Evaluation of Cool-Season Extratropical Cyclones in a Multimodel Ensemble for Eastern North America and the Western Atlantic Ocean

NATHAN G. KORFE AND BRIAN A. COLLE

School of Marine and Atmospheric Sciences, Stony Brook University, State University of New York, Stony Brook, New York

(Manuscript received 19 March 2017, in final form 23 October 2017)

ABSTRACT

This paper evaluates the extratropical cyclones within three operational global ensembles [the 20-member Canadian Meteorological Centre (CMC), 20-member National Centers for Environmental Prediction (NCEP), and 50-member European Centre for Medium-Range Weather Forecasts (ECMWF)]. The day-0–6 forecasts were evaluated over the eastern United States and western Atlantic for the 2007–15 cool seasons (October–March) using the ECMWF's ERA-Interim dataset as the verifying analysis. The Hodges cyclone-tracking scheme was used to track cyclones using 6-h mean sea level pressure (MSLP) data. For lead times less than 72 h, the NCEP and ECMWF ensembles have comparable mean absolute errors in cyclone intensity and track, while the CMC errors are larger. For days 4–6 ECMWF has 12–18 and 24–30 h more accuracy for cyclone intensity than NCEP and CMC, respectively. All ensembles underpredict relatively deep cyclones in the medium range, with one area near the Gulf Stream. CMC, NCEP, and ECMWF all have a slow along-track bias that is significant from 24 to 90 h, and they have a left-of-track bias from 120 to 144 h. ECMWF has greater probabilistic skill for intensity and track than CMC and NCEP, while the 90-member multimodel ensemble (NCEP + CMC + ECMWF) has more probabilistic skill than any single ensemble. During the medium range, the ECMWF + NCEP + CMC multimodel ensemble has the fewest cases (1.9%, 1.8%, and 1.0%) outside the envelope compared to ECMWF (5.6%, 5.2%, and 4.1%) and NCEP (13.7%, 10.6%, and 11.0%) for cyclone intensity and along- and cross-track positions.

1. Introduction

a. Background

Extratropical cyclones during the cool season impact millions of people. Along the U.S. East Coast heavy snow, mixed precipitation, and damaging winds can interrupt and halt public services and transportation over this region. For example, a nationwide survey found that unfavorable driving conditions during winter storms account for approximately 3000 deaths and 1.4 million accidents every year across the continental United States (Goodwin 2003). Numerous studies have concluded that improved forecasts of the timing and location of winter storms can decrease traffic volume and improve public safety during these cases (Hanbali and Kuemmel 1993; Knapp et al. 2000).

During the cool season months (October–March), extratropical cyclones frequently develop near the U.S. East Coast as approaching upper-level disturbances amplify through interaction with a low-level baroclinic zone (Miller 1946). For example, a coastal cyclone on 8–9 February 2013 resulted in up to 1 m of snowfall along the coastal areas of New York and Connecticut (Picca et al. 2014). This cyclone was relatively well forecast 2–3 days in advance, with the ECMWF model (but not other models) providing key information into the cyclone's development 6 days in advance (R. Grumm 2013, personal communication). Other cases, such as the surprise 25 January 2000 cyclone (Zhang et al. 2002) along the mid-Atlantic coast, had considerable forecast uncertainty from the medium range (days 4–6) to the short range (days 1–3). Research focusing on improving the short- and medium-range forecasts of extratropical cyclones will increase public safety and aid emergency managers in planning their responses to such events.

b. Ensemble cyclone forecast skill

Several studies have focused on the tracking and verification of ensemble prediction systems (EPSs) to help improve our understanding of cyclone forecast skill (Froude et al. 2007; Charles and Colle 2009; Froude 2010).

Corresponding author: Dr. Brian A. Colle, brian.colle@stonybrook.edu

DOI: 10.1175/WAF-D-17-0036.1

Charles and Colle (2009) verified the intensity and position of cyclones around North America and its adjacent oceans using the National Weather Service (NWS) Short-Range Ensemble Forecast system (SREF) for the 2004–07 cool seasons (October–March). The 15-member SREF mean provided a better overall forecast than its various subgroups for cyclone displacement and cyclone central pressure along the East Coast and the western Atlantic. The SREF intensity bias along the U.S East Coast from hours 3–15 is slightly negative (~0.5 hPa), but there exists a gradual positive trend for the rest of the forecast with no bias from hours 45 to 51, and a slight positive bias noted from hours 57 to 63. The SREF was also found to have slightly greater skill than the blended Global Forecast System (GFS) and the North American Mesoscale Forecast System (NAM) for central pressure; however, the SREF was found to be overdispersed on average, especially early in the forecast. It was also determined that using a multimodel ensemble of the GFS, NAM, and SREF mean did not add value to the forecasts of cyclone central pressure compared to the individual multimodel ensemble components as a result of outlier members or a skewed distribution around the observed pressure value.

Froude et al. (2007) verified extratropical cyclone tracks in the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble and the Global Ensemble Forecast System (GEFS) run at the National Centers for Environmental Prediction (NCEP) between January and April 2005. The ECMWF EPS consisted of 50 perturbed members with a spectral resolution of T255L40, while the GEFS EPS (hereafter referred to as NCEP) consisted of 10 perturbed members with a resolution of T126L28. Because of the lower resolution of the EPSs during this period and poor confidence in the models, the authors excluded cyclones that developed after day 3 of the forecast period. The results showed that the ECMWF ensemble had a slightly higher level of accuracy than the NCEP ensemble for cyclones in the Northern Hemisphere (NH), while in the Southern Hemisphere (SH) the NCEP ensemble had significantly better accuracy for cyclone intensity. Overall, the ECMWF ensemble mean had greater skill than its control member however, after day 3 of the forecast the control member forecast had an additional 12–24 h of accuracy compared with the perturbed members.

Froude (2010) conducted a detailed study using nine different EPSs from the TIGGE archive to analyze the prediction of extratropical cyclones in the NH for a 6-month period (February–July) in 2008. The cyclones were identified and tracked using the 850-hPa relative vorticity and verified using the ECMWF operational analysis. Results show that the ensemble mean errors for cyclone position, intensity, and propagation speed were the lowest in the ECMWF ensemble. The NCEP and Canadian Meteorological Centre (CMC) ensembles were shown to have 1 day less of accuracy for the position of cyclones throughout the 7-day forecast range. Froude (2010) also found that the ensemble mean provides an advantage over the control member for all EPSs in cyclone position; however, cyclones were found to propagate too slowly in all EPSs and control forecasts. Additionally, a majority of the EPSs underpredict cyclone intensity, excluding ECMWF, early in the forecast from days 1 to 3. All the EPSs are more underdispersive for cyclone intensity than cyclone position, as was also recognized by Froude et al. (2007), with ECMWF performing best for intensity and CMC performing best for propagation speed.

c. *Motivation*

A survey by Novak et al. (2008) indicates that many operational forecasters in the NWS want to better utilize ensembles to forecast high-impact weather such as coastal storms, but more information on how these ensembles perform is warranted so the forecast uncertainty can be better assessed. Furthermore, there has been limited research on the verification of extratropical cyclones using different EPSs for the U.S. East Coast and western Atlantic after 2010, as well as limited determination of the benefits of using multimodel ensembles to forecast extratropical cyclones. This analysis is imperative to East Coast forecasters as they attempt to understand the track and evolution of these cyclones and predict the sensible weather impacts that are associated with these storms. Charles and Colle (2009) focused on verifying a short-range EPS system, and Froude (2010) binned all NH cyclones together to conduct verification over a short 6-month period. There has not been adequate research done to verify medium-range ensemble forecasts of cool-season cyclones, especially on a regional scale using multiple EPSs used by operational forecasters. This paper will address this by completing a multiyear assessment of ensemble cyclone errors across the U.S. East Coast and western Atlantic for the day-0–6 forecast period.

There have been few studies analyzing the forecasts of U.S. East Coast cyclones using the large and comprehensive TIGGE archive for the purposes of conducting a long-term verification of ensemble-predicted cyclones. A long-term ensemble verification dataset of cyclones will aid in our knowledge of the predictability of cyclone strength and position over different forecast periods for successive ensemble upgrades and improvements. This will allow for interannual comparisons of EPS

performance and also the ability to identify potential short- and medium-range biases.

This research will address the following questions:

1) What are the cyclone position and intensity forecast errors in cool-season extratropical cyclone forecasts from days 1 to 6 in operational ensembles?
2) How do cyclone position and intensity errors and biases vary spatially across the U.S. East Coast and western Atlantic Ocean?
3) Have cyclone forecasts in the short and medium range improved over the past several years?
4) Is there any probabilistic skill for cyclone intensity and position in the operational ensembles?
5) Can a multimodel ensemble help improve the deterministic and probabilistic skill scores compared to a single model ensemble?

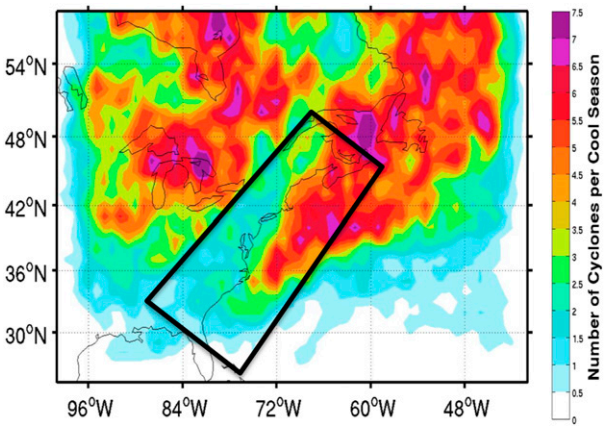

FIG. 1. Cyclone track density for the ERA-Interim analyses showing the number of cyclones per cool season (October–March) per 50 000 km^2 for 2007–15. The black box shows the verification domain used for cyclones passing within this box during the forecast.

2. Data and methods

a. Data description

The mean sea level pressure (MSLP) ensemble forecast and analysis data were obtained for the 0000 UTC cycle from the TIGGE (Bougeault et al. 2010) archive (http://apps.ecmwf.int/datasets/data/tigge/) in 6-hourly increments from 1 October 2007 to 31 March 2015 for the cool-season period (1 October–31 March). The 2007 start time was selected because it is the first date when the full ensemble datasets are available. This study examines the cyclone performance of three operationally employed EPSs: the 20-member CMC, 20-member NCEP, and 50-member ECMWF. The control member data of each EPS were used for comparison with the ensemble mean and to check for consistency with previous studies. In addition to single-model ensemble performance, 40-member and 90-member multimodel ensemble blends (NCEP + CMC and NCEP + CMC + ECMWF, respectively) were analyzed by combining the forecasts of the individual EPSs. The combination of CMC + NCEP is often referred to as the North American Ensemble Forecast System (NAEFS).

The ERA-Interim reanalysis (Uppala et al. 2005) at T255 spectral grid resolution and 60 vertical levels was used to verify and evaluate cyclone properties for each EPS. Previous studies have also used ECMWF analyses to verify cyclones (Froude et al. 2007; Froude 2010). The GFS analyses were also utilized in our study in order to assess the sensitivity of the results to different analyses, but little sensitivity (<5%) was found past hour 12 of the forecast (Korfe 2016). To allow the use of an automated cyclone-tracking scheme on a latitude–longitude grid, all the analysis and ensemble data were linearly interpolated onto a 1.0° latitude–longitude grid using a simple 2D spatial interpolation technique. This resolution allowed for efficient use of the long-term ensemble dataset and a fair comparison of results for each EPS.

b. Cyclone verification approach

The details pertaining to the cyclone-tracking process and cyclone-matching methodology are described in this section. The following cyclone-tracking approach closely resembles that of Colle et al. (2013), and the verification approach closely resembles that of research conducted by Froude (2010), with some noted changes to better match forecast cyclones in the medium range.

1) CYCLONE TRACKING

The Hodges cyclone-tracking scheme (Hodges 1994, 1995) was used to obtain the cyclone tracks by tracking 6-hourly anomaly MSLP data from the models and analyses. The domain for the study is from 25° to 65°N and 100° to 40°W over eastern North America and the western Atlantic, with the East Coast and west Atlantic (ECWA) box embedded for identifying U.S. East Coast cyclones (Fig. 1). Only cyclones that pass through the ECWA box were analyzed unless otherwise noted. All EPSs were tracked using the 0000 UTC model cycle only for the day-0–6 forecasts using 6-hourly output increments. The following data-processing and cyclone-tracking methods using MSLP data has been utilized in other studies (Hoskins and Hodges 2002; Colle et al. 2013). First, the data were preprocessed using a spectral bandpass filter (Anderson et al. 2003). The planetary scales with wavenumbers below 6 are removed from the dataset, as well as wavenumbers above 70 for easier

cyclone detection. Similar to Colle et al. (2013), the cyclone-tracking scheme was set to retain only cyclones that lasted a minimum of four time steps (24 h) and traveled at least 1000 km. To allow for more accurate feature point detection, the Hodges approach tracks the anomaly pressure, the MSLP field after it has undergone a discrete cosine transform. Consequently, to find the minimum MSLP values for each cyclone, the lowest pressure within 5° of the identified pressure anomaly was applied. Colle et al. (2013) also manually evaluated and hand tracked cyclones for 11 months (2286 cyclones) using the same criteria and found that the probability of detection (POD) was ~92%, and the false alarm rate was ~5%, so the uncertainty in the automated tracking results is likely between 5% and 15% of the total number of storms.

Figure 1 shows the average cool-season cyclone density (storms per 50 000 km^2) using the ERA-Interim analyses. There are two clusters of high cyclone track density (more than six cyclones per cool season) extending northward from around the Great Lakes toward Hudson Bay and another over the western Atlantic to the north of the Gulf Stream and extending northward to the North Atlantic. Areas in the North Atlantic tend to be areas of cyclolysis, where cyclones propagate very slowly, occlude, and can persist for days (Hoskins and Hodges 2002).

2) Cyclone-matching methodology

To validate the ensemble forecast cyclone tracks against the analysis cyclone tracks, a systematic approach matched the forecast tracks to the appropriate analysis track using criteria similar to those of Froude (2010). The two tracks must meet predefined spatial and temporal criteria:

1) The pairing distance d of each point in an individual forecast track to each point in the analysis track, which coincides in time with the analysis track, was less than the maximum pairing distance d_{max}, which will be described later. The distance d is calculated at every 6-h time period within the analysis track:

$$d \le d_{max}. \tag{1}$$

2) As we show below, at least T% of the points in the forecast track coincide with the analysis track, where N_A and N_F denote the total number of points in the analysis track and forecast tracks, respectively, and N_M denotes the number of points in the analysis track that coincide in time with the forecast track:

$$100 \times \left(\frac{2N_M}{N_A + N_F}\right) \ge T. \tag{2}$$

Each analysis point has a corresponding storm identification (ID) number of the closest forecast track that satisfies Eq. (1). If the matched forecast ID for a particular analysis cyclone is the same for greater than or equal to T% of the lifetime of the analysis cyclone, the forecast track is matched with the analysis track.

Sensitivity tests were conducted to determine the best values for d_{max} and T, with preferred values determined by matching results and previous studies. The d_{max} value of 1200 km and T value of 60% were used based on this assessment (Korfe 2016). If the d_{max} value, which can be described as a radius around the analysis point, is too large (>1600 km), there will more inaccurate matching of forecast tracks in the medium range or for cases where the model does not identify a cyclone as it should for a particular region. If the d_{max} value is too small (<800 km), valuable forecasts in the medium range that have cyclone formation may be unmatched along the U.S. East Coast because the cyclone position may extend beyond the smaller radius value. Similarly, Froude (2010) used a T value of 60%, but they used a technique in which only the first four track points are compared with the analysis track to determine matching. Conversely, our study takes into account the entire track length to help improve matching during the medium range. After this matching process has concluded, any remaining unmatched analysis (forecast) cyclones were considered misses (false alarms).

The number of ensemble members that match a given analysis track will vary for different forecast start times and different EPSs. Also some forecasts can have different track lengths, which will cause differences in the number of data points for each model. As lead time increases, the number of cyclones matched will decrease; therefore, a minimum number of ensemble members must be identified for an ensemble mean calculation to be performed. Only cases with greater than or equal to 40% of the ensemble members (as in Buckingham et al. 2010) matched in all three EPSs were included. Therefore, NCEP and CMC require 8 minimum members, while ECMWF requires 20 members. This is a 15% increase in membership compared with some previous studies (Froude 2010). The ECMWF + NCEP + CMC ensemble requires all three EPSs to satisfy their matching criteria for a case to be included in the dataset (e.g., 8 + 8+ 20 = 36 members). This approach allows for enough members to be included to calculate probabilistic metrics, but since it does not include events with fewer members tracked, the results in this study may make the cyclones more predictable than in reality.

c. *Ensemble metrics*

The ensemble mean cyclone position can be defined as the average of each perturbation member's forecast position at a given lead time:

$$\mathbf{x}_{\bar{e}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \tag{3}$$

where $\mathbf{x}_i$ represents the position of the ith member of the ensemble, and N is the total number of ensemble members. Equation (3) indicates the ensemble mean is inferred by finding the average position of the perturbation members at a given lead time in vector space. The ensemble spread is defined by Goerss (2000) as simply the average distance of the perturbation members to the ensemble mean.

All distances between track points are calculated using the great-circle distance between two points, as in Froude et al. (2007), such that potential biases caused by projections will be avoided. The track vector is an average of the track vectors during the 6 h prior to and 6 h after the comparison time. The absolute track error is computed for each perturbation member and can be described as the distance between the forecast and observed cyclone positions. The absolute track can be decomposed into the along- and cross-track components. The cross-track error is defined as the error normal to the observed track. The along-track error is then defined as the distance between the observed cyclone position and the intersection of the cross-track line coinciding with the observed cyclone track (Froude et al. 2007). For statistical purposes, the cross-track error is positive (negative) when a cyclone is forecast to the right (left) of the observed cyclone, and the along-track error is positive (negative) when a forecast cyclone is ahead of (behind) the observed cyclone. This allows for the calculation of the forecast bias by taking the average error in the along- and cross-track directions.

Some of the commonly used definitions and metrics used to discuss ensemble performance are utilized in this study, such as mean error (ME), mean absolute error (MAE), Brier score (BS; Wilks 1995), Brier skill score (BSS), and probability within spread (PWS; Buckingham et al. 2010). The ME and MAE scores were calculated using either those selected matched members or the cyclone position and intensity values from the calculated ensemble mean. The BS quantifies some of the probabilistic performance, with its reliability (REL) component evaluating the mean difference between the ensemble's probability forecast and the actual probability forecast. The BSS assesses the probabilistic skill relative to some reference forecast, in which the probability of exceeding a particular threshold can only be 0 or 1. PWS estimates the likelihood of an observed cyclone falling within the dispersion of the ensemble by considering varying distances from the ensemble mean (Buckingham et al. 2010). PWS can be described as

$$\mathrm{PWS} = \frac{1}{N}\sum_{n=1}^{N}\begin{cases} 0: s_{\mathrm{obs}} > k(\sigma)_n \\ 1: s_{\mathrm{obs}} \le k(\sigma)_n \end{cases}, \tag{4}$$

where k is an integer, n is an integer, N is the total number of forecasts at a given lead time, s_{obs} is the distance from the ensemble mean to the observed cyclone, and σ is the spread of the ensemble. Assuming the members are sampled from a normal distribution with standard deviation σ, PWS [Eq. (4)] should have values of 0.68, 0.95, and 0.997 corresponding to 1σ, 2σ, and 3σ.

To test for statistical significance when computing ensemble statistics, a bootstrapping method was utilized to resample the data and determine proper confidence intervals around the ensemble mean errors (Zwiers 1990). The bootstrapping technique makes no assumptions about the overall distribution of the data. To test if two different ensemble means are significantly different at a given lead time, sample values would be drawn and resampled from the original data, allowing repeated selections. This process was completed 1000 times to find the 90th percentile confidence intervals around the ensemble means of the resampled values. If the confidence intervals for two values do not overlap, those values are considered significantly different at the 90% level.

3. Deterministic verification

a. *Cyclone intensity*

Figure 2a shows the MAE for intensity (cyclone central pressure) versus forecast lead time calculated for each individual ensemble member and averaged over the domain region shown in Fig. 1. For the 0–60-h forecast period, ECMWF and NCEP have similar errors for intensity (3.9- and 4.1-hPa errors by 60 h, respectively), while CMC has a larger error (~4.7 hPa at 60 h) that is significantly larger than ECMWF and NCEP at the 90% level. For 84–120-h lead times, the NCEP error growth increases more rapidly than ECMWF, resulting in ECMWF (NCEP) having less (more) error of ~7 hPa (~8 hPa) by 120 h. Overall, ECMWF has the smallest MAEs for intensity after 72 h, while CMC has the largest intensity errors throughout the medium-range period (72–144 h). This results in ECMWF having 12–18 h more lead-time accuracy than NCEP and a 24–30-h advantage over CMC. When analyzing the ensemble mean by taking the average MSLP field using all members and verifying the mean track, the models rank similarly (Fig. 2a), the NCEP + CMC ensemble mean decreases the cyclone intensity MAE by 0.25–0.5 hPa relative to the NCEP mean for 96–144 h (Fig. 2a), and this ensemble blend is more comparable to ECMWF than either NCEP or CMC individually. The

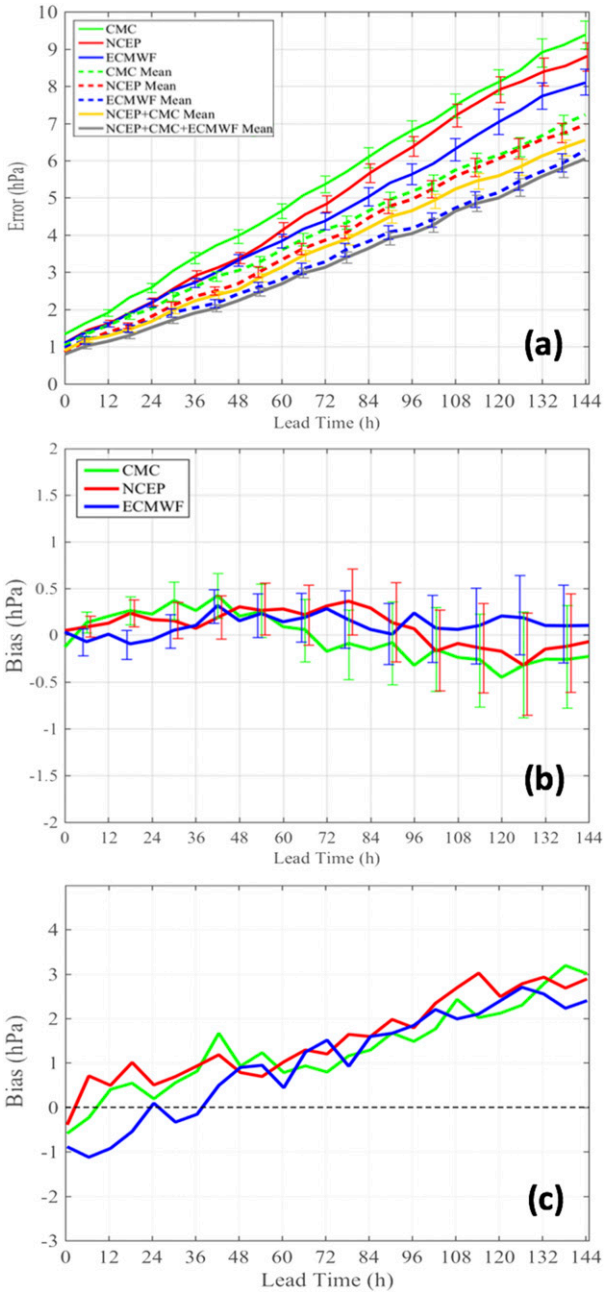

FIG. 2. (a) MAE for cyclone intensity (central pressure) averaged for all individual ensemble members and the ensemble mean. (b) As in (a), but for ME but only for the averaged ensemble members. (c) As in (b), but for relatively deep (greater than one standard deviation) cyclones in the analysis or any ensemble member.

ECMWF mean still has smaller errors than the NCEP mean and CMC mean that is significant at the 95% level starting at hour 48. Also, we explored whether the lower MAEs for ECMWF were the result of ECMWF having 50 members, while there are only 20 members for NCEP and CMC. The MAE was recalculated for ECMWF but using 20 random members, and the MAEs were nearly identical before hour 72 and less than 5% different after hour 72 (not shown). This is consistent with Majumdar and Finocchio (2010), who verified 2008 Atlantic tropical cyclones and found little difference between the results when using 20 random members of ECMWF versus all 50 members.

To assess the model intensity bias, the ME was calculated using the difference in the mean intensity of each ensemble member track and the matched analysis track (Fig. 2b). Thus, a positive bias is associated with a cyclone underprediction. After a small positive bias early in the forecast period, the CMC bias changes from positive to negative (~0.40 hPa) at 60 h, and this pattern persists throughout the medium range. Meanwhile, NCEP develops a slight negative bias (−0.25 hPa) after 96 h. In contrast, ECMWF maintains a small (~0.20 hPa) positive bias throughout the medium range.

For relatively deep cyclone events, in which the cyclone intensity for the ensemble mean forecast or observed cyclones was more than 1.0 standard deviation below the mean cyclone intensity threshold for the U.S. East Coast (mean ~994 hPa; standard deviation ~14 hPa), there is intensity underprediction beginning in the short range (days 1–3) and growing above 2 hPa by the medium range (days 4–6) in all three EPSs (Fig. 2c). These deep cyclone events represent ~22% of the total cases in the ECWA domain. Figure 3 shows the tracks of the NCEP control member with large positive (negative) intensity error above (below) the mean error at 96 h. This member was used since there is no ensemble mean member in the dataset, and each member should share the same systematic pressure bias as the ensemble mean. Cyclones with the largest forecast errors frequently develop along the East Coast, with errors maximizing as the cyclones pass the Gulf of St. Lawrence. For the large positive cyclone errors at 96 h (>7.6 hPa), the cyclone tracks are generally more offshore near the edge of the Gulf Stream and extend toward the northeast, while the 96-h negative pressure error cyclones (<−7.9 hPa) are clustered more over the eastern Great Lakes and near Newfoundland. This would imply that the model struggles at times to produce the necessary cyclonic amplification offshore where there are strong low-level baroclinic zones and potentially large surface fluxes, while there is too much cyclone amplification as the cyclones mature over the Great Lakes and Newfoundland regions.

Additional evidence for underdeepening of cyclones offshore can be obtained by comparing the difference in spatial cyclone track density per cool season from the ECMWF control member (Fig. 4a) and NCEP control member (Fig. 4b) day-4–6 forecasts with the ERA-Interim reanalysis in Fig. 1. This approach does not

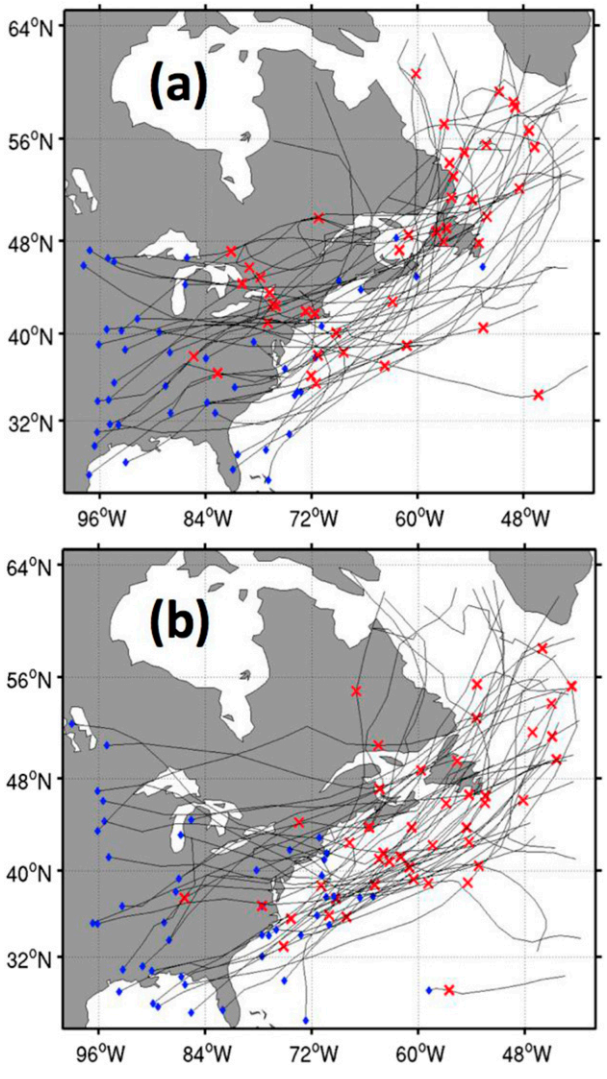

FIG. 3. Tracks of NCEP control member forecasts for cyclones with (a) negative and (b) positive intensity errors at hour 96 that are more than 1.5 standard deviations above (below) the mean error of all cyclones (7.6 hPa for positive error and −7.9 hPa for negative error). The red exes indicate the locations of the cyclones at forecast hour 96 when the large error is occurring.

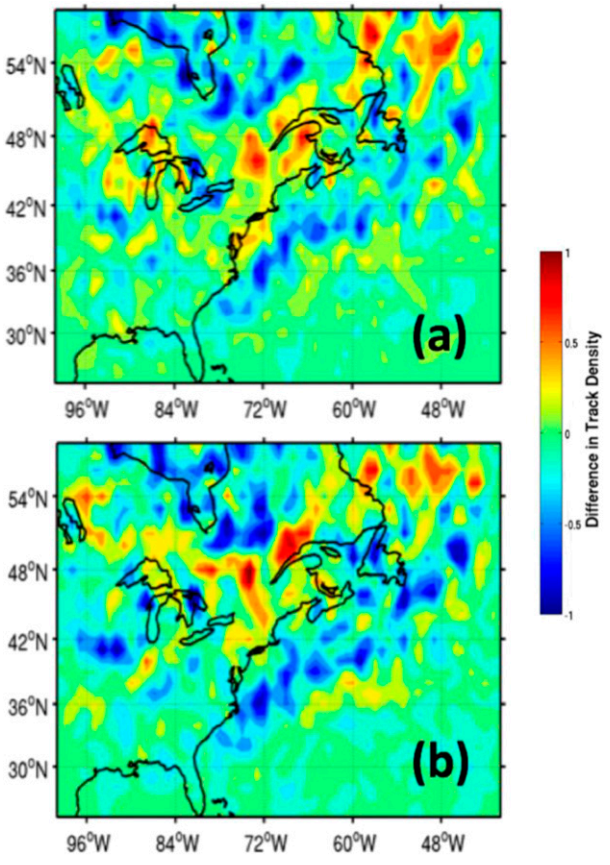

FIG. 4. Difference in cyclone track density for the (a) ECMWF control member and (b) NCEP control member compared with the ERA-Interim analyses showing this per cool season (October–March) per 50 000 km^2 for 2007–15.

require cyclone matching and its associated uncertainties. There is an underprediction of 0.5–1.0 cyclones per cool season (5%–10%) off the U.S. East Coast extending northeastward along the active cyclone track region into the Atlantic. Since this signal is present in both the ECMWF and NCEP control member density fields, the underforecasting of cyclones is a common issue during the medium-range period.

b. *Cyclone track*

Figure 5a shows the MAEs for cyclone displacement versus lead time using each individual ensemble member separated into the along- and cross-track components as described in section 2c. ECMWF has the smallest MAEs for total displacement after 60 h, but only the smaller errors for hours 84–120 are statistically significant compared to the other forecasts. Beginning at 84 h, the CMC absolute track error growth slows and by 144 h CMC has a ~700-km average error that is within the confidence intervals of ECMWF and NCEP (640- and 670-km errors, respectively). The along- and cross-track error evolutions are similar to the total track error evolution. The along-track component contributes more to the total track error than the cross-track component, with nearly ~60% (~55%) of track error at 144 h (72 h). During the short-range period (0–72 h), ECMWF and NCEP have similar along-track errors. From 96 to 144 h, NCEP and CMC have similar along-track errors, while ECMWF has a significantly smaller along-track error.

From 12 to 48 h, the ECMWF cross-track absolute errors are significantly less than CMC (Fig. 5a), but not significantly less than NCEP. From 60 to 144 h, both

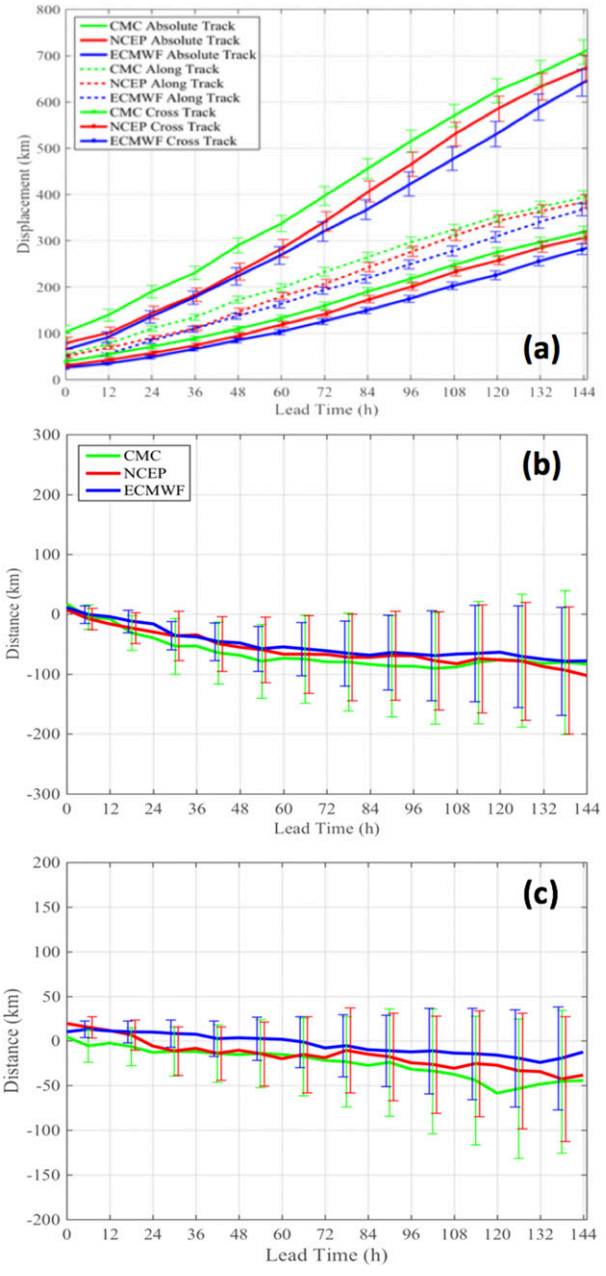

FIG. 5. (a) Average MAE (km) for absolute (total) track and cross- and along-track directions for all members tracked separately and the different ensemble systems (NCEP, CMC, and ECMWF). (b),(c) As in (a), but for ME (km) for the along-track and cross-track directions, respectively. Along-track error 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. Confidence intervals at the 90% significance level are given by the vertical bars.

CMC and NCEP have significantly larger cross-track errors than ECMWF. Cross-track errors at 144 h are 280, 300, and 310 km for ECMWF, NCEP, and CMC, respectively, indicating a smaller error spread among the ensembles during the medium-range period. The combination of the NCEP + CMC ensembles decreases the along-track MAE by ~40 km in the medium range (not shown), so this ensemble blend is more comparable to ECMWF than either the NCEP or CMC ensembles individually. However, the ECMWF + NCEP + CMC and ECMWF ensembles have superior cross-track MAE scores before the medium range, with nearly 30 km less error than the NCEP + CMC ensemble (not shown).

All three EPSs have a negative along-track bias (0–100 km) from past 12 h to the end of the forecast period, which is statistically significantly different than zero from 24 to 90 h (Fig. 5b). This implies that the ensemble cyclones move too slowly on average. The negative along-track bias ranges from 20–75 km during most of the short range (12–72 h) to 50–100 km in the medium range. Because of the large variability in the forecasts, this result is not significant from 96 to 144 h, with the confidence interval ranging from 10 to −200 km at hour 144. Cross-track biases also show a (5–50 km) negative bias (left of track bias) throughout the forecast period for CMC and NCEP (Fig. 5c), while ECMWF has no bias through the short range and a small negative bias (5–30 km) in the medium range. Although not statistically significant, ECMWF, NCEP, and CMC have negative cross-track biases of 25, 35, and 50 km from 120 to 144 h, respectively.

Figure 6 shows the percentage best and worst for short-range (0–72 h) and medium-range (72–144 h) forecasts for the cyclone intensity (hPa), along-track (km), and cross-track (km) errors. For medium-range (72–144 h) cyclone intensity forecasts (Fig. 6), the ECMWF mean is best (worst) approximately 32% (24%) of the time, NCEP 24% (38%), CMC 23% (36%), and NCEP + CMC + ECMWF 12% (1%). While the multimodel ensemble blends, especially the ECMWF + NCEP + CMC ensemble, are not the most likely to be the best forecast out of the five models in the short and medium ranges, they are very rarely the worst EPS and often are the second- or third-best EPS for intensity and along- and cross-track forecasts.

c. *Spatial distribution of errors*

To illustrate the spatial distribution of MEs for cyclone intensity and displacement (along and cross track) for the medium range (72–144 h), the errors were spatially interpolated onto a 1.0° latitude–longitude grid and averaged over the eight cool seasons. A minimum threshold of 20 cyclone data points for each 1° box is required to be included in the spatial display, which is ~2.5 cyclones per cool season per grid point. There represents a 4–6-hPa underdeepening bias in the North

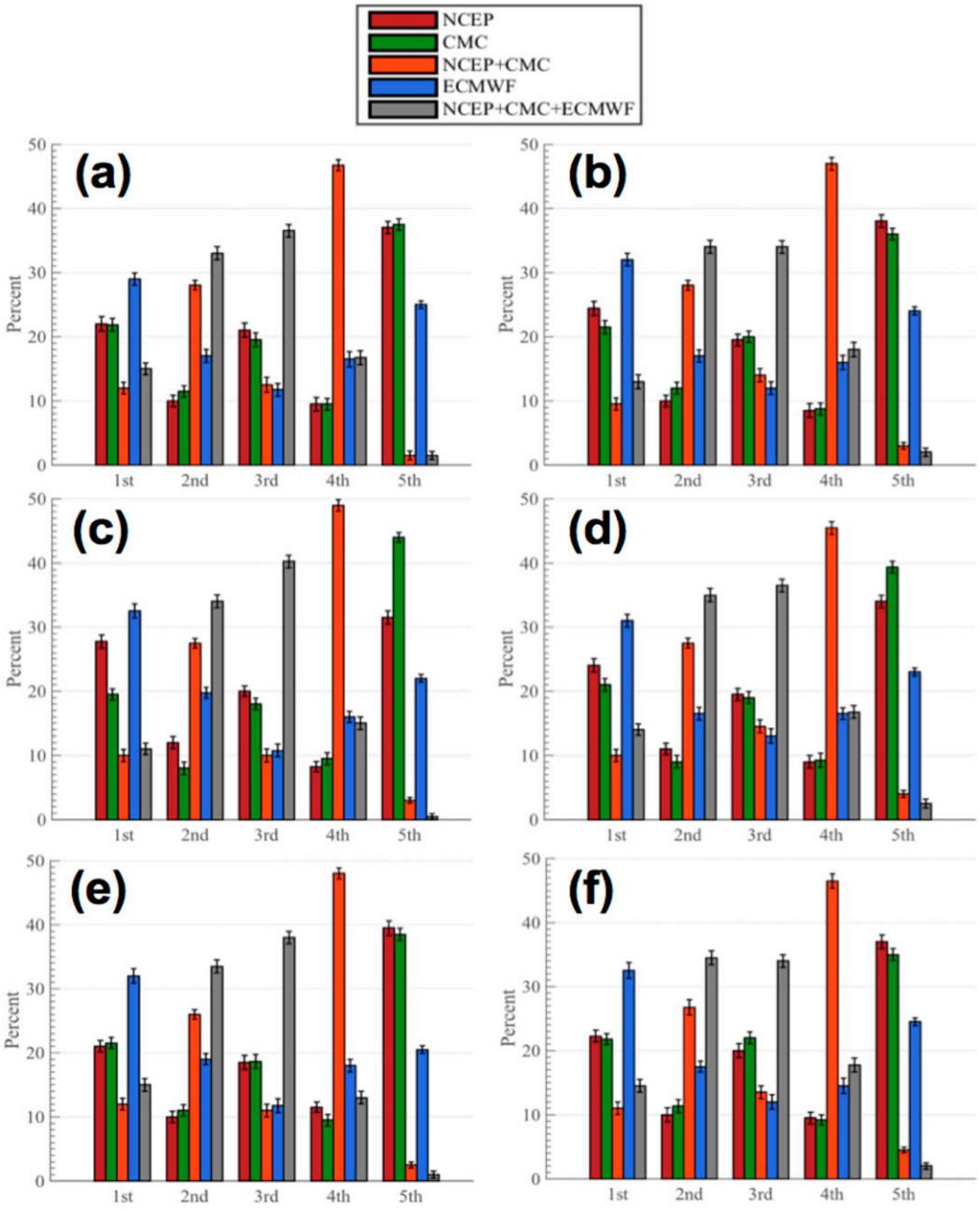

FIG. 6. EPS ensemble mean percentage from best (first) to worst (fifth) grouped bar charts for the short-range (0–72 h) (a) cyclone intensity (hPa), (c) along-track (km), and (e) cross-track (km) forecasts and for the medium-range (72–144 h) (b) cyclone intensity (hPa), (d) along-track (km), and (f) cross-track (km) forecasts. Confidence intervals at the 90% significance level are given by the vertical bars.

Atlantic in all three EPSs (Figs. 7a–c), while the over-deepening along the East Coast is 2–4 hPa. The standard deviation is 5–8 hPa in these areas for all ensembles, with CMC showing larger standard deviation values across the domain.

The spatial distribution of the medium-range along-track error has a well-defined negative (slow) bias across the domain for all EPSs (Fig. 8). Across the western Atlantic and active cyclone track region, there is a 100–150-km slow bias, while across the eastern United States there is a small positive along-track bias (50–100 km) that is more prevalent in NCEP (Fig. 8b). A larger standard deviation of ~150 km is more widespread with CMC (Fig. 8a) than the ~100-km values in ECMWF and

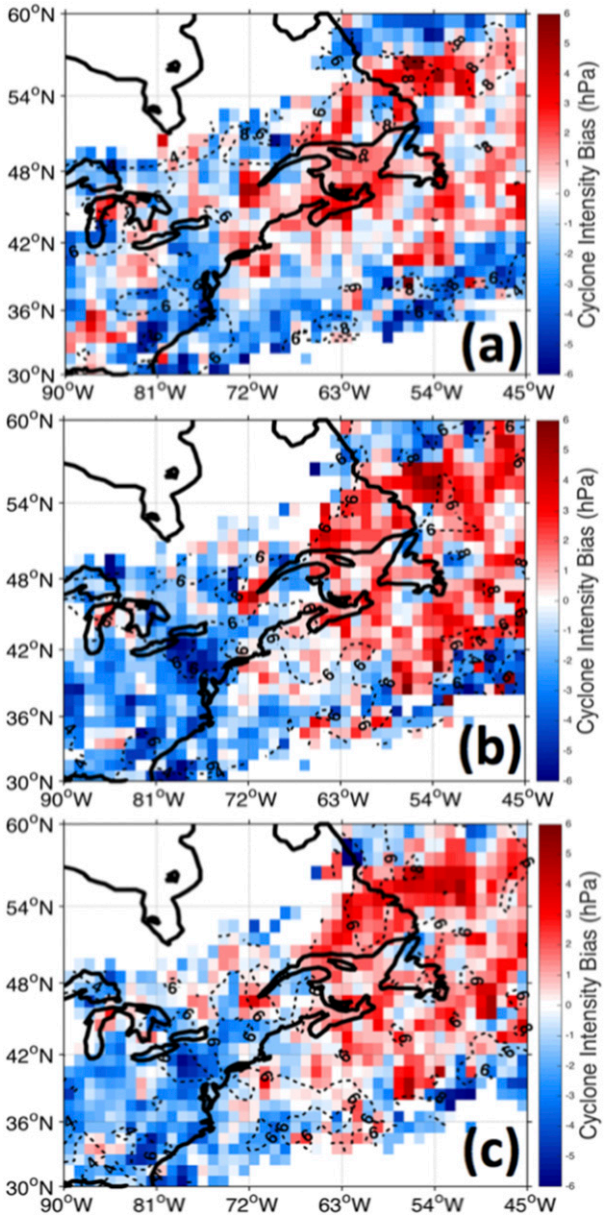

FIG. 7. Medium-range forecast (72–144 h) spatial distribution of cyclone intensity mean error (shaded; hPa) for (a) CMC, (b) NCEP, and (c) ECMWF. Dashed contoured values indicate the standard deviation of the error (every 2 hPa).

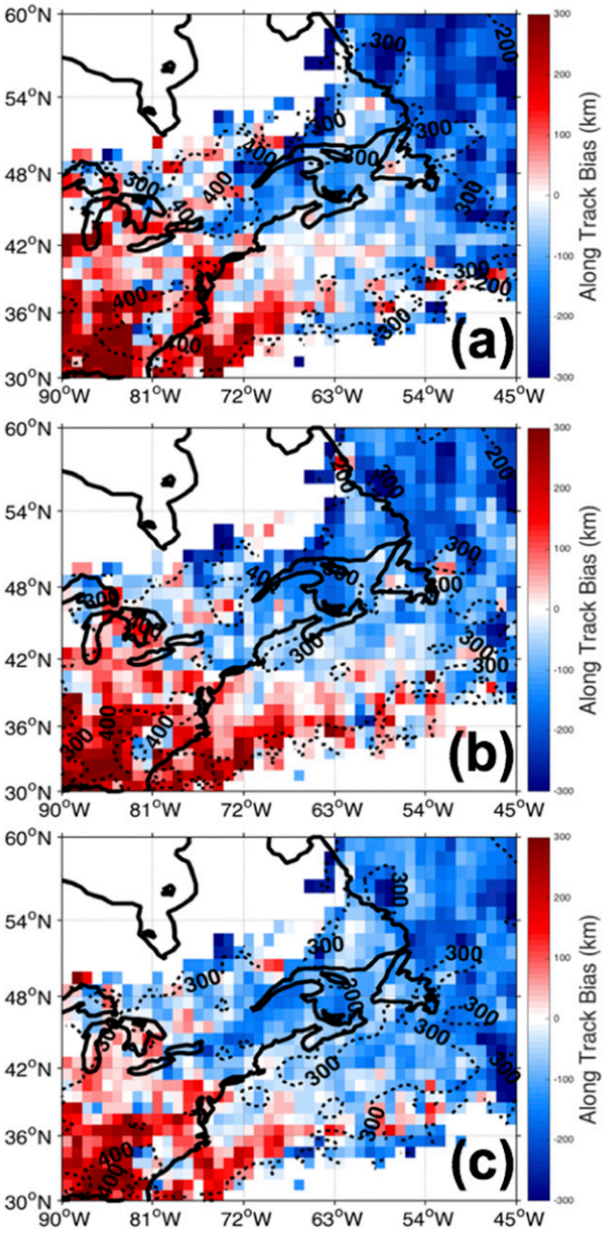

FIG. 8. Medium-range forecast (72–144 h) spatial distribution of along-track mean errors (shaded; km) for (a) CMC, (b) NCEP, and (c) ECMWF. Along-track error is positive (negative) when the forecast lies ahead of (behind) an observed cyclone. Dashed contoured values indicate the standard deviation (dashed every 100 km).

NCEP (Fig. 8c). Over the western Atlantic a negative cross-track bias (100–150 km) exists, suggesting a left-of-track bias for this area (not shown). The largest negative cross-track errors occur over the southeastern United States and have a larger magnitude for NCEP and CMC (250–300 km) compared with ECMWF (200–250 km). It was shown above that these storms are overdeepened over the southeast United States, so it is hypothesized that these storms are slow and too far west since the upper trough is too amplified. Once again, large standard deviation values (200–300 km) over this region indicate a large range of errors.

When comparing the spatial distribution of large biases for intensity and displacement in the medium range, there are some noteworthy overlaps. The large area of underdeepening (~3–5 hPa) in the North

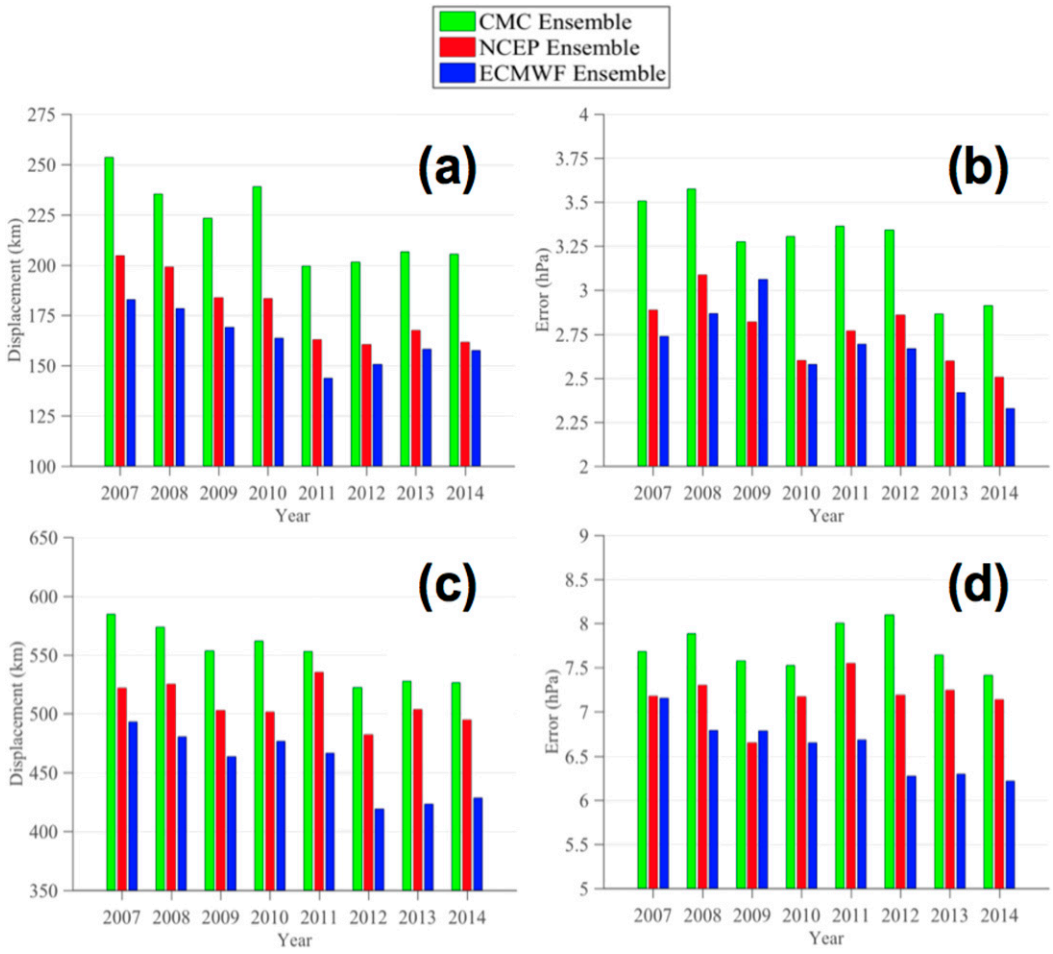

FIG. 9. Individual cool season (2007–15) grouped bar charts of short-range (0–72 h) cyclone (a) intensity (hPa) and (b) displacement (km) MAEs and medium-range (72–144 h) cyclone (c) intensity (hPa) and (d) displacement (km) MAEs for CMC, NCEP, and ECMWF.

Atlantic (Fig. 7) is also an area of large negative along-track bias (~200–250 km) (Fig. 8). Large discrepancies in model and observed cyclones intensity could be occurring because the forecast cyclones are propagating too slowly. Weaker synoptic-scale flow and associated weaker baroclinicity with the system could cause this underdeepening and slow bias. Since this is a common region for rapid cyclone intensification, this result requires further investigation of the physical processes. Additionally, large areas of overdeepening bias (~2–4 hPa) in the eastern United States coincide with areas of large positive along-track error (~250–300 km) and moderate negative cross-track errors (~200 km).

d. Intraseasonal cyclone errors

To determine the change in yearly cyclone intensity and displacement MAEs, the errors were binned into short- and medium-range groups and displayed as grouped yearly bar charts (Fig. 9). Each EPS shows a trend of decreasing MAEs in yearly cyclone intensity for the short range, while the first 5 years show a decrease in cyclone track MAEs and more similar errors over the last 3 years. During the medium range, there is large interannual variability, with only ECMWF showing a gradual decline in MAEs for cyclone intensity (~1 hPa) over the 8-yr period. Medium-range cyclone absolute-track MAE values have improved for each ensemble, with ECMWF and CMC demonstrating less variability then NCEP on a yearly basis. Once again ECMWF has the largest track error reduction during the period, especially between 2011 and 2012. There were upgrades to the ECMWF model in May 2012 and June 2012 that may have led to this improvement (https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model/ifs-documentation).

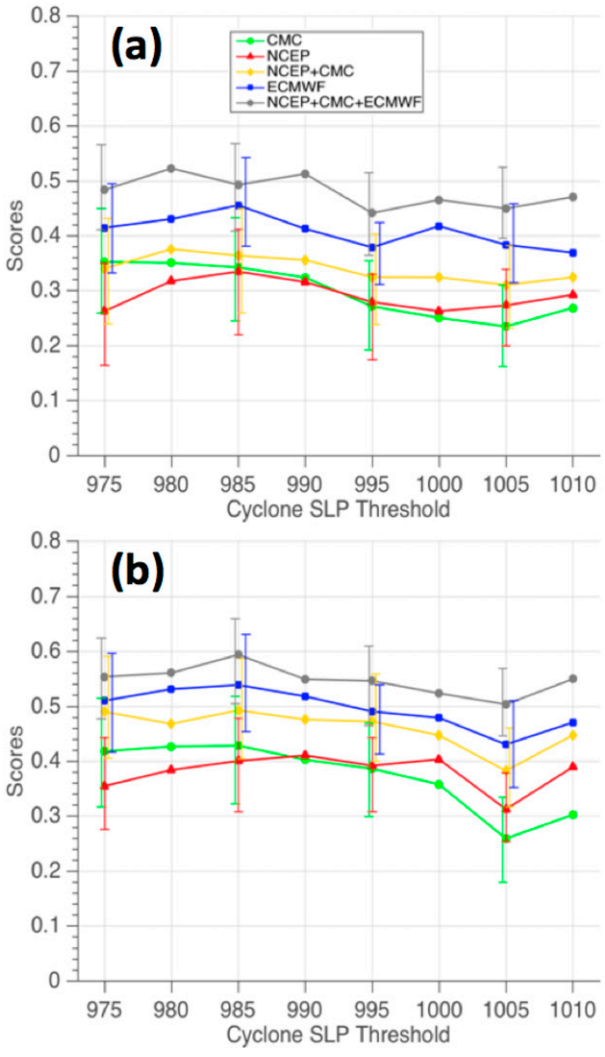

FIG. 10. BSS calculated over a range of intensity thresholds during the (a) short range (0–72 h) and (b) medium range (72–144 h) for the ensemble systems labeled in the inset box. Confidence intervals at the 90% significance level are given by the vertical bars.

4. Probabilistic verification

a. Brier skill score

To evaluate if an ensemble shows any improvement over another ensemble or deterministic model, the BSS was calculated. For this study, the NCEP control member is used as the reference model. A BSS of 1 indicates that the ensemble gives a perfect forecast compared to the reference score, while a BSS of 0 would indicate that the ensemble shows no improvement over the reference score, and a negative value shows that the ensemble is less skillful than the reference model.

Figure 10 shows the BSS calculated using the ECMWF, NCEP, and CMC models, as well as two multimodel ensembles, at various intensity thresholds. All three EPSs and multimodels are more skillful than the reference NCEP control member for all thresholds in the short and medium ranges. For the short range (Fig. 10a), the ECMWF + NCEP + CMC ensemble has the largest BSS scores, is significantly more skillful than NCEP for all thresholds from 975 from 1010 hPa, and shows significantly more skill than CMC for thresholds of 995–1010 hPa. ECMWF has slightly lower BSS values than the ECMWF + NCEP + CMC ensemble scores. The NCEP + CMC blend has slightly better BSS scores than either NCEP or CMC for all thresholds; however, this result is not significant. Figure 10b shows the medium-range BSS for cyclone intensity. The ECMWF + NCEP + CMC ensemble continues to demonstrate superior probabilistic skill across all thresholds, and it is significantly better than NCEP for all thresholds and significantly better than CMC for the weaker thresholds (1000–1010 hPa). The NCEP + CMC ensemble has BSS values similar those of ECMWF, indicating some added benefit, as ECMWF is significantly better than CMC at weaker thresholds (1000–1010 hPa) and shows much greater skill than NCEP at stronger thresholds (975–985 hPa).

The BSS was also calculated for each EPS over a range of displacement thresholds using the absolute-track displacement for each ensemble member (Fig. 11). This is a comparison of the cyclone displacement errors for the NCEP control member and the ensembles by showing the probability of the ensemble members being forecast within the given displacement radius thresholds. The displacement BSSs for the short range are similar for the ECMWF + NCEP + CMC ensemble and ECMWF for all thresholds. The largest probabilistic skill is for the <150- and 300-km displacement bins, with values ranging from 0.52 to 0.48. For larger displacement thresholds (750 and 900 km), the ECMWF + NCEP + CMC ensemble and ECMWF are significantly better than the other three EPSs. The NCEP + CMC ensemble has a 0.05–0.1 greater BSS than either CMC or NCEP for all thresholds, and it is comparable to ECMWF at the smallest thresholds (<150, 300, and 450 km). CMC and NCEP have similar BSS values between 0.10 and 0.30, with the lowest values (0.10 and 0.15) occurring at the 600-km-distance threshold. The performance of each EPS is similar for the medium range (Fig. 11b); however, the BSS values slightly increase as the displacement thresholds increase. Unlike the short range, the NCEP + CMC ensemble is comparable to the ECMWF + NCEP + CMC ensemble and ECMWF for all thresholds. Overall, the BSS scores are higher in the medium range compared with the short range for all thresholds, indicating more skill than the reference model.

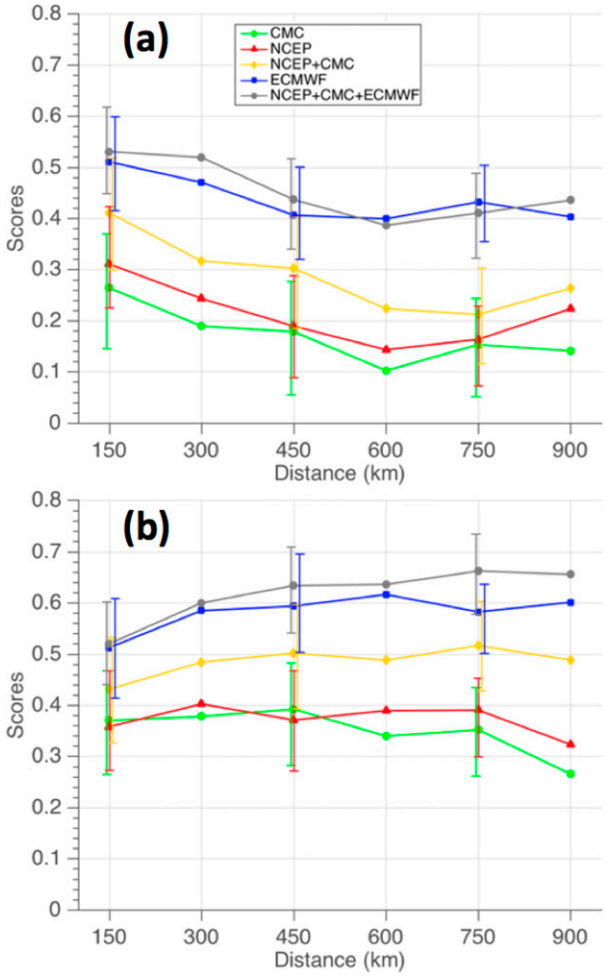

FIG. 11. As in Fig. 10, but for cyclone displacement thresholds.

b. *Reliability*

To visually identify how often a forecast probability actually occurs, reliability diagrams at various intensity thresholds were created for the medium range (Fig. 12). For this calculation, all events with at least 40% membership for all three EPSs were verified using all forecast and matched cyclones to ensure a fair comparison. Missing members in those cases were used; that is to say, the forecast probabilities are always calculated based on 20 members for NCEP and CMC and 50 members for ECMWF. Since minimum ensemble membership assumptions were used for the reliability calculation, these results may overestimate the skill in the medium range. A perfectly reliable ensemble is defined by the solid 1:1 line in Fig. 12, the dashed sloped lines indicates no skill, and the dashed horizontal and vertical lines signify no resolution (climatology). Values above the 1:1 line signify the ensemble is underconfident, while slopes below the 1:1 line signify the ensemble is overconfident. The

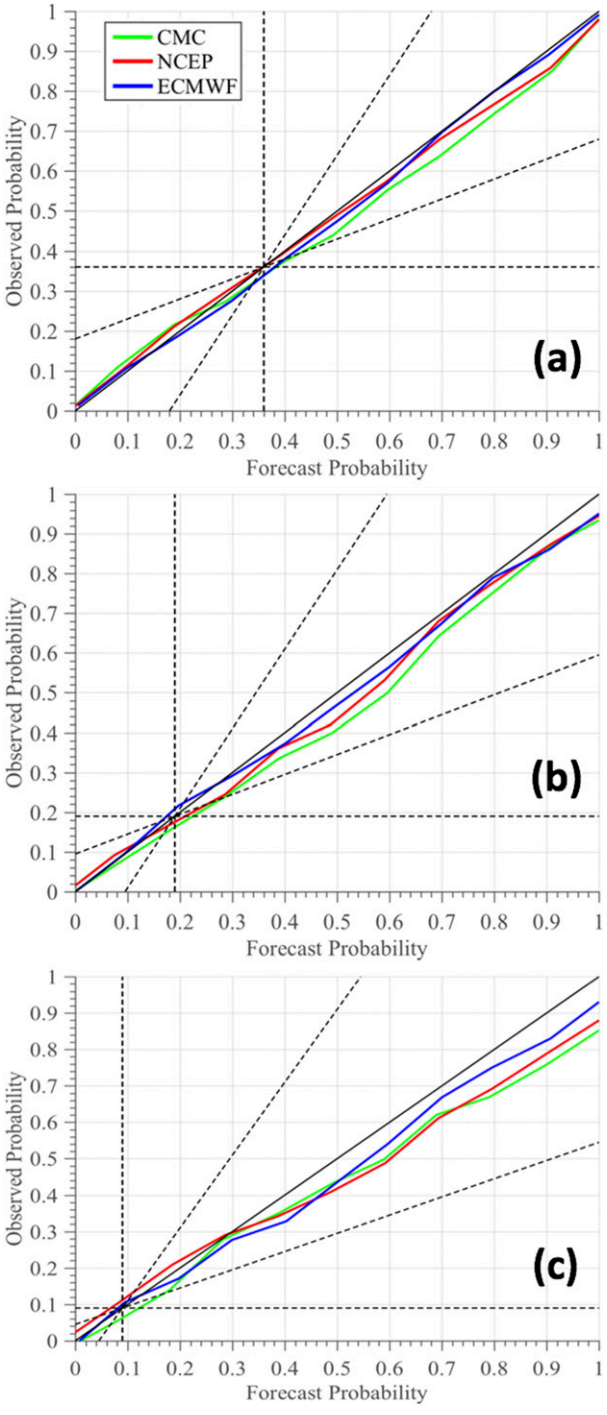

FIG. 12. Reliability diagrams with respect to (a) average cyclone intensity, (b) 1.0 standard deviation below the average cyclone intensity, and (c) 1.5 standard deviations below the average cyclone intensity for the medium range (72–144 h). The reliability is represented by the solid curve. A perfect ensemble forecast is shown by the 1:1 solid black line. The sloped dashed lines indicate an ensemble with no skill, and the dashed horizontal and vertical lines indicate no resolution.

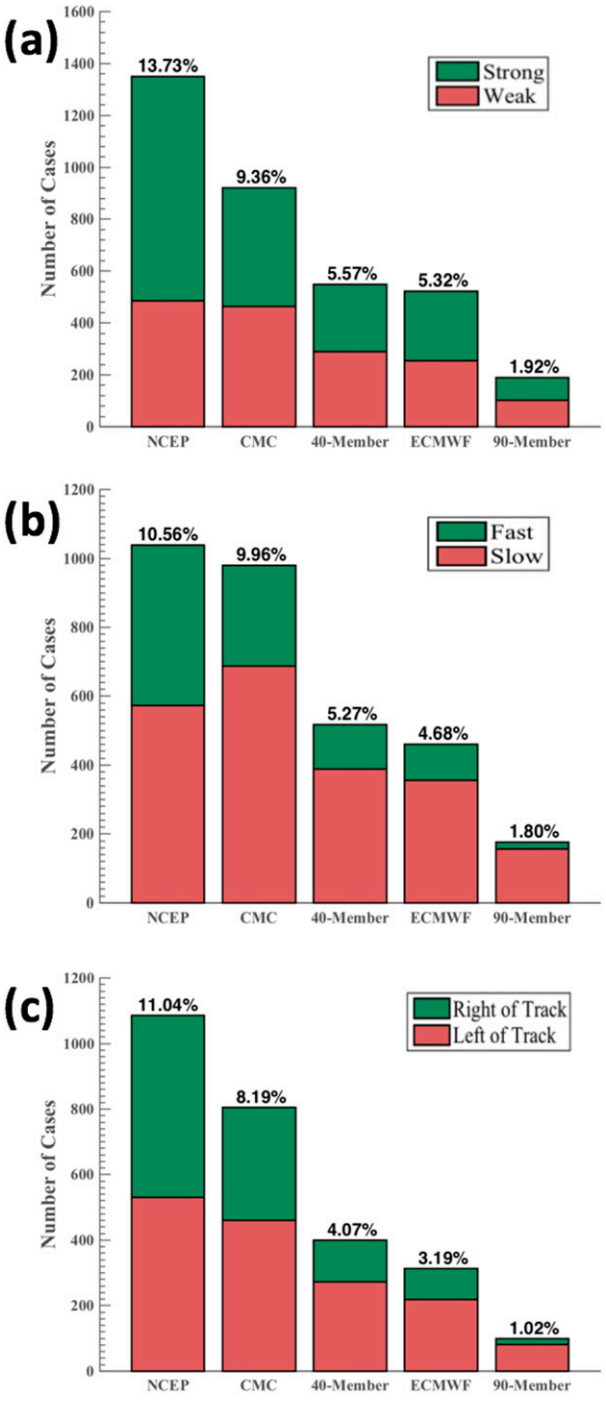

FIG. 13. Bar charts showing the numbers of cases where the observed cyclone is outside the ensemble envelope during the medium range (72–144 h) for (a) cyclone intensity, (b) along-track, and (c) cross-track forecasts. Color shading indicates how the ensemble performs for these cases compared to the observed. The total percentage outside the envelope is noted above each bar.

probabilities are defined as whether or not the cyclone intensity (hPa) is less than or equal to the average minimum central pressure (Fig. 12a), 1.0 standard deviation below the average minimum central pressure (deep cyclones) (Fig. 12b), or 1.5 standard deviations below the average minimum central pressure (Fig. 13c). NCEP and CMC are somewhat underconfident at lower probabilities during the medium range (72–144 h) for average cyclones (0.10–0.30) by 3%–5% and overconfident at higher probabilities (>0.70) by 3%–5% while ECMWF is very reliable during these periods. For deep cyclones (Fig. 12c), NCEP and CMC are largely overconfident for nearly all probabilities (0.20–1.0), while ECMWF is also overconfident for most probabilities (0.30–1.0). CMC notably overforecasts deep cyclone events by 10% for moderate forecast probabilities (0.50–0.60), while for the same probabilities NCEP overforecasts by 5% and ECMWF by 3%. During events where each ensemble gives a 100% chance of the cyclone being of deep cyclone intensity, only ~94% verify. For the much deeper cyclones (Fig. 12c), all three EPSs are greatly overconfident for moderate probabilities (0.40–0.60) by 5%–10% and at higher probabilities (>0.70) by 5%–8%, 10%, and 12%–15% for ECMWF, NCEP, and CMC, respectively. Overall, each EPS has less reliability for forecasting deep cyclones than average cyclones for moderate to high probabilities, while lower probabilities (0.10–0.30) show more variability for each ensemble.

c. *Ensemble consistency*

1) OUTSIDE THE ENVELOPE

A useful metric when investigating whether the spread of an ensemble is appropriate is calculating the number of cyclone cases that fall outside of the ensemble envelope. These cases occur because the ensemble is underdispersed, and there is not enough ensemble spread introduced when using all the matched members. Outside the envelope cases are found by ordering the matched ensemble member intensity, along-, and cross-track values from lowest to highest for a particular forecast time period and counting the cases where the observed value is above or below all ensemble member values. After the total number of outside the envelope cases is found, the percentage of outside the envelope cases can be determined by dividing by the total number of cases for the forecast period. Assuming any distribution with limited members, there is a finite probability that the observed value lies outside the envelope. Therefore, when the ensemble size increases, the probability of cases outside the envelope should decrease.

Figure 13 shows the total number of outside-the-envelope cases in the medium range for each EPS for

cyclone intensity and track. ECMWF has slightly lower outside-the-envelope frequency when compared to the NCEP + CMC ensemble for cyclone intensity (5.3%–5.6%), along-track (4.7%–5.2%), and cross-track (3.2%–4.1%) results. The ECMWF + NCEP + CMC ensemble has the least amount of cases (1.9%, 1.8%, and 1.0%) outside the envelope. For cyclone intensity, most of the outside-the-envelope cases tend to have forecasts that are too strong (Fig. 13a), while the along-track cases tend have forecasts that are too slow (Fig. 13b), and the cross-track sample has a less noticeable, but slight tendency to have cases that are to the left of track (Fig. 13c). These results are consistent with the cyclone intensity, along-, and cross-track biases from the previous section. Overall, there are adequate differences between the 20-member CMC and NCEP EPSs, as well as similarities between the NCEP + CMC and 50-member ECMWF, with very few cases outside the envelope for the ECMWF + NCEP + CMC ensemble.

2) PROBABILITY WITHIN SPREAD

Another quantity of interest when investigating the consistency and dispersion of an ensemble is the PWS. Figure 14 shows the PWS for distances of 1σ, 2σ, and 3σ from the ensemble mean cyclone intensity. The dashed lines show what the expected probabilities should be, but these should be used as a reference because the ensembles are not completely Gaussian. Values above (below) the expected probability would indicate overdispersion (underdispersion) calculated by considering the likelihood the observed cyclone falls within a set dispersion of the ensemble calculated by varying standard deviations σ from the ensemble mean. For 1σ, PWS is less than 0.68 for all lead times (0–144 h) and ensembles. The ECMWF + NCEP + CMC ensemble has the largest values between 0.67 and 0.62 from 0 to 60 h, while NCEP has the lowest from 0.55 to 0.57 over the same period. For lead times of 0–24 and 120–144 h, the ECMWF + NCEP + CMC ensemble has a significantly larger PWS than the individual single-model ensembles, while the NCEP + CMC ensemble also has larger values than each single-model ensemble, albeit not significant. For 2σ and 3σ the multimodel ensembles have larger PWS results across all lead times (0–144 h), while ECMWF, NCEP, and CMC have similar values that are comparable.

PWSs for along-track distances from the ensemble mean show the ECMWF + NCEP + CMC ensemble and NCEP + CMC ensemble are above the expected 0.68 and 0.95 values for 1σ and 2σ distances for all lead times (0–144 h), while the ECMWF + NCEP + CMC ensemble has significantly larger PWSs for 3σ distances compared with each single-model ensemble (Fig. 15). For 1σ distances, ECMWF, NCEP, and CMC are all at the expected 0.68

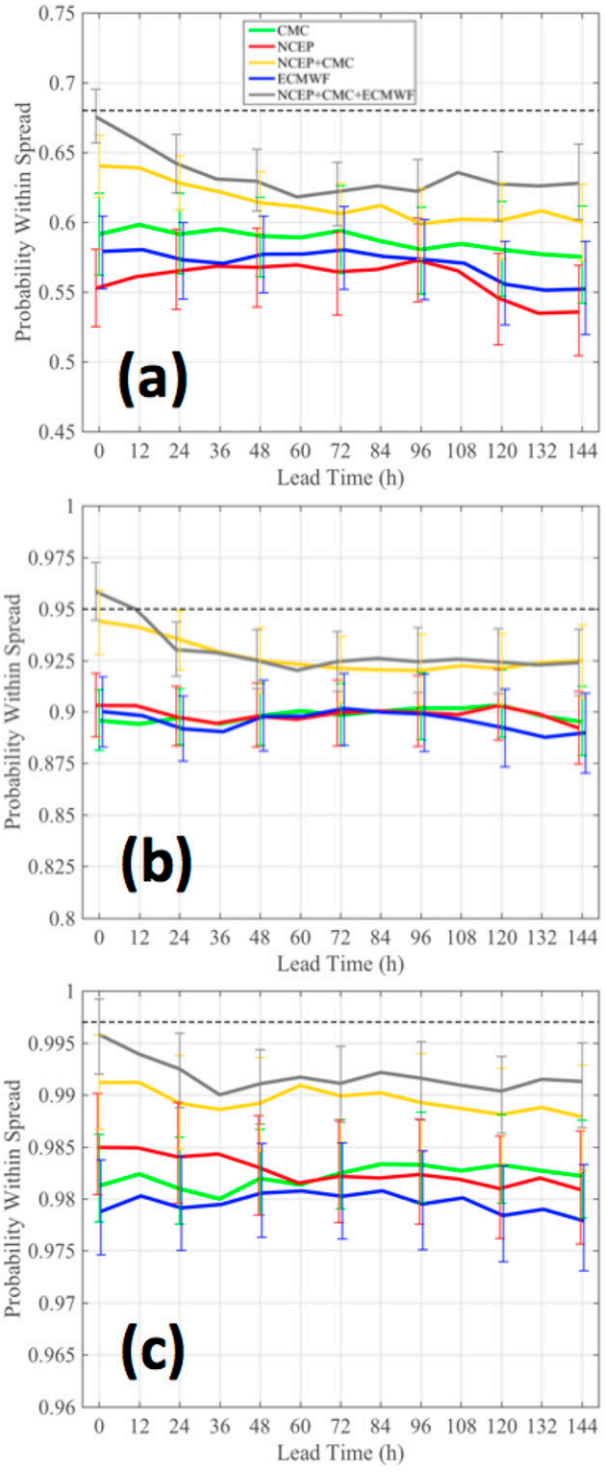

FIG. 14. PWS diagrams for cyclone intensity (hPa) for (a) 1σ, (b) 2σ, and (c) 3σ. The dashed line denotes the expected probabilities, assuming a normal distribution. Confidence intervals at the 90% significance level are given by the vertical bars.

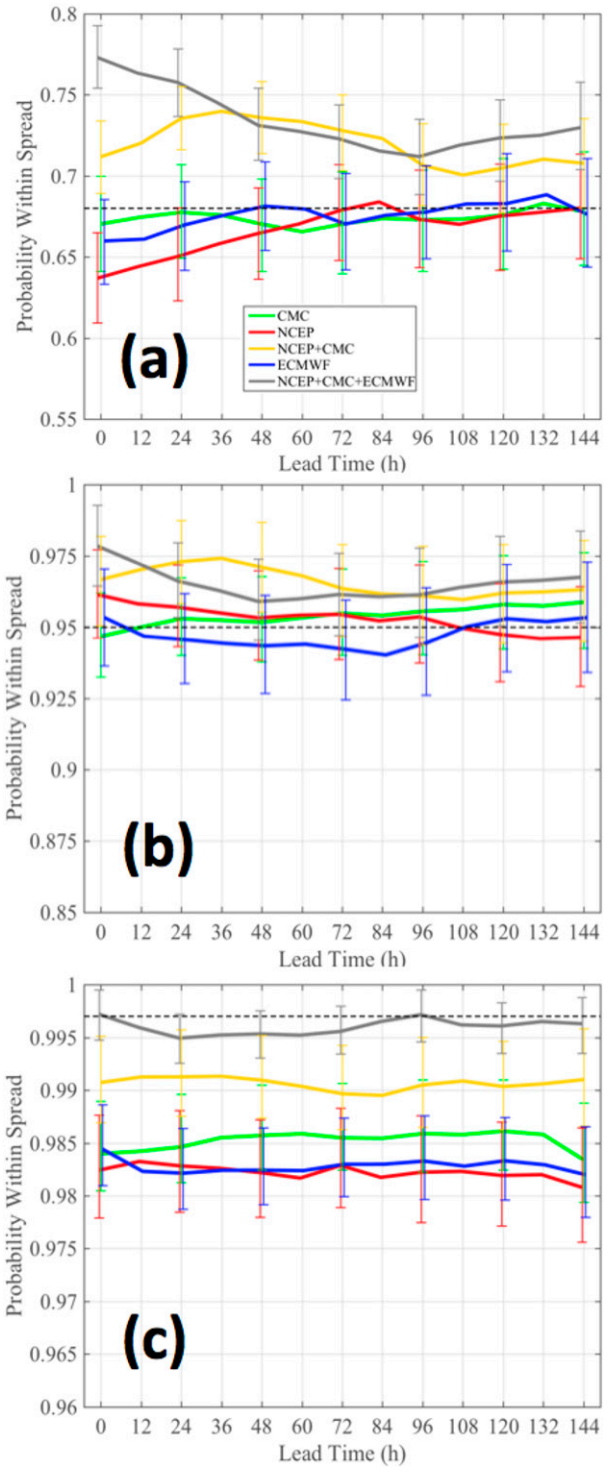

FIG. 15. As in Fig. 14, but for along-track displacement (km).

value from 72 to 144 h and within 0.01 of the expected 0.95 values for 2σ distances and comparable to the multimodel ensembles. This would indicate that the multimodel ensembles are slightly overdispersed for 1σ distances from the mean, compared to the single-model ensembles. The PWSs for the cross-track results show that only NCEP falls below the 0.68 expectation for 1σ and has a significantly lower PWS than both of the multimodel ensembles from 0 to 108 h (Fig. 16). Additional cross-track probabilities between 2σ and 3σ are similar to the along-track probabilities. Overall, the PWS for the multimodel ensembles shows significant benefits over single-model ensembles, especially early (0–24 h) and late (120–144 h) in the forecast period.

5. Summary and conclusions

The primary goal of this study was to complete a comprehensive verification of cool-season extratropical cyclones from an ensemble perspective using the ECMWF, NCEP, and CMC ensembles from 2007 to 2015 for eastern North America and the western Atlantic Ocean. The cyclone verification is binned into different groups according to forecast lead time, cyclone intensity, and range of cyclone errors at various lead times. Individual member analysis shows ECMWF has the lowest intensity MAEs after 72 h, while CMC has the largest intensity error throughout the medium-range period (72–144 h). ECMWF has 12–18 h more accuracy than NCEP and a 24–30-h advantage over CMC. Yearly analysis indicates ECMWF has gained 18–24 h of lead time in the short range over the past 8 years, while NCEP and CMC have gained ~12 h of lead time. This improvement over the years is likely from an increase in model resolution and better data assimilation.

In contrast, the medium range has much more interannual variability for cyclone intensity. The ECMWF mean has significantly less MAE compared with the NCEP and CMC means from 42 to 144 h. Additionally, during the medium range the ECMWF mean maintains ~1 hPa more skill than the NCEP and CMC means. At these medium-range time scales, the initial upper-level disturbance is often over the northern Pacific (Zheng et al. 2013), thus suggesting that the ECMWF may be assimilating observations over this region. Spatially, during the medium range the magnitude of the underdeepening bias for all ensemble systems in the North Atlantic increases to 4–6 hPa and the overdeepening along the East Coast is 2–4 hPa with standard deviation values ranging from 5 to 8 hPa. The overdeepening suggests that the models are either too strong with the low-level baroclincity over this coastal land region or there is too much latent heating from precipitation. On the other hand, underdeepening over the Atlantic may reflect too little surface heat and moisture fluxes over the water. These hypotheses need to be tested in future work.

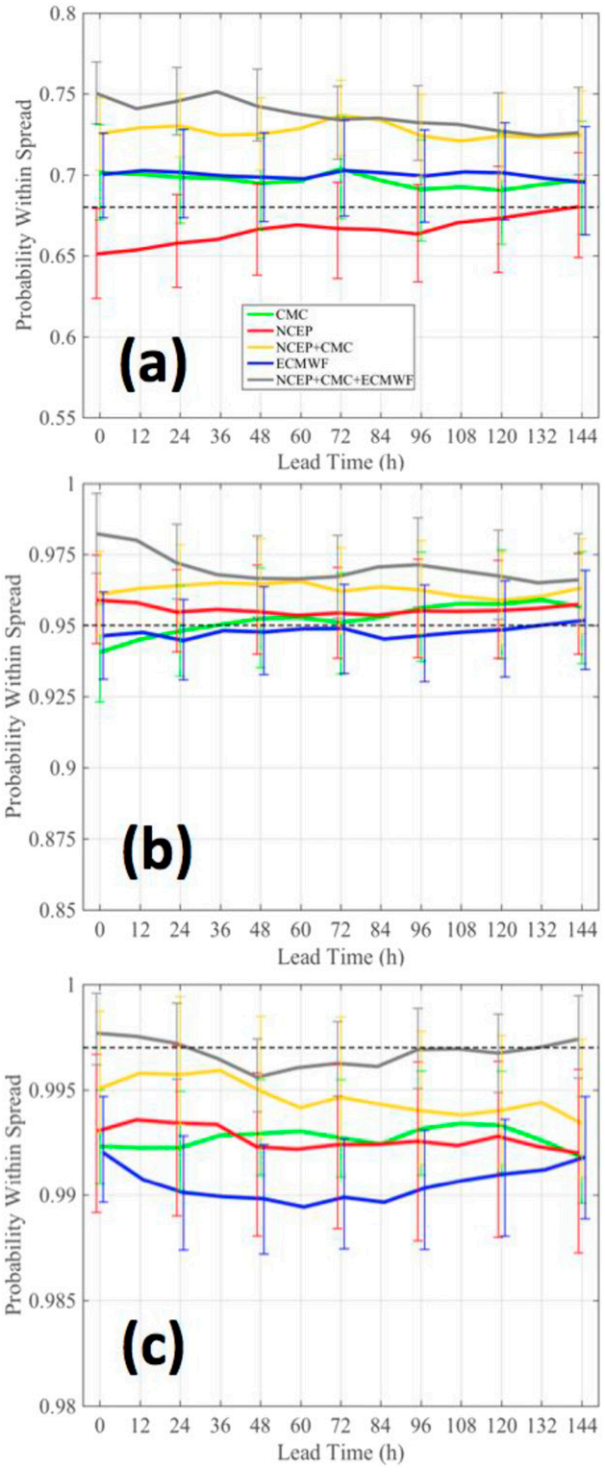

FIG. 16. As in Fig. 14, but for cross-track displacement (km).

For cyclone displacement from 0- to 144-h lead times, ECMWF has the smallest absolute-track MAE after 60 h, but only hours 84–120 are statistically significant. Beginning at 84 h, the CMC absolute-track error growth slows and by 144 h CMC has a ~700-km average error that is comparable to those of ECMWF and NCEP. The along-track component contributes more to the total track errors than the cross-track component, with nearly ~60% (~55%) of the track error at 144 h (72 h) for each EPS. Cross-track errors at 144 h vary between 280 and 310 km for all three EPSs, indicating a smaller error spread among the ensembles during the medium-range period. The annual variability in absolute-track error shows improvement, especially in the medium range, with the ECMWF having the greatest improvement and smallest errors. All three EPSs have a negative (slow) along-track bias (0–100 km) from past 12 h to the end of the forecast period, and the results are significant from 24 to 90 h. The slow bias is consistent with the overdeepened cyclones and associated upper-level troughs over the eastern United States, which as a result progress eastward more slowly. Cross-track biases also show a (5–50 km) negative bias (left-of-track bias) throughout the forecast period for NCEP and CMC, while ECMWF has no bias through the short range and a small negative bias (5–30 km) in the medium range.

The combination of the NCEP + CMC ensemble decreases cyclone intensity (along track) MAE by ~0.5 hPa (~40 km) in the medium range, so this ensemble blend is more comparable to ECMWF than either NCEP or CMC individually. For medium-range cyclone intensity forecasts, the ECMWF, NCEP, and CMC means are the best (worst) 32% (24%), 24% (38%), and 23% (36%) of the time, respectively. While the multimodel ensemble blends, especially the ECMWF + NCEP + CMC ensemble, are often not the best forecast out of the five models in the short and medium ranges, they are very rarely the worst EPS and are often the second- or third-best EPS for intensity, along-, and cross-track forecasts. This highlights the importance of using multimodel ensembles to avoid potential forecast busts.

ECMWF has the greatest probabilistic skill when compared to NCEP and CMC; however, on average the multimodel ensembles have better probabilistic skill than all single-model ensembles. Intensity BSS calculations show that the ECMWF + NCEP + CMC ensemble demonstrates superior probabilistic skill across all thresholds, and it is significantly better than NCEP for all thresholds and significantly better than CMC for the weaker thresholds (1000–1010 hPa) during the short and medium ranges. The NCEP + CMC ensemble has BSS values similar to ECMWF, indicating some added benefit. Displacement BSSs calculated over a range of displacement thresholds using the absolute-track

displacement for each member show that for larger displacement thresholds (750 and 900 km), the ECMWF + NCEP + CMC ensemble and ECMWF are significantly better than the other three EPSs during the short range. The NCEP + CMC ensemble has a 0.05–0.1 greater BSS than either NCEP or CMC for all thresholds, and it is comparable to ECMWF at the smallest thresholds (<150, 300, and 450 km). Thus, there is an advantage in using this NAEFS (CMC + NCEP) ensemble rather than each of the ensembles individually. The performance of each EPS is similar for the medium range; however, the BSS values slightly increase as the displacement thresholds increase. Overall, the BSS scores are higher in the medium range compared with the short range for all thresholds, indicating more skill than the reference model for larger lead times.

Ensemble consistency was evaluated by calculating the cases outside the ensemble envelope and the probability within spread (PWS). During the medium range, the ECMWF + NCEP + CMC multimodel ensemble has the least amount of cases (1.9%, 1.8%, and 1.0%) outside the envelope compared with ECMWF (5.6%, 5.2%, and 4.1%) and NCEP (13.7%, 10.6%, and 11.0%) for cyclone intensity, along-, and cross-track forecasts. The PWS for multimodel ensembles shows significant benefits over single-model ensembles, especially early (0–24 h) and late (120–144 h) in the forecast period.

Overall, this study improves the understanding and utility of ensembles for forecasting extratropical cyclones along the East Coast. This paper emphasizes the importance of using multimodel ensembles for forecasting these events. The ECMWF ensemble delivers the best performance of any one ensemble in the medium range, but utilizing all three ensembles is best. There are important biases for forecasters to consider such as the along-track slow bias and an underprediction bias in the medium range. Future work will explore some of the potential reasons for these systematic track and intensity biases. Also, one limitation of our approach is that our verification is limited to those forecast members that have cyclone tracks, thus ignores those members with no cyclones. We are developing other approaches to validate cyclone events by including all ensemble members, such as clustering within a principle component phase space, as in Zheng et al. (2017), to analyze the number of members within the analysis cluster.

Acknowledgments. The authors thank the TIGGE data archives for the availability of the ensemble data. This work is supported by NOAA-CSTAR (NA13NWS4680002). We thank the three anonymous reviewers for their constructive comments to improve the manuscript.

REFERENCES

Anderson, D., K. I. Hodges, and B. J. Hoskins, 2003: Sensitivity of feature-based analysis methods of storm tracks to the form of background field removal. *Mon. Wea. Rev.*, **131**, 565–573, https://doi.org/10.1175/1520-0493(2003)131<0565:SOFBAM>2.0.CO;2.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

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.

Charles, M. E., and B. A. Colle, 2009: Verification of extratropical cyclones within the NCEP operational models. Part II: The Short-Range Ensemble Forecast system. *Wea. Forecasting*, **24**, 1191–1214, https://doi.org/10.1175/2009WAF2222170.1.

Colle, B. A., Z. Zhang, K. A. Lombardo, K. M. Chang, P. Lui, and M. Zhang, 2013: Historical evaluation and future prediction of eastern North American and western Atlantic extratropical cyclones in the CMIP5 models during the cool season. *J. Climate*, **26**, 6882–6903, https://doi.org/10.1175/JCLI-D-12-00498.1.

Froude, L. S. R., 2010: TIGGE: Comparison of the prediction of Northern Hemisphere extratropical cyclones by different ensemble prediction systems. *Wea. Forecasting*, **25**, 819–836, https://doi.org/10.1175/2010WAF2222326.1.

——, L. Bengtsson, and K. I. Hodges, 2007: The prediction of extratropical storm tracks by the ECMWF and NCEP ensemble prediction systems. *Mon. Wea. Rev.*, **135**, 2545–2567, https://doi.org/10.1175/MWR3422.1.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

Goodwin, L. C., 2003: Weather-related crashes on U.S highways in 2001. Mitretek Systems, Rep. prepared for U.S. Department of Transportation, 8 pp.

Hanbali, R. M., and D. A. Kuemmel, 1993: Traffic volume reductions due to winter storm conditions. *Third Int. Symp. on Snow Removal and Ice Control Technology*, Minneapolis, MN, National Transportation Board, 159–164, https://trid.trb.org/view.aspx?id=379645.

Hodges, K. I., 1994: A general method for tracking analysis and its application to meteorological data. *Mon. Wea. Rev.*, **122**, 2573–2586, https://doi.org/10.1175/1520-0493(1994)122<2573:AGMFTA>2.0.CO;2.

——, 1995: Feature tracking on the unit sphere. *Mon. Wea. Rev.*, **123**, 3458–3465, https://doi.org/10.1175/1520-0493(1995)123<3458:FTOTUS>2.0.CO;2.

Hoskins, B. J., and K. I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.*, **59**, 1041–1061, https://doi.org/10.1175/1520-0469(2002)059<1041:NPOTNH>2.0.CO;2.

Knapp, K. K., L. D. Smithson, and A. J. Khattak, 2000: The mobility and safety impacts of winter storm events in a freeway environment. *Proc. Mid-Continent Transportation Symp.*, Ames, IA, National Transportation Board, https://trid.trb.org/view.aspx?id=654630.

Korfe, N., 2016: Evaluation of cool season extratropical cyclones in a multi-model ensemble for eastern North America and the western Atlantic Ocean. M.S. thesis, School of Marine and Atmospheric Sciences, Stony Brook University, State

University of New York, Stony Brook, NY, 110 pp., https://ir.stonybrook.edu/xmlui/handle/11401/77756.

Majumdar, S. J., and P. M. Finocchio, 2010: On the ability of global ensemble prediction systems to predict tropical cyclone track probabilities. *Wea. Forecasting*, **25**, 659–680, https://doi.org/10.1175/2009WAF2222327.1.

Miller, J. E., 1946: Cyclogenesis in the Atlantic coastal region of the United States. *J. Meteor.*, **3**, 31–44, https://doi.org/10.1175/1520-0469(1946)003<0031:CITACR>2.0.CO;2.

Novak, D. R., D. Bright, and M. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084, https://doi.org/10.1175/2008WAF2222142.1.

Picca, J. C., D. M. Schultz, B. A. Colle, S. Ganetis, D. R. Novak, and M. Sienkiewicz, 2014: The value of dual-polarization radar in diagnosing the complex microphysical evolution of an intense snowband. *Bull. Amer. Meteor. Soc.*, **95**, 1825–1834, https://doi.org/10.1175/BAMS-D-13-00258.1.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012, https://doi.org/10.1256/qj.04.176.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale predictability of the ''surprise'' snowstorm of 24–25 January 2000. *Mon. Wea. Rev.*, **130**, 1617–1632, https://doi.org/10.1175/1520-0493(2002)130<1617:MPOTSS>2.0.CO;2.

Zheng, M., K. Chang, and B. A. Colle, 2013: Ensemble sensitivity tools for assessing extratropical cyclone intensity and track predictability. *Wea. Forecasting*, **28**, 1133–1156, https://doi.org/10.1175/WAF-D-12-00132.1.

——, E. Chang, B.A. Colle, Y. Luo, and Y. Zhu: 2017: Applying fuzzy clustering to a multimodel ensemble for U.S. East Coast winter storms: Scenario identification and forecast verification. *Wea. Forecasting*, **32**, 881–903, https://doi.org/10.1175/WAF-D-16-0112.1.

Zwiers, F. W., 1990: The effect of serial correlation on statistical inferences made with resampling procedures. *J. Climate*, **3**, 1452–1461, https://doi.org/10.1175/1520-0442(1990)003<1452:TEOSCO>2.0.CO;2.