The Use of Ensemble Clustering on a Multimodel Ensemble for Medium-Range Forecasting at the Weather Prediction Center

WILLIAM S. LAMBERSON,[a,b] MICHAEL J. BODNER,[b] JAMES A. NELSON,[b] AND SARA A. SIENKIEWICZ[c,b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/NWS/NCEP/Weather Prediction Center, College Park, Maryland*
[c] *I.M. Systems Group, Inc., Rockville, Maryland*

(Manuscript received 1 September 2022, in final form 20 December 2022, accepted 22 January 2023)

ABSTRACT: This article introduces an ensemble clustering tool developed at the Weather Prediction Center (WPC) to assist forecasters in the preparation of medium-range (3–7 day) forecasts. Effectively incorporating ensemble data into an operational forecasting process, like that used at WPC, can be challenging given time constraints and data infrastructure limitations. Often forecasters do not have time to view the large number of constituent members of an ensemble forecast, so they settle for viewing the ensemble's mean and spread. This ignores the useful information about forecast uncertainty and the range of possible forecast outcomes that an ensemble forecast can provide. Ensemble clustering could be a solution to this problem as it can reduce a large ensemble forecast down to the most prevalent forecast scenarios. Forecasters can then quickly view these ensemble clusters to better understand and communicate forecast uncertainty and the range of possible forecast outcomes. The ensemble clustering tool developed at WPC is a variation of fuzzy clustering where operationally available ensemble members with similar 500-hPa geopotential height forecasts are grouped into four clusters. A representative case from 15 February 2021 is presented to demonstrate the clustering methodology and the overall utility of this new ensemble clustering tool. Cumulative verification statistics show that one of the four forecast scenarios identified by this ensemble clustering tool routinely outperforms all the available ensemble mean and deterministic forecasts.

SIGNIFICANCE STATEMENT: Ensemble forecasts could be used more effectively in medium-range (3–7 day) forecasting. Currently, the onus is put on forecasters to view and synthesize all of the data contained in an ensemble forecast. This is a task they often do not have time to adequately execute. This work proposes a solution to this problem. An automated tool was developed that would split the available ensemble members into four groups of broadly similar members. These groups were presented to forecasters as four potential forecast outcomes. Forecasters felt this tool helped them to better incorporate ensemble forecasts into their forecast process. Verification shows that presenting ensemble forecasts in this manner is an improvement on currently used ensemble forecast visualization techniques.

KEYWORDS: Empirical orthogonal functions; Ensembles; Forecasting techniques; Numerical weather prediction/forecasting; Operational forecasting

1. Introduction

Advances in computational resources have allowed most global Numerical Weather Prediction (NWP) centers to produce ensemble forecasts. Due to the chaotic nature of the atmosphere and the amplification of initial condition errors with time, ensemble forecasts are indispensable in the creation of medium-range (3–7 day) forecasts. This is because ensemble forecasts give the best approximation of the entire range of possible forecast outcomes. At the National Centers for Environmental Prediction's (NCEP's) Weather Prediction Center (WPC), global ensemble forecasts from NCEP, the European Centre for Medium-Range Weather Forecasts (ECMWF), and the Canadian Meteorological Centre (CMC) are routinely consulted in the preparation of medium-range forecasts. Each EPS alone does not have enough spread to sufficiently simulate the true range of possible forecast outcomes (Buizza et al. 2005), so the available EPSs are routinely combined to form a multimodel ensemble.

Forecasters working under strict time constraints do not necessarily have time to thoroughly interrogate and investigate the wealth of information contained in this multimodel ensemble. Traditionally, forecasters have used multimodel ensemble forecasts by viewing the ensemble mean and spread (standard deviation) or by viewing spaghetti plots (described by Inness and Dorling 2012) that show the forecasted location of a given meteorological parameter contour (e.g., the 5800-m contour of 500-hPa geopotential height) for each ensemble member. This gives a forecaster a quick approximation of the uncertainty of the forecast and a cursory idea of the range of possible forecast outcomes, but a host of useful information is ignored. To alleviate the data overload problem, the National Weather Service (NWS) has developed the National Blend of Models (NBM; Craven et al. 2020). The NBM intelligently postprocesses and combines deterministic and ensemble forecasts to create one

Lamberson's current affiliation: Extreme Event Solutions at Verisk, Boston, Massachusetts.

Sienkiewicz's current affiliation: Karen Clark and Company, Boston, Massachusetts.

Corresponding author: William S. Lamberson, bill.lamberson@gmail.com

DOI: 10.1175/WAF-D-22-0154.1

consistent and accurate forecast for use as a common starting point across the NWS. This is essential for consistent messaging across NWS organizations concerning high-impact weather events. However, forecasters have expressed concerns about solely relying on the NBM. Forecasters want tools that will show them the underlying data the NBM used to arrive at its forecast as well as plausible alternatives to the NBM forecast. This is especially needed for medium-range forecasts where there is still significant forecast uncertainty. Thus, a tool that allows forecasters to efficiently and objectively get a more detailed view of the most prevalent forecast scenarios contained in the multimodel ensemble forecasts that are incorporated into the NBM is needed and would greatly enhance the forecast process.

A tool that could accomplish these goals is an ensemble clustering tool. Ensemble clustering is a well-established method of grouping ensemble members with similar forecasts together, thereby reducing a large set of ensemble forecasts down into the most prevalent forecast scenarios. Different methods of ensemble clustering have been used successfully to address various operational forecasting challenges. In the short range (1–3-day forecasts), Johnson et al. (2011) used a hierarchical cluster analysis on a convection-allowing model to show different forecast scenarios for short-range precipitation events. In the medium range (4–10-day forecasts), Ferranti and Corti (2011) use empirical orthogonal functions (EOFs) and the k-means clustering algorithm to cluster medium-range forecasts over Europe from the ECMWF Ensemble (ENS). Last, in the subseasonal range (10–30-day forecasts), Palmer et al. (1990) used EOF analysis and a Ward hierarchical clustering algorithm on ECMWF ENS forecasts to produce different 10–30-day forecast scenarios over the Northern Hemisphere, noting that one of the clusters usually had a more accurate forecast than the ensemble mean forecast.

WPC's own most recent development work on the topic of ensemble clustering is documented in Brill et al. (2015). Brill et al. (2015) developed a divisive clustering method that applied a one-dimensional discrete Fourier transformation to ensemble member deviations from the ensemble mean 500-hPa geopotential heights in a truncated zonal band over North America. Brill et al. (2015) applied this clustering method to a combination of the NCEP Global Ensemble Forecast System (GEFS) and the ECMWF ENS. WPC medium-range forecasters found these cluster forecasts plausible and of value to their forecast process. However, the largest clusters produced by this method were not able to consistently outperform the ECMWF ENS mean. In addition, this method reduces the geopotential data to one dimension, potentially ignoring important information about two-dimensional forecast variability. Thus, a clustering algorithm that can produce cluster forecasts of comparable accuracy to the available EPS means and takes into account two-dimensional variability is desired.

An ensemble clustering methodology that could satisfy these desired requirements is EOF analysis in conjunction with fuzzy k-means cluster analysis, hereafter referred to as ensemble fuzzy clustering (EFC). EFC was developed by Harr et al. (2008) who used the technique on ensemble forecasts of western Pacific typhoons undergoing extratropical transition to learn what factors drive the predictability of the extratropical transition process. More recently, Zheng et al. (2017) has used EFC to determine forecast scenarios for high-impact U.S. East Coast winter storms from a large multimodel ensemble. The authors of Zheng et al. (2017) made a version of their EFC product available for testing in WPC's Winter Weather Experiment (WWE) during the 2017/18 winter season. WWE participants were intrigued by this tool and its potential to assist the forecast process. When WPC launched its Extended Range Forecast Experiment (ERFE) in 2017, WPC developers set out to develop an ensemble clustering tool that could reveal the most prevalent medium-range (3–7 day) forecast scenarios of temperature and precipitation over the contiguous United States (CONUS) and could serve as a prototype for the needed and desired complement to the NBM. Development of this tool began in 2017 and it has been refined through testing during ERFE sessions. Over time, the use of the tool has spread from WPC to forecasters throughout the NWS. This paper details WPC's EFC tool which is geared toward forecasting temperature and precipitation over the CONUS during days 3–7 and demonstrates its utility through a case study.

The remainder of this paper proceeds as follows. Section 2 discusses the datasets and details the ensemble clustering methodology. Section 3 provides a case study that demonstrates the utility of WPC's EFC tool for creating medium-range forecasts. Statistical verification of the cluster forecasts is presented in section 4. Section 5 provides a summary and key conclusions.

2. Data and methodology

a. Data

The ensemble data used in this study are 0.5° resolution versions of the CMC Global Ensemble Prediction System (GEPS; 20 members; Lin et al. 2019), the NCEP GEFS (30 members; Zhou et al. 2022), and the ECMWF ENS (50 members; ECMWF 2021). These data were used because they are what is available to forecasters at WPC to consult in their preparation of medium-range forecasts. These three ensemble prediction systems (EPSs) also make up the bulk of the data the NBM combines to create 3–7-day forecasts. The NBM applies a postprocessing technique (Hamill et al. 2017) to these 0.5° resolution ensemble member forecasts of several surface-based variables, including maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (QPF). This is done to improve their accuracy and fine-scale detail. To be a true compliment to the NBM, our EFC tool would have used these postprocessed forecasts. However, at the time of development, they were not available so we used the available raw versions. The multimodel ensemble comprised of these three global EPSs contains 100 members and should reasonably include most possible forecast outcomes. The scope of ERFE was limited to creating experimental forecasts for 500-hPa geopotential height (Z500), TMAX, TMIN, and QPF. Consequently, these are the only ensemble forecast parameters used in our prototype ensemble clustering tool. These data were accessed and available in real-time at NCEP but may be downloaded from ECMWF's THORPEX Interactive

Grand Global Ensemble (TIGGE; Bougeault et al. 2010; Buizza 2014; Swinbank et al. 2016) archive.

b. EOF analysis

The first step of the EFC methodology demonstrated by Harr et al. (2008), and modified for use in this study, is to perform EOF analysis on an ensemble forecast parameter of interest. EOF analysis is a common data analysis technique used in climatological studies to determine the spatial patterns that most efficiently explain the variability of a multivariate dataset with time (Richman 1986). EOF analysis can be calculated across the model ensemble member dimension instead of the more common time dimension. In this case, the resulting modes of the EOF analysis will show the dominant patterns of the difference between individual ensemble members and the ensemble mean. The leading principal components (PCs) derived from this type of EOF analysis represent the projections of the dominant EOF patterns onto the difference between each of the ensemble members and the mean of all ensemble members. The EOF patterns together with the PC values for each ensemble member will efficiently detail the dominant two-dimensional modes of variability in the ensemble forecast and how a given ensemble member's forecast differs from the mean of all ensemble members. Accordingly, a clustering algorithm can be applied to the PCs to group together members that have similar forecasts for the parameter of interest. As in Harr et al. (2008) and Zheng et al. (2017), each PC is normalized to have a variance of 1, and only the first two leading EOFs and their attendant PCs are considered.

The most important part of EFC is selecting the proper forecast variable and domain on which to perform the EOF analysis. What variable and domain is proper is dictated by the end goal of EFC. The goal of this study is to objectively create forecast scenarios of temperature and precipitation over regions of the CONUS for each day of the 3–7-day forecast period. This requires performing EOF analysis on a single variable contained in the ensemble forecasts that will succinctly capture and efficiently describe what the weather will be like on a given day. A variable we felt satisfies these requirements and is familiar to forecasters is Z500. The Z500 pattern usually governs the location of the storm track (areas of high precipitation) as well as areas of anomalous TMIN and TMAX. Thus, Z500 is one variable that will give a forecaster some information about the TMIN, TMAX, and QPF forecast. For our ensemble clustering tool, ensemble clusters for a given day were derived from EOF analysis of the 24-h-averaged Z500 field spanning that day. Sets of clusters were provided for a CONUS-wide domain as well as three predefined regions of CONUS (east, west, and central). The boundaries for the domains over which the EOF analysis is conducted are listed in Table 1.

TABLE 1. EOF domains.

Domain name	Domain area
CONUS	22°–72°N, 40°–155°W
EAST	25°–55°N, 60°–100°W
CENTRAL	25°–55°N, 75°–115°W
WEST	25°–55°N, 95°–135°W

c. Cluster analysis

The second step of EFC is applying the fuzzy k-means clustering algorithm to the EOF PCs to group ensemble members with similar forecasts together. Traditional k-means clustering (Lloyd 1982) is a popular and widely used method of clustering data points. In traditional k-means clustering, each data point is assigned to only one cluster. In fuzzy k-means clustering, each data point can belong to more than one cluster. Each data point is assigned a membership coefficient for each cluster that describes the degree to which it belongs to that cluster. This information can be used to create more nuanced clusters. Harr et al. (2008) used the membership coefficient to leave ensemble members on the border between two clusters unclassified. Alternatively, Zheng et al. (2017) assigned each ensemble member to the cluster for which it had the highest membership coefficient. This effectively creates a clustering scheme that is likely identical to traditional k-means clustering. For our ensemble clustering tool, traditional k-means clustering was utilized as potentially excluding large numbers of ensemble members from classification was not desired. However, excluding large outliers from classification was desired, so ensemble members with at least one normalized PC of greater than 3.5 were filtered out before applying the k-means clustering algorithm. In practice, most of the time no ensemble members are filtered out before applying the k-means clustering algorithm.

As detailed in Zheng et al. (2017), the steps of the k-means clustering algorithm are as follows:

1) place a predefined number of clusters (initial guess) in the EOF PC1–PC2 phase space,
2) assign each ensemble member represented by its pair of PCs to the nearest group center,
3) compute new centers by minimizing an objective function that evaluates the distance from each point to each new cluster,
4) reexamine each point relative to the updated cluster centers, and
5) repeat steps 2–4. If no points can be reassigned because they lie closer to another center, the iterations stop.

In practice, steps 2–4 are usually repeated less than 15 times. The other critical aspect of both fuzzy and traditional k-means clustering is deciding the number of predefined clusters. As detailed in Harr et al. (2008), objectively determining the optimal number of clusters is very difficult and is often immaterial. The goal of ensemble clustering in this application, as in Harr et al. (2008) and Zheng et al. (2017), is to partition the data into an adequate subdivision of similar groups. Harr et al. (2008) took a subjective approach to determining the adequate number of clusters. Zheng et al. (2017) took an objective approach and used the Rand index (Yeung and Ruzzo 2001) to determine the optimal number of clusters, finding that the optimal number of clusters was often 3–5. Subjectively determining the optimal number of clusters was not an option as our tool would be automated. For simplicity,

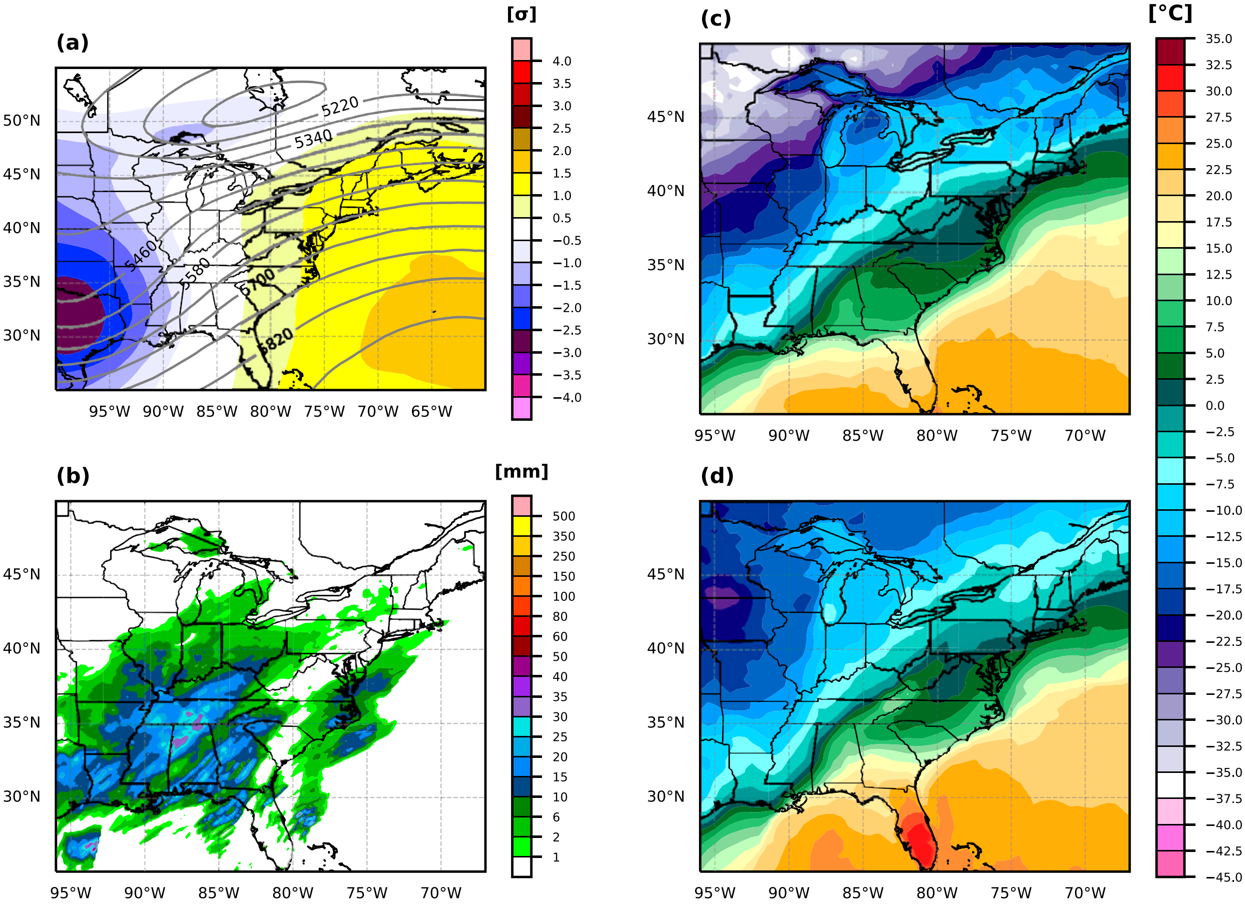


FIG. 1. (a) GFS analysis 24-h-averaged Z500 (gray contours; m) and standardized anomalies (shading; σ) with respect to CFSR climatology ending at 0000 UTC 16 Feb 2021. (b) CCPA 24-h precipitation (shading; mm) ending at 0000 UTC 16 Feb 2021. (c) Lowest ERA5 2-m temperature between 0000 and 1200 UTC 15 Feb 2021 (shading; °C). (d) Highest ERA5 2-m temperature between 1200 UTC 15 Feb 2021 and 0000 UTC 16 Feb 2021 (shading; °C).

we decided to keep the number of clusters fixed at four clusters, the middle of the optimal range found by Zheng et al. (2017). However, this decision may need to be revisited as some users have voiced a preference for the number of clusters to vary.

d. Cluster display

WPC forecasters and other end users were provided with a simple website to view the clusters. For each day in the 3–7-day forecast period and for each region, this website displayed the two leading 24-h-averaged Z500 EOFs and a phase space that showed each ensemble member's PC values and to which cluster it belonged. The website also displayed cluster mean forecasts for 24-h-averaged Z500, daily TMAX, daily TMIN, and 24-h QPF. Cluster mean forecasts were displayed as the difference between the cluster mean and the multimodel mean of all 100 ensemble members. This was done so that forecasters could quickly see how each cluster was different from the multimodel ensemble mean. The cluster forecasts were presented to participants in the order of their ranked size (i.e., cluster 1 was comprised of the largest number of ensemble members). The utility of the WPC clustering tool will now be demonstrated through a case study.

3. 15 February 2021 eastern U.S. case

February of 2021 was an active weather month over the central and eastern United States. The middle of the month was particularly active with the region being impacted by a notable cold air outbreak and several impactful precipitation events. The 15 February 2021 case is in the middle of this active period and will be used to demonstrate the utility of WPC's EFC tool as a means to better understand the range of possible forecast outcomes.

a. Synoptic overview and verifying analyses

Before discussing the ensemble clustering output, a brief overview of the verifying analyses for the parameters of interest (Z500, QPF, TMIN, and TMAX) is presented. The 24-h-averaged Z500 analysis ending at 0000 UTC 16 February from the GFS (Fig. 1a) shows a closed 500-hPa low centered over James Bay in Canada. Troughing extends southwestward

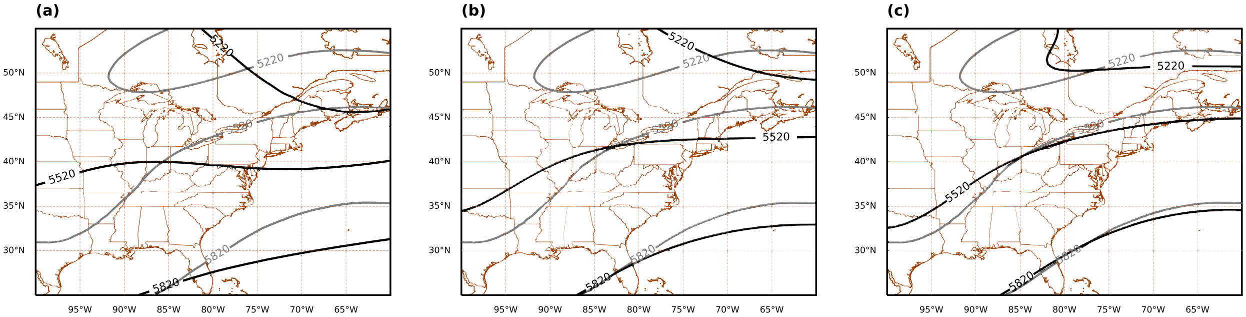


FIG. 2. 24-h-averaged Z500 verifying analysis (gray contours; m) and multimodel ensemble mean (black contours; m) from ensemble forecast initialized at (a) 0000 UTC 8 Feb 2021 (7-day forecast), (b) 0000 UTC 10 Feb 2021 (5-day forecast), and (c) 0000 UTC 12 Feb 2021 (3-day forecast) and valid at 0000 UTC 16 Feb 2021.

and then southward from this feature over the central United States, with the most anomalous troughing located over Texas. Downstream from the trough, there is an expansive ridge over the east coast of the United States and the western Atlantic Ocean. Between the central United States trough and downstream ridge, a strong front is draped from the lower Mississippi Valley to the Tennessee Valley. Morning TMINs from ECMWF's ERA5 dataset (Hersbach et al. 2020) were as low as −15°C in western Tennessee and as high as 5°C in eastern Tennessee (Fig. 1c). Afternoon TMAXs stayed below −7.5°C in western Tennessee and surged above 15°C in eastern Tennessee (Fig. 1d). According to The Environmental Modeling Center's Climatology-Calibrated Precipitation Analysis (CCPA) dataset (Hou et al. 2014), areas along this front experienced heavy precipitation of up to 35 mm (Fig. 1b) with some of this precipitation falling as snow, sleet, and freezing rain. Overall, this was a high-impact weather day for the central and eastern United States.

b. *Ensemble forecasts at different lead times*

Figure 2 shows the multimodel ensemble mean 24-h-averaged Z500 forecast and verifying analysis for this case at different lead times. The 7-day multimodel ensemble mean forecast (Fig. 2a) was poor and featured weak toughing over the eastern United States and the western Atlantic Ocean where ridging occurred. Over the central United States, the 7-day multimodel ensemble mean forecast has weak ridging where troughing occurred. This serves to highlight the perils of only using an ensemble mean forecast. By 5 days prior to the event, the multimodel ensemble mean forecast had improved (Fig. 2b). The multimodel mean 24-h-averaged Z500 field is relatively zonal and does not adequately depict the magnitude of troughing over the central United States or ridging over the eastern United States and western Atlantic that occurred, but it no longer forecasts troughing where a ridge verified and ridging where are trough verified. The 3-day multimodel ensemble mean forecast (Fig. 2c) showed further improvement. While the amplitude of the 24-h-averaged Z500 features was underdone, the forecast had the central U.S. trough and the eastern U.S. ridge in the right locations. The following subsections will focus on the application of WPC's EFC methodology to the 7- and 5-day ensemble forecasts for this high-impact event.

c. *7-day ensemble clustering forecast output*

Figure 3a shows the 7-day multimodel mean and spread forecast for the 24-h-averaged Z500 field initialized at 0000 UTC 8 February 2021. As previously noted, this forecast was poor. However, the ensemble spread in 24-h-averaged Z500 is greatest over eastern Canada and the northeastern United States. This suggests that the ensembles contain a wide range of forecast solutions, with perhaps some individual solutions being more accurate than the multimodel mean. A traditional spaghetti plot of the 24-h-averaged Z500 is shown in Fig. 3b, but like most spaghetti charts, it is difficult to interpret and useful patterns are hard to determine.

The first step of our ensemble clustering methodology is to evaluate the first two EOF patterns for the 7-day 24-h-averaged Z500 field over the EAST domain defined in Table 1. Figures 3c and 3d show these first two leading EOFs, which explain 50.6% and 27.3% of the variance in the 24-h-averaged Z500 field, respectively. The first EOF (EOF1; Fig. 3c) is a monopole pattern with positive values collocated with the weak multimodel ensemble mean troughing over the northeastern United States. Thus, members with positive (negative) PCs for EOF1 have higher (lower) heights and stronger ridging (troughing) in this area. Meanwhile, the second leading EOF (EOF2; Fig. 3d) is a dipole with negative values over the Upper Midwest and Ontario and positive values centered just south of Nova Scotia. This indicates that members with positive (negative) PCs for EOF2 have increased troughing (ridging) when compared to the multimodel ensemble mean over the Upper Midwest and Ontario and increased ridging (troughing) when compared to the multimodel ensemble mean over Atlantic Canada.

Figure 4 shows the partition of the 100 multimodel ensemble members into the four clusters as determined by the clustering method. The 39 members that comprise cluster 1 are generally in the lower right quadrant of the phase space, with positive values for PC1 and negative values for PC2. Accordingly, cluster 1 contains the ensemble members that have more ridging over the central and eastern United States and troughing over Atlantic Canada. The 32 members that comprise cluster 2 are located on the left side of the phase space, with negative values for PC1. Thus, cluster 2 contains the ensemble members that have enhanced troughing over the eastern United States. The 23 ensemble members that comprise

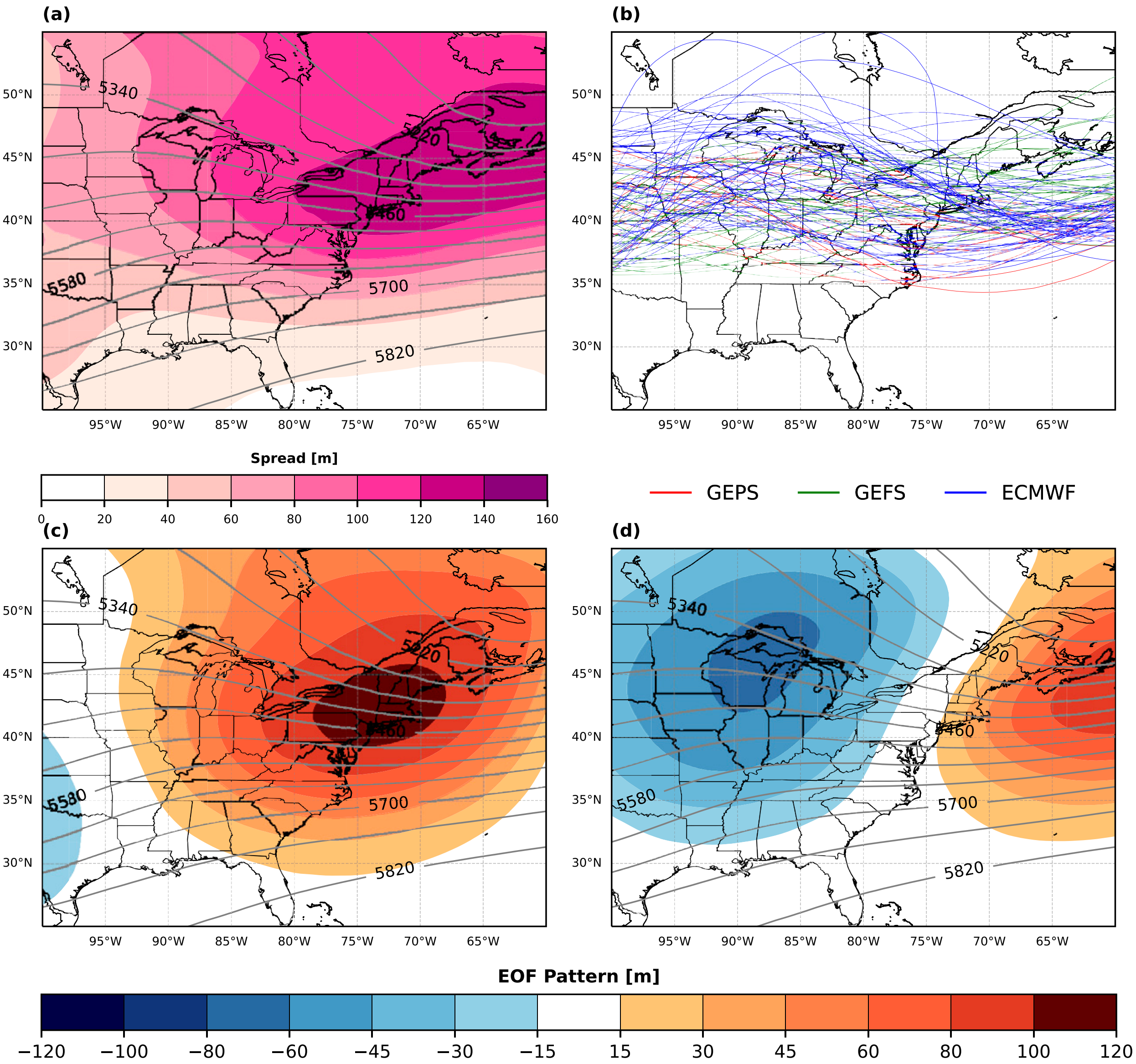


FIG. 3. (a) 24-h-averaged Z500 multimodel ensemble mean (contours; m) and spread (shading; m) initialized at 0000 UTC 8 Feb 2021 and valid for the 24-h period ending at 0000 UTC 16 Feb 2021. (b) Spaghetti plot of 5460-m 24-h-averaged Z500 contour for the 100 multimodel ensemble members (GEPS ensemble members in red, GEFS members in green, and ENS members in blue). (c) The regressed pattern of 24-h-averaged Z500 corresponding to EOF PC1 (shading; m) and multimodel ensemble mean 24-h-averaged Z500 (contours, m) ending at 0000 UTC 16 Feb 2021. (d) As in (c), but for the EOF2 pattern.

cluster 3 are in the upper right quadrant of the phase space. These members have positive values for PC1 and PC2. Thus, cluster 3 contains the ensemble members with enhanced troughing over the central United States and enhanced ridging over the eastern United States, Quebec, and Atlantic Canada. Last, the 6 members that comprise cluster 4 are located in the upper portion of the PC1–PC2 phase space. These ensemble members have large positive values for PC2 and represent the ensemble members with the strongest troughing over the central and eastern United States and the strongest ridging over Atlantic Canada.

The phase space shown in Fig. 4 is exactly as it would have looked in real time to ERFE participants, with one notable exception. The magenta cross in Fig. 4 shows the position of the GFS analysis 24-h-averaged Z500 field over the EAST domain in the PC1–PC2 framework. The verification presented in the PC1–PC2 framework is located far from the multimodel ensemble mean, which, by definition, is located at the origin of the PC1–PC2 phase space. It is also located far from the mean of each EPS (filled circles). However, it is located on the outer edge of the group of members that forms cluster 3, indicating that this cluster had the most accurate forecast for

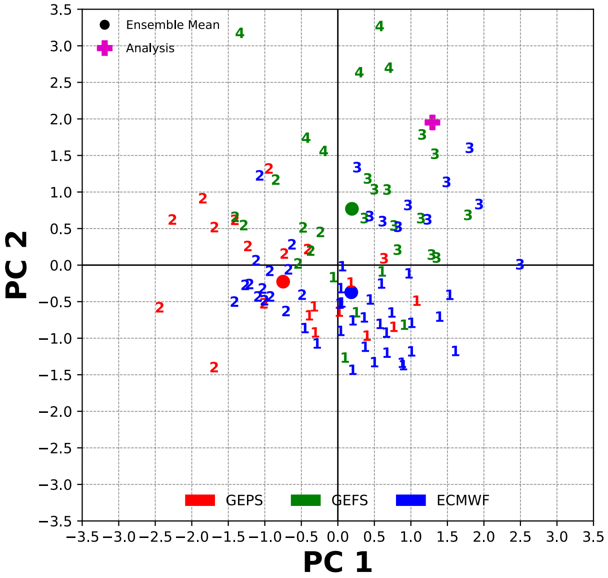


FIG. 4. The PC1–PC2 phase space with the location of each ensemble member and the cluster it belongs to for the 7-day forecast initialized at 0000 UTC 8 Feb 2021 and valid for the 24-h period ending at 0000 UTC 16 Feb 2021. Members in cluster 1 are labeled with “1”, members in cluster 2 with “2,” members in cluster 3 with “3,” and members in cluster 4 with “4.” The larger filled circles denote the location of the three EPS means. By definition, the multimodel ensemble mean is located at the origin. Ensemble members in red are GEPS members, members in green are GEFS members, and members in blue are ENS members.

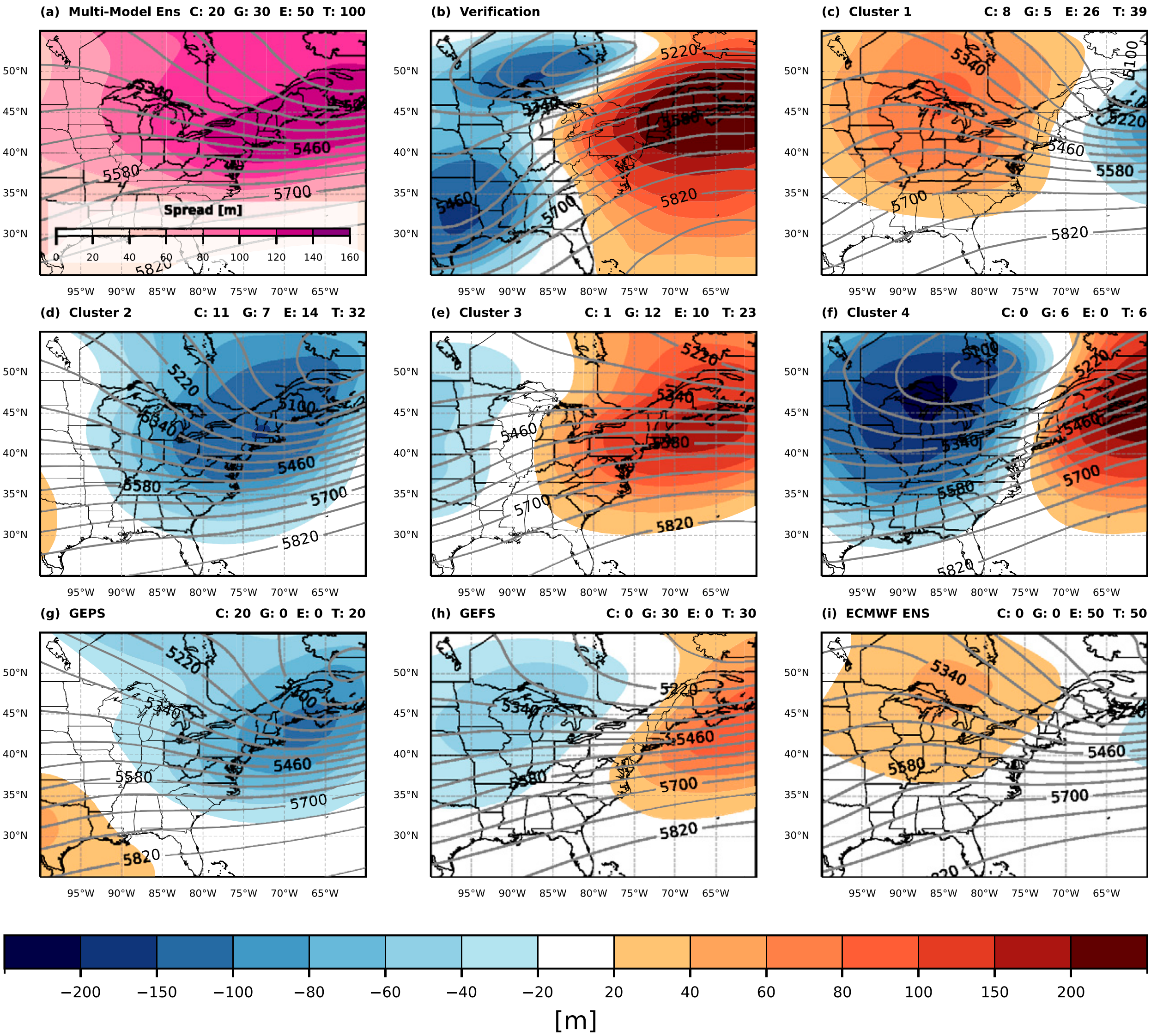
this case. The members that comprise cluster 4 are also relatively close to the verification location, indicating that members in cluster 4 also had relatively good forecasts. This could not have been known in advance, but it demonstrates how the multimodel ensemble mean and the mean of each EPS does not always adequately reflect the range of possible forecast outcomes. Further, it demonstrates how clustering ensemble members with similar forecasts can do a better job of accomplishing this goal.

Determining the physical interpretation of each cluster from the EOFs and the PC1–PC2 phase space can be tricky for those who are unfamiliar with EOF analysis. But, the ensemble clustering tool can be effectively used without viewing the EOFs or the PC1–PC2 phase space. The physical interpretation of each cluster will become apparent when the forecasts from each cluster are viewed. Figure 5 shows the mean forecast for 24-h-averaged Z500 for the members that comprise each cluster (cluster mean) and how they differ from the multimodel ensemble mean. Unlike during ERFE sessions, Fig. 5 includes the verifying analysis so that the reader can get a better sense of the relative accuracy of each cluster forecast. Each EPS mean forecast is also included so that the reader may see how the cluster means compare to the EPS means and if they offer any value over viewing just the EPS means, as is traditionally done in operational forecasting. Indeed, the physical interpretation of each cluster forecast and how it differs from the multimodel ensemble mean becomes obvious when the cluster forecasts are viewed in this manner. For example, it is clear that cluster 1 (Fig. 5c) is the forecast scenario with enhanced ridging over the central and eastern United States and enhanced troughing over Atlantic Canada.

Figure 5a shows the multimodel ensemble mean and spread forecast, with Fig. 5b showing the verifying analysis and how it differs from the multimodel ensemble mean. Together, Figs. 5a and 5b show that the multimodel ensemble mean drastically underforecast the strength of the troughing extending south from James Bay and forecasted a trough over Atlantic Canada and the northeastern United States where ridging was observed. In addition to the visual verification provided by Fig. 5, the anomaly correlation (AC) over the EAST domain for the multimodel ensemble mean forecast, cluster mean forecasts, and EPS mean forecasts were evaluated and are provided in Table 2. These AC values are centered ACs [described by Wilks (2006, p. 311)] based on climatological means and standard deviations derived from the 0.5° NCEP Climate Forecast System Reanalysis (CFSR) dataset (Saha et al. 2010) for a 30-yr period (1980–2010). The verifying analysis is the GFS analysis that is available on the same 0.5° grid as the ensemble data. The 24-h-averaged Z500 AC of 0.11 for the multimodel mean is poor for a 7-day forecast. However, it is better than the forecasts predicted by clusters 1 and 2 (Figs. 5c,d). Cluster 2 in particular had a poor forecast (AC of −0.33) because it forecast a trough over the eastern United States and Atlantic Canada where the verifying analysis had a ridge. By contrast, clusters 3 and 4 (Figs. 5e,f) had relatively good forecasts. Both forecasts correctly predicted enhanced troughing relative to the multimodel ensemble mean from James Bay southwestward and enhanced ridging over the eastern United States and Atlantic Canada. Cluster 3 correctly predicted these features would be located slightly further west than cluster 4, resulting in cluster 3 having a better forecast than all other cluster or EPS means.

Clustering ensemble members that have similar forecasts together inevitably produces clusters with a wide range of forecast accuracy. The most frequent question we received from ERFE forecast session participants focused on how they could identify which cluster forecast would be most accurate. We stressed to participants that selecting the most accurate cluster was beside the point. This tool aims to help forecasters better understand and communicate the range of possible forecast outcomes contained in the global ensembles. Currently, in operational forecasting, the GEPS, GEFS, and ENS means are often treated as de facto ensemble clusters. However, they often do not adequately represent the range of possible forecast outcomes. That was precisely the case for this event. The visual verification presented in Fig. 5 shows that the ensemble means for the GEPS (Fig. 5g), GEFS (Fig. 5h), and ECMWF ENS (Fig. 5i) were less different from the multimodel ensemble mean than the mean of each cluster was from the multimodel ensemble mean. This case illustrates how the EFC methodology we propose can effectively identify and visualize the different forecast outcomes contained in the global ensemble forecasts.

In addition to producing experimental forecasts for Z500, ERFE was tasked with creating experimental forecasts for TMAX, TMIN, and QPF. To assist participants in the preparation



(a) Multi-Model Ens C: 20 G: 30 E: 50 T: 100

(b) Verification

(c) Cluster 1 C: 8 G: 5 E: 26 T: 39

(d) Cluster 2 C: 11 G: 7 E: 14 T: 32

(e) Cluster 3 C: 1 G: 12 E: 10 T: 23

(f) Cluster 4 C: 0 G: 6 E: 0 T: 6

(g) GEPS C: 20 G: 0 E: 0 T: 20

(h) GEFS C: 0 G: 30 E: 0 T: 30

(i) ECMWF ENS C: 0 G: 0 E: 50 T: 50

FIG. 5. (a) Multimodel ensemble mean (contours) and spread (shading) for 24-h-averaged Z500 initialized at 0000 UTC 8 Feb 2021 and valid at 0000 UTC 16 Feb 2021. The numbers on top of each panel indicated how many ensemble members from each EPS are included in that panel's mean forecast, "C" before the number of CMC GEPS ensemble members, "G" before the number of GEFS members, "E" before the number of ECWMF ENS members, and "T" before the total number of ensemble members. (b) GFS analysis of 24-h-averaged Z500 (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021. (c)–(i) Cluster or EPS mean 24-h-averaged Z500 (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021.

of these forecasts, the TMAX, TMIN, and 24-h QPF cluster mean forecasts from the clusters derived from the 24-h-averaged Z500 EOF analysis were presented. We were unsure if clusters based on Z500 would create useful forecasts for variables other than Z500. More rigorous verification of the cluster forecasts over a 6-month period that will answer this question will be shown in section 4, but for this case, the cluster forecasts based on the 24-h-averaged Z500 EOF analysis did produce useful forecasts for TMIN, TMAX, and 24-h QPF. Similar to Fig. 5, the multimodel mean, cluster means, and individual EPS means for TMAX and 24-h QPF for this case are shown in Figs. 6 and 7, respectively.

The verifying TMAX analysis (Fig. 6b) shows much warmer TMAXs than the multimodel ensemble mean out ahead of the cold front over the eastern United States and much colder TMAXs behind the front over the central United States. This is not surprising as the verifying 24-h-averaged Z500 analysis (Fig. 5b) had stronger ridging over the eastern United States and stronger troughing over the central United States than the multimodel ensemble mean. It is also clear from the TMAX verification that cluster 3 (Fig. 6e) had the most accurate forecast for TMAX, with cluster 4 not far behind (Fig. 6f). Both clusters correctly predicted warmer TMAXs than the multimodel ensemble mean ahead of the front over the eastern

TABLE 2. The 24-h-averaged Z500 centered anomaly correlation over the EAST domain for the ensemble forecasts initialized at 0000 UTC 8 Feb 2021 and valid at 0000 UTC 16 Feb 2021.

Forecast	24-h-averaged Z500 anomaly correlation
Multimodel mean	0.11
Cluster 1	−0.07
Cluster 2	−0.33
Cluster 3	0.82
Cluster 4	0.63
GEPS mean	−0.34
GEFS mean	0.61
ENS mean	0.02

United States and colder TMAXs behind the front. Though, for both clusters 3 and 4, their exact location of the front is slightly off. Cluster 3 is too far west with the front placement, and cluster 4 is too far east. This is consistent with each cluster's 24-h-averaged Z500 forecast. Overall, the TMAX forecasts from the clusters based on the 24-h-averaged Z500 EOF analysis for this case make sense physically and participants found them useful.

The verifying 24-h precipitation analysis (Fig. 7b) featured higher precipitation amounts over much of the central and eastern United States than the multimodel ensemble mean predicted for this case. It is also clear from the 24-h precipitation verification that cluster 3 (Fig. 7e) had the most accurate forecasts for 24-h precipitation. Cluster 3 correctly predicted higher 24-h precipitation amounts than the multimodel ensemble mean over the central and eastern United States. In contrast, cluster 2 (Fig. 7d), which had one of the poorest 24-h-averaged Z500 forecasts, also had one of the poorest 24-h precipitation forecasts. Cluster 2 predicted less precipitation over the central and eastern United States than the multimodel ensemble mean when more precipitation than the multimodel ensemble mean forecast is what verified. As with the TMAX forecasts from the clusters based on the 24-h-averaged Z500 EOF analysis, the 24-h QPF forecasts for this case make sense physically and participants found them useful.

d. 5-day ensemble clustering forecast output

Only the 5-day cluster forecasts for 24-h-averaged Z500 are included for brevity. The 24-h-averaged Z500 multimodel ensemble spread was lower for the 5-day forecast than the 7-day forecast but it was still maximized over eastern Canada and the northeastern United States. This indicates that these areas were still where the forecast was most uncertain. In addition, the 24-h-averaged Z500 EOFs for the 5-day forecast were broadly similar to those of the 7-day forecast (Figs. 3c,d). This indicates that the dominant modes of uncertainty had not changed. Another reason we are not including the 24-h-averaged Z500 EOF and PC1–PC2 analysis for the 5-day forecast is that many forecasters do not consult them. Personal communication with forecasters who use the tool has revealed this. The authors maintain that useful information about the forecast can be gleaned from consulting the 24-h-averaged Z500 EOF analysis and accompanying PC1–PC2 phase space, it is not required to effectively use the tool. To emphasize this, we have only included the 24-h mean Z500 cluster and EPS mean forecasts, which are shown in Fig. 8.

Figure 8a shows that the multimodel ensemble mean no longer forecasts troughing over Atlantic Canada and the northeastern United States and now forecasts weak ridging. This is an improvement, but Fig. 8b shows that the degree of ridging in this location is still underforecast by the multimodel ensemble mean. The strength of the troughing extending south from James Bay is also underforecast by the multimodel mean. The 31 members that comprise cluster 1 (Fig. 8c) had ridging over Ontario and troughing over Atlantic Canada, resulting in a poor forecast. The 26 ensemble members that comprise cluster 2 (Fig. 8d) had a trough from Atlantic Canada, extending southwestward over the Atlantic coast of the United States. This was also a poor forecast. The 23 members that comprised cluster 3 (Fig. 8e) had troughing over Ontario and extending southward over the central United States. Cluster 3 also had ridging over the eastern United States and eastern Canada. This was a successful forecast. Finally, the 20 members that comprise cluster 4 (Fig. 8f) had troughing over the central United States and ridging over Ontario and eastern North America. However, this ridging did not extend further eastward to Atlantic Canada as the verification shows to have occurred. Overall, the three EPS mean forecasts (Figs. 8g–i) are not as different from the multimodel mean forecast as the four ensemble cluster mean forecasts are from the multimodel mean. This indicates that the cluster forecasts identified by WPC's EFC tool again did a better job than the EPS mean forecasts of showing forecasters the range of possible forecast solutions contained in the multimodel ensemble.

4. Limited statistical verification

The preceding case study demonstrates how the WPC EFC tool can be a more instructive method for forecasters to incorporate ensemble forecast information than the traditional method of viewing each EPS mean. However, this is just one case study, and more rigorous verification over an extended period is required to ascertain whether the WPC EFC tool routinely provides a better assessment of the range of possible forecast outcomes.

It is well established that individual ensemble members can outperform the ensemble mean for individual cases (e.g., Palmer et al. 1990), but do not outperform the ensemble mean for a large collection of cases (Toth and Kalnay 1993). Brill et al. (2015) performed statistical verification on 8 months' worth of cases for their clustering method. Their analysis compared the accuracy of the forecasts from the two largest clusters to the accuracy of forecasts from the deterministic GFS and ECMWF, as well as the GEFS and ECMWF ENS means. Brill et al. (2015) found that the two largest clusters did not outperform the ECMWF ENS mean over their large sample of cases. It is important to note that the clustering method of Brill et al. (2015) tended to produce clusters that were comprised of a small number of members (typically between 4 and 13 members). The largest clusters produced by their clustering methodology are smaller than the smallest clusters produced

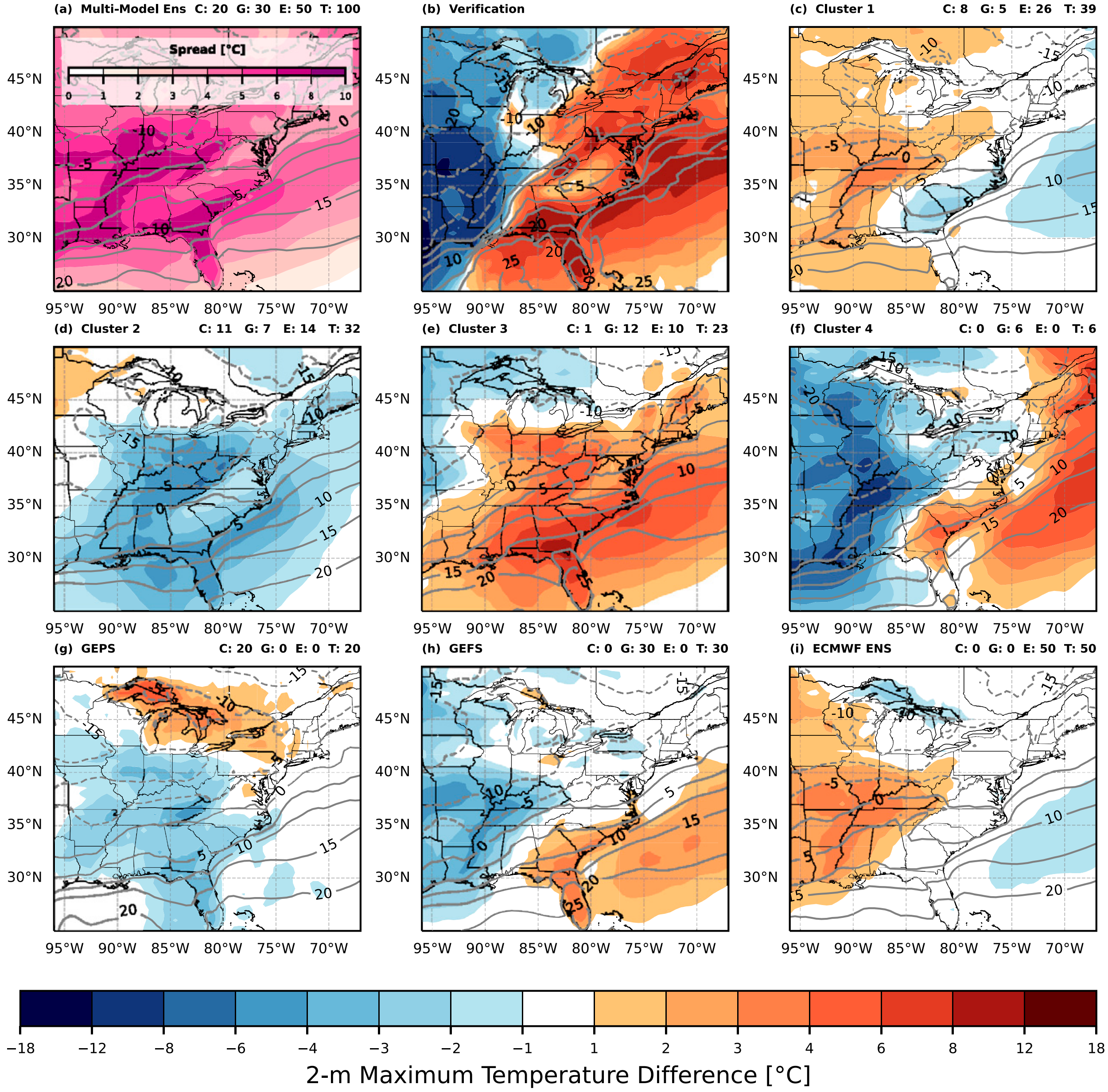

(a) Multi-Model Ens C: 20 G: 30 E: 50 T: 100
Spread [°C]
(b) Verification
(c) Cluster 1 C: 8 G: 5 E: 26 T: 39
(d) Cluster 2 C: 11 G: 7 E: 14 T: 32
(e) Cluster 3 C: 1 G: 12 E: 10 T: 23
(f) Cluster 4 C: 0 G: 6 E: 0 T: 6
(g) GEPS C: 20 G: 0 E: 0 T: 20
(h) GEFS C: 0 G: 30 E: 0 T: 30
(i) ECMWF ENS C: 0 G: 0 E: 50 T: 50
45°N
40°N
35°N
30°N
95°W 90°W 85°W 80°W 75°W 70°W
−18 −12 −8 −6 −4 −3 −2 −1 1 2 3 4 6 8 12 18
2-m Maximum Temperature Difference [°C]

FIG. 6. (a) Multimodel ensemble mean (contours) and spread (shading) for 24-h TMAX initialized at 0000 UTC 8 Feb 2021 and valid at 0000 UTC 16 Feb 2021. The numbers on top of each panel indicated how many ensemble members from each EPS are included in that panel's mean forecast, "C" before the number of CMC GEPS ensemble members, "G" before the number of GEFS members, "E" before the number of ECWMF ENS members, and "T" before the total number of ensemble members. (b) ERA5 analysis of 24-h TMAX (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021. (c)–(i) Cluster or EPS mean 24-h TMAX (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021.

by our methodology. Table 3 shows the median cluster size of the largest through the smallest cluster produced by our clustering method for days 3–7 over the 6 months spanning from 15 October 2020 to 15 April 2021. The results are pooled over the four regions as each region had similar cluster sizes. Table 3 shows that our ensemble clusters are larger and are roughly comparable to the sizes of the three constituent EPSs. If ensemble cluster accuracy is a function of cluster size, as hypothesized by Brill et al. (2015), then our clusters should compare favorably to the accuracy of the EPS means over a large sample of cases.

Our statistical verification will mirror that of Brill et al. (2015) with one key difference, we rank the ensemble cluster means, EPS means, and deterministic forecasts for each case and then perform verification on the ranked forecasts. This will determine if the best-performing cluster mean outperforms the

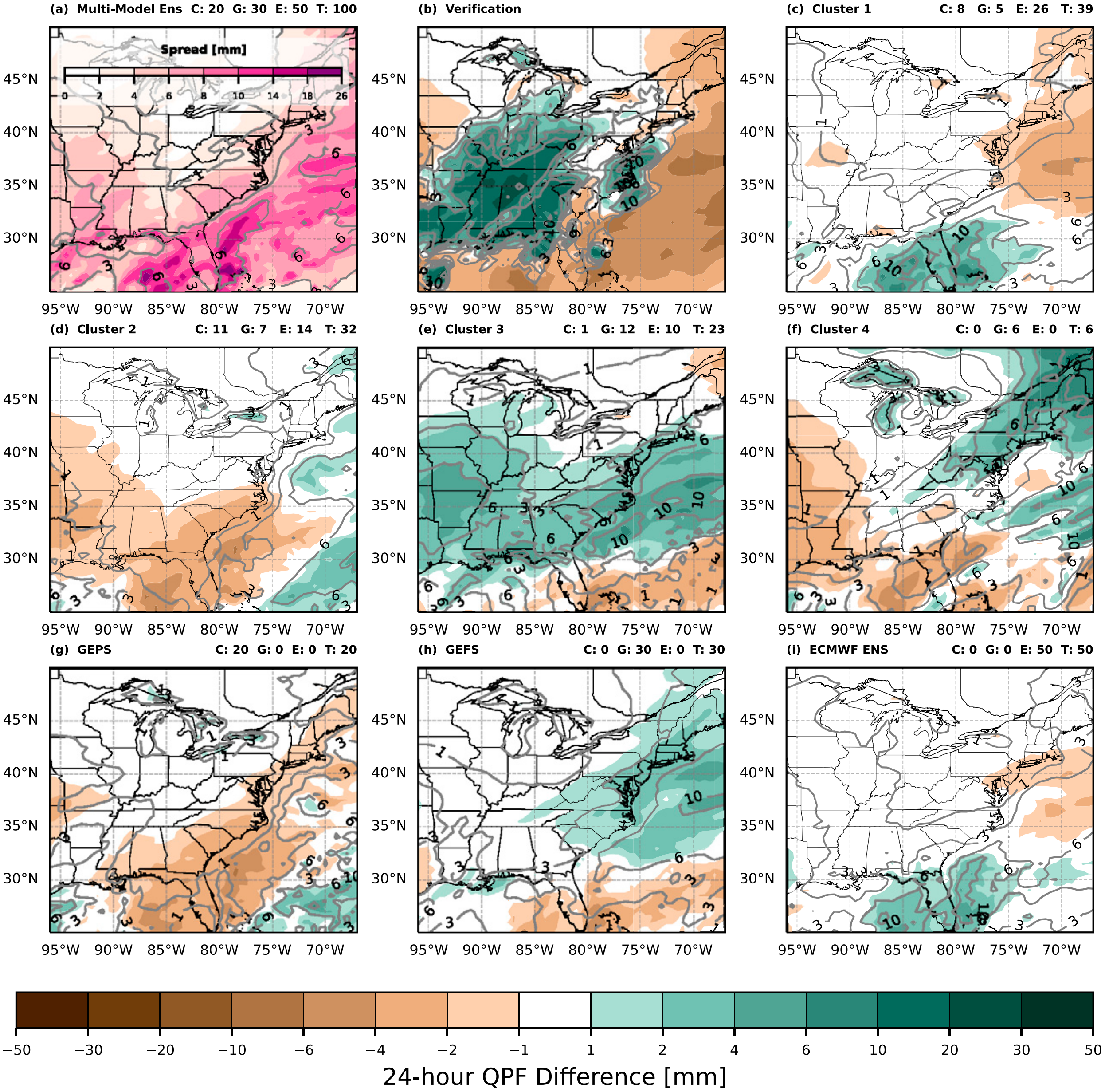

(a) Multi-Model Ens C: 20 G: 30 E: 50 T: 100
Spread [mm]
0 2 4 6 8 10 14 18 26
(b) Verification
(c) Cluster 1 C: 8 G: 5 E: 26 T: 39
(d) Cluster 2 C: 11 G: 7 E: 14 T: 32
(e) Cluster 3 C: 1 G: 12 E: 10 T: 23
(f) Cluster 4 C: 0 G: 6 E: 0 T: 6
(g) GEPS C: 20 G: 0 E: 0 T: 20
(h) GEFS C: 0 G: 30 E: 0 T: 30
(i) ECMWF ENS C: 0 G: 0 E: 50 T: 50
45°N
40°N
35°N
30°N
95°W 90°W 85°W 80°W 75°W 70°W
−50 −30 −20 −10 −6 −4 −2 −1 1 2 4 6 10 20 30 50
24-hour QPF Difference [mm]

FIG. 7. (a) Multimodel ensemble mean (contours) and spread (shading) of 24-h QPF initialized at 0000 UTC 8 Feb 2021 and valid at 0000 UTC 16 Feb 2021. The numbers on top of each panel indicated how many ensemble members from each EPS are included in that panel's mean forecast, "C" before the number of CMC GEPS ensemble members, "G" before the number of GEFS members, "E" before the number of ECWMF ENS members, and "T" before the total number of ensemble members. (b) CCPA analysis of 24-h QPF (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021. (c)–(i) Cluster or EPS mean 24-h QPF (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021.

best-performing EPS mean or deterministic forecast over a large collection of cases. We feel that performing the verification in this manner will establish if the WPC EFC tool consistently provides a more accurate solution than any of the available EPS means or deterministic forecasts. For our verification, verified EPS means include the CMC GEPS, NCEP GEFS, and ECWMF ENS. Verified deterministic forecasts include the GFS and ECMWF. Our limited statistical verification period spans 183 days (15 October 2020–15 April 2021) and clusters were generated for the 0000 and 1200 UTC initializations on each day. This provides 366 sample forecasts for each day in days 3–7 from which to draw statistical conclusions.

Similar to Brill et al. (2015), we apply the resampling method of Hamill (1999) to obtain the distribution of differences in performance metrics for pairs of forecast sources

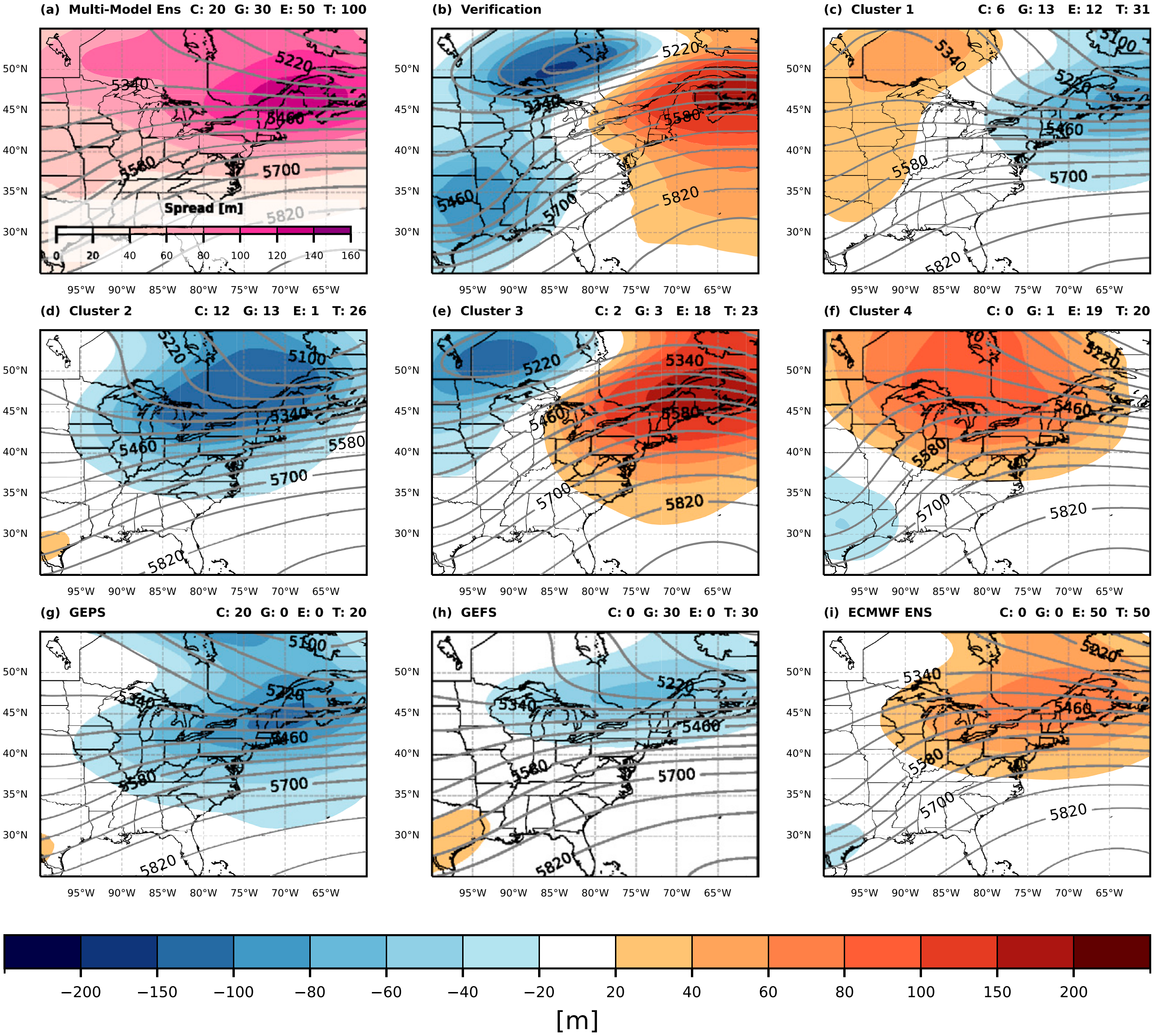

(a) Multi-Model Ens C: 20 G: 30 E: 50 T: 100
(b) Verification
(c) Cluster 1 C: 6 G: 13 E: 12 T: 31
(d) Cluster 2 C: 12 G: 13 E: 1 T: 26
(e) Cluster 3 C: 2 G: 3 E: 18 T: 23
(f) Cluster 4 C: 0 G: 1 E: 19 T: 20
(g) GEPS C: 20 G: 0 E: 0 T: 20
(h) GEFS C: 0 G: 30 E: 0 T: 30
(i) ECMWF ENS C: 0 G: 0 E: 50 T: 50
Spread [m]
0 20 40 60 80 100 120 140 160
50°N 45°N 40°N 35°N 30°N
95°W 90°W 85°W 80°W 75°W 70°W 65°W
5100 5220 5340 5460 5580 5700 5820
−200 −150 −100 −80 −60 −40 −20 20 40 60 80 100 150 200
[m]

FIG. 8. (a) Multimodel ensemble mean (contours) and spread (shading) for 24-h-averaged Z500 initialized at 0000 UTC 10 Feb 2021 and valid at 0000 UTC 16 Feb 2021. The numbers on top of each panel indicated how many ensemble members from each EPS are included in that panel's mean forecast, "C" before the number of CMC GEPS ensemble members, "G" before the number of GEFS members, "E" before the number of ECWMF ENS members, and "T" before the total number of ensemble members. (b) GFS analysis of 24-h-averaged Z500 (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021. (c)–(i) Cluster or EPS mean 24-h-averaged Z500 (contours) and difference from the multimodel ensemble mean (shading) valid at 0000 UTC 16 Feb 2021.

(e.g., best-performing ensemble cluster versus best-performing EPS mean). As in Brill et al. (2015), the error bars (barred line segments) associated with histogram bars in our verification (Figs. 9–12) show confidence intervals obtained from the distribution of paired differences for the forecast source represented by that histogram bar compared to the forecast source of the first histogram bar (the best-performing ensemble cluster forecast). All forecast sources are compared with this best-performing ensemble cluster forecast. The vertical extent of each error bar is determined by the level of the statistical significance test (0.05) and depicts the 95% confidence interval (ranging from the 2.5th percentile to the 97.5th percentile value in the order statistics of resampled differences). Each error bar is plotted with the zero value of the distribution of randomly resampled differences aligned with the value along the ordinate of the performance metric of the best-performing ensemble cluster forecast, a position consistent with the null hypothesis of no difference. Therefore, an error bar partially overlapping the color of the underlying histogram bar indicates no statistically significant difference between that forecast source and the best-performing ensemble cluster forecast. An error bar completely overlapping a color bar or completely

TABLE 3. Median size in number of constituent members for the four clusters for forecast days 3–7.

Forecast day	Day 3	Day 4	Day 5	Day 6	Day 7
Median size of largest cluster	35	35	34	34	34
Median size of second largest cluster	28	28	27	27	27
Median size of third largest cluster	22	22	22	22	22
Median size of smallest cluster	15	15	16	16	16

clear of a color bar indicates a statistically significant difference. The number of samples used in the bootstrap resampling is 5000.

The 24-h-averaged Z500 verification metric is the centered AC described earlier. Accordingly, the "best cluster" is defined as the cluster mean with the highest Z500 AC. The best ensemble and best deterministic models are defined similarly. Figure 9 shows results for the 24-h-averaged Z500 verification over the EAST domain. The centered standardized anomaly correlation is positively oriented, meaning higher values indicate better performance. For days 3–7, the best-performing ensemble cluster outperforms the best-performing EPS mean forecast and the best-performing deterministic forecast and these results are statistically significant. This result was consistent for the three other domains for which cluster forecasts were made available. The difference between the best-performing cluster mean forecast and the best-performing EPS mean or deterministic forecast is small (though statistically significant) on day 3. However, the difference grows to be more substantial by days 6 and 7. These results show that our EFC tool often provides a more accurate 24-h-averaged Z500 forecast than any of the available EPS means or deterministic forecasts for days 3–7, with the utility of the WPC EFC tool increasing as forecast lead time increases.

The other question that needs to be answered by our statistical verification is whether ensemble clusters derived from Z500 can provide accurate forecasts for the other variables of interest. To determine this, the TMAX, TMIN, and QPF forecasts for the clusters were verified similarly to the 24-h-averaged Z500 forecasts over the same 6-month period. For each initialization, the TMAX, TMIN, and QPF forecasts from the ensemble cluster with the best 24-h-averaged Z500 AC were compared against the EPS mean forecast and deterministic forecast with the best 24-h-averaged Z500 AC. The verification metric for TMAX, TMIN, and QPF is the mean absolute error (MAE) over the domain of interest (Table 1). The verifying dataset for TMAX and TMIN is ECMWF's ERA5 dataset (Hersbach et al. 2020). The verifying dataset for QPF is CCPA (Hou et al. 2014). This verification over the EAST domain for TMAX, TMIN, and QPF is shown in Figs. 10–12, respectively.

For TMAX and TMIN (Figs. 10 and 11), the best-performing ensemble cluster had a lower average MAE than the best-performing EPS mean or deterministic forecast for every day in days 3–7. This result was statistically significant at the 95% level. As with the Z500 verification (Fig. 9), the difference between the average MAE of the best-performing cluster and the best-performing ensemble mean or deterministic forecast was small on day 3 but grew throughout the 3–7-day period. This lends further credence to the assertion that the WPC EFC tool provides the most value at the end of the 3–7-day period. The TMAX and TMIN forecast results closely mirroring those of 24-h-averaged Z500 lend support for choosing Z500 as the variable to determine the clusters. For QPF, the verification results were more mixed. The QPF MAE for the best-performing cluster was lower than that of the best-performing EPS mean or best perform deterministic model for every day of days 3–7; however, the difference was small and not always statistically significant at the 95% level. These results indicate that the

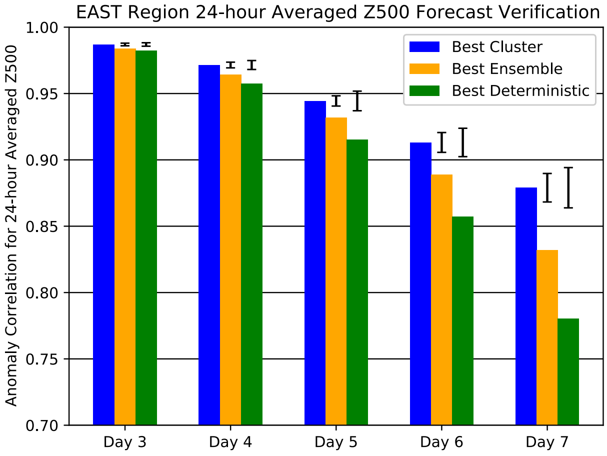


FIG. 9. The 24-h-averaged Z500 centered anomaly correlations as a function of forecast day (indicated along the abscissa) for the EAST region. The color key for the histogram bars is given above the graph. See text for interpretation of error bars.

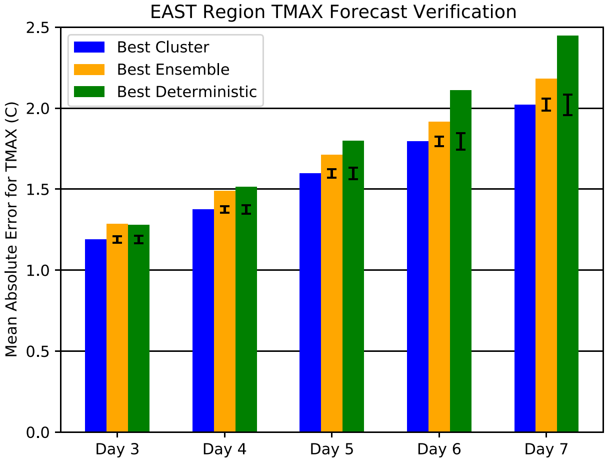


FIG. 10. As in Fig. 9, but for TMAX.

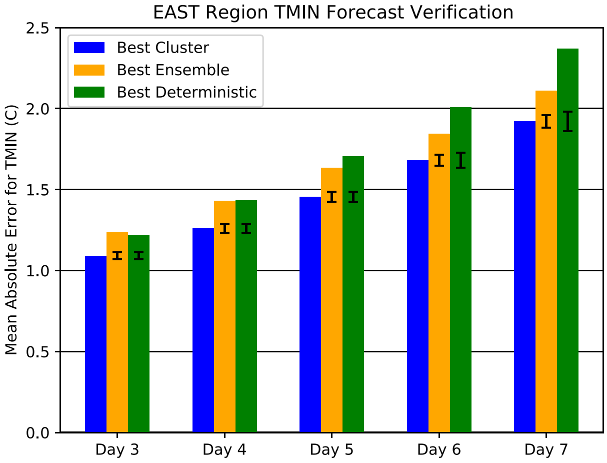


FIG. 11. As in Fig. 9, but for TMIN.

WPC EFC tool provides less informative QPF scenarios than it does TMAX and TMIN forecast scenarios. The verification for TMAX, TMIN, and QPF presented here for the EAST region was also consistent across the other regions of the WPC EFC tool (not shown).

Our verification results show that the best-performing ensemble cluster typically outperforms the best-performing ensemble mean. This result is at least partially due to an ensemble mean's tendency to destroy the structure and reduce the amplitude and variability associated with partially predictable features like the placement and amplitude of large-scale troughs and ridges (Feng et al. 2020). Ensemble cluster means should be less susceptible to this source of forecast degradation as they are comprised of members with similar forecasts for large-scale features. Another implication of this is that the probability of one of the ensemble cluster forecasts outperforming all three operational EPS mean forecasts increases as the number of clusters increases. The verification shows that WPC's EFC tool and its four clusters are an improvement over using the three operational EPS mean forecasts as de facto forecast scenarios.

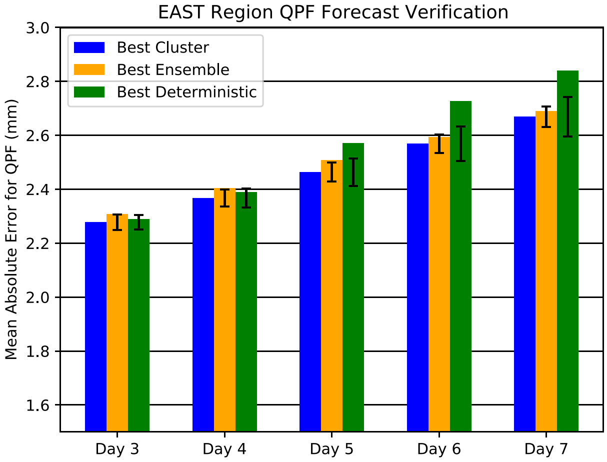


FIG. 12. As in Fig. 9, but for QPF.

But what is the ideal number of clusters? A larger number of clusters could be beneficial, but there will be a point when the number of clusters becomes too much for a forecaster to quickly view, process, and incorporate into their forecast duties. More work needs to be done to determine the ideal number of clusters that will provide forecasters with a good approximation of the range of possible forecast solutions, but not overwhelm them with data.

5. Summary and conclusions

This study detailed an ensemble clustering tool recently developed at WPC to assist NWS forecasters in their preparation of medium-range (3–7 day) forecasts. The clustering method employed by this tool is a variation of fuzzy clustering. Four clusters are generated from a 100-member multimodel ensemble comprised of the CMC GEPS, NCEP GEFS, and ECMWF ENS for each forecast day in days 3–7. These ensemble clusters are generated from an EOF analysis of Z500 over one of four predefined domains. For each day in days 3–7, forecasters can view the cluster mean 24-h-averaged Z500 forecasts and cluster mean forecasts for TMAX, TMIN, and 24-h QPF. The WPC EFC tool was initially developed to help forecasters participating in WPC's ERFE. The tool was well received by ERFE participants who felt the tool helped them to better incorporate ensemble forecasts into their forecast process. Since its debut in WPC's ERFE in 2017, WPC's EFC tool has gained greater adoption by forecasters throughout the NWS.

The 15 February 2021 case study demonstrates why forecasters have found the tool valuable. The day-7 and day-5 forecasts from the four ensemble clusters gave a better approximation of the range of possible forecast outcomes than viewing the mean of each EPS, which is how ensemble forecasts have typically been incorporated into the forecast process. One cluster in particular outperformed the three available EPS means. To ascertain whether the positive results of the 15 February 2021 case study were typical or atypical for WPC's EFC tool, a limited statistical verification was conducted over a 6-month period from 15 October 2020 to 15 April 2021. The limited statistical verification proved that the best-performing cluster based on 24-h-averaged Z500 AC usually outperforms the best-performing EPS mean or deterministic forecast for the variables of 24-h-averaged Z500, TMAX, and TMIN. Results for 24-h QPF were more mixed, indicating that Z500 is not necessarily the best variable for picking out precipitation forecast scenarios.

One major drawback of the WPC EFC tool is that it relies on coarse (0.5°) resolution versions of the global ensembles. This can preclude the tool's utility in regions where fine-scale details are important, like the high terrain of the western United States. To alleviate this problem, we are actively working with the NBM developers to gain access to the aforementioned postprocessed version of each ensemble member's forecast for incorporation into WPC's EFC tool. This would allow WPC's EFC tool to fulfill our ultimate vision of becoming a true companion to the NBM.

Work is also ongoing to improve the underlying clustering methodology. As shown by the QPF verification, solely using

Z500 to generate the cluster forecasts does not always create useful cluster forecasts for all variables of interest. We are investigating ways that variables in addition to Z500 can be incorporated into the cluster generation methodology. We are also investigating whether a time component should be added to the clustering methodology. One of the most frequent criticisms from forecasters using the WPC EFC tool is that the forecast scenarios it presents usually have no continuity from day to day. This can lead to confusion when creating forecasts for events spanning multiple days. We are also investigating how many clusters the WPC EFC tool should present to forecasters. In its current configuration, the tool presents four clusters. Allowing the tool to present more than four clusters could be beneficial, but it will take collaboration with NWS forecasters to determine the ideal number.

Acknowledgments. We thank the participants of WPC's ERFE for the insightful comments and suggestions for improvement during the WPC EFC tool's development. The authors appreciate the encouragement of WPC's management to pursue the further development of this tool outside of ERFE. The authors are grateful to the anonymous reviewers whose comments and suggestions led to significant improvements in the manuscript. This work was supported in part by NOAA's Cooperative Agreement with CIRES, NA17OAR4320101.

Data availability statement. The data used in this study were accessed and available in real time at NCEP. However, they are publicly accessible and may be downloaded from ECMWF's TIGGE archive (https://apps.ecmwf.int/datasets/data/tigge)

REFERENCES

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.

Brill, K. F., A. R. Fracasso, and C. M. Bailey, 2015: Applying a divisive clustering algorithm to a large ensemble for medium-range forecasting at the Weather Prediction Center. *Wea. Forecasting*, **30**, 873–891, https://doi.org/10.1175/WAF-D-14-00137.1.

Buizza, R., 2014: The TIGGE global, medium-range ensembles. ECMWF Tech. Memo. 739, 53 pp., https://www.ecmwf.int/en/elibrary/7529-tigge-global-medium-range-ensembles.

——, 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.

Craven, J. P., D. E. Rudack, and P. E. Shafer, 2020: National blend of models: A statistically post-processed multi-model ensemble. *J. Oper. Meteor.*, **8**, 1–14, https://doi.org/10.15191/nwajom.2020.0801.

ECMWF, 2021: IFS documentation CY47R3—Part V: Ensemble prediction system. ECMWF Doc., 23 pp., https://www.ecmwf.int/en/elibrary/81272-ifs-documentation-cy47r3-part-v-ensemble-prediction-system.

Feng, J., J. Zhang, Z. Toth, M. Peña, and S. Ravela, 2020: A new measure of ensemble central tendency. *Wea. Forecasting*, **35**, 879–889, https://doi.org/10.1175/WAF-D-19-0213.1.

Ferranti, L., and S. Corti, 2011: New clustering products. *ECMWF Newsletter*, No. 127, ECMWF, Reading, United Kingdom, 6–11, https://www.ecmwf.int/sites/default/files/elibrary/2011/14596-newsletter-no127-spring-2011.pdf.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

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.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Inness, P., and S. Dorling, 2012: *Operational Weather Forecasting*. John Wiley and Sons, 248 pp., https://doi.org/10.1002/9781118447659.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, https://doi.org/10.1175/MWR-D-11-00016.1.

Lin, H., and Coauthors, 2019: Global Ensemble Prediction System (GEPS): Update from version 5.0.0 to version 6.0.0. Tech. Note, Environment and Climate Change Canada, 72 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-600_20190703_e.pdf.

Lloyd, S., 1982: Least squares quantization in PCM. *IEEE Trans. Info. Theory*, **28**, 129–137, https://doi.org/10.1109/TIT.1982.1056489.

Palmer, T. N., and Coauthors, 1990: The European Centre for Medium-Range Weather Forecasts (ECMWF) program on extended-range prediction. *Bull. Amer. Meteor. Soc.*, **71**, 1317–1330, https://doi.org/10.1175/1520-0477(1990)071<1317:TECFMR>2.0.CO;2.

Richman, M. B., 1986: Rotation of principal components. *J. Climatol.*, **6**, 293–335, https://doi.org/10.1002/joc.3370060305.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

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.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Yeung, K. Y., and W. L. Ruzzo, 2001: 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. M. 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, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.