Evaluating U.S. East Coast Winter Storms in a Multimodel Ensemble Using EOF and Clustering Approaches📎

MINGHUA ZHENG

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

EDMUND K. M. CHANG AND BRIAN A. COLLE

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

(Manuscript received 16 February 2018, in final form 14 February 2019)

ABSTRACT

Empirical orthogonal function (EOF) and fuzzy clustering tools were applied to generate and validate scenarios in operational ensemble prediction systems (EPSs) for U.S. East Coast winter storms. The National Centers for Environmental Prediction (NCEP), European Centre for Medium-Range Weather Forecasts (ECMWF), and Canadian Meteorological Centre (CMC) EPSs were validated in their ability to capture the analysis scenarios for historical East Coast cyclone cases at lead times of 1–9 days. The ECMWF ensemble has the best performance for the medium- to extended-range forecasts. During this time frame, NCEP and CMC did not perform as well, but a combination of the two models helps reduce the missing rate and alleviates the underdispersion. All ensembles are underdispersed at all ranges, with combined ensembles being less underdispersed than the individual EPSs. The number of outside-of-envelope cases increases with lead time. For a majority of the cases beyond the short range, the verifying analysis does not lie within the ensemble mean group of the multimodel ensemble or within the same direction indicated by any of the individual model means, suggesting that all possible scenarios need to be taken into account. Using the EOF patterns to validate the cyclone properties, the NCEP model tends to show less intensity and displacement biases during 1–3-day lead time, while the ECMWF model has the smallest biases during 4–6 days. Nevertheless, the ECMWF forecast position tends to be biased toward the southwest of the other two models and the analysis.

1. Introduction

The U.S. East Coast and the adjacent ocean are favorable for cool-season extratropical cyclone activity (Miller 1946). Intense extratropical cyclones in this region often have large socioeconomic impacts on transportation (e.g., road, aviation, and marine), human health, and property with their strong winds, heavy precipitation, or storm surges (Mather et al. 1964; Davis and Dolan 1993; Novak et al. 2008; Chang 2013; Booth et al. 2015; Colle et al. 2015; Ma and Chang 2017). Considering the high population density of the eastern United States, accurate forecasts of these storms are crucial to reduce their impact, including both human and economic losses.

The predictability of extratropical cyclones is related to uncertainties in initial conditions (ICs) and the forecast model characteristics (e.g., resolution, physics). As a result, an ensemble approach has been demonstrated to improve forecast skills in general when compared with a single-model (deterministic) approach (Tracton and Kalnay 1993; Molteni et al. 1996; Buizza 1997) by using a variety of ICs, physical parameterizations, and/or models (Toth and Kalnay 1993; Molteni et al. 1996; Buizza et al. 1999). Several recent studies have demonstrated the value of probabilistic information for medium-range forecasts of severe weather events, including the high-impact winter storms (Hewson et al. 2014; Matsueda and Nakazawa 2015; Swinbank et al. 2016).

The characteristics of ensemble forecasts for cyclones vary in different operational models depending on the variations in ICs and model physics (Du et al. 1997;

📎 Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-18-0052.s1.

Corresponding author: Minghua Zheng, ming.h.zheng@gmail.com

DOI: 10.1175/MWR-D-18-0052.1

Stensrud et al. 1999; Froude et al. 2007; Charles and Colle 2009; Froude 2009; Colle and Charles 2011; Froude 2010). Froude et al. (2007) investigated the cyclone tracks in the 50-member European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (EPS) and 10-member National Centers for Environmental Prediction (NCEP) EPSs between 6 January and 5 April 2005 using an objective feature tracking methodology to identify and track cyclones along the forecast trajectories. They found that the ECMWF ensemble on average has a higher forecast skill than the NCEP ensemble for both cyclone intensity and position in the Northern Hemisphere (NH) while the NCEP ensemble has smaller errors for cyclone intensity in the Southern Hemisphere (SH). Both EPSs indicate a higher level of forecast skill for cyclone position than intensity. The propagation speed of cyclones is generally too slow in the ECMWF EPS. Both ECMWF ensemble mean and the best ensemble member had greater accuracy than the control forecast for both the position and intensity of the cyclones, although the ECMWF ensemble was underdispersed. Froude (2010) further analyzed the predictions of extratropical cyclones in the NH by nine EPSs from The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010; Swinbank et al. 2016) archive between 1 February and 31 July 2008. They showed that the ECMWF ensemble has a higher predictive skill for all aforementioned cyclone properties. However, the ECMWF model consistently overpredicts cyclone intensity, although the bias is small. The Japan Meteorological Agency (JMA), Met Office (UKMO), NCEP, and Canadian Meteorological Centre (CMC) have 1 day less skill for the position of cyclones throughout the forecast lead times. The NCEP model has larger errors for cyclone intensity than for position. It was also found that cyclones in all EPSs propagate too slowly.

Charles and Colle (2009) comprehensively verified the strengths and positions of storms around North America and the adjacent oceans within the NCEP Short Range Ensemble Forecast (SREF; Du and Tracton 2001; Du et al. 2003) system during 2004–07. They found that the SREF has slightly more probabilistic skill over the eastern United States and the western Atlantic than the western potions of the domain for cyclone central pressure. The 15-member SREF mean for both cyclone position and central pressure on average has a smaller error than its own five-member subgroups and the North American Mesoscale Forecast System (NAM) model have in many regions, but not the Global Forecast System (GFS) model for many lead times. The SREF probabilities are fairly reliable, although it is overconfident at higher probabilities in all regions. Colle and Charles (2011) showed that the cyclones for 72–120 h are too weak on average by 2–3 hPa near the U.S. East Coast. These cyclones move too fast in the GFS deterministic model in the medium range while they are too slow and too far west for the short range. Korfe and Colle (2017) evaluated the extratropical cyclones within the CMC, ECMWF, and NCEP EPSs using a cyclone-tracking scheme. They found that the NCEP EPS has comparable forecast skill with the ECMWF EPS for lead times less than 72 h while the ECMWF EPS has more accuracy for cyclone intensity than the other two EPSs for days 4–6. All three models have a significant slow along-track bias for lead times 24–90 h, and they have a left of track bias for lead times 120–144 h.

Long-term verification of cyclone forecasts in ensemble models have so far mainly employed the tracking and matching methods (e.g., Froude 2010; Charles and Colle 2009; Korfe and Colle 2017), which only verify cases in which several ensemble member cyclones can be tracked and matched to the observed cyclone. Hence, these verifications exclude a significant fraction of cyclone forecasts especially in the medium range, and thus the results for metrics such as the mean errors/absolute errors, the anomaly correlation, and the rank histogram are expected to be biased. Therefore, a complementary method utilizing the forecasts from all ensemble members is desirable for comparison.

In this work, we propose to evaluate the ensemble forecasts of cyclones in a novel way by using the scenario-based method (Zheng et al. 2017), which includes all ensemble members by using an empirical orthogonal function (EOF) and fuzzy clustering methodology. Keller et al. (2011) compared the forecast scenarios associated with 10 extratropical transition (ET) cases in the TIGGE data by applying fuzzy clustering analysis and found that some EPSs are confined to a few scenarios while others contribute to almost all scenarios. The benefit of multimodel ensemble over a single model has been demonstrated in some literatures (e.g., Du et al. 2003; Zhou and Du 2010) for different weather phenomenon. Other studies (e.g., Harr et al. 2008; Grams et al. 2011) have also shown that fuzzy clustering is a suitable diagnostic method to detect the physical processes associated with different weather systems. Zheng et al. (2017) applied this method to interpret scenarios in a set of ensemble forecasts. They have shown that the EOF/fuzzy clustering method can classify different scenarios associated with the uncertainty in extratropical cyclone intensity and position. This work will utilize the

EOF/fuzzy clustering method to assess the scenarios within the CMC, ECMWF, and NCEP EPSs. The statistics of analysis scenario will provide useful guidance for operational forecasters to interpret outputs from multiple models. It also initiates a scenario-based study for model developers to diagnose the problems in forecasting cyclone scenarios in operational models.

Overall, this study provides a recent snapshot of the performance of the aforementioned three operational ensemble models. To be more specific, this work will address the following motivational questions:

- Which model performs better in capturing the analysis scenario in predicting winter storms using a multimodel ensemble?
- What are the significant errors and bias in forecasting the winter storms in different models for all observed cyclone cases?
- What are the benefits of using multimodel combination and the multimodel ensemble mean?

2. Data and methodology

The ensemble data and techniques employed for this paper are described in this section. This work involves cyclone cases for calculating model statistics and bias/errors; therefore, how the cases were selected will be briefly presented.

a. The dataset

Ensemble data for this study were retrieved from the TIGGE data archive. TIGGE was established since 2005 to support a range of THORPEX research activities by providing operational ensemble forecast data to the international research community (Bougeault et al. 2010; Swinbank et al. 2016). Forecast data from 10 participating centers are available with a lag of 48 h. We used data from three centers—NCEP, CMC, and ECMWF—mainly because they are popular ensembles used by operational forecasters.

This study analyzes the ensemble forecasts in cool seasons (November–March in the following year) from 2007/08 to 2014/15. The three EPSs comprise a large set of ensemble forecasts with 90 members (50 from ECMWF, 20 each from NCEP and CMC) in 12-h forecast intervals and interpolated onto 1° latitude × 1° longitude grid. Mean sea level pressure (MSLP) was chosen to investigate cyclone intensity and tracks. Note that the control forecasts are not included in the multimodel ensemble for a fair comparison among ensemble members. The NCEP operational analysis data for MSLP are used to verify ensemble forecasts. The ECMWF operational analysis was also investigated as an alternate analysis data, and the results are found to be very similar.

TABLE 1. The number of cyclone cases used for each forecast day.

Lead time (days)	1	2	3	4	5	6	7	8	9
Case number	164	176	180	158	170	185	170	168	178

b. Winter storm case selection

To evaluate the performance of different operational models in forecasting East Coast cyclones, we first tracked the analysis cyclones across the verification regions over the East Coast and western Atlantic (region 1, 32°–45°N, 79°–62°W). A bigger region (region 2, 30°–50°N, 95°–65°W) was also investigated, including part of the central United States and the East Coast. The cyclone cases included only analyzed cyclones with the minimum central pressure less than 1005 hPa crossing these two regions determined by the tracking scheme developed by Hodges (1994, 1995, 1999). A minimum pressure of 1005 hPa is chosen because it has been shown that over two-thirds of the East Coast cyclones have minimum pressure less than 1005 hPa (Hirsch et al. 2001), and we are most interested in the significant cyclones. To increase the sample size for model evaluations, one or two verification times (VTs) were included if they are within the verified region. We chose one or two based on the following criteria: 1) there must be at least one time step with minimum pressure <1005 hPa within the investigation domain; 2) if there is only one time step with minimum pressure <1005 hPa, this will be chosen as the unique VT; 3) if there are two time steps with minimum pressure <1005 hPa, both will be chosen as the VT; and 4) if there are more than two time steps with minimum pressure <1005 hPa, two time steps closest to the center of the domain will be chosen as two VTs.

We evaluated the 1–9-day forecasts for the observed cyclones. Due to the missing data for some cases in the TIGGE archive, we have chosen 158–185 cases (Table 1) for the scenario-based model evaluation calculations. Intensity distribution for each lead time is shown in Fig. S1 in the online supplemental material. We have also tested similar calculations using unique cases (102 and 135) for 3- and 6-day forecasts by only selecting one VT for each cyclone track and the results are very consistent.

c. Evaluation using EOF/fuzzy clustering method

EOF analysis is a statistical method to condense the information of a large dataset in order to examine its variability (Hannachi et al. 2007; Wilks 2011). In terms of ensemble forecast, there are regions of low and high

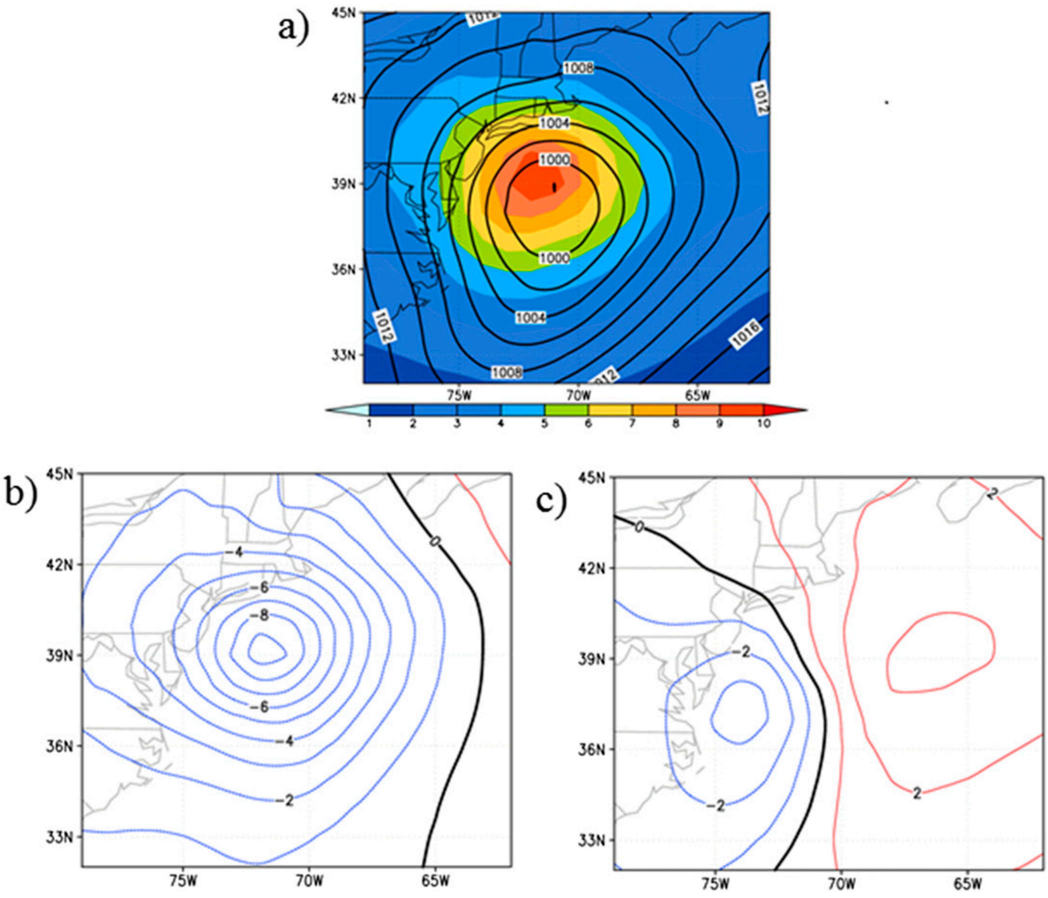

FIG. 1. (a) Ensemble mean MSLP (contours, hPa) and the spread (shading, hPa); (b) MSLP EOF1 pattern (hPa); and (c) MSLP EOF2 pattern (hPa). Valid time (VT): 0000 UTC 30 Dec 2012; initial time (IT): 0000 UTC 27 Dec 2012. The ensemble is based on a combined 90-member ensemble from NCEP, CMC, and ECMWF models. The EOF1 and EOF2 pattern explained 62.1% and 22.0% of the total ensemble variance, respectively.

forecast uncertainty where low and high spread occurs, respectively. The EOF method can summarize the spread information or the dominant differences between individual members while removing the redundant information among them.

We use an EOF analysis across the ensemble members at the VT to determine the dominant patterns of variations in ensemble MSLP forecasts over a verification region. Following Zheng et al. (2017), the principal components (PCs) corresponding to the leading two EOF patterns are used as a base to perform fuzzy clustering on the ensemble MSLP forecasts over the verification region. Each member is assigned a weight that identifies its relative strength of membership to each of the five clusters depending on its distance from the cluster mean in the PC phase space. A member is assigned to the cluster with the largest weight. Note that Du and Zhou (2011) also used the distance-based weight calculations for the rank of the ensemble members. Zheng et al. (2017) has discussed the application of EOF/fuzzy clustering method to diagnose the forecast scenarios in an ensemble. More details of this approach are described in the supplemental material. This work will focus on its application to model evaluations. We will only present verification over region 1 as the results from two regions are consistent.

Figure 1 shows an example of the leading two EOF patterns associated with an East Coast cyclone verified at 0000 UTC 30 December 2012. EOF1 shows a monopole structure, representing the intensity uncertainty in the three models while EOF2 shows a dipole, representing the location uncertainty along the west-southwest to east-northeast direction. Figure 2 shows the five-cluster solution for the 90-member combined ensemble using fuzzy clustering method.

After the analysis is available, the scenario closest to the analysis can be used to verify the ensemble forecasts (Zheng et al. 2017), hence we call this method ‘‘scenario-based ensemble verification.’’ Since our cluster analysis

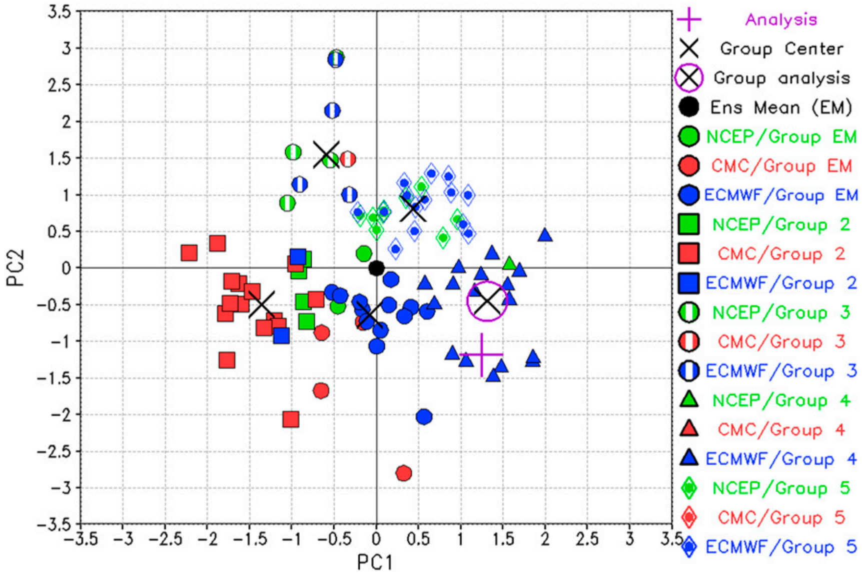

FIG. 2. Five-cluster solution based on EOF PC1 and PC2 metrics of 90-member ensemble forecasts for MSLP. Purple plus sign is the position of projected analysis on EOF1 and EOF2. Group 4 is the analysis group or "Group ANA." The IT and VT are same with Fig. 1. Note that the CMC model is not contributing to Group 4 so there is no red triangle on the plot.

is based on EOF PC space, the analysis scenario is also defined based on the projection (von Storch 1999) of the analysis onto the leading EOFs. The analysis anomaly A' relative to the ensemble mean at the VT is projected onto the leading two EOF patterns (E_1 and E_2) by using the following equation:

$$\alpha_i = \frac{\mathrm{cov}(A', E_i)}{\mathrm{var}(E_i)}, \quad i = 1, 2, \tag{1}$$

where α_i stands for the projection coefficient of the analysis anomaly onto the EOF 1 or 2 patterns. The "cov" in Eq. (1) means covariance between two scalars while the "var" represents the variance of a scalar. This is based on the property that the EOF patterns are all orthogonal to each other. Therefore, the verifying analysis at the VT is translated onto the EOF PC1–PC2 phase space by adding the pair of projection coefficient α_i ($i = 1, 2$) to the PC1–PC2 scatterplot (e.g., purple plus in Fig. 2). The cluster with the center having the shortest distance to the analysis point is considered to represent the analysis more closely than the remaining clusters and is defined as "Group ANA" or the "analysis scenario." Note that the analysis point is not included in the clustering procedure to prevent modifications of the cluster assignments.

Zheng et al. (2017) have shown that the analysis scenario defined this way can represent the analyzed cyclone as well as its associated precipitation better than the remaining scenarios, including the ensemble mean scenario. The analysis scenario tends to have fewer errors than other subgroups when verified by conventional metrics such as root-mean-square error and pattern correlation coefficient, not only at the VT, but also over a period of time (can be up to 3.5 days or more) centered at the VT. The selection of the analysis group is hence reliable to be applied to model evaluations for cyclone forecasts.

d. Verification based on EOF PC metrics

When using the projection of the analysis onto the PC space and the definition of the analysis group as a verification tool in ensemble verification, an assumption of this application is that the forecast errors do project primarily onto the leading two EOF patterns. To examine whether this holds true, the fraction of squared error (error here is defined as ensemble mean minus the analysis) explained by each EOF pattern is calculated following Eq. (2):

$$\text{fraction of squared error} = \gamma_i \times \mathrm{cov}(\mathrm{Err}, E_i) \Big/ \sum_{i=1}^{i=M} [\gamma_i \times \mathrm{cov}(\mathrm{Err}, E_i)], \tag{2}$$

where $\gamma_i = \mathrm{cov}(\mathrm{Err}, E_i)/\mathrm{var}(E_i, E_i)_{,}\ i = 1, 2, \ldots, M$. The M in Eq. (2) is the ensemble member size 90. The Err in Eq. (2) is the forecast error pattern over the domain and E_i is the ith EOF pattern. The fraction of squared error is plotted in Fig. 3 based on day 3 and day 6 forecasts for the 102 and 135 unique cyclone cases over region 1. For day 3 forecasts, the median fraction of explained squared error by the leading two EOF patterns is around 70%,

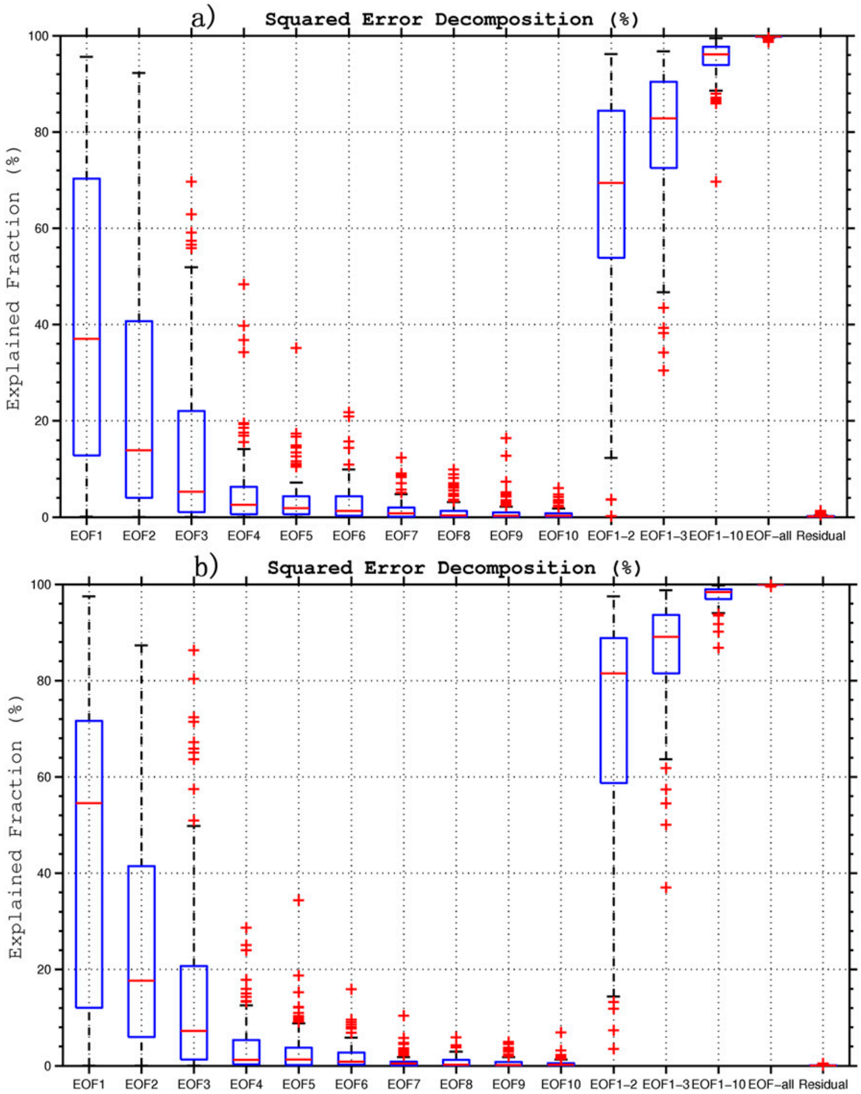

FIG. 3. The squared error fraction explained by EOF1 to EOF10 patterns labeled by EOF1, EOF2, . . . , EOF10 at the x axis, and the accumulated fraction explained by the leading 2 (EOF1–2), 3 (EOF1–3), 10 (EOF1–10), all 90 EOF patterns (EOF-all), and the residual (Residual) for (a) day 3 forecast and (b) day 6 forecast. The red center line of each boxplot shows the median value of the fraction. The blue box's bounds show the interquartile range, and the whiskers outside the bounds show the most extreme values not considered as outliers within 1.5 times the interquartile range. The plus symbols are the outliers.

suggesting that over two thirds of the squared error can be explained by the leading two EOF patterns. This value reaches 81% for day 6, suggesting that most of the forecast errors do project on the leading two patterns. There are a few cases in which other EOFs (i.e., EOF 3 or 4) explain a large fraction of the error. Unfortunately, we find that there is no correlation between the amount of error explained by a particular high-order EOF and the amount of variance explained by that EOF, thus there is no simple way to adaptively include additional EOFs into the EOF/clustering method without a priori knowledge of the analysis. Hence we have decided to focus on EOFs 1 and 2 in this study.

3. Results

a. *Scenario-based statistics*

1) OUTSIDE-OF-ENVELOPE (OOE) CASES

When analyzing the ensemble forecasts, there can be cases in which the analysis is out of the multimodel ensemble envelope. Figure 4 compares two cases with

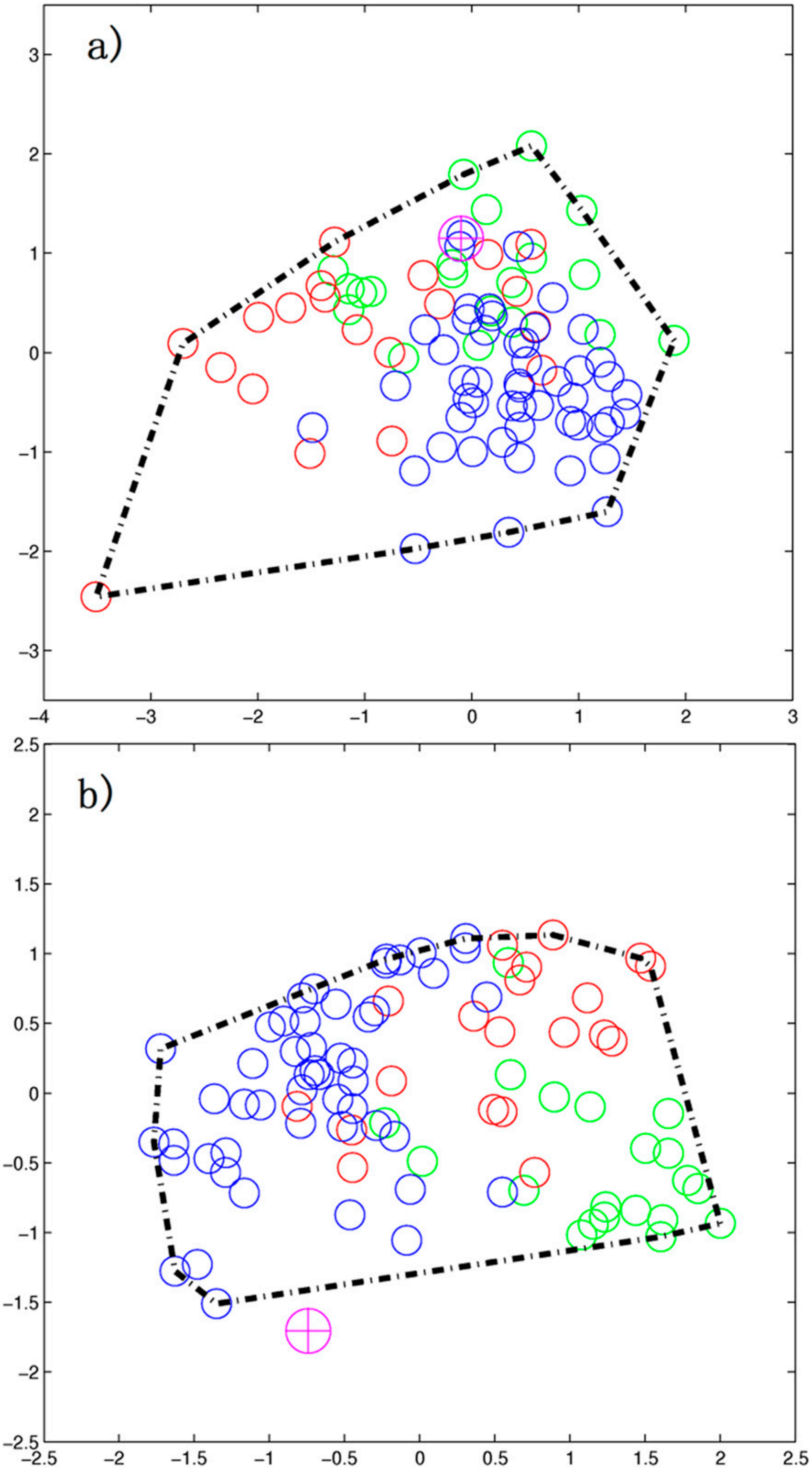

FIG. 4. (a) Inside-of-envelope and (b) outside-of-envelope examples. Green, red, and blue open circles represent members from the NCEP, CMC, and ECMWF models, respectively. Black dashed represents the outside envelope of the multimodel ensemble. Magenta circle with a plus sign denotes the analysis point. (a) 3-day MSLP forecast initialized at 1200 UTC 24 Jan 2015, and (b) 3-day MSLP forecast initialized at 1200 UTC 9 Dec 2008.

the analysis inside and outside of the forecast envelope. The analysis is located within the cluster on the PC1–PC2 space for the case shown in Fig. 4a; in contrast, the analysis is clearly out of the region enclosed by the dashed line in Fig. 4b. Quantitatively, the outlier cases or OOE cases are defined by the following criteria: 1) the analysis is outside the boundary defined by the line segments joining the vertices on the PC1–PC2 coordinate, and 2) the distance between the analysis and the closest member on the PC space is larger than the average distance between any two members plus 1 standard deviation. Among the cases we examined, there are 4 outlier cases for 3-day forecast, 16 for 6-day forecast, and 19 for 9-day forecast. For most of the statistics discussed in section 3a(2), which are dependent on the existence of an analysis group, we have excluded the corresponding OOE cases for forecast at each lead time considering that the outlier cases do not really have an analysis group and those cases should not be included in the related statistics. However, the OOE cases are still included in calculating error/spread relation statistics (section 3b) since these statistics do not directly depend on the existence of an analysis group when evaluating the multimodel ensemble to avoid biasing the results.

Given that the OOE cases are selected based on PC1–PC2 space, it is important to confirm that the case selection criteria make sense on the corresponding physical atmospheric field. Figure 5a shows the spaghetti plot for the OOE case shown in Fig. 4b. It is clear that the analysis (black dashed line) is more south-westward than most of the ensemble contours; except for three members from the ECMWF model. However, even these closest three members are much deeper and extend more west-northwest than the analysis. Figure 5b depicts the corresponding group means for the five-cluster solution, which clearly shows that the analysis is quite distinct from the mean of any of the five clusters. Therefore, the definition of the OOE case for this case appears reasonable in the physical space. We have visually examined the other OOE cases determined based on the PC1–PC2 phase space and confirmed that most of these cases are real outliers based on the physical fields (not shown).

Figure 6 shows the fraction of OOE cases for each lead time. During the lead times of 1–3 days, there are only less than 3% of OOE cases, which might indicate the multimodel ensemble is not underdispersed. The fraction increases to 5% on day 5, reaches 8% for days 6 and 7, and further increases to ~14% on days 8 and 9. Overall, the OOE fraction increases with lead time in medium to extended range, which could be associated with an increasing error-spread underdispersion (see discussions below) due to the decreasing spread or growing model biases in the forecast or both.

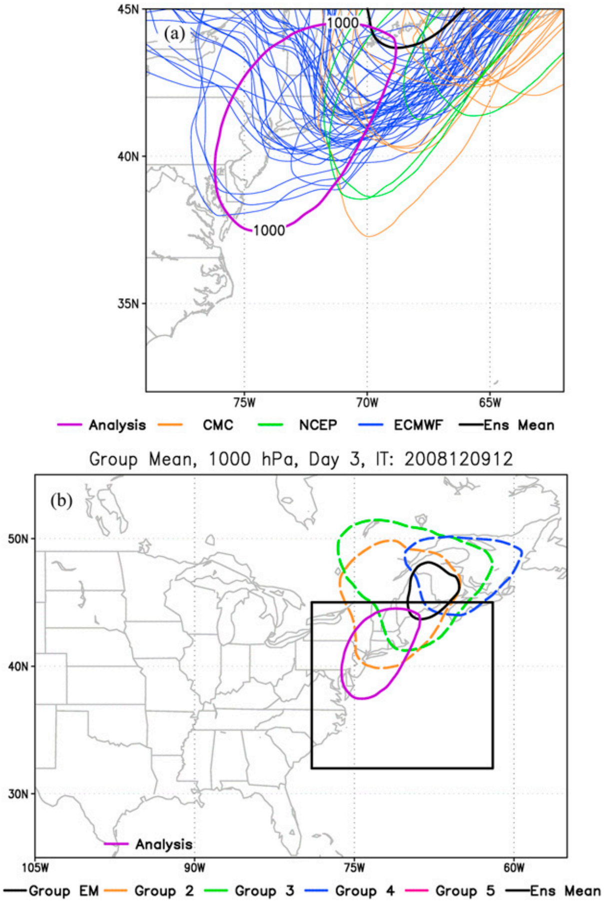

FIG. 5. (a) Spaghetti plot of 1000-hPa contour line for the 3-day forecasts initialized at 1200 UTC 9 Dec 2008: the purple contour is the analysis; black contour is the multimodel ensemble mean valid at 1200 UTC 12 Dec 2008; blue, red, and green represent members from the ECMWF, CMC, and NCEP centers, respectively. (b) Cluster mean plot for the 1000-hPa contour line for the five clusters. Black solid is the multimodel ensemble mean, and the purple contour is the analysis.

2) HIT AND MISS RATE FOR EACH MODEL

The analysis group has been determined based on the method discussed in section 2c. Note that the OOE cases have been excluded from the calculation in this subsection. We have examined the hit rate and the miss rate for each ensemble model and the multimodel. The NCEP, CMC, and ECMWF have a total of 20, 20, and 50 ensemble members, respectively. The hit rate is a percentage calculated by counting the number of each model members in the analysis group and dividing it by the total member number of each model. The average percentage is then calculated based on all the historical cyclone cases for each lead time. The miss rate is

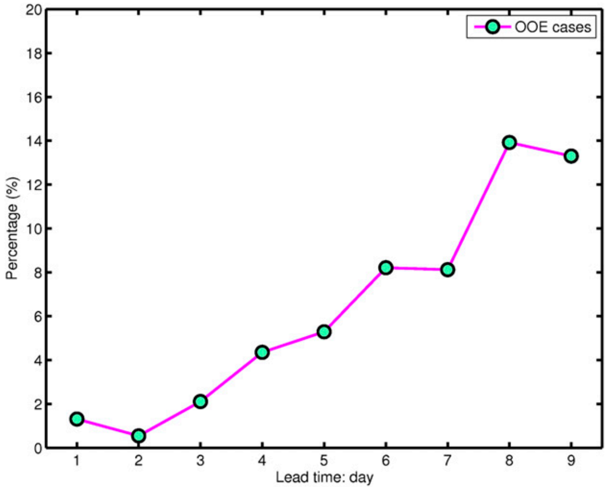

FIG. 6. The fraction of OOE cases to all observed cyclone cases for each forecast lead time.

calculated by counting the cases with zero members in the analysis group and dividing it by the total case number at each lead time. Figure 7a shows the average hit rate of the ensemble members contributing to the analysis group with respect to the total number of ensemble members for NCEP, CMC, and ECMWF over region 1 at lead times of 1–9 days. For short-range forecast (days 1–2), there are around 26% of NCEP members contributing to the analysis group. The highest percentage of members among the three EPSs, suggests that the NCEP is better in capturing the analysis scenario in short-range forecast. In contrast, the CMC model has the lowest percentage (~18%) contributing to the analysis scenario among the three models. The hit rate for the ECMWF model (~22%) is significantly higher than the CMC model but slightly less than the NCEP model. The combined CMC and NCEP ensemble, often referred to as the North American Ensemble Forecasting System (NAEFS), has comparable percentage for hit rate with the ECMWF model. Note that the conclusions for short-range forecast are not changed by using different operational analysis or including the control members. For example, Table 2 shows the comparisons of day 1 forecast verified by the ECMWF and NCEP analyses, respectively. The NCEP EPS still has the highest hit rate among the three EPSs when using the ECMWF analysis with a slightly lower value than using the NCEP analysis.

For medium-range forecast (days 3–6), the ECMWF model has the highest percentage (~23%) of members assigned to the analysis group, suggesting that it does the best job in capturing the analysis scenario in medium-range forecasts. The CMC model still has the lowest percentage (~17%), but they are not substantially lower than that for the NCEP model (~17%–22%) except for day 3. The average percentage of the NCEP model drops 10% from day 2 to day 6, suggesting its decreasing forecast skill in the medium range in capturing the analysis scenario. The NAEFS has the hit rate (~17%–19%) same with the average of the NCEP and CMC models for each forecast time.

For extended-range forecasts (days 7–9), the NCEP model's percentage increases by ~5% since day 6 and becomes the highest one among the three models. Meanwhile, the average percentage for the ECMWF

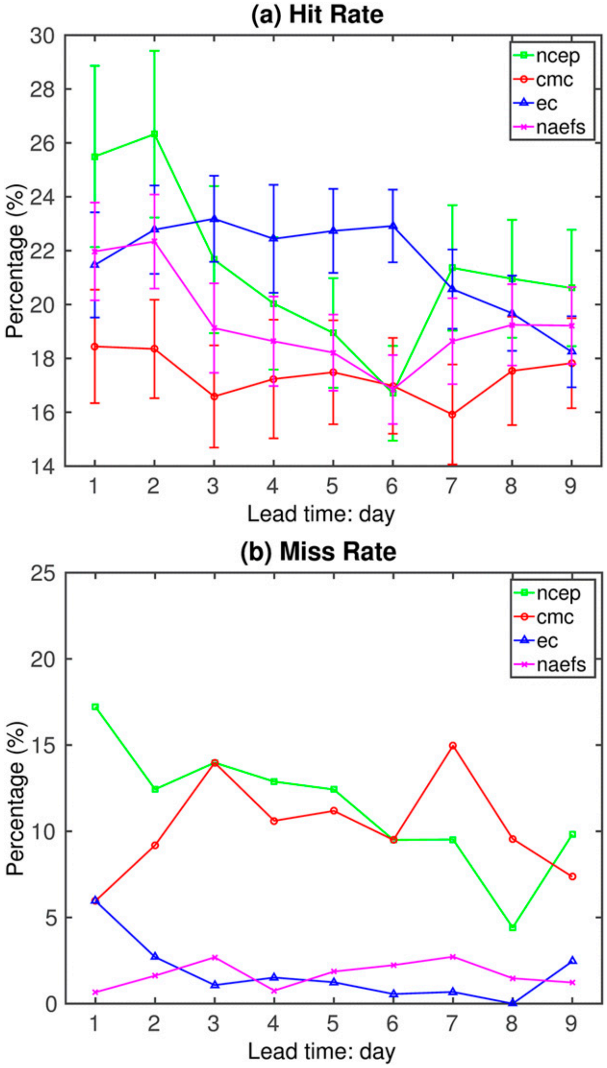

FIG. 7. (a) Average percentages (hit rate) of members in the analysis group for each EPS and NAEFS at each lead time. The vertical bars represent the 95% confidence level for each lead time. (b) The percentage (miss rate) for each EPS and NAEFS that misses the analysis group based on all cases at each lead time.

TABLE 2. Hit rate and miss rate of each ensemble model for day 1 forecast using the ECMWF analysis and the NCEP analysis for day 1 forecast, respectively. The control members from three EPSs are also included when calculating the statistics. The values in parentheses for each hit rate are the corresponding standard error.

	NCEP	CMC	ECMWF	NAEFS
Hit rate (ECMWF analysis)	26.1% (3.5%)	19.4% (2.0%)	20.9% (2.0%)	22.7% (2.0%)
Hit rate (NCEP analysis)	27.3% (3.5%)	19.2% (2.1%)	20.3% (2.1%)	23.2% (2.0%)
Miss rate (ECMWF analysis)	15.1%	2.2%	4.3%	1.4%
Miss rate (NCEP analysis)	10.7%	2.2%	5.8%	0.72%

model decreases by ~4% since day 6, but it is not significantly lower than the NCEP model. The average percentage for the CMC model is still the lowest one among the three, and it increases ~2% from day 6 to day 9. On day 9, the CMC is not significantly different from the ECMWF model.

In some cases, one or two EPS(s) can have zero number of members in the analysis group. In other words, the EPS(s) fails to predict that particular analysis scenario. We define a miss rate for each EPS to represent the fraction of cases that the EPS fails to predict the analysis scenario. Figure 7b shows the miss rate for each EPS and the NAEFS. For short-range forecasts (days 1–2), the NCEP model has the largest missing rate (>12%) among the three models. The ECMWF model miss rate is ~6% on day 1, but it reduces to <3% on day 2. The CMC model also has a miss rate value of ~6% on day 1, but it increases to around 9% on day 2. During the medium-range (days 3–6), the NCEP and CMC models have comparable missing rates (9%–14%), which are much larger than that of the ECMWF model (~1%). For the extended range (days 7–9), the CMC has larger miss rate (~8%–15%) than the NCEP model (~5%–10%) except for day 9, while the ECMWF model still has the smallest miss rate (<4%) among the three models. One thing worth noting is that the NAEFS has comparable miss rate (~2%–3%) as the ECMWF model, suggesting one benefit of combining the NCEP and CMC ensembles is to significantly reduce the miss cases in forecasting winter storms.

b. *PC metric-based statistics*

1) ERROR-SPREAD RELATION

One widely accepted measure of the utility of an EPS is the relationship between its forecast accuracy and ensemble spread. From Fig. 7, we have seen the capability of different EPSs in capturing the analysis scenario. The performance of an EPS is partly depending on if the EPS could provide reliable forecast variability in simulating East Coast storms' forecast error. Here, the error-spread relation in different individual models as well as the multimodel is examined under the framework of EOF PCs. Note that the OOE cases are included in calculating the following statistics in this and the next subsections to avoid biasing the results.

To investigate the dispersion characteristic of each EPS, the RMSE is calculated from the differences between the magnitude of the projection of the individual ensemble member onto an EOF of the multimodel ensemble (i.e., the PC) and the magnitude of the analysis onto that EOF. Then the spread is calculated by getting the PC magnitude from each individual member, calculating the distances relative to the ensemble mean and taking the variance of these distances. The error–spread ratio is calculated for each ensemble model and the multimodel ensemble by dividing the RMSE by the spread calculated from all cases. Figures 8a and 8b show the error–spread ratios for the leading two EOF PCs based on the observed cyclone cases at each lead time for all three EPSs and the multimodel ensemble. Figures 8c and 8d show the same results with the bias removed, which we will discuss in next subsection. A value of 1 represents the perfect relation. A model is considered to be underdispersed if the ratio is greater than 1; otherwise, it is overdispersed. For short-range forecasts, both PC metrics suggest that the NCEP ensemble is severely underdispersed (ratio >1.55), which partly explains the larger miss rate that is shown in Fig. 7b. The CMC model is the least underdispersed among three individual EPSs on day 1 for both PCs. The ECMWF model is less underdispersed than the other two EPSs during days 2–3 for both PCs. For medium-range forecasts, both the NCEP and CMC models are more underdispersed than the ECMWF model for PC1 metric (Fig. 8a). The NCEP ensemble shows the highest underdispersion in PC2 (Fig. 8b) while the other two show comparable underdispersion. As for extended-range forecasts, all three models are severely underdispersed in PC1 (Fig. 8a) with comparable error-spread ratios. The ECMWF model exhibits less underdispersion than the other two EPSs in PC2 (Fig. 8b).

For PC1 (Figs. 8a,b), the multimodel ensemble shows the least underdispersion in short- to medium-range forecasts for PC1 and in all lead times for PC2. The underdispersion is increasing with lead time for both PCs except for that it reaches a secondary peak on day 4 for PC2. For the extended range, the error–spread skills for all EPSs converge toward too low of a spread, which is consistent with previous studies (e.g.,

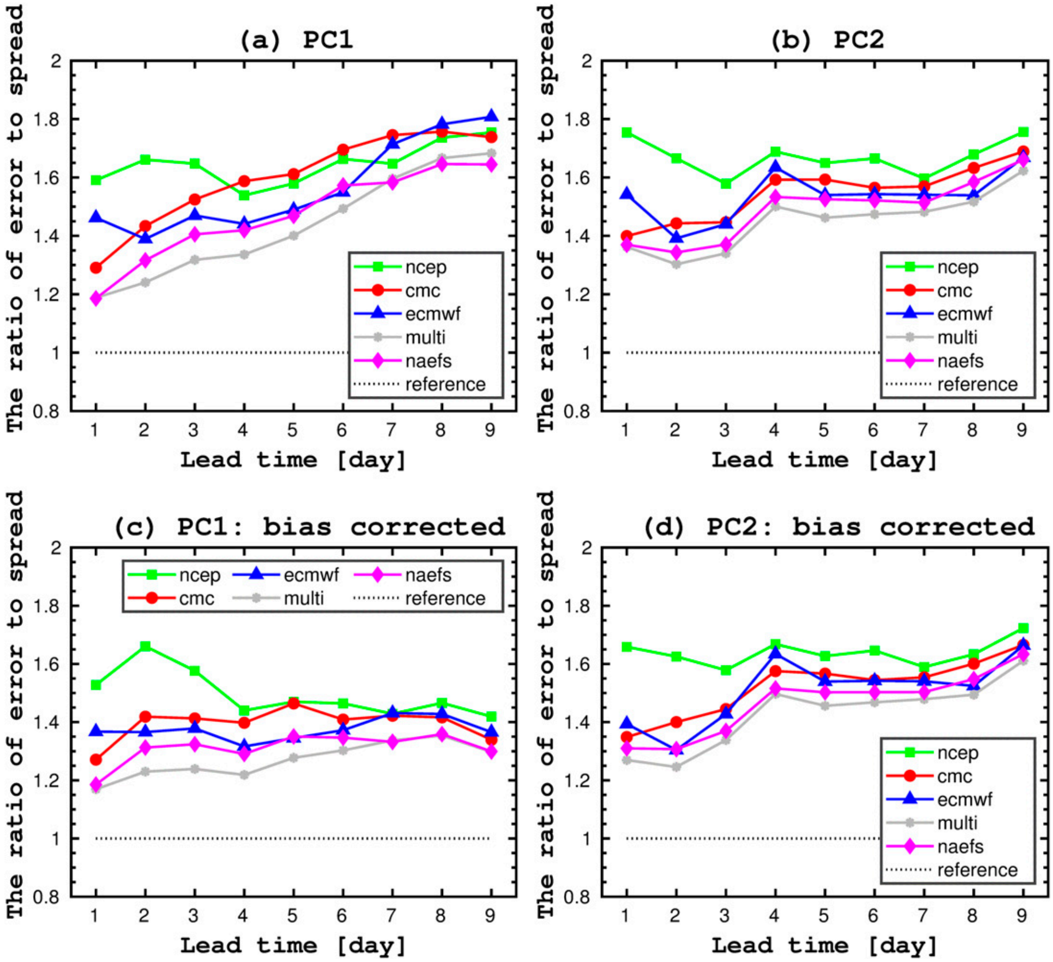

FIG. 8. The ratio of forecast error to the spread of (a) PC1 and (b) PC2 at each lead time for each EPS. (c),(d) As in (a) and (b), but the forecast bias at each lead time is removed when calculating the forecast errors.

Park et al. 2008). The NAEFS ensemble outperforms the ECMWF in short- to extended-range forecasts for PC1 and short- to medium-range forecasts for PC2. Overall, the multimodel ensemble and the NAEFS show a lower underdispersion for short- and medium-range forecasts than all the individual model ensembles, suggesting the benefit of combining different EPSs in the short and medium range.

Since all three models experienced updates during the time period we examined, the error-spread ratios have been recalculated for two different periods: the first half (2007/08–2010/11) and the second half period (2011/12–2014/15), separately. Figure S2 shows the error–spread ratio for the two periods. For the PC1 metric, the NCEP model becomes even more underdispersed in the second half period for day 1–2 lead times. Beyond day 2 (5), the NCEP (ECMWF) EPS shows improved error-spread relation in the second period while the CMC becomes more underdispersed beyond day 3. For PC2, all three models overall show improvement on day 5–7 during the second period. The NCEP model shows the largest improvement among the three models for PC2.

To sum up, the NCEP ensemble is severely underdispersed in the short range for both PC metrics, suggesting the NCEP ensemble may not have enough ensemble dispersion. Connecting this with its higher percentage of members contributing to the analysis scenario (Fig. 7), a preliminary conclusion is that the NCEP ensemble has smaller forecast errors but a very narrow ensemble spread in the short range. In contrast, the CMC model is less underdispersed during the short range. Since it has lower chance to be included in the analysis group, the CMC model seems to have large forecast errors but also a broad ensemble. The ECMWF model has better error–spread relationship during the medium range, demonstrating its superior performance in the medium range which is consistent with Fig. 7. Note that during the earlier forecast time (1–3 days), the differences in the error–spread skills for the three models are larger, which could be due to their differences in generating ensemble perturbations. The NCEP EPS mainly employed the ensemble transformation with rescaling (ETR) to generate the initial perturbations during the investigated period. Zhou et al. (2016) showed that the spread from ETR is smaller than that from the ensemble Kalman filter, which was employed in the CMC ensemble. Lewis et al. (2017) found that the NCEP ensemble is also strongly underdispersive for the quantitative precipitation forecast.

The multimodel (NCEP + CMC + ECMWF) and NAEFS ensemble shows overall smaller underdispersion in short- to medium-range forecasts, suggesting the benefit of combining different EPSs to provide more forecast variability in the short- to medium-forecast range. Above results are consistent with the findings by Hagedorn et al. (2012) who analyzed forecasts of 850-hPa temperature, 2-m temperature, and 500-hPa geopotential in the extratropics. Zhou and Du (2010) also showed the benefit of combing multiple regional models in fog prediction.

2) Model uncertainty and bias based on EOF patterns

One benefit of using EOF analysis is the orthogonal property of the EOF patterns, which can also be used to perform uncertainty and error decompositions/combinations. To decompose the forecast uncertainty for the multimodel ensembles, we first computed each member's anomaly with respect to the ensemble mean for each case and repeated this calculation for all cases. An EOF analysis was then calculated on the combined set of anomalies for all members and all cases. The leading patterns represent the dominant forecast uncertainties in all cyclone forecasts.

Figure 9 shows the dominant uncertainty patterns in all ensemble members in forecasting the observed cyclone cases from day 1 to day 6 forecasts. The leading pattern (EOF1, Fig. 9) tends to have a monopole structure at all forecast times, which is a bit northward (westward for day 6) of the ensemble mean cyclone (purple dot). This pattern suggests the main uncertainty in forecasting East Coast storms is the intensity together with a shift (~150–220 km) of the center in the north–south direction except day 6. The variance explained varies from 35.9% (for day 1) to 56.6% (for day 6). The second pattern (EOF2, Fig. 10) is a dipole pattern representing a southwest and northeast shift in cyclone position. The explained forecast variance by the leading two EOF patterns increases from 53.8% for day 1 to 76.7% for day 6. Therefore, the leading two patterns represent most of the forecast uncertainties.

The systematic errors on each EOF direction can also be calculated. On the above EOF PC1–PC2 space, each member occupies one point for one case. When the analysis is projected onto the corresponding EOF1 and 2 patterns, it also occupies one point for each case. The difference in the abscissa or the ordinate between each member and the analysis reflects the relative error for that member in EOF PC1 or PC2. By averaging the differences between the members from one EPS and the analysis, the systematic error associated with EOF1/2 pattern corresponding to that EPS can be evaluated. The original error fields can be approximately reconstructed from the reduced set of the two leading PCs and their associated EOF patterns. The reconstructed error fields can be further decomposed into cyclone intensity (minimum pressure) and displacement errors. The above decompositions of forecast errors associated with the leading two patterns are combined and summarized in Figs. 11 and 12 to provide a physical understanding of the model errors based on the EOF PC1–PC2 space.

Figure 11 illustrates the minimum pressure error for each EPS as well as the combined EPSs from day 1 to day 6 forecasts derived from the leading two EOF PC errors. Figure 12 depicts the displacement errors for each EPS associated with the leading two patterns. For the short range (1–2 days, Figs. 11 and 12a,b), the ECMWF model shows slightly larger positive intensity errors (cyclones too weak) than the other models for days 1 and 2. In contrast, the NCEP model has negative intensity errors than the other models. The CMC, NAEFS, and the three ensembles combined have the errors between the above two. For the medium range (3–6 days, Figs. 11 and 12c–f), all models have positive intensity errors from 2 to 7 hPa, suggesting an underprediction of the observed East Coast winter storms. Among them, the CMC forecasts have the largest intensity errors while the NCEP forecasts have the smallest errors for days 3–5 and the ECMWF model has the smallest errors for day 6 forecasts.

Note that during days 1–4 (Figs. 12a–d), the ECMWF has the largest displacement errors toward the west-southwest or south-southwest while the NCEP model has the least errors and is furthest to the north or northeast among all models. For days 5–6, both NCEP and CMC are biased toward the east-southeast of east and have larger error than ECMWF. When comparing with the other models, the ECMWF model is always to the west-southwestward direction, suggesting that when forecasting East Coast winter

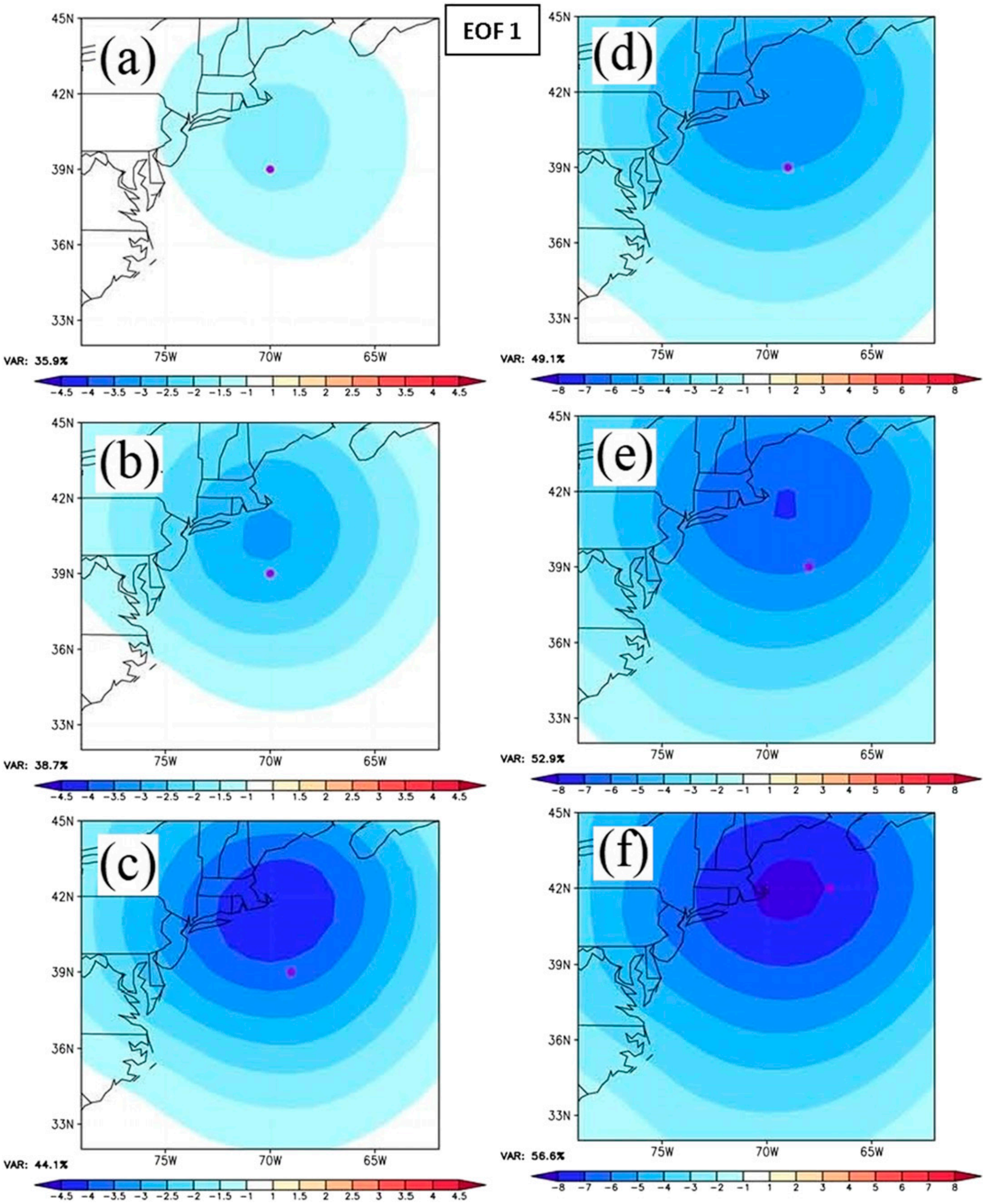

FIG. 9. (a)–(f) The first EOF pattern (shaded) in the multimodel ensemble from day 1 to day 6 forecast. The purple dot represents the ensemble mean cyclone position averaged over all cases for each lead time. The explained variance for this pattern (see bottom-left numbers on each panel) increases from 35.9% for day 1 lead time to 56.6% for day 6 lead.

storm, one tendency of the ECMWF model is to forecast them too west-southwestward than the other two models. This tendency likely causes larger west-southwestward displacement errors in the ECMWF model for days 1–4 than the other two models. This could be partly due to the slow propagation in the ECMWF model as suggested by Froude (2010).

When taking the bias into consideration, the error–spread skills could be changed. Bias correction is not straightforward here since error–spread is examined in the PC space instead of physical space. Nevertheless, the discussions above suggest that the leading EOF patterns for most cases are similar, and that the models exhibit significant bias in the PC space. Therefore, we have also

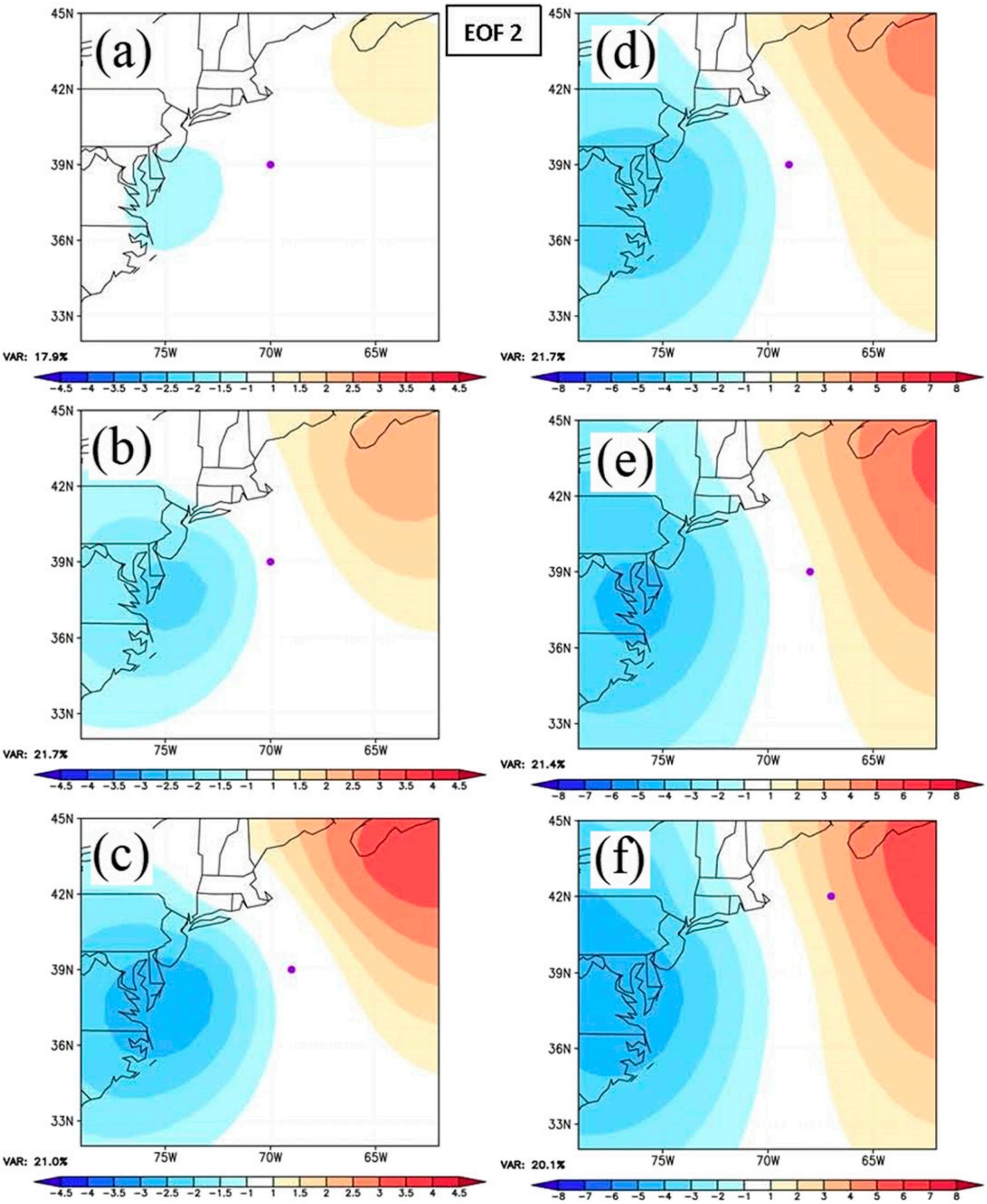

FIG. 10. As in Fig. 9, but for EOF2 pattern. The explained variance for this pattern varies from 17.9% to 21.7% (see bottom-left numbers on each panel).

recalculated the error–spread ratio after removing the bias for each EPS. For each case, the mean bias over all cases for each PC is removed from each EPS after ensuring that the EOF for that case projects positively on the dominant EOF pattern for each PC shown in Figs. 9 and 10. Figures 8c and 8d show the bias-corrected error–spread ratio for each PC. The ratios for all EPSs are reduced when compared with those calculated with bias. The skill of each EPS or combined ensemble becomes stable for PC1 during the medium and extended ranges. Other results are consistent with the original error–spread ratio plots (Figs. 8a,b). The bias-corrected ratios for two time periods are also presented in Figs. S2c and S2d. Comparations among different EPSs are consistent with the results from Figs. S2a and S2b, except for that the increasing trend of underdispersion with lead time does not hold true for PC1. Here, bias is corrected as a posteriori, hence guaranteeing a reduction in the error.

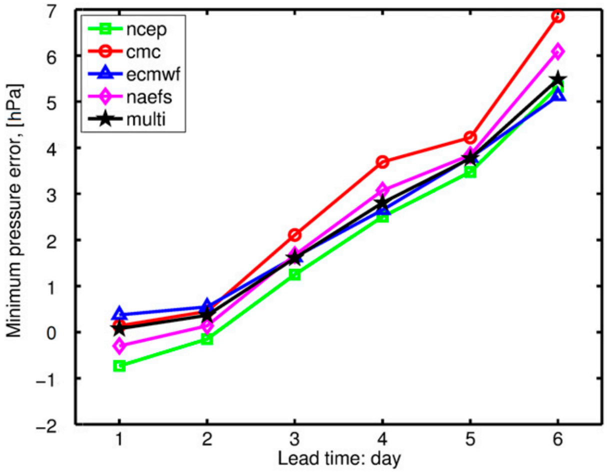

FIG. 11. Mean minimum pressure error (hPa) of each EPS and the combined model for PC1 and PC2 errors from day 1 to day 6 forecast.

In future work it would be of interest to examine whether a priori bias correction in PC space in the forecast based on historical bias can also lead to reduction in errors.

Note that in calculating the forecast biases (as well as for the error–spread relation discussed in the previous subsection), all observed cyclone cases and ensemble members are included in the calculations, hence we believe that the evaluations conducted in this study can provide results that are complementary to those provided by cyclone matching (e.g., Froude 2010). The cyclone matching can directly verify forecast errors in intensity and displacement but can only verify cases in which many ensemble members can be matched to the observed cyclone, as well as verifying only the ensemble members that can be matched, which represent a decreasing fraction of all cyclone forecasts as the lead time is increased. However, since only observed cyclone cases were used in this analysis, and not those cases simulated by some members but not observed, the errors here may not represent the full ensemble bias.

4. Discussions

a. *Utilizing the ensemble mean clusters*

In general, the ensemble mean forecast when averaged over many cases is assumed to be closer to the truth than any of the individual forecasts in an ensemble (Leith 1974; Murphy 1988; Whitaker and Loughe 1998). As a result, the ensemble mean is most widely used to represent the best available estimate of the future state of the atmosphere. Du and Zhou (2011) proposed an ensemble ranking method based on the distance between members to ensemble mean under the assumption of no model bias. However, the difficulty to use ensemble-mean based method is that model has bias in reality. If a model has bias, it is expected that the ensemble mean will not be verified the best. How model bias impacts ensemble verification has been recently investigated by Wang et al. (2018), who show that an ensemble verification could be completely misleading with model bias. Therefore, the ensemble mean is insufficient to be useful guidance for a forecaster when there is a pending severe winter storm. For example, for the January 2015 case discussed by Zheng et al. (2017), the ensemble mean had the cyclone closer to the U.S. East Coast while a subset of ensemble members suggested a cyclone more to the northeast. If the forecaster relied too heavily on the ensemble mean and ignored the other scenarios a wrong forecast could be made. Therefore, a verification of the ensemble mean is necessary to investigate whether the ensemble mean is really better than other subsets of an ensemble. Scenarios based on PC1–PC2 phase space provide a new perspective to verify the ensemble mean. Since the analysis group is a representation of the analysis scenario while the ensemble mean group (Group EM) is often a representation of the ensemble mean scenario, it is straightforward to verify whether the ensemble mean group is similar to the analysis group. If the ensemble mean group includes the analysis scenario more often than the other scenario groups, it will suggest that the ensemble mean does have a better skill than the other subsets of the ensemble in terms of the capability of including real development scenarios.

Figure 13a shows the fraction of cases in which the ensemble mean group is the same as the analysis group. For short-range forecast (1–2 days), the ensemble mean group does include the analysis scenario more often (~40%) than the expected average chance (20%) of each group. This indicates that the ensemble mean scenario is more reliable than the other ensemble groups. However, beyond day 2, the percentage is only slightly higher than the expected percentage in the medium range (days 3–6) and even less than the expected average in the extended range. This suggests that the ensemble mean group has no advantage when compared with the other groups for the medium and extended range forecasts. Even in the short-range forecast, there are still around 60% of cases with the analysis group being different from the ensemble mean group. Therefore, focusing on the ensemble mean too much could be misleading in most cases given model biases.

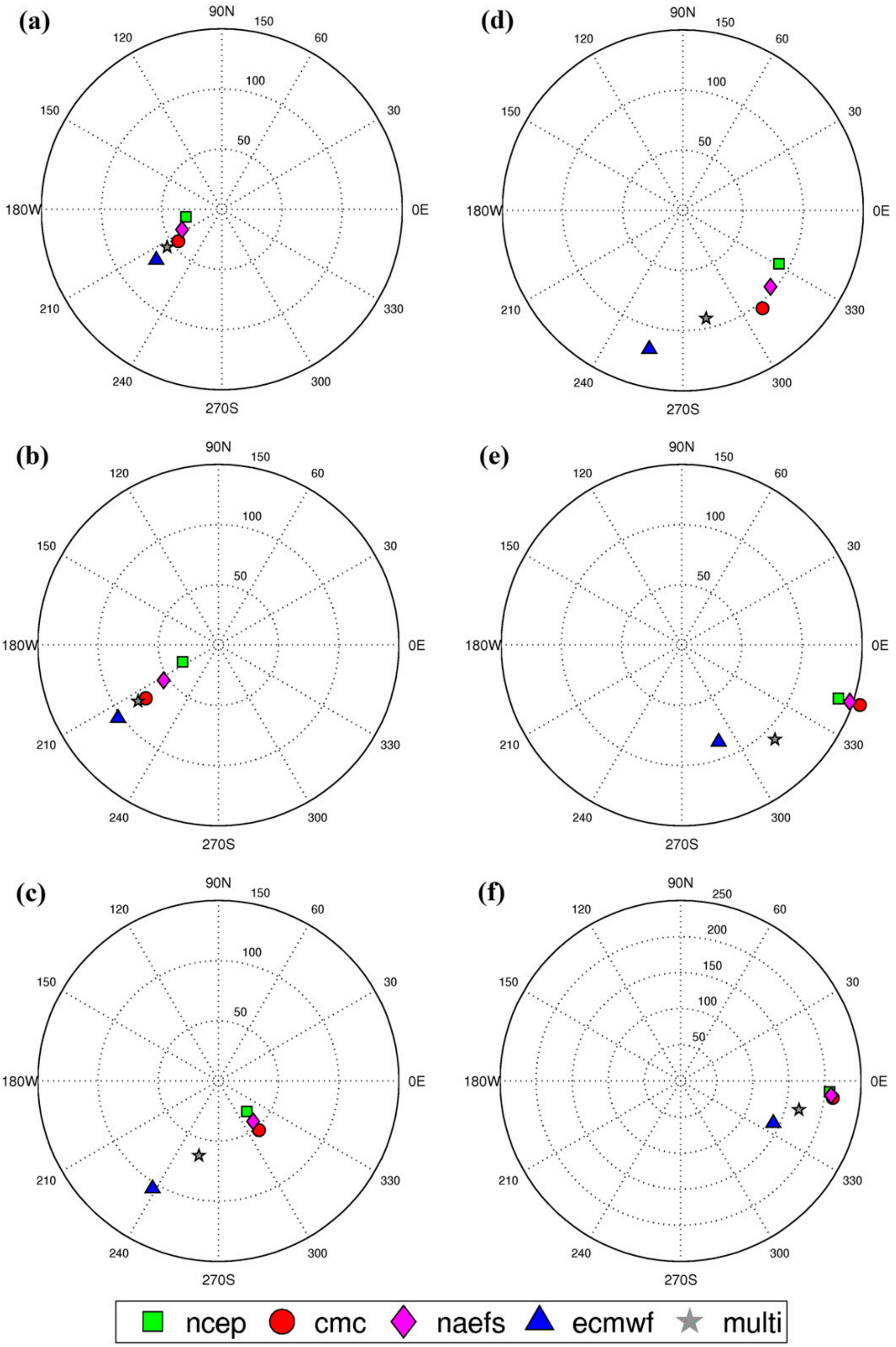

FIG. 12. The displacement errors corresponding to the combined PC1 and PC2 errors for each EPS as well as the combined models for (a)–(f) day 1 to day 6 forecast. The radius has unit of kilometers. Note that (f) has a radius of 250 km instead of 150 km.

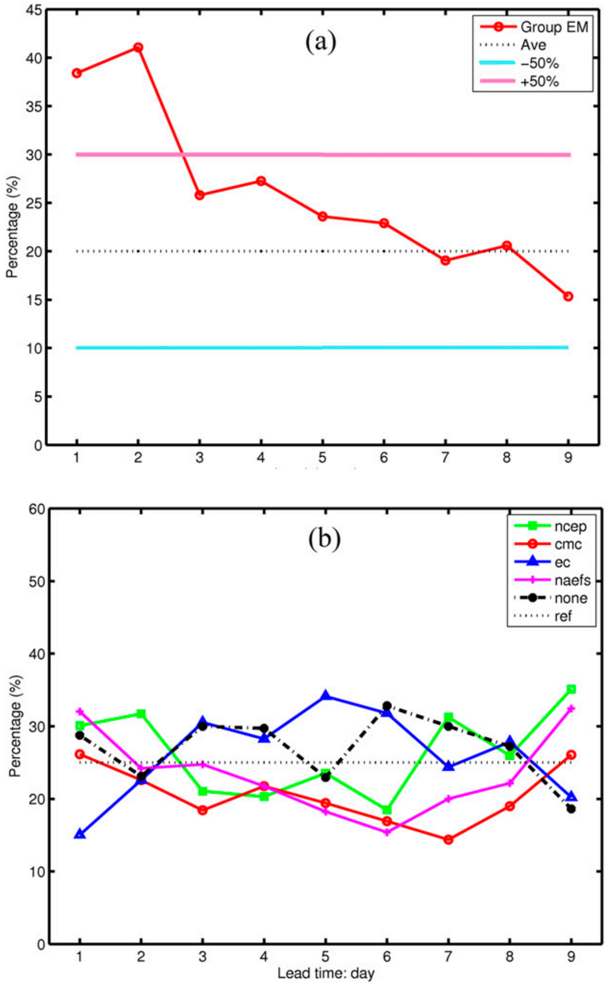

FIG. 13. (a) The fraction of cases (red line with dot) with the Group EM same with Group ANA for all cases at each lead time. For the five-cluster solution, each group is supposed to have 20% (black dotted line) chance (average chance) to be similar with the Group ANA. The cyan and magenta solid lines represent 50% less and 50% more chance than the average chance to include analysis group, respectively. (b) The fraction of cases in which the analysis falls into the same quadrant as NCEP, CMC, ECMWF, and NAEF means, as well as in a quadrant in which none of the three EPS means are in, which is defined as ''none'' in the legend.

In operational forecasting, the individual model means are also used a lot, especially the ECMWF mean. Previous studies suggested that the ECMWF model is the ''best'' global ensemble model (e.g., Buizza et al. 2005; Keller et al. 2011). Operational forecasters have the tendency to hedge toward the ECMWF forecast direction when the forecast uncertainties are large. Figure 13a has shown that in the medium range the analysis scenario is not more likely to be in the Group EM. Here, we examine how often the analysis tends to be in the direction of individual model means on the PC1–PC2 space. We investigate this question by examining whether the projection of the analysis on the PC1–PC2 space is in the same quadrant as the different EPS means. Figure 13b shows the percentage of cases in which the analysis falls within the same quadrant as each EPS mean at each lead time. The percentage that the analysis is located outside of any of the EPS mean quadrants is also shown for comparison. Since there are four quadrants, if everything is random, each EPS mean should have a probability of 1/4 (25%) of being within the same quadrant as the analysis. As can be seen from Fig. 13, the ECMWF model does show higher chance during the medium range to be in the analysis quadrant. However, even the highest percentage (for day 5) is ~35%. The NCEP and NAEFS show slightly higher chance for days 1–2 and day 9 to be in the analysis quadrant. One thing worth noting is that there are a moderate number (~20%–30%) of cases in which the analysis lies in a quadrant in which none of the three EPS means fall. The quadrant statistics in Fig. 13b suggest that although the ECMWF ensemble shows a slightly higher chance to be in the quadrant in which the analysis falls for the medium range, it also misses the analysis cluster in around two-thirds of the cases. These results have practical implications in operational forecasting, suggesting that it is not a good practice for forecasters to hedge toward the ECMWF ensemble mean solution, even though the ECMWF might be the best ensemble. Note that in cases of high-amplitude, relatively small-scale features, such as a strong east coast cyclone, taking the ensemble mean will result in a smoother, larger, and weaker system than any of the members individually because of spatial differences in cyclone center among the members. Thus, the ensemble mean may reduce overall RMSE compared to individual members (e.g., Du and Zhou 2011), but the storm structure could also be changed or distorted. Results shown in this subsection indicate that all scenarios must be taken into consideration in the formulation of a forecast and the model bias needs to be corrected first before an ensemble is used to produce forecast products.

b. *Sensitivity to number of clusters and fuzziness information*

The fuzzy clustering algorithm is often considered sensitive to the choice of cluster numbers as well as the fuzziness parameters. In this work, we have used a fixed number 5 for the clustering process. However, we have tested the results for two- to eight-cluster solutions using 100 random seedings for each case. The adjusted Rand index (ARI; Yeung and Ruzzo 2001), which measures pairwise cluster partition agreement based on the contingency tables, is used to validate the consistency among the random runs. A larger ARI means a higher agreement between two partitions. We have found that five-cluster solutions tend to be the most stable solution for most cases with the highest ARI. In addition, the second ARI, six-cluster solutions, has shown consistent statistics with the five-cluster solutions. Therefore, we consider the choice of using five clusters to be representative of the reasonable groupings in most cases.

Another issue during the clustering process is the fuzziness features of each cluster. In this application, we assign a member to one unique cluster, thus ignoring the fuzziness possibility of groupings. In other words, if a member is located in the boundary regions of two clusters, we still assign it to a slightly closer cluster. To investigate if the fuzziness could impact our results, we have examined the impact of adding a threshold to define if a member is significantly assigned to a cluster. The threshold is the difference between the mean of the membership weights of all members within a cluster minus the standard deviation of the weight. A confident member in a cluster has the membership weight greater than the threshold. Otherwise, the member is not assigned to any cluster. The significant cases (or confident cases) are selected when the weights of the analysis for the analysis group is higher than the mean minus one standard deviation of all members' strongest weights within a cluster. A total of 54 and 72 cases were chosen for day 3 and day 6, respectively.

Figure 14 compares the previous scenario statistics shown in Fig. 7a with all cluster members and with the confident members only for day 3 and day 6. For day 3 forecast, the ECMWF and NCEP models are comparable in capturing the analysis scenario when only the confident members are used for calculations, which is consistent with the results using all members in the analysis group. The CMC model again shows the lowest percentage in the three models in analysis group. For day 6, the percentage of ECMWF members in analysis group is significantly higher than the other two models, which is also consistent with the results using all members. Thus, the results for analysis members with and without insignificant members mentioned above are consistent, suggesting that the overall results in this study are not sensitive to the choice of the fuzziness parameter.

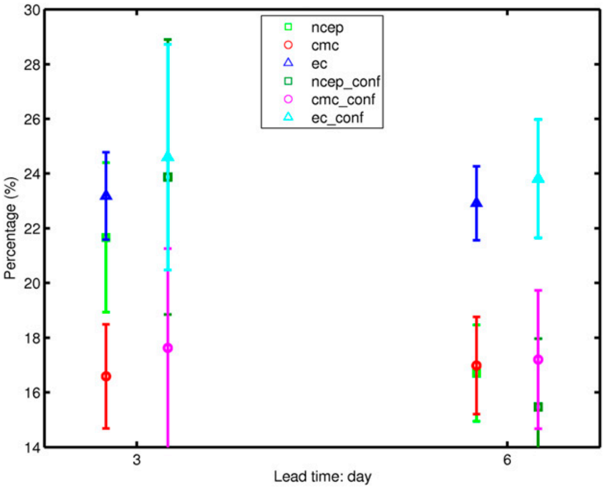

FIG. 14. As in Fig. 7a, but for significant cases of 3- and 6-day forecasts when the analysis falls clearly within the analysis group. The dark green square, magenta circle, and cyan triangle (end with _conf in legend) are for NCEP, CMC, and ECMWF models in significant cases while the green, red, and blue markers are the same with Fig. 7a for the purpose of comparisons.

5. Summary

A scenario-based ensemble verification method is applied to examine the capability of different EPSs in capturing the analysis scenarios for 158–185 historical East Coast cyclone cases at lead times of 1–9 days. The scenario separation uses the PCs of the two leading EOFs computed from a set of ensemble forecasts as a base. Fuzzy clustering is then performed to separate the ensemble into a fixed number of 5 clusters. By projecting the analysis onto the PC coordinates, the analysis group is found by searching for the closest cluster.

The OOE cases are defined by the analysis falling outside the boundary occupied by the ensemble members in the PC phase space. We find that there are less than 3% of OOE cases for short-range forecasts (days 1–3), 5%–8% for medium range (days 4–7), and 8%–14% for extended range (days 8–9). The OOE fraction increases with the leading time from the medium to the extended range, possibly indicating increasing underdispersion of the multimodel ensemble.

With the OOE cases excluded, we have computed the performance of the three models regarding hit rate and miss rate for the analysis scenarios. The NCEP ensemble members have the highest probability to be included in the analysis group for the short-range forecasts; however, this ensemble also has the highest miss rate. The ECMWF ensemble shows the best performance in the medium range with the highest percentage contributing to the analysis group and the lowest missing rate among the three EPSs, suggesting its superiority in medium-range forecasts of East Coast storms. The CMC model overall shows the smallest percentage of members contributing to the analysis group and a relatively higher miss rate, suggesting that it is less reliable in capturing the analysis scenario. However, it has lower miss rate than the NCEP model in the short range. The combination of the CMC and NCEP models can reduce the miss rate significantly, demonstrating the value of combining the two models, which is consistent with the conclusion of Zhou and Du (2010).

The ensemble mean of a multimodel or individual models (in particular the ECMWF model) is widely used by operational forecasters to represent the best available estimate of the future state of the atmosphere. It was found that in the majority of cases (>60%), the analysis is not within the ensemble mean group for the multimodel ensemble. Meanwhile, the quadrant statistics suggest that the ECMWF model misses the analysis direction in a majority of past storms due to model bias although it shows a slightly higher chance to be in the analysis quadrant in the medium range than the other two EPSs.

To measure the quality of the ensemble, we have calculated the error-spread ratio using EOF PC metrics. The NCEP model is severely underdispersed in the short range for both PC metrics, suggesting the NCEP model may not have enough dispersion. Considering its higher percentage of members contributing to the analysis scenario, we conclude that the NCEP model has less forecast errors but a narrow ensemble spread in the short range. In contrast, the CMC model is the least underdispersed on day 1. Since the CMC model has a lower chance to be included in the analysis group, this model seems to have large forecast errors but also a broad ensemble at short range. The ECMWF model has better error-spread relationship during the medium range, demonstrating its superior performance in the medium range. On the other hand, the multimodel (NCEP + CMC + ECMWF) and the NAEFS ensemble shows less underdispersion in short- and medium-range forecasts than any individual model does, suggesting the benefit of combining different EPSs to provide more forecast variability in the medium and extended range.

The model uncertainty and biases for 1–6-day forecasts have been decomposed for the leading two EOF patterns of MSLP forecasts. The results show that for all lead times, the first EOF pattern is associated with the intensity uncertainty for the multimodel ensemble for all lead times, while the second pattern is associated with cyclone position uncertainty along either the west–east or southwest–northeast direction. The NCEP model tends to better represent the leading two EOF patterns by showing less intensity and displacement biases during 1–4 days. The ECMWF model has the smallest biases in both patterns during 5–6 days. The CMC model shows moderate biases for days 1–2 and the largest biases for days 3–6. We have also found that the East Coast cyclone in the ECMWF forecast, which is often considered as ‘‘best’’ among global EPSs, tend to be toward the southwest of the other two models in representing the leading two patterns, which suggests that the ECMWF model may have a tendency to show a closer-to-shore solution in forecasting East Coast winter storms. The error-spread skill problem is revisited after removing the model bias for each PC. The underdispersion is significantly smaller in the medium and extended range for PC1 and the increasing trend with lead time does not hold true. Otherwise, the comparisons among different EPSs are consistent with the results calculated with the model bias included.

Note that all the calculations in this work are based on the orthogonal EOF patterns. We are aware of the reduced information when only using the leading two EOF patterns. However, this work provides a complementary and novel way to verify ensemble outputs in forecasting East Coast storms compared to existing verifications based on cyclone matching. Scenario-based verification could be more efficient and intuitive than the traditional matching methods. One future direction is to apply advanced machine learning methods (e.g., the convolutional neural networks; Längkvist et al. 2016) to separate clusters and verify them with analyses.

Acknowledgments. The authors thank the TIGGE data archives for the availability of the ensemble data. We thank Zhenhai Zhang for setting up the cyclone tracking scheme. This work is supported by NOAA-CSTAR (NA13NWS4680002). Drs. Edmund K. M. Chang and Brian A. Colle are also supported by NOAA-CSTAR NA17NWS4680004.

REFERENCES

Booth, J. F., H. E. Rieder, D. E. Lee, and Y. Kushnir, 2015: The paths of extratropical cyclones associated with wintertime high-wind events in the northeastern United States. *J. Appl. Meteor. Climatol.*, **54**, 1871–1885, https://doi.org/10.1175/JAMC-D-14-0320.1.

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.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.*, **125**, 99–119, https://doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

——, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Chang, E. K., 2013: CMIP5 projection of significant reduction in extratropical cyclone activity over North America. *J. Climate*, **26**, 9903–9922, https://doi.org/10.1175/JCLI-D-13-00209.1.

Charles, M. E., and B. A. Colle, 2009: Verification of extratropical cyclones within the NCEP operational models. Part I: Analysis errors and short-term NAM and GFS forecasts. *Wea. Forecasting*, **24**, 1173–1190, https://doi.org/10.1175/WAF2222169.1.

Colle, B. A., and M. E. Charles, 2011: Spatial distribution and evolution of extratropical cyclone errors over North America and its adjacent oceans in the NCEP global forecast system model. *Wea. Forecasting*, **26**, 129–149, https://doi.org/10.1175/2010WAF2222422.1.

——, J. F. Booth, and E. K. M. Chang, 2015: A review of historical and future changes of extratropical cyclones and associated impacts along the US East Coast. *Curr. Climate Change Rep.*, **1**, 125–143, https://doi.org/10.1007/s40641-015-0013-7.

Davis, R. E., and R. Dolan, 1993: Nor'easters. *Amer. Sci.*, **81** (5), 428–439.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting (SREF) at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes*, Fort Lauderdale, FL, Amer. Meteor. Soc., P4.9, http://ams.confex.com/ams/pdfpapers/23074.pdf.

——, and B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev.*, **139**, 3284–3303, https://doi.org/10.1175/MWR-D-10-05007.1.

——, S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

——, G. DiMego, M. S. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: Multi-IC, multi-model and multi-physics approach. J. Cote, Ed., Research Activities in Atmospheric and Oceanic Modelling Rep. 33, WMO/TD 1161, 5.09–5.10.

Froude, L. S., 2009: Regional differences in the prediction of extratropical cyclones by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **137**, 893–911, https://doi.org/10.1175/2008MWR2610.1.

——, 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.

Grams, C. M., and Coauthors, 2011: The key role of diabatic processes in modifying the upper-tropospheric wave guide: A North Atlantic case-study. *Quart. J. Roy. Meteor. Soc.*, **137**, 2174–2193, https://doi.org/10.1002/qj.891.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, https://doi.org/10.1002/qj.1895.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson, 2007: Empirical orthogonal functions and related techniques in atmospheric science: A review. *Int. J. Climatol.*, **27**, 1119–1152, https://doi.org/10.1002/joc.1499.

Harr, P. A., D. Anwender, and S. C. Jones, 2008: Predictability associated with the downstream impacts of the extratropical transition of tropical cyclones: Methodology and a case study of Typhoon Nabi (2005). *Mon. Wea. Rev.*, **136**, 3205–3225, https://doi.org/10.1175/2008MWR2248.1.

Hewson, T. D., L., Magnusson, O. Breivik, F. Prates, I. Tsonevsky, and J. W. de Vries, 2014: Windstorms in northwest Europe in late 2013. *ECMWF Newsletter*, No. 139, ECMWF, Reading, United Kingdom, 22–28, https://www.ecmwf.int/en/elibrary/17343-windstorms-northwest-europe-late-2013.

Hirsch, M. E., A. T. Degaetano, and S. J. Colucci, 2001: An East Coast winter storm climatology. *J. Climate*, **14**, 882–899, https://doi.org/10.1175/1520-0442(2001)014<0882:AECWSC>2.0.CO;2.

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.

——, 1999: Adaptive constraints for feature tracking. *Mon. Wea. Rev.*, **127**, 1362–1373, https://doi.org/10.1175/1520-0493(1999)127<1362:ACFFT>2.0.CO;2.

Keller, J. H., S. C. Jones, J. L. Evans, and P. A. Harr, 2011: Characteristics of the TIGGE multimodel ensemble prediction system in representing forecast variability associated with extratropical transition. *Geophys. Res. Lett.*, **38**, L12802, https://doi.org/10.1029/2011GL047275.

Korfe, N. G., and B. A. Colle, 2017: Evaluation of cool-season extratropical cyclones in a multimodel ensemble for eastern North America and the Western Atlantic Ocean. *Wea. Forecasting*, **33**, 109–127, https://doi.org/10.1175/WAF-D-17-0036.1.

Längkvist, M., A. Kiselev, M. Alirezaie, and A. Loutfi, 2016: Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sens.*, **8**, 329, https://doi.org/10.3390/rs8040329.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lewis, W. R., W. J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western United States. *Wea. Forecasting*, **32**, 1007–1028, https://doi.org/10.1175/WAF-D-16-0179.1.

Ma, C. G., and E. K. Chang, 2017: Impacts of storm-track variations on wintertime extreme weather events over the Continental United States. *J. Climate*, **30**, 4601–4624, https://doi.org/10.1175/JCLI-D-16-0560.1.

Mather, J. R., H. Adams III, and G. A. Yoshioka, 1964: Coastal storms of the eastern United States. *J. Appl. Meteor.*, **3**, 693–706, https://doi.org/10.1175/1520-0450(1964)003<0693:CSOTEU>2.0.CO;2.

Matsueda, M., and T. Nakazawa, 2015: Early warning products for severe weather events derived from operational medium-range ensemble forecasts. *Meteor. Appl.*, **22**, 213–222, https://doi.org/10.1002/met.1444.

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.

Molteni, F., and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.*, **119**, 269–298, https://doi.org/10.1002/qj.49711951004.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–493, https://doi.org/10.1002/qj.49711448010.

Novak, D. R., B. A. Colle, and S. E. Yuter, 2008: High-resolution observations and model simulations of the life cycle of an intense mesoscale snowband over the northeastern United States. *Mon. Wea. Rev.*, **136**, 1433–1456, https://doi.org/10.1175/2007MWR2233.1.

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, https://doi.org/10.1002/qj.334.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446, https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

von Storch, H., 1999: Spatial patterns: EOFs and CCA. *Analysis of Climate Variability*, H. von Storch and A. Navarra, Eds., Springer, 231–263.

Wang, J., J. Chen, J. Du, Y. Zhang, Y. Xia, and G. Deng, 2018: Sensitivity of ensemble forecast verification to model bias. *Mon. Wea. Rev.*, **146**, 781–796, https://doi.org/10.1175/MWR-D-17-0223.1.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302, https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Yeung, K. Y., and W. L. Ruzzo, 2001: Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774, https://doi.org/10.1093/bioinformatics/17.9.763.

Zheng, M., E. K. 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.

Zhou, B., and J. Du, 2010: Fog prediction from a multimodel mesoscale ensemble prediction system. *Wea. Forecasting*, **25**, 303–322, https://doi.org/10.1175/2009WAF2222289.1.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074, https://doi.org/10.1175/WAF-D-16-0109.1.