates of confidence intervals for two lead times are shown. Plots at 72-h lead time illustrate the convergence of the confidence intervals with increasing sample number when a large number of samples are available. Calculations at 168-h lead time demonstrate the bootstrap method applied to the most limiting case—the dataset and lead time with the fewest number of cases. The number of samples used in the bootstrap estimate was progressively increased from three to the maximum number of samples available, and confidence intervals were estimated at the 95% level. Samples were chosen at random. At 72-h lead time the maximum number of samples is 359 while at 168-h lead time this number is 53.

Looking at confidence intervals of average track error at 72-h lead time, one observes that values vary significantly within the first 25 samples but nominally approach an equilibrium when 50 or more samples are included. The confidence intervals of the average track errors when more than 50 samples are included are characterized by standard deviations of about 9.6, 10.3, and 7.7 n mi for the absolute, along-, and cross-track errors, respectively. We believe these standard deviations are tolerable for the present study. Figure A1 also illustrates the variation of the confidence intervals for 168-h lead time. Inspection of confidence intervals of average absolute track error when 0–20 samples are included reveals large variability in the upper bound. This may be simply an artifact of the bootstrap method since it has the potential to replicate outlying values in synthesized datasets and, consequently, to bias the confidence intervals. However, all average track errors show convergence beyond 32 samples.

Thus, while associated with some variability, confidence intervals estimated using the bootstrap method provide approximate bounds on average track errors, despite the correlation between samples. For the present data, plots of the confidence intervals show convergence of intervals when at least 50 samples are present. Thus, in this study, datasets and lead times for which less than 50 cases exist are not considered. This method of estimating confidence levels was also applied to the variance, missing rate error, and probability within spread.

REFERENCES

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.,* **3,** 396–409.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Campana, K., and Coauthors, cited 2009: Technical Procedures Bulletin for the T382 Global Forecast System. [Available online at http://www.emc.ncep.noaa.gov/gc_wmb/Documentation/TPBoct05/T382.TPB.FINAL.htm.]

Carr, L. E., III, and R. L. Elsberry, 2000: Dynamical tropical cyclone track forecast errors. Part II: Midlatitude circulation influences. *Wea. Forecasting,* **15,** 662–681.

Eckel, F. A., and C. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Efron, B., 1979: Bootstrap methods: Another look at the jackknife. *Ann. Stat.,* **7,** 1–26.

——, and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. CRC Press, 436 pp.

Elsberry, R., and L. E. Carr III, 2000: Consensus of dynamical tropical cyclone track forecasts—Errors versus spread. *Mon. Wea. Rev.,* **128,** 4131–4138.

Emery, W. J., and R. E. Thompson, 2001: *Data Analysis Methods in Physical Oceanography*. 2nd. ed. Elsevier, 638 pp.

Froude, L. S. R., L. Bengtsson, and K. Hodges, 2007: The prediction of extratropical storm tracks by the ECMWF and NCEP ensemble prediction systems. *Mon. Wea. Rev.,* **135,** 2545–2567.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.,* **128,** 1187–1193.

——, 2007: Prediction of consensus tropical cyclone track forecast error. *Mon. Wea. Rev.,* **135,** 1985–1993.

Jones, S. C., and Coauthors, 2003: The extratropical transition of tropical cyclones: Forecast challenges, current understanding, and future directions. *Wea. Forecasting,* **18,** 1052–1092.

Marchok, T., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology,* San Diego, CA, Amer. Meteor. Soc., P1.13. [Available online at http://ams.confex.com/ams/pdfpapers/37628.pdf.]

McClung, T., 2009: Amended Date: Global Ensemble Forecast System/North American Ensemble Forecast System changes: Effective February 23, 2010. NWS Tech. Implementation

Notice 09-34. [Available online at http://www.weather.gov/os/notification/tin09-34aab_gefs.txt.]

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.,* **134,** 2029–2050.

Payne, K. A., R. L. Elseberry, and M. A. Boothe, 2007: Assessment of western North Pacific 96- and 120-h track guidance and present forecastability. *Wea. Forecasting,* **22,** 1003–1015.

Rappaport, E. N., and Coauthors, 2009: Advances and challenges at the National Hurricane Center. *Wea. Forecasting,* **24,** 395–419.

Rice, J. A., 1995: *Mathematical Statistics and Data Analysis*. 2nd ed. Duxbury Press, 602 pp.

Sampson, C. R., J. A. Knaff, and E. M. Fukada, 2007: Operational evaluation of a selective consensus in the western North Pacific basin. *Wea. Forecasting,* **22,** 671–675.

Talagrand, O., R. Vautrard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus,* **60A,** 60–79.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Williford, C. E., T. N. Krishnamurti, R. Correa Torres, S. Cocke, Z. Christidis, and T. S. Vijaya Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.,* **131,** 1878–1894.

Yamaguchi, M., R. Sakai, M. Kyoda, T. Komori, and T. Kadowaki, 2009: Typhoon ensemble prediction system developed at the Japan Meteorological Agency. *Mon. Wea. Rev.,* **137,** 2592–2604.

Zhang, Z., and T. N. Krishnamurti, 1997: Ensemble forecasting of hurricane tracks. *Bull. Amer. Meteor. Soc.,* **78,** 2785–2795.

——, and ——, 1999: A perturbation method for hurricane ensemble predictions. *Mon. Wea. Rev.,* **127,** 447–469.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus,* **52A,** 280–299.