Comparison of Clustering Approaches in a Multimodel Ensemble for U.S. East Coast Cold Season Extratropical Cyclones

BENJAMIN M. KIEL[a] AND BRIAN A. COLLE[a]

[a] *School of Marine and Atmospheric Sciences, Stony Brook University, State University of New York, Stony Brook, New York*

(Manuscript received 29 January 2023, in final form 2 January 2024, accepted 4 January 2024)

ABSTRACT: Several clustering approaches are evaluated for 1–9-day forecasts using a multimodel ensemble that includes the GEFS, ECMWF, and Canadian ensembles. Six clustering algorithms and three clustering spaces are evaluated using mean sea level pressure (MSLP) and 12-h accumulated precipitation (APCP) for cool-season extratropical cyclones across the Northeast United States. Using the MSLP cluster membership to obtain the APCP clusters is also evaluated, along with applying clustering determined at one lead time to cluster forecasts at a different lead time. Five scenarios from each clustering algorithm are evaluated using displacement and intensity/amount errors from the scenario nearest to the MSLP and 12-h APCP analyses in the NCEP GFS and ERA5, respectively. Most clustering strategies yield similar improvements over the full ensemble mean and are similar in probabilistic skill except that 1) intensity displacement space gives lower MSLP displacement and intensity errors; and 2) Euclidean space and agglomerative hierarchical clustering, when using either full or average linkage, struggle to produce reasonably sized clusters. Applying clusters derived from MSLP to 12-h APCP forecasts is not as skillful as clustering by 12-h APCP directly, especially if several members contain little precipitation. Use of the same cluster membership for one lead time to cluster the forecast at another lead time is less skillful than clustering independently at each forecast lead time. Finally, the number of members within each cluster does not necessarily correspond with the best forecast, especially at the longer lead times, when the probability of the smallest cluster being the best scenario was usually underestimated.

SIGNIFICANCE STATEMENT: Numerical weather prediction ensembles are widely used, but more postprocessing tools are necessary to help forecasters interpret and communicate the possible outcomes. This study evaluates various clustering approaches, combining a large number of model forecasts with similar attributes together into a small number of scenarios. The 1–9-day forecasts of both sea level pressure and 12-h precipitation are used to evaluate the clustering approaches for a large number of U.S. East Coast winter cyclones, which is an important forecast problem for this region.

KEYWORDS: Extratropical cyclones; Empirical orthogonal functions; Ensembles; Postprocessing; Clustering

1. Introduction

Winter extratropical cyclones along the U.S. East Coast pose a variety of hazards, including coastal flooding (Blake et al. 2013), strong winds (Booth et al. 2015), and heavy precipitation (Colle et al. 2013). The impacts of these cyclones are highly dependent on storm-track variations (Ma and Chang 2017) and can be difficult to predict. One such tool to help forecast these storms and other types of extreme weather is model ensembles, which have been widely used in operational forecasting since the 1990s (Molteni et al. 1996; Houtekamer et al. 1996; Toth and Kalnay 1993); however, the common products for interpreting model ensemble output, such as the ensemble mean, spaghetti plots, and probability plots, are not easily applicable for Impact-Based Decision Support Services (IDSS) of the National Weather Service (Uccellini and Hoeve 2019). In addition, U.S. East Coast extratropical cyclones have recently shown general intensity underprediction and left-of-track biases at the medium range, and bias toward slower progression of cyclones in the short range, but with a high degree of variation from case to case (Korfe and Colle 2018).

One method to generate possible outcomes from model ensembles is to cluster ensemble members with similar forecasts and then evaluate the "scenario," or the ensemble mean of each cluster. The Met Office (UKMO) has used *k*-means clustering (KMC) in the climatology of extratropical cyclones to identify several circulation patterns and then match ensemble members to the climatological circulations (Neal et al. 2016). Using climatological and forecast frequencies in combination allows UKMET to derive the probabilities of a certain circulation pattern emerging and allows forecasters to become familiar with these patterns.

Zheng et al. (2017, hereafter ZH17) developed a "fuzzy" clustering (FZC) method using a 90-member multimodel ensemble, including the Global Ensemble Forecast System (GEFS) (Toth and Kalnay 1993), the Canadian Meteorological Center Ensemble (CMC) (Molteni et al. 1996), and European Centre for Medium-Range Weather Forecasts (ECMWF) (Houtekamer et al. 1996) and applied it to two U.S. East Coast extratropical cyclones. The ZH17 method clustered mean sea level pressure (MSLP) from ensemble members using the leading two principal components (PCs) of a principal component analysis, analyzing modes of variability across the model ensemble. The two

Corresponding author: Benjamin M. Kiel, benjaminmkiel@gmail.com

DOI: 10.1175/WAF-D-23-0017.1

© 2024 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

case examples in ZH17 demonstrated that the method could produce well-separated clusters with distinct probabilities for MSLP and 500-hPa geopotential height. Furthermore, ZH17 found lower root-mean-square errors and higher correlations for the scenario nearest to analysis (SNA) compared to the 90-member ensemble mean. Zheng et al. (2019, hereafter ZH19) applied the two-PC (2PC) FZC method to a set of 180 East Coast winter cyclone events from 2007 to 2015 using the 24–216-h lead times. On average the mean of the SNA performed better than the 90-member ensemble mean. ZH19 also found that each of the ensemble systems were generally underdispersed at lead times of 72 h and later, while clustering the multimodel 90-member ensemble resulted in the least underdispersion. ZH19 also detailed an analysis of the influence that the dynamical cores had on clustering and found that the ECMWF members had a higher probability of being in the cluster nearest to the analysis than members of either the CMC or GEFS. Nevertheless, ZH19 still found that the SNA performed better than the ECMWF ensemble mean and the multimodel ensemble mean.

Recently, the Weather Prediction Center (WPC) has utilized the ZH17 clustering approach on the same global multimodel ensemble but generates the four clusters using 500-hPa geopotential heights and a KMC approach (Lamberson et al. 2023, hereafter L23). WPC's clustering tool showed forecasting improvements over the ensemble mean when using the "best cluster," with lower mean absolute errors for forecasts of daily maximum and minimum temperature and precipitation. Furthermore, the clustering method was well received by forecasters.

The ZH17, ZH19, and WPC results are encouraging and the 2PC KMC method is now operational within the National Weather Service forecast offices using the DESI (Dynamic Ensemble-based Scenarios for IDSS) software (J. Nelson 2022, personal communication). However, the NWS and prior studies only utilize the 2PC KMC approach, only cluster the mass field (e.g., MSLP or 500-hPa height), and clusters are done independently for each forecast lead time. Therefore, our study evaluates six clustering algorithms, each in three clustering spaces, to determine the accuracy and probabilistic skill of each method for the most likely SNA. The list of clustering algorithms tested include 1) FZC, 2) KMC, 3) agglomerative hierarchical clustering (AHC) using Ward's linkage, 4) AHC using average linkage, 5) AHC using full linkage, and 6) clustering by self-organizing map. The list of clustering spaces tested includes 1) the 2PC space, 2) the Euclidean space, and 3) the intensity displacement (IDISP) space. The details of the clustering algorithms and the clustering spaces are discussed in section two. Direct application to model ensembles in previous studies, along with current operational use, motivate the comparison of FZC, KMC, and the 2PC space. AHC was selected due to the ease of use and understanding of the approach. Self-organizing map was selected for being a distinct alternative to contrast the other algorithms tested. Euclidean space was selected to demonstrate the importance of isolating modes of variability. IDISP space presented a new option not directly tied to modes of variability in contrast with the Euclidean and 2PC spaces.

While AHC and self-organizing map have not been used for ensemble clustering, other meteorological applications exist for these algorithms. Lopes and Machado (2015) gathered information about the temporal changes in tornadic activity, and then used AHC to categorize the yearly tornadic activity and then find the common weather patterns for each cluster. Dolan and Davis (1992) used a two-stage intensity-based clustering technique with AHC to evaluate Nor'easter climatology. Finally, a technique to identify and classify the flow patterns around extratropical cyclones was used in Hart et al. (2015), clustering each flow pattern by identifying a positional vector representing the flow trajectory around the cyclone.

A self-organizing map was used in Ohba and Sugimoto (2019) to identify synoptic patterns for the mei-yu–baiu front in China. The self-organizing map was given daily climatological zonal and meridional winds from the mei-yu–baiu region as inputs. This generated representative patterns which could be used to identify the synoptic and mesoscale behaviors along the front. Rousi et al. (2015) fed a self-organizing map daily 500-hPa geopotential height anomalies to identify teleconnection patterns. Reusch et al. (2007) used a SOM for analysis of monthly MSLP climate variability patterns for the North Atlantic Oscillation.

While a diversity of clustering algorithms exist, with many having meteorological applications, little work exists to compare the effects of clustering approaches on model ensembles. The main goal of our work is to fill this gap by comparing several combinations of clustering approaches. Additional approaches which do not appear to be addressed in the literature include evaluating the efficacy of using precipitation scenarios obtained using mass field (e.g., MSLP) and the impact of using the same cluster membership from earlier or later model runs of a case. This study addresses the following questions:

- What is the most skillful clustering approach as a function of forecast lead time?
- How well does using the clustering at one lead time to cluster forecasts at different lead times compare to their synchronized counterparts?
- How does the skill of clustering 12-h accumulated precipitation (APCP) using MSLP clusters compare to clustering APCP itself?

2. Datasets

The same 90-member ensemble was used as in ZH19, consisting of the 50 ECMWF, 20 CMC, and 20 GEFS members, respectively. The Global Forecast System (GFS) analysis was used for the verification for MSLP to be consistent with ZH19. ECMWF Reanalysis v5 (ERA5) data from the Copernicus Climate Data Store (Hersbach et al. 2019) was used for precipitation verification since it has 1-h precipitation accumulation, and it has agreed well with observations over the Northeastern United States (Crossett et al. 2020). All ensemble and reanalysis data were bilinearly interpolated to a 1° × 1° grid after obtaining them from The Observing System Research and Predictability Experiment's Interactive Grand Global Ensemble Archive (TIGGE) (Bougeault et al. 2010; Swinbank et al. 2016).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

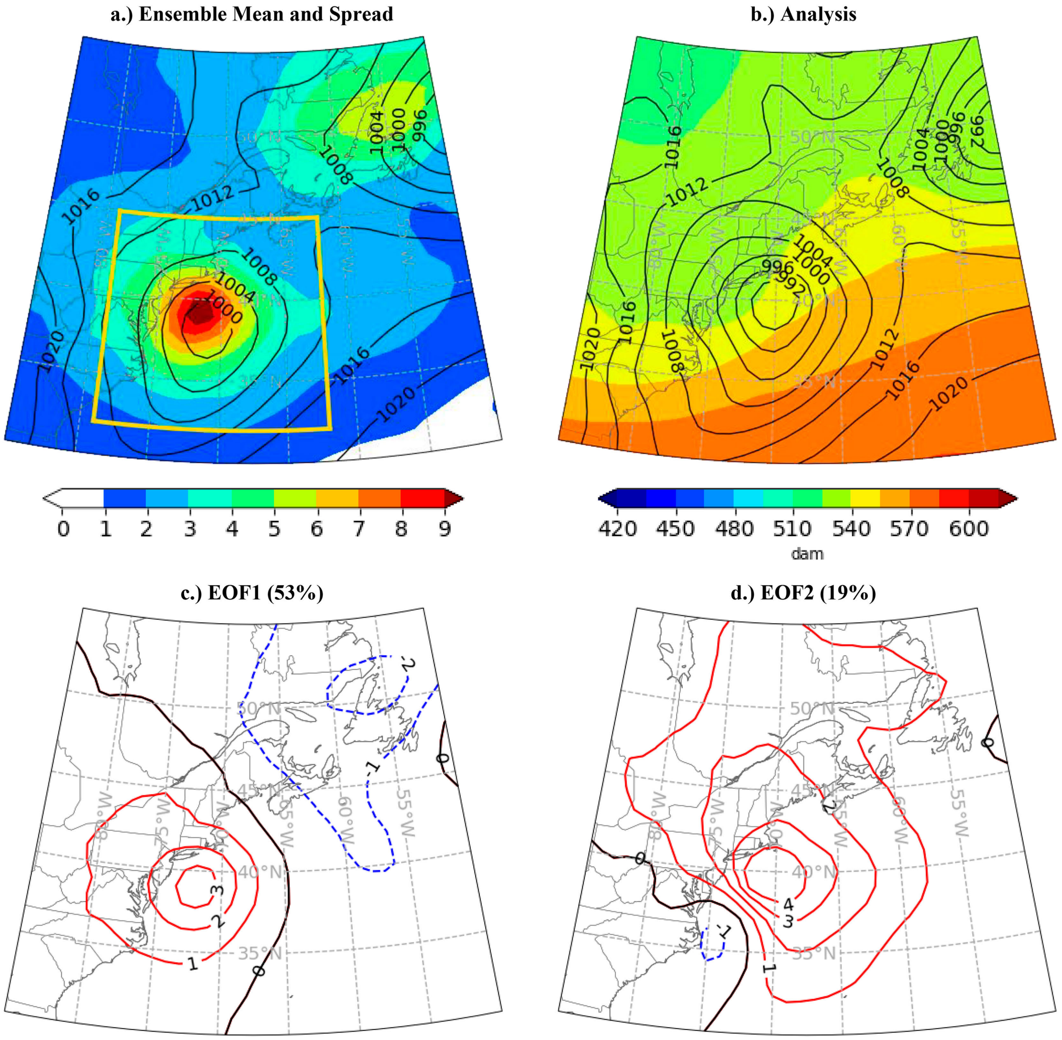


FIG. 1. (a) Outline of region 1 (solid yellow), and the 90-member ensemble mean MSLP (every 4 hPa, solid black) and spread of MSLP (shaded, every one standard deviation) for a valid time (VT) at 0000 UTC 30 Dec 2012, with an initial time (IT) at 0000 UTC 27 Dec 2012. (b) NCEP analysis MSLP (solid every 4 hPa) and 500-hPa height (color shaded every 15 dam). (c) First EOF of PCA (red positive and blue negative) calculated for MSLP for this ensemble valid time. (d) As in (c), but for the second EOF.

As in ZH19, this study utilized the same 180 winter extratropical cyclones from November through March from 2007 to 2015 that attain an analysis minimum pressure $<$ 1005 hPa within the region from 32°–45°N to 79°–62°W (region 1, yellow box in Fig. 1a) at any point in its lifetime. If a cyclone obtains the pressure minimum multiple times while within region 1, the two valid times that are nearest to the center of region 1 were both added to the dataset. For each cyclone, each ensemble forecast was analyzed every 24 h from a day-1 (24 h) lead time to a day-9 (216 h) lead time. The results were combined into short (24, 48, 72 h), medium (96, 120, 144 h), and long (168, 192, 216 h) lead times.

Because of TIGGE data availability, the total number of cases analyzed at any given forecast hour is less than 180. In addition, data for APCP must also have the corresponding MSLP data available so that cross-variable clustering could be performed. The percentage of cyclones evaluated for each lead time for MSLP are 70%, 66%, and 64% for the short, medium, and long lead times, respectively, and 64%, 61%, and 59% for APCP.

While region 1 was used to identify cyclones as in ZH19 (Fig. 1a), the evaluation of model ensemble clusters for this paper was done on a larger region 2 (30°–55°N, 85°–50°W, the entire region depicted in Fig. 1a). If the MSLP analysis pressure minimum fell exactly on the boundary of region 2, the case was excluded to reduce instances where the nontargeted cyclones were present. This filtering removed 52 (29% of) cyclones. A domain larger than region 2 was not considered or tested in our study, since a larger domain increases the risk that the first PCs do not fully capture the variability of the targeted cyclone.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

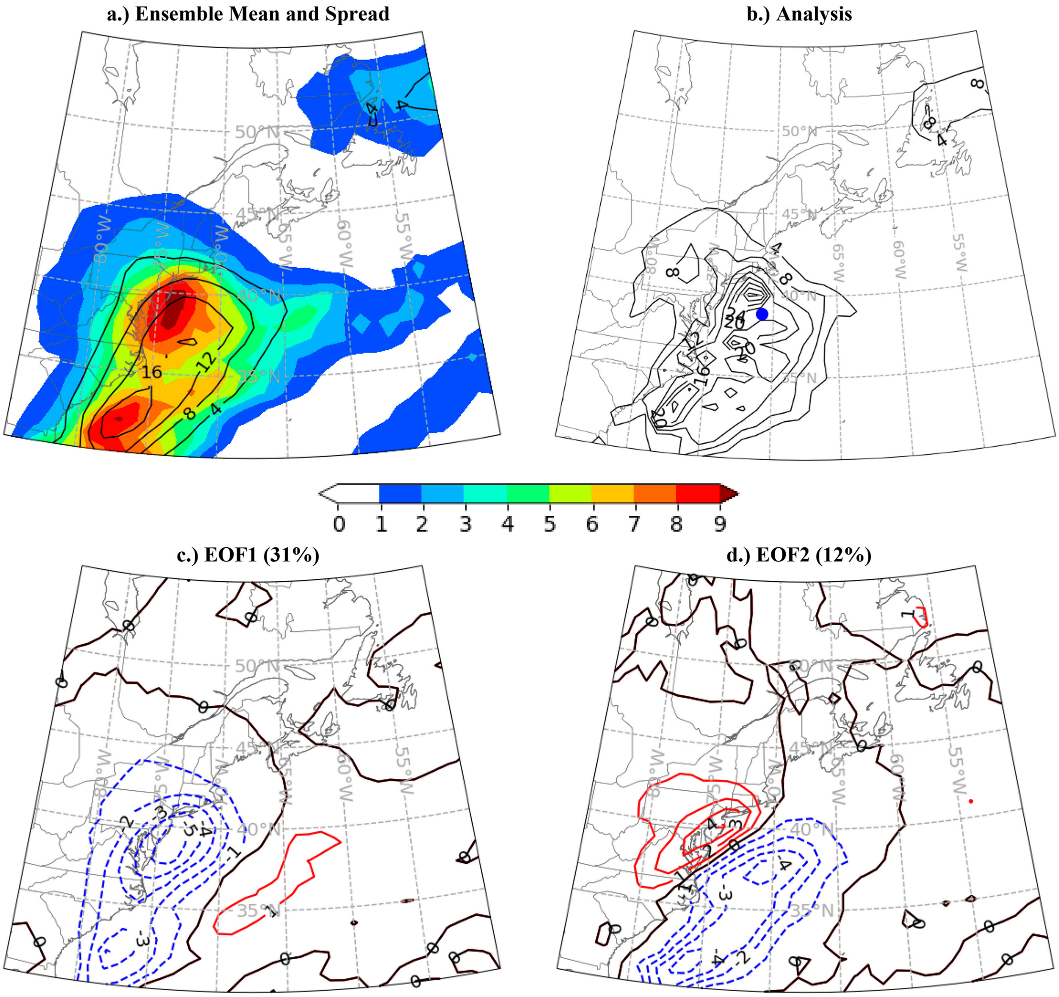


FIG. 2. (a) The 90-member ensemble mean of 12-h APCP (every 4 mm, solid black) and spread (shaded, every one standard deviation) at the same VT and IT as in Fig. 1. (b) ERA5 analysis of 12-h APCP (every 4 mm, solid black) and center of mass (blue dot). (c) First EOF of PCA calculated for 12-h APCP for this model run. (d) As in (c), but for the second EOF.

3. Clustering approaches

Clustering requires two decisions: algorithm and space. The tested spaces included: 1) 2PC space, 2) Euclidean space, and 3) an "intensity displacement" (IDISP) space. The tested algorithms include: 1) *k*-means clustering (KMC), 2) fuzzy clustering (FZC), 3) agglomerative hierarchical clustering (AHC) using Ward's linkage, 4) AHC using average linkage, 5) AHC using full linkage, and 6) clustering by self-organizing map (SOM). Therefore, 18 combinations were evaluated. For all clustering approaches, the number of clusters was fixed to 5 as in ZH17 and ZH19. ZH19 had tested a number of clusters ranging from 2 to 8, with 5–6 clusters yielding the most stable and skillful solution. Furthermore, WPC settled on 4 clusters for the forecasters (L23), which is similar to our number.

a. Clustering spaces

This study used the same 2PC space method as described in ZH17, ZH19, and L23. First, a PC analysis was performed across model members in region 2 for either MSLP or APCP at a particular lead time. PC analysis begins with the construction of a matrix containing the ensemble data, with each column representing an ensemble member and each row representing a grid point. The EOFs are the eigenvectors $\mathbf{v}_n$ of the covariance matrix. The first and second eigenvectors, $\mathbf{v}_1$ (EOF1) and $\mathbf{v}_2$ (EOF2), represent the first and second greatest modes of variability within the ensemble. The scalar product between EOF1 and EOF2 and the ensemble members gives the PCs representing the greatest (PC1) and second greatest (PC2) modes of variability. PC1 and PC2 serve as the phase space to perform clustering. ZH19 showed that the interquartile range of the variance explained by each PC ranges for the 180 cyclone cases from 12% to 66% for PC1 and 6%–42% for PC2, thus the total variance explained within the 2PC space can be low for some cases. However, adding more PCs to the clustering space does not significantly change results (Kiel 2021); while additional PCs may

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

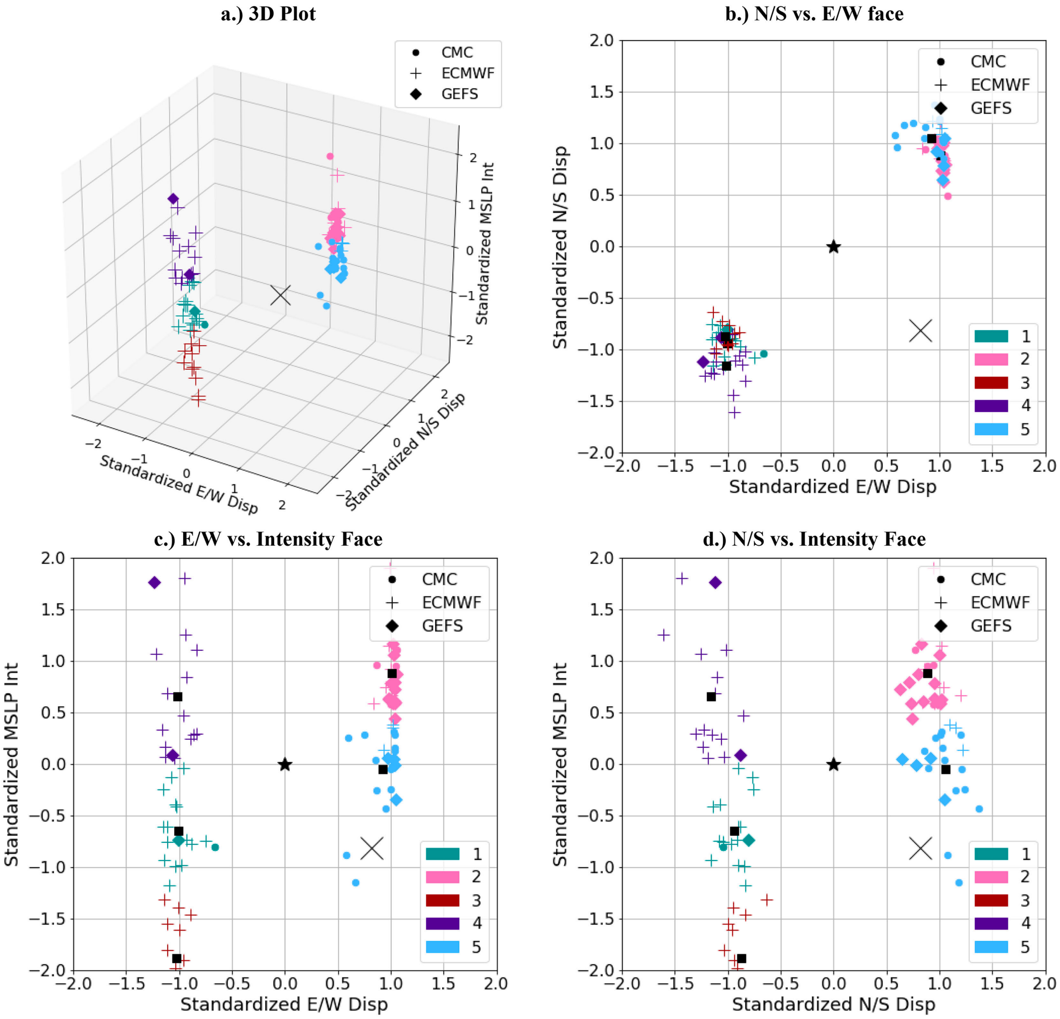


FIG. 3. (a) Plot of the 3D intensity displacement space for the December 2012 case (Fig. 1) using FZC. (b) The N/S vs E/W face of the space. (c) The E/W vs intensity face of the space. (d) The N/S vs intensity face of the space. The NCEP analysis (black X) is marked in all plots. In (b)–(d), cluster means (black squares), and the 90-member ensemble mean (black star) are also marked.

improve clustering in a small number of cases, often a low variance explained indicates a high degree of noise (North et al. 1982).

Figures 1 and 2 show an example of a PC analysis for MSLP and 12-h APCP, respectfully, for a 72-h forecast valid at 0000 UTC 30 December 2012. There are two cyclones within region 2, with the southwest (SW) cyclone more relevant to northeastern U.S. forecasters. The ensemble mean and spread of MSLP (Fig. 1a) and 12-h APCP (Fig. 2a) demonstrate the large variance in ensemble member solutions.

The first two EOFs demonstrate the spatial patterns in MSLP that explain 53% and 19% of the ensemble spread and variance in this case. For MSLP, EOF1 is dominated by the strength of the SW cyclone and along-track spatial variations in the northeast (NE) cyclone (Fig. 1c). EOF2 variance is dominated by along-track variance in the SW cyclone (Fig. 1d). For 12-h APCP, EOF1 (Fig. 2c) represents intensity variations in the precipitation shield and explains 31% of the variance (Fig. 2c), while EOF2 represents an east/west shift in the precipitation shield and explains 12% of the variance (Fig. 2d).

The next space evaluated was the Euclidean space. Each ensemble member was organized as a two-dimensional latitude/longitude grid (X_m, Y_n). The Euclidean space transforms the two-dimensional grid into a one-dimensional vector (0, $\mathbf{Z}_{m+n}$, in which $m = 18$ and $n = 14$). Clustering was done by comparing the Euclidean distances of each of these one-dimensional vectors. Alternatively, the Euclidean space can be considered equivalent to using all of the PCs of a PC space. While some of the results from clustering using the Euclidean space are shown below, the Euclidean space itself cannot be visualized due to its high dimensionality.

The final space is a proposed 3D space: the IDISP space. For clustering weather features such as cyclones, IDISP is defined by an intensity parameter on one axis and the displacement

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

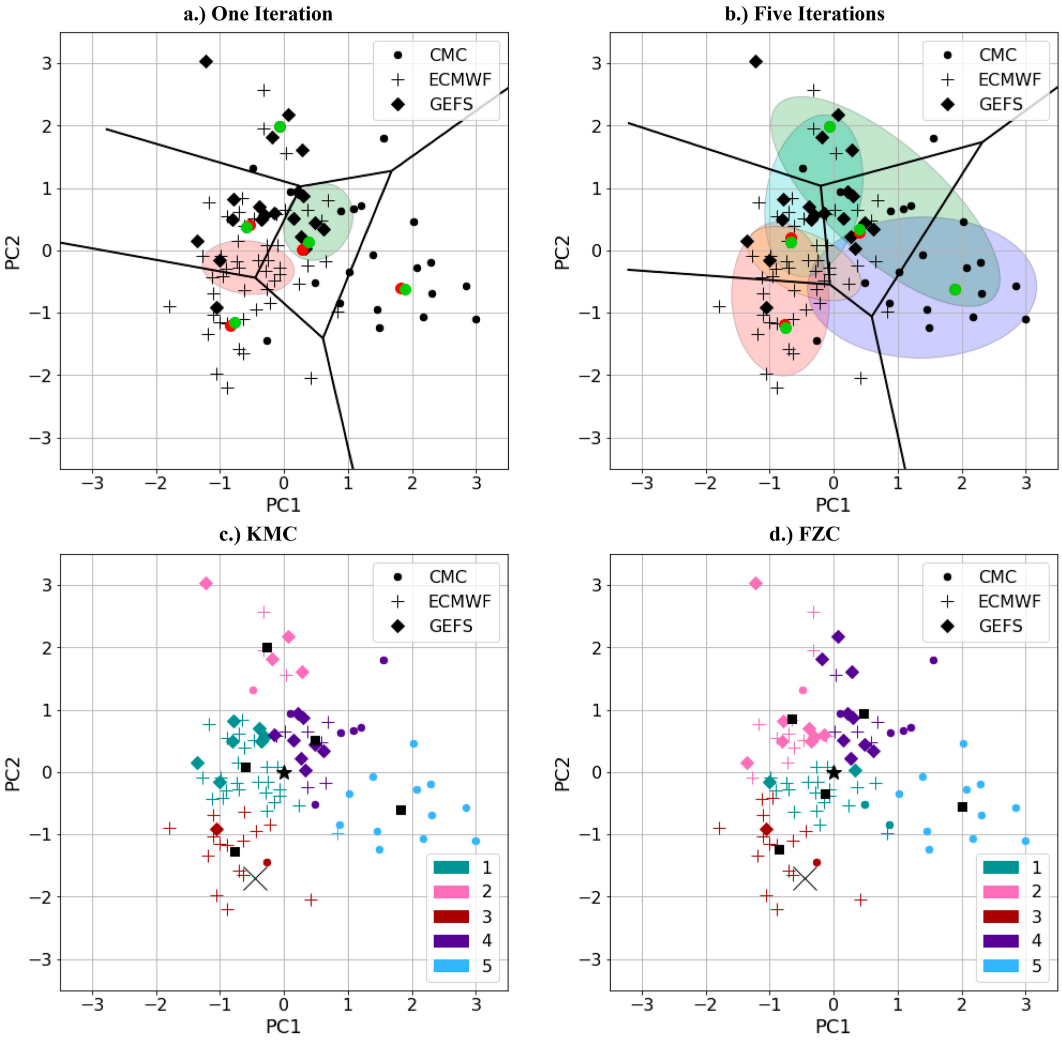
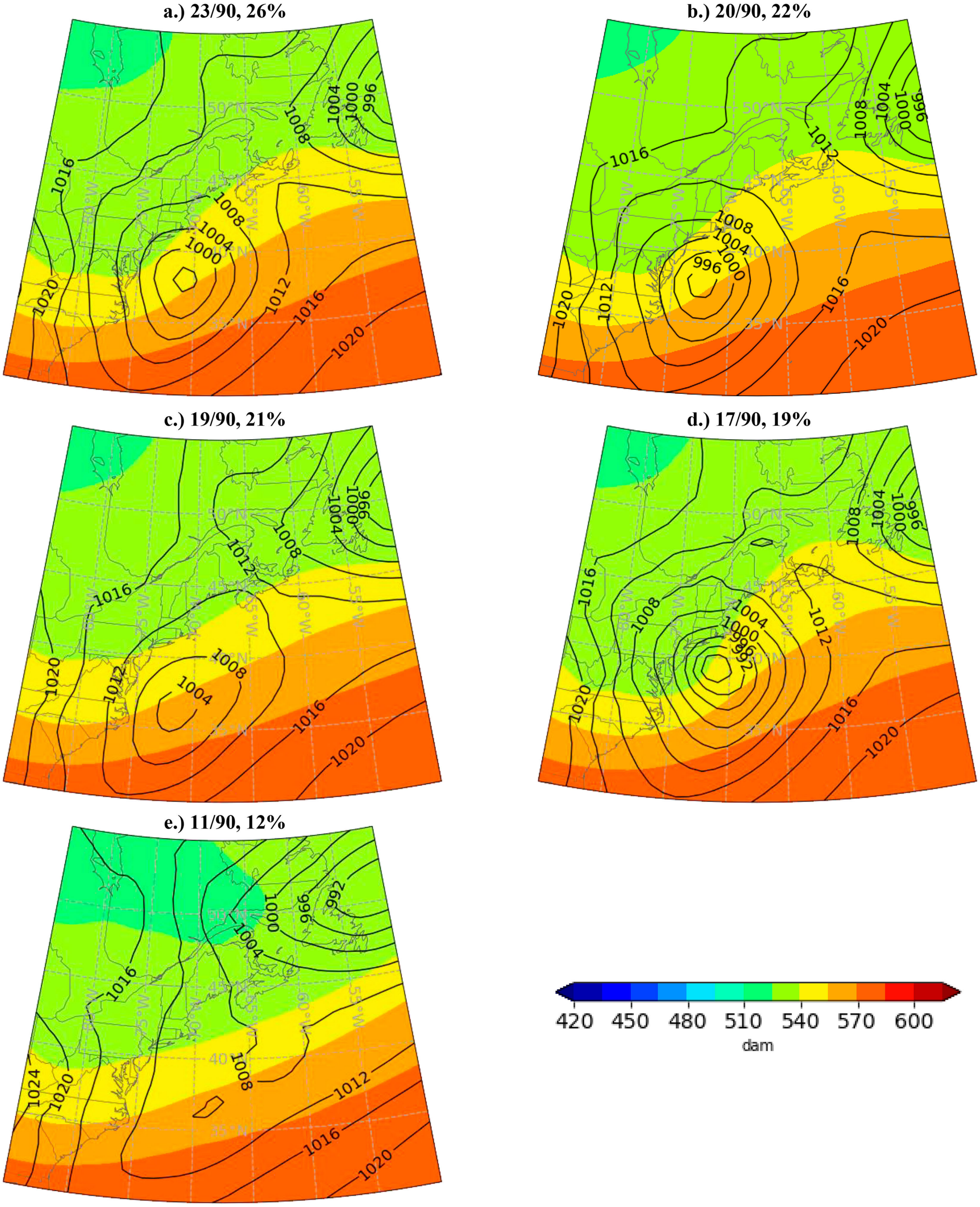


FIG. 4. KMC and FZC of 2PC space of the December 2012 case (same VT as Fig. 1) after (a) one and (b) four iterations. KMC is represented by preiterative cluster centers (red), the cell where each cluster center is nearest (solid black lines), and new cluster centers (green) after averaging all phase points within the cell l. FZC is represented by bivariate ellipses around where each phase point's probability of belonging to a given cluster exceeds 25% (shaded). Next, color-coded clusters (see legend 1–5) are shown after convergence for (c) KMC (8 iterations) and (d) FZC (82 iterations). Also plotted in (c) and (d) are the NCEP analysis (black X), 90-member ensemble mean (black star), and mean PC1/PC2 values of clusters (black squares).

variations of the intensity parameter on the other two axes (Fig. 3). For MSLP, the intensity component is the minimum pressure value, and the two displacement components are the north/south (N/S; latitude) and east/west (E/W; longitude) displacements of the minimum pressure for a cyclone relative to the ensemble mean. The MSLP IDISP approach was highly sensitive to the location of the pressure minimum, and in scenarios where at least two cyclones were present within the clustering region, bimodal distributions occurred as the pressure minimum flipped between cyclones (Fig. 3). Given the large variability in precipitation across ensemble members, the intensity for APCP is simply the average of all precipitation at each grid point in region 2, and displacement is defined as the N/S and E/W displacement of the center of mass of APCP in this region. The center of mass gives a reference location of the precipitation shield (e.g., Bytheway and Kummerow 2015; Duda and Gallus 2013) and it is determined by weighing each grid point latitude/longitude value by the precipitation amount and averaging the weighted values together. All MSLP and APCP intensities and displacements were found for each ensemble member and standardized. The standardized value is calculated as the raw value minus the ensemble mean and the difference divided by the ensemble standard deviation. By definition, IDISP is the only space that does not use the general variance of the field to cluster, instead using the variance of a specific feature within the field (e.g., variance of pressure minimum) to cluster.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC



a.) 23/90, 26%

b.) 20/90, 22%

c.) 19/90, 21%

d.) 17/90, 19%

e.) 11/90, 12%

FIG. 5. Resulting scenarios plotted as in Fig. 1b, or the "cluster ensemble mean" of the ensemble members represented by each cluster, ordered from (a) largest to (e) smallest, of Fig. 4d (2PC EOF FZC), plotted as in Fig. 1b.

b. Clustering algorithms

KMC attempts to minimize the sum of the squared distances between all points representing the ensemble members within a given clustering space (hereafter, phase point) and the nearest cluster center (MacQueen 1967). KMC is conducted in several steps. First, an initial set of clusters is generated using the *k*-means plus algorithm (Arthur and Vassilvitskii 2007) (Fig. 4a). Next, the Elkan algorithm (Elkan 2003), which uses the triangle inequality to accelerate KMC, is used to converge toward a local minimum. Finally, the algorithm is rerun 10 times. The solution with the lowest sum of squared intracluster distances was taken (Fig. 4c).

FZC is similar to KMC, except that each element is assigned a probability of belonging to each cluster (Dunn 1973). Each intracluster distance is weighted by the average of its probability. Each probability was initially assigned randomly.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

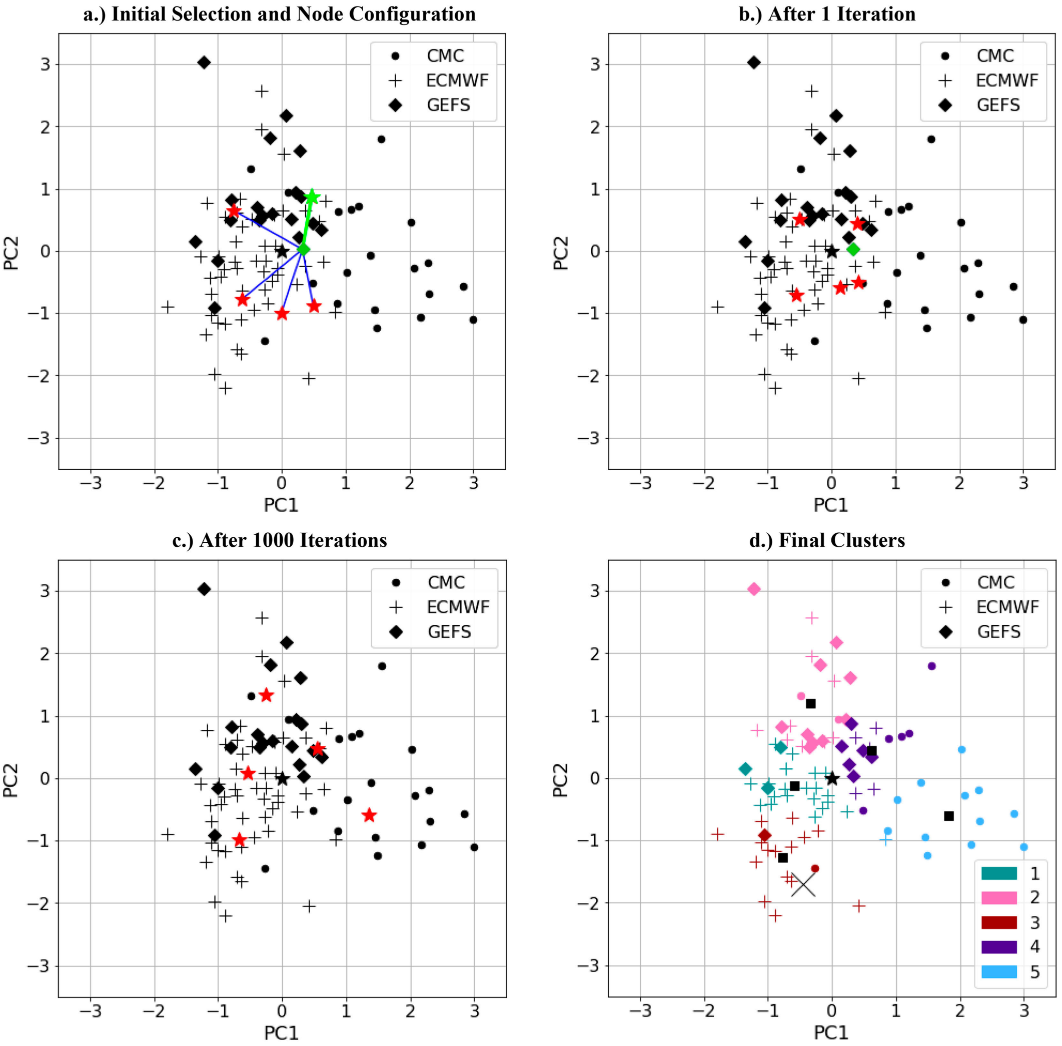


FIG. 6. SOM for 2PC space for the December 2012 case. (a) Plot of the randomly selected ensemble member (green diamond), SOM nodes (red stars), and the lines representing the Euclidean distances between the ensemble member and the SOM nodes (blue lines), and the phase point closest to the node (green star, green line). (b) Plot of the SOM nodes after they are adjusted toward the green star. (c) Plot of SOM nodes after 1000 iterations. (d) The final color-coded clusters, NCEP analysis (black X), 90-member ensemble mean (black star), and mean PC1/PC2 values of clusters (black squares).

The Ross algorithm (Ross 2010), which applies a weight to the "fuzziness" of clusters, was used to iterate from the random assignments toward the local minimum solution efficiently (Fig. 4a). After FZC was completed, each member was assigned to the cluster which has the highest probability of representing it (Fig. 4b). The iterations continued until a local minimum was reached (Fig. 4d). Sensitivity from the initialization of the clustering algorithm using various random seeds is insufficient to significantly change our results. Figure 5 shows the scenarios from FZC along with the number of members which contributed to each scenario. For FZC of the Euclidean space, an additional behavior is noted where the algorithm sometimes fails to generate a complete set of five clusters even when they are specified. The behavior is a result of assigning each phase point to the cluster with the highest probability of belonging to that cluster. If every phase point has a higher chance of being a part of one of four clusters than the fifth cluster, the fifth cluster is excluded and only four clusters form. This behavior is only observed with Euclidean space; FZC in the other spaces always leads to five clusters. In addition, the loss of the intended number of clusters is not an issue for any discrete clustering algorithms.

AHC works by iteratively merging the phase points/clusters one at a time until the desired number of clusters is reached. Clusters are merged on each iteration based on a linkage criterion. This study tests 1) Ward's linkage, 2) full linkage, and 3) average linkage. Ward's linkage finds the two phase points or group of phase points that when merged results in the lowest increase in the sum of squared Euclidean distances between all members within the same cluster (Ward 1963). Full

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

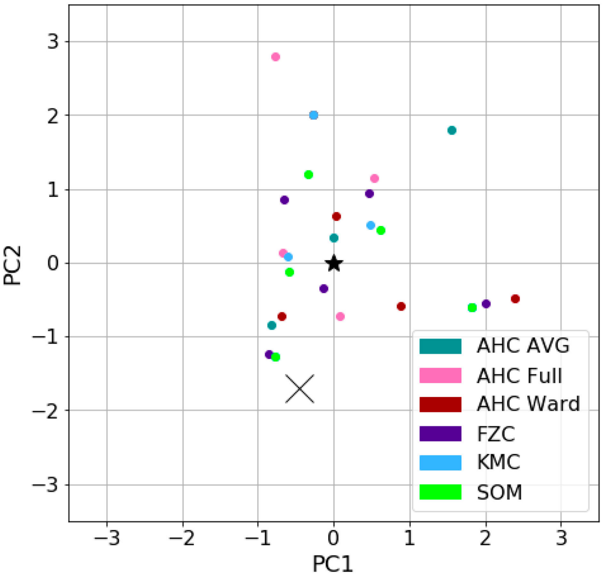


FIG. 7. 2PC cluster means of each of the clustering algorithms tested (colored dots), NCEP analysis (black X), and 90-member ensemble mean (black star).

linkage finds the minimum Euclidean distance between the furthest members of each phase point/group (Defays 1977). Average linkage finds the average distance between all possible phase point/group combinations and merges the two phase points/groups with the smallest average. (Sokal and Michener 1958).

The largest scenario of AHC in a 2PC space gave a larger first cluster of 35/90 (39%), compared to 23/90 (26%) for KMC, but MSLP scenarios for AHC Ward's linkage maintain visual similarity to the 2PC FZC scenarios.

SOM clustering is done by iteratively evaluating the distance between each phase point in the clustering space against randomly located nodes (Kohonen 1982). Random locations, or "nodes," equivalent to the number of clusters desired, were selected across the clustering space. On each iteration, one phase point was selected at random. The node which was nearest to the selected phase point by the Euclidean distance was defined as the best matching unit (BMU). The BMU and the other nodes were moved closer to the phase point using an inverse function of distance from the BMU, where the further the nodes were from the BMU, the less they were adjusted toward the phase point. Rousi et al. (2015) shows the equations used in this study for the SOM. SOM is demonstrated by showing the identified phase point (green diamond) and BMU (green star, connected by green solid line to green diamond) in the first iteration (Fig. 6a), followed by the node locations after that iteration (Fig. 6b). After repeating for 1000 iterations (Fig. 6c), the resulting clusters (Fig. 6d) were retrieved. SOM scenarios only show slight deviations from FZC (Figs. 7 and 8).

c. Alternative clustering strategies

Two "alternative" tests using 2PC fuzzy clustering were considered: 1) Using the MSLP clusters to cluster 12-h APCP or "cross-field" clustering; and 2) using an alternative lead time, or applying cluster memberships obtained at one lead time to ensemble members at a different lead time. The first method checks if the clusters remain consistent across different forecast parameters. The second method tests if clusters remained consistent across lead times and the viability of utilizing the same cluster members across many lead times, which would help one to animate fields smoothly through the forecast.

4. Evaluation of clustering approaches

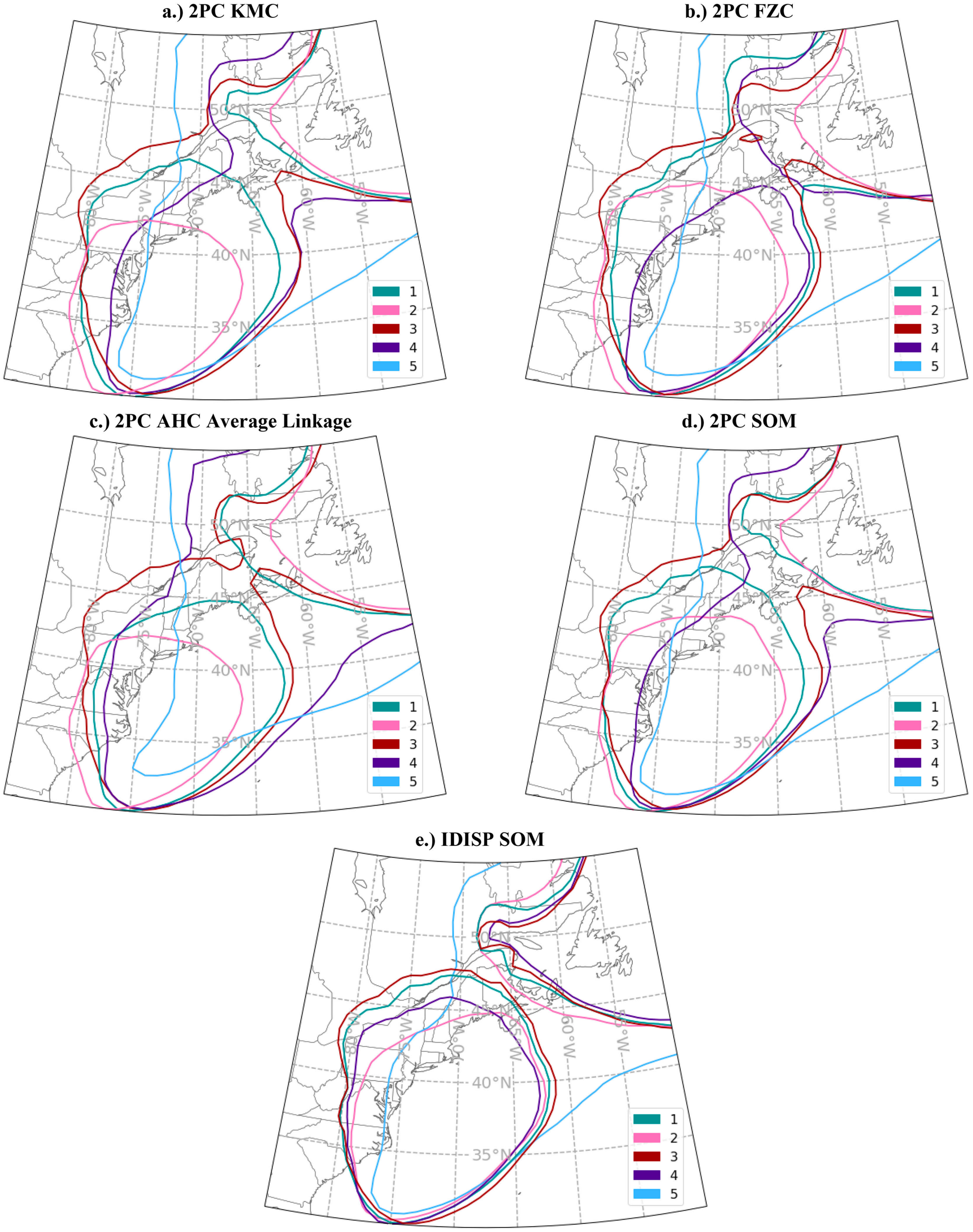
The statistics used to compare clustering approaches include: 1) the mean magnitude of displacement error of the SNA and mean magnitude of intensity/amount error of the SNA, 2) the silhouette score, 3) the weighted index distribution, 4) the mean adjusted Rand index (ARI) between clusters, and 5) the Brier skill score based on probabilistic cluster membership. All statistical analyses were evaluated over the appropriate 95% bootstrapped confidence interval.

Some statistics relied on the identification of the SNA, defined as the scenario whose mean location in the clustering space has the minimum Euclidean distance from the projection of the analysis onto the clustering space. The analysis "projection" is where the analysis would exist in the clustering space if it was a member of the ensemble. For the Euclidean space, the projection replicates the same process that was done with the ensemble members, with the analysis data flattened from a 2D array to a 1D vector. The projection of the IDISP space analysis is also the same process as was done with the ensemble members: finding the analysis pressure minimum and N/S and E/W position for MSLP, finding the N/S and E/W position of the center of mass for APCP, and weighting each relative to the respective IDISP space. For 2PC, adding a phase point to the EOF analysis changes the structure of the PC space itself. An additional projection formula is required to obtain the theoretical location that the analysis would occur if it were an ensemble member. The projection equation that was used is the same one that was formulated by von Storch (1999), with the formula described by and used in ZH19 and section a of the appendix.

a. Magnitude of intensity/amount and displacement error

Magnitude of displacement and intensity errors were determined differently for MSLP and APCP. The mean intensity error for MSLP was calculated as the average MSLP difference between each member of the SNA and the analysis in a 3° × 3° box centered on the analysis pressure minimum. A 3° × 3° square was used to get a more representative general intensity of the extratropical cyclone. The displacement error for MSLP was found by taking the Euclidean distance between the location of the cyclone pressure minimums of the analysis and each scenario. The intensity error for APCP was calculated by finding the 10 highest precipitation amounts from each member of each scenario, averaging them, and then finding the difference between this result and its analysis analog. The process is also repeated for the 90-member ensemble as well. The displacement error of APCP was found

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC


a.) 2PC KMC

b.) 2PC FZC

c.) 2PC AHC Average Linkage

d.) 2PC SOM

e.) IDISP SOM

FIG. 8. The 1012-hPa spaghetti plots representing scenarios for a sample of clustering approaches for the December 2012 case, including (a) 2PC KMC, (b) 2PC FZC, (c) 2PC AHC average linkage, (d) 2PC SOM, and (e) IDISP SOM.

using the Euclidean distance between the center of masses for the analysis and each scenario for the entire precipitation field within region 2.

b. Silhouette score

An important aspect of cluster evaluation is determining the value of cluster discreteness. Ideally, each cluster should be well separated with clear visual gaps between all groups. The silhouette score quantifies discreteness. It compares the mean intracluster distance, or average Euclidean distance between all pairs of phase points within a cluster, and the mean nearest-cluster distance, or average distance between all pairs of phase points of the two nearest clusters (Rousseeuw 1987). For example, a silhouette score of "0" indicates that there are

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

TABLE 1. Average size of the smallest and largest clusters for each clustering algorithm over all spaces.

Cluster	FZC	KMC	SOM	AHC-Ward's	AHC-Full	AHC-Avg
Largest	25/90 (28%)	28/90 (31%)	23/90 (26%)	31/90 (34%)	38/90 (42%)	52/90 (58%)
Smallest	11/90 (12%)	8/90 (9%)	12/90 (13%)	7/90 (8%)	4/90 (4%)	2/90 (2%)

no discernable boundaries between clusters, such as if clustering is attempted on a set of phase points uniformly distributed across a grid. The silhouette score would be 0 on a uniform grid no matter the choice of clustering algorithm. A silhouette score of "1," in contrast, occurs if all pairs of phase points within every cluster have a distance of zero between them, with clusters having a nonzero distance between each other. A formula for the silhouette score may be found in section b of the appendix.

c. *Weighted index distribution*

The weighted index distribution evaluates if the size of each cluster represents the sample probability of being the SNA and is similar to the reliability component of a Brier score. Ideally, the size of the cluster would represent the probability that it will be the SNA. For example, the largest cluster should be the most likely cluster to be the SNA.

To begin, each scenario was assigned an index, ranked from 1 to 5 from the greatest to the least number of members. Next, the index of the SNA was identified. For example, using 2PC FZC from the December 2012 case, the index values and the number of members of the five MSLP scenarios are: [1: 23, 2: 20, **3: 19**, 4: 17, 5: 11]. The SNA is marked in bold. After identifying the SNA for every case, the number of times each index is identified to be the SNA is divided by the total number of tested cases to get the sample probability of any index being the SNA. However, the cluster with most members will intrinsically be more likely to be closest to analysis, as the largest cluster will cover the greatest range of a reasonable distribution of potential analysis outcomes. The results were adjusted by first finding the average size of each index across all cases. For example, the average of all index 1 values would be the average size of the largest cluster for the entire dataset. After finding each index's average, the number of members in each scenario is divided by that value to generate the weighted index distribution. If, after weighting, a cluster has a higher probability to be the SNA than expected, the cluster would show evidence of forecast skill greater than what the number of members would suggest. In contrast, evidence of lower skill than the number of members would suggest that the largest cluster has a lower forecast skill than expected.

d. *Adjusted Rand index*

The Rand index (RI) is used to determine if two cluster algorithms behave differently from each other by evaluating the pairwise similarity between clusters (Rand 1971). To find the RI, the number of times that the algorithms agree or disagree on whether a pair of phase points belong to the same cluster or not were summed. The RI describes the ratio of the number of agreements over the sum of agreements and disagreements. A score of 1 indicates perfect similarity and a score of 0 indicates perfect dissimilarity. Perfect similarity means that two clustering approaches result in the exact same cluster memberships, whereas perfect dissimilarity means there is not a single instance in which the clustering approaches agree that a given pair of members belong to the same cluster. The adjusted Rand index (ARI) adjusts the RI to account for random chance (Hubert and Arabie 1985). A value lower than 0 can exist for ARI, in which case, agreement is less than what would be expected to exist at random. Otherwise, a score of 0 means any agreements would be coincidental (completely random), and a score of 1 indicates a perfect cluster agreement with no degree of randomness. Details on the formulas used to calculate the ARI may be found in Vinh et al. (2009), and they are provided in section c of the appendix as well.

e. *Brier skill score by cluster probability*

The Brier skill score (BSS) is a probabilistic skill metric used to compare an ensemble forecast to a reference forecast (Brier 1950). In this study BSS was calculated by assigning each cluster a probability equal to the number of members in each cluster divided by the total number of members (90). For verification, the cluster nearest analysis was compared to a "yes" "observation (probability of 1.0) while the other clusters were compared to a "no" observation (probability of 0.0). For the reference score, each cluster was given the same probability of occurrence (random selection of a cluster).

5. Results

All the clustering strategies were evaluated by first showing differences for two case examples and then evaluating them for the entire dataset.

a. *MSLP cluster comparisons*

Changing a clustering algorithm, without changing the clustering space, usually does not result in much variation across the scenarios aside from the shuffling of the cluster index number (e.g., Figs. 8a,b). However, an exception is that the AHC algorithm generally produces a larger first cluster and a smaller fifth cluster (Table 1). AHC average linkage in particular frequently places outliers into a separate cluster, with the average size of the smallest cluster being 2/90 (2%), with the behavior being less apparent in AHC full linkage (4/90, 4%). While Ward's and full linkage may still be useful when more uneven cluster sizes are desired, the outlier sensitivity produced by AHC average linkage can be problematic when trying to avoid having a majority of phase points belonging to a single cluster.

Of the clustering spaces tested, Euclidean space tends to place the majority, and sometimes nearly all, of the ensemble members into a single cluster. For instance, the average cluster size of the largest cluster for Euclidean clustering using AHC average

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

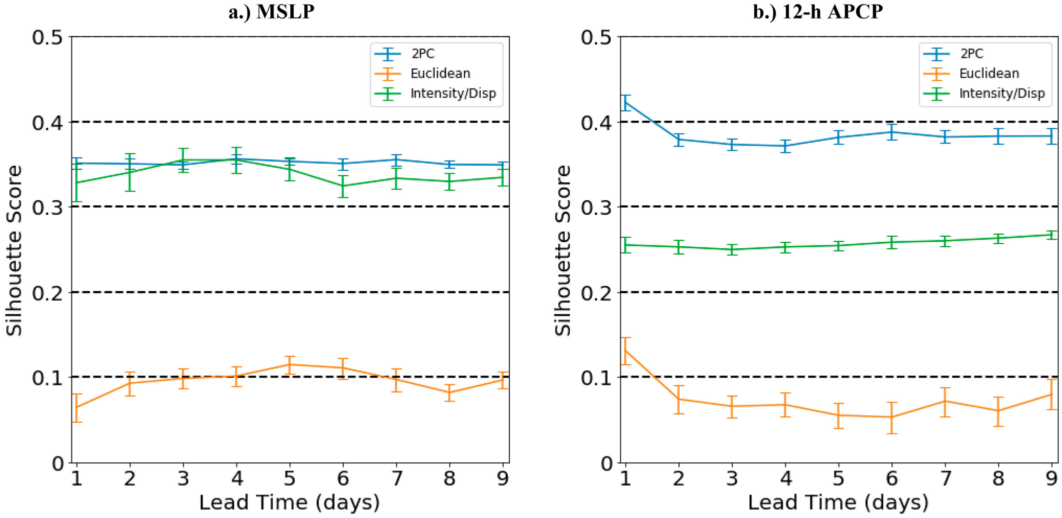


FIG. 9. Mean silhouette scores and 95% bootstrapped confidence intervals for spaces using FZC, for (a) MSLP and (b) 12-h APCP, by lead time day. Higher silhouette values indicate higher discreteness.

linkage is 66/90 (68%), higher than the average cluster size for 2PC of 46/90 (51%) and IDISP of 48/90 (53%). Euclidean space suffers from a lack of discreteness, with phase points within the Euclidean space being nearly equidistant from each other. The behavior is demonstrated when plotting the silhouette scores (Fig. 9). The rather low overall average silhouette score of 0.15 for Euclidean space compared to other spaces (e.g., 0.34 for 2PC) demonstrates the poor discreteness of Euclidean space making it a challenge to cluster in that space effectively.

The average ARI for MSLP is 0.51 ± 0.02 and 0.31 ± 0.02 between clustering algorithms and clustering spaces, respectively. The lower average ARI value for clustering spaces indicates that there are larger differences in cluster assignment when changing the clustering space than when changing the clustering algorithm. Therefore, for the algorithms and spaces tested in this study, the choice of clustering space is a more important factor than choice of clustering algorithm in determining how model ensembles are clustered.

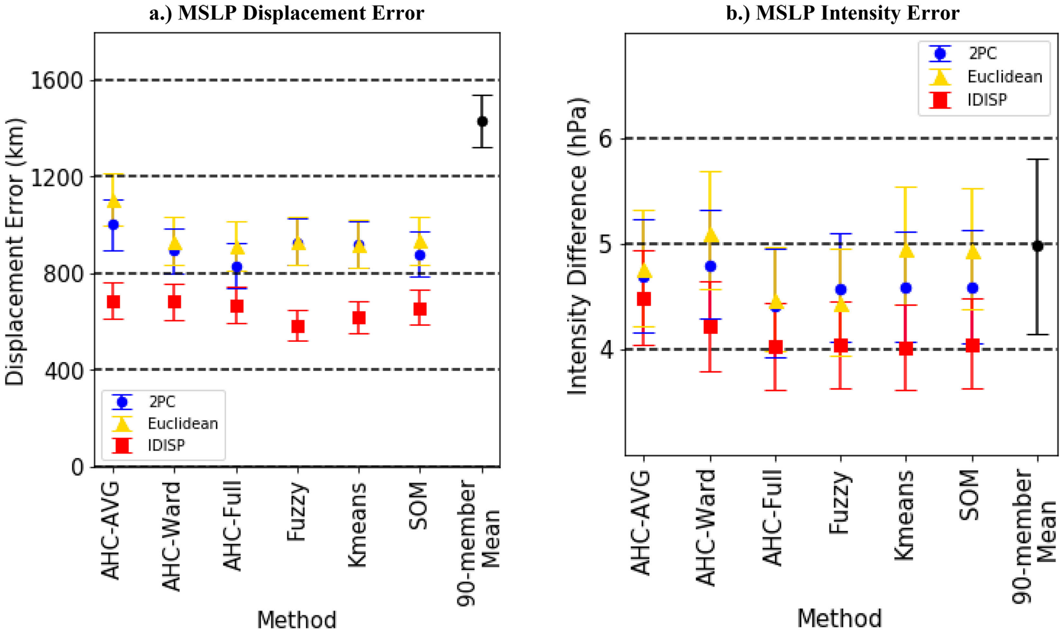


FIG. 10. MSLP long lead time (days 7–9) mean (a) magnitude of displacement and (b) magnitude of intensity errors for the SNA of all spaces.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

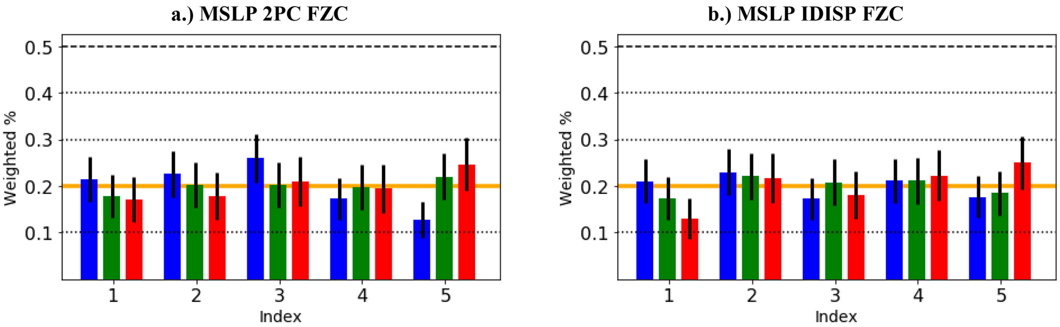


FIG. 11. MSLP weighted index distribution for FZC for (a) 2PC space and (b) IDISP space. Each index is separated, from left to right, into short (blue), medium (green), and long (red) lead times, along with 95% confidence intervals (black line), and a 20% threshold (orange line). Values below the solid red line indicate that the probability of the index containing the SNA is less likely than random, and above the solid orange line, it is more likely than random.

The ARI did not vary significantly across lead times (Kiel 2021). For example, the average ARI between all possible pairs of clustering algorithms for 2PC space at short, medium, and long lead times is 0.49 ± 0.03, 0.53 ± 0.03, and 0.52 ± 0.02, respectively. Furthermore, the ARI has a narrow distribution across clustering algorithms. ARIs generally remain consistent across clustering algorithms when fixing clustering spaces, and clustering spaces when fixing clustering algorithms. There are

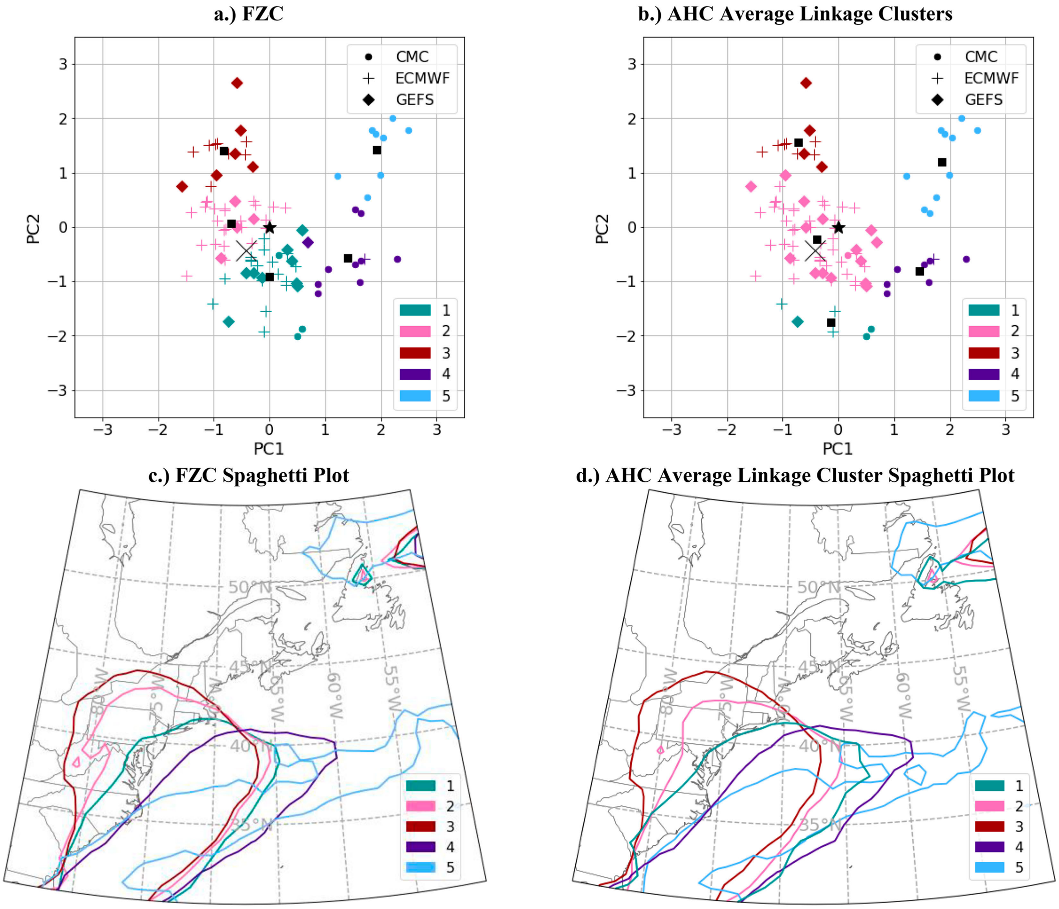


FIG. 12. The 2PC clustering (see label bar shading) of 12-h APCP for the December 2012 case (same VT as Fig. 1) when using (a) FZC or (b) AHC clustering using average linkage. (c),(d) Spaghetti plot for 4 mm 12-h APCP for the five scenarios.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

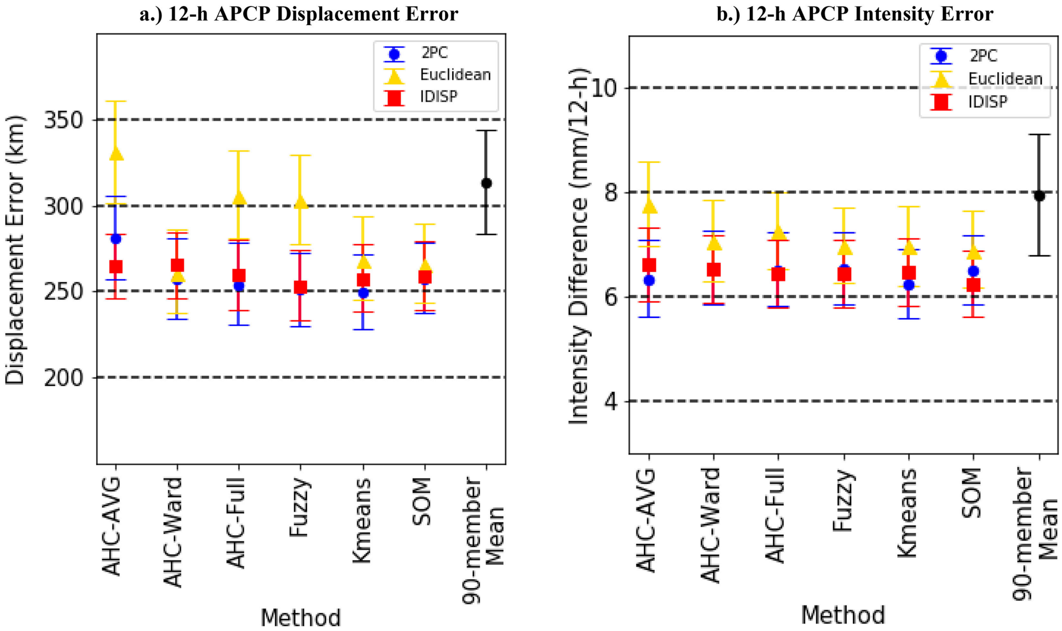


FIG. 13. The 12-h APCP long lead time (days 7–9) mean (a) magnitude of displacement and (b) magnitude of intensity errors for the SNA of all spaces.

two instances where ARI shows deviant behaviors. First, the ARI between KMC and FZC are much greater than the average between all clustering algorithms, with an average across all lead times of 0.72 ± 0.03. KMC and FZC are computationally similar, so the high degree of similarity between the two clustering algorithms is expected. Second, the average ARI of IDISP against all other clustering spaces for FZC is 0.18 ± 0.02, lower than the overall average between clustering spaces.

Both displacement and intensity errors increase with lead time. The average displacement error magnitude for the SNA of all of the 2PC algorithms is 294, 565, and 879 km at short, medium, and long lead times, respectively, while the average magnitude of intensity error is 1.7, 3.5, and 4.5 hPa. The relative behavior between each clustering strategy does not change with lead time, and long lead times are used to demonstrate them (Fig. 10). Between clustering algorithms, displacement and intensity errors lie within the margin of error of each other within any fixed clustering space. Between clustering spaces, IDISP gives lower displacement error magnitudes than the other spaces, significant at the 95% CI (Fig. 10a). Specifically, IDISP space gives average displacement errors of 195, 391, and 649 km at short, medium, and long lead times, respectively. A reason for the lower displacement errors in IDISP is its high sensitivity to the location of the pressure minimum. For example, in the December 2012 case, there are two cyclones in the analysis (Fig. 1b), but the location of each displacement component depends on which cyclone has the lower pressure. The split distribution ultimately influences how ensembles are clustered (Fig. 3). High sensitivity is also observed in closed low versus open trough situations (not shown). This sensitivity may contribute to why the SNA has the lowest displacement errors for IDISP space.

The weighted index distribution reveals that the number of members within a cluster does not necessarily correspond with the probability of that scenario being the SNA. Since the number of clusters is fixed to five, 20% represents the expected

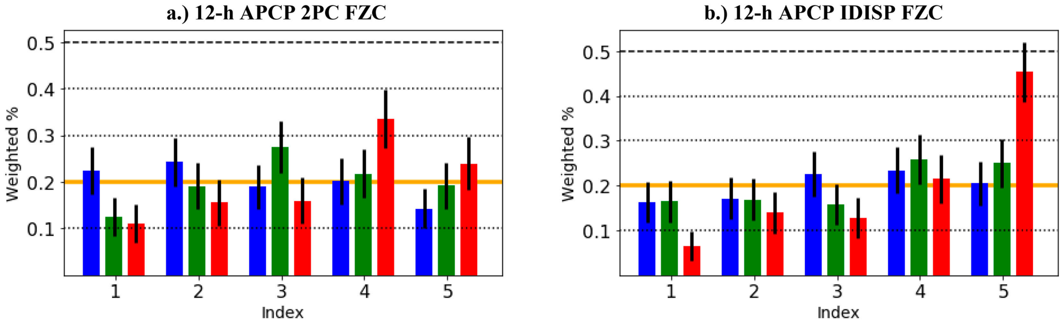


FIG. 14. As in Fig. 11, but for 12-h APCP instead of MSLP.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

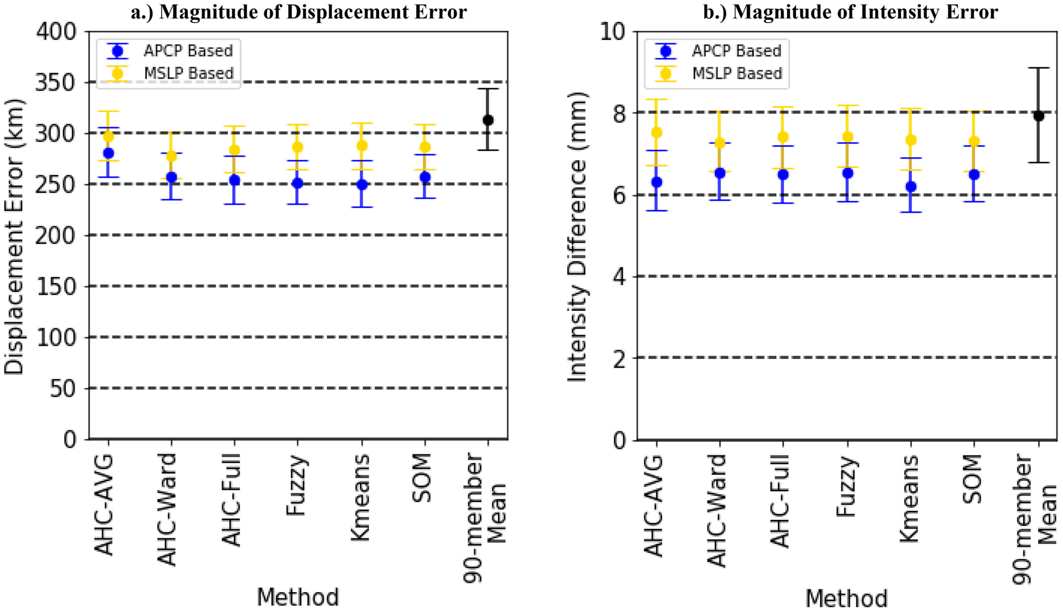


FIG. 15. Comparison of mean SNA results for 12-h APCP when using 12-h APCP-based clusters (blue) vs MSLP-based clusters (yellow), plotted with 95% confidence intervals, along with the 90-member ensemble mean (black). Results are for (a) magnitude of displacement error and (b) magnitude of intensity error.

probability of selecting a cluster based solely on the cluster size. If the index probability is >20% (<20%), then the index is more (less) likely to be the SNA than what the cluster size would suggest. For both 2PC and IDISP FZC of the smallest cluster (index 5) is <20% for short lead times (Figs. 11a,b). For example, the smallest cluster of the December 2012 case (Fig. 5) would suggest a probability of being the SNA in less than the expected probability of 11/90 (12%). One possible explanation for this behavior is that the smallest cluster is most likely to contain a set of outliers. The smallest cluster (index 5) at short lead times is <20% at a 95% confidence interval for 2PC for AHC full linkage and KMC; for Euclidean for AHC Ward's linkage, AHC full linkage, and KMC; and for IDISP for KMC and SOM. Another behavior, observed in IDISP space long lead times, is that the largest cluster (index 1) is <20% (Fig. 11b). The largest cluster (index 1) at long lead times is less likely to be the SNA than what is suggested by its membership count. The weighted index distribution for the largest cluster (index 1) is <20% for long lead times for all clustering approaches, except in Euclidean for FZC, KMC, and SOM, and in IDISP for AHC average linkage and FZC. As lead times lengthen, the sample probability of the largest cluster being the SNA decreases, while the sample probability of the smallest cluster being the SNA increases. Forecasters should be cautious about interpreting the number of members within a cluster as a probability of occurrence, as these percentages often do not correspond with the sample probability unless at medium lead times.

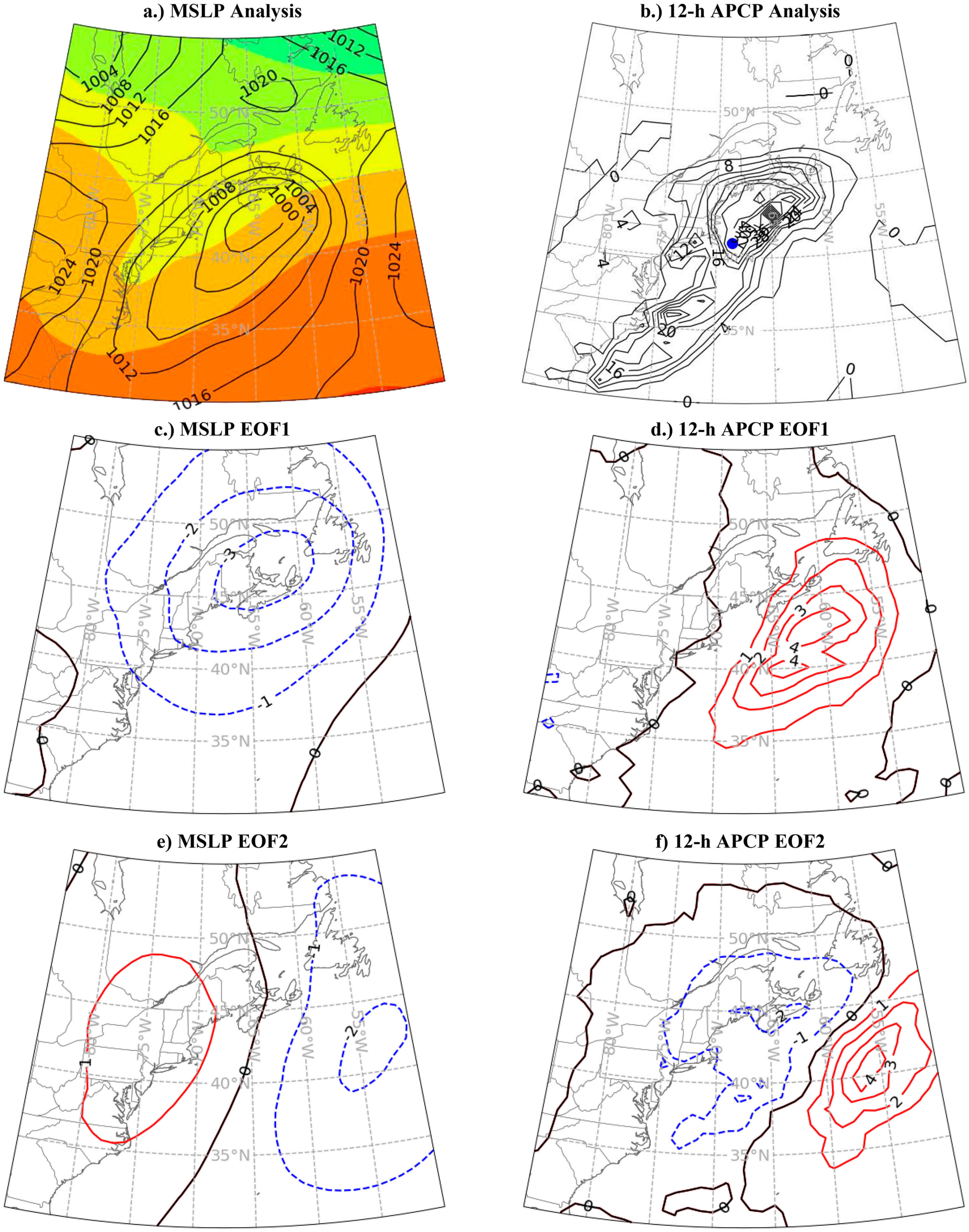
b. *12-h APCP cluster comparisons*

As with MSLP, 12-h APCP AHC tends to produce a larger first cluster than the other algorithms (Figs. 12a,b). The average size of the largest cluster for AHC in 2PC are 48/90 (60%) for average linkage, 32/90 (36%) for Ward's linkage, and 38/90 (39%) for full linkage. In contrast, the average size of the largest cluster in 2PC FZC is 26/90 (30%). Using Euclidean space generates even larger clusters, with average size of the largest cluster near or exceeding one-half of ensemble members for AHC average linkage (75/90, 84%), AHC Ward's linkage (35/90, 35%), AHC full linkage, (51/90, 57%), and FZC (44/90, 49%). These large cluster sizes are a result of the combined effects of a very low overall discreteness lower than that of MSLP (0.12, Fig. 9b), and the tendency for AHC to cluster outliers by themselves.

TABLE 2. Percent of cases that 12-h APCP-based clusters instead of cross-field clusters give a lower displacement or intensity error for 2PC space, along with 95% bootstrapped confidence intervals. The values are the averaged result of all clustering algorithms tested in 2PC space.

	Magnitude of displacement error	Magnitude of intensity error
Short lead time	62% ± 6%	57% ± 6%
Medium lead time	59% ± 6%	64% ± 6%
Long lead time	57% ± 6%	62% ± 6%

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC



a.) MSLP Analysis

b.) 12-h APCP Analysis

c.) MSLP EOF1

d.) 12-h APCP EOF1

e) MSLP EOF2

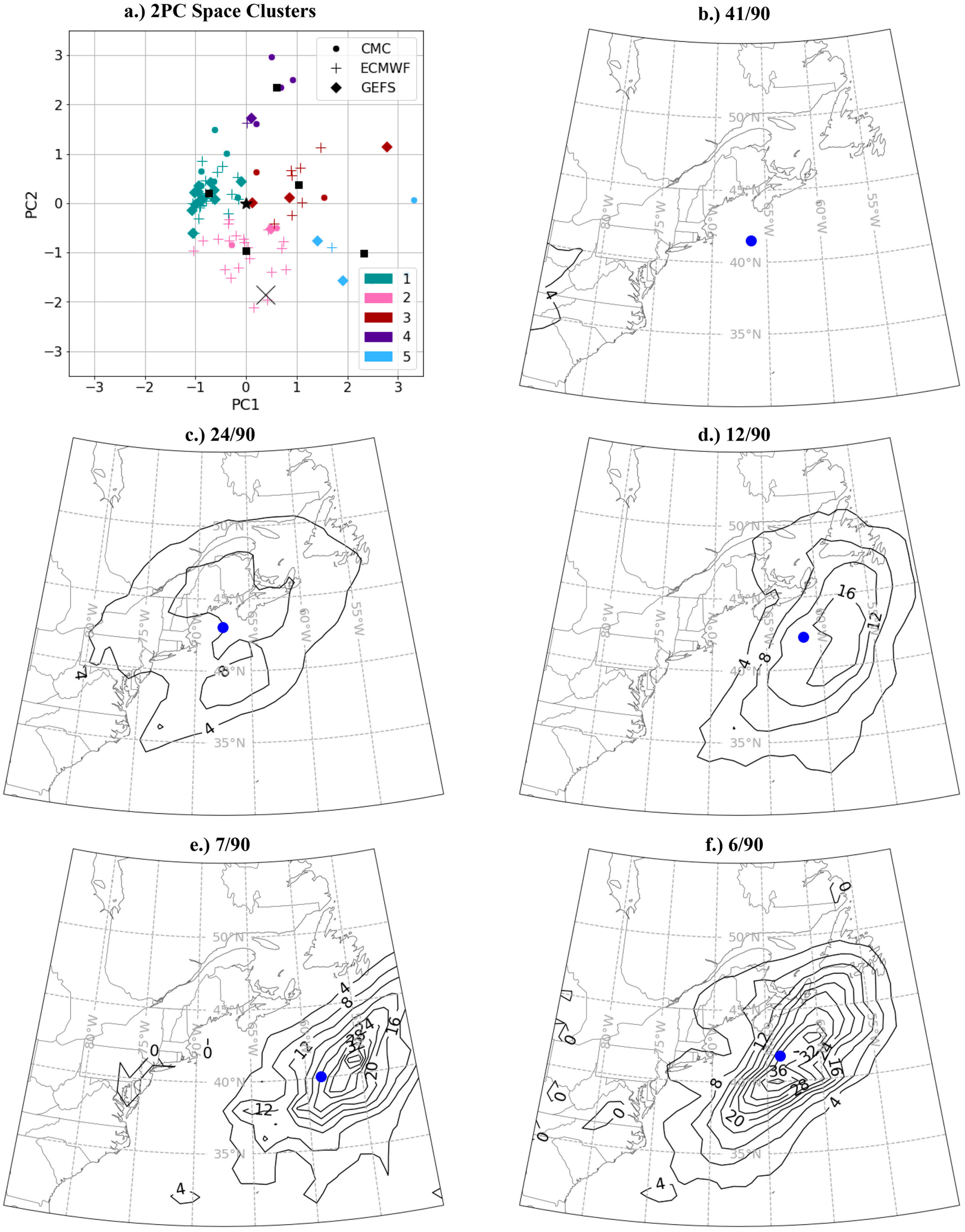
f) 12-h APCP EOF2

FIG. 16. (a) MSLP analysis and 500-hPa height and (b) 12-h APCP analysis (solid every 4 mm) along with center of mass (blue dot); (c) MSLP and (d) 12-h APCP EOF1s; and (e) MSLP and (f) 12-h APCP EOF2s, for VT at 0000 UTC 24 Nov 2011 and IT at 0000 UTC 16 Nov 2011. MSLP analysis is plotted as in Fig. 1b, APCP analysis is plotted as in Fig. 2b, and EOFs are plotted as in Figs. 1c and 1d.

Effectively, AHC clusters in Euclidean space act to detect outliers instead of evenly clustering the ensemble members.

The ARI results for 12-h APCP are like MSLP, with ARIs once again not varying across lead times. In addition, averages remain within the margin of uncertainty of each other across clustering algorithms within clustering spaces, and clustering spaces when testing clustering algorithms (not shown). Average ARIs between clustering algorithms (e.g., 0.53 ± 0.02 for 2PC) remain higher than average ARI between clustering spaces (e.g., FZC is 0.33 ± 0.02). As with MSLP, KMC and

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC


a.) 2PC Space Clusters

b.) 41/90

c.) 24/90

d.) 12/90

e.) 7/90

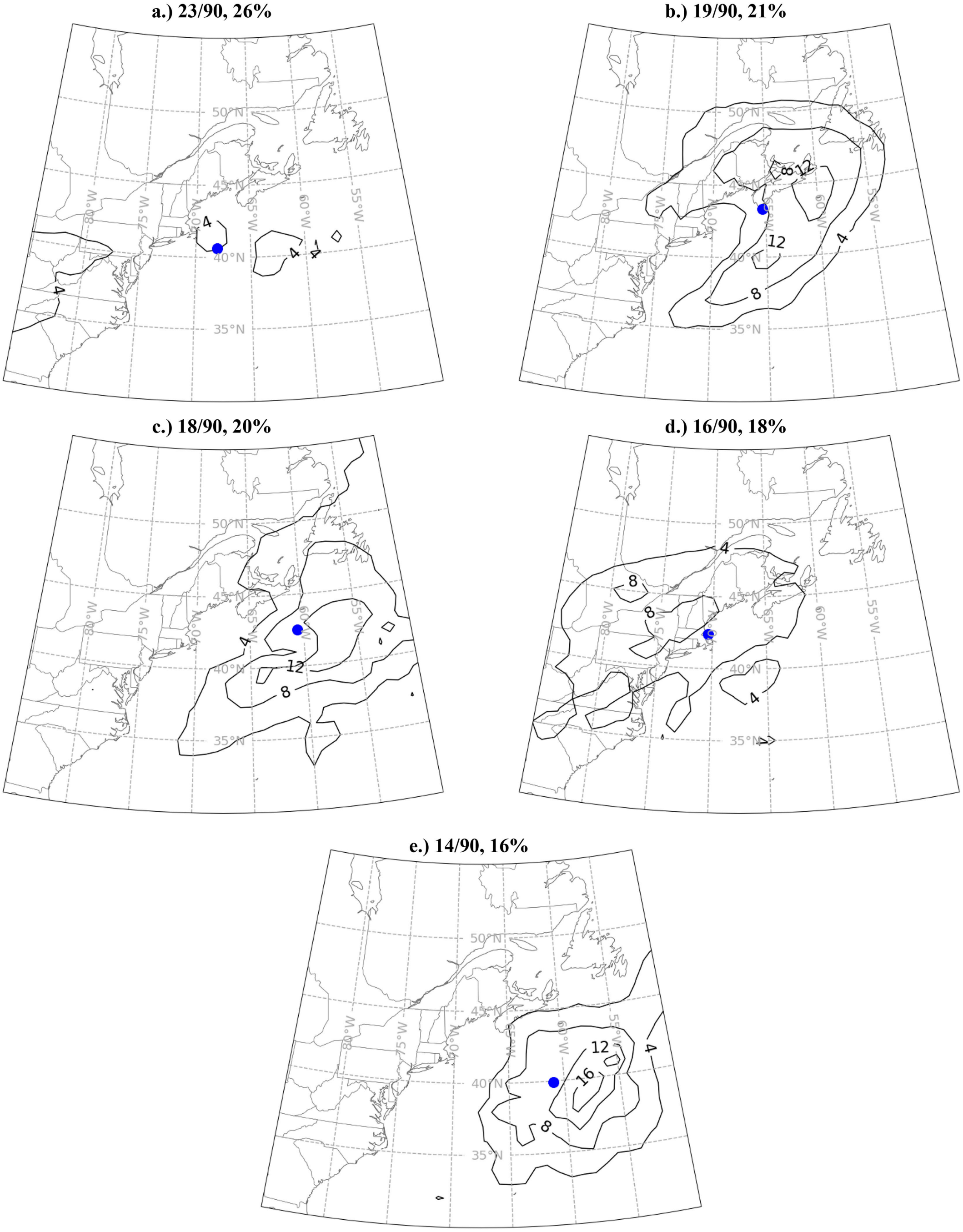
f.) 6/90

FIG. 17. (a) The 2PC FZC of 12-h APCP in the 12-h APCP clustering spaces. (b)–(f) Resulting 12-h rainfall amounts for each scenario for the November 2011 case, as in Fig. 16b.

FZC are the most similar (e.g., 0.67 ± 0.03 in 2PC space). IDISP remains the most dissimilar from other clustering spaces, with the average ARI of IDISP against all other clustering spaces for FZC of 0.18 ± 0.02.

As with MSLP, the 12-h APCP SNA displacement and amount errors increase with lead time as well. The displacement errors on average are 71, 157, and 268 km for short, medium, and long lead times, respectively, while the amount errors are 3.70, 5.17, and 6.78 mm (12 h)$^{-1}$. As with MSLP, the displacement and amount errors increase proportionately by a multiplicative factor with lead time, so the long lead time displacement (Fig. 13a) and amount (Fig. 13b) also represent an amplified version of short and medium lead time results. The SNA in the Euclidean space for AHC average and full linkage

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC


a.) 23/90, 26%

b.) 19/90, 21%

c.) 18/90, 20%

d.) 16/90, 18%

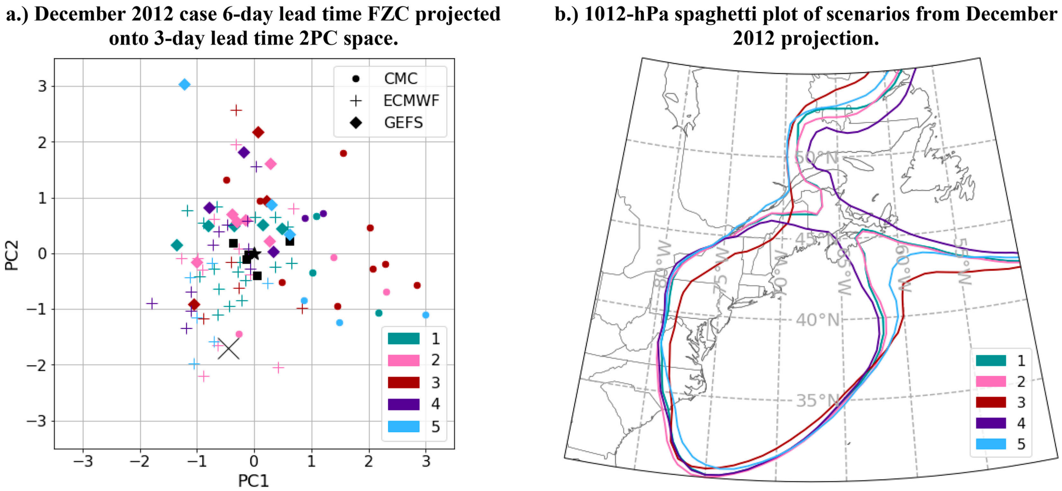
e.) 14/90, 16%

FIG. 18. The 12-h APCP scenarios showing the precipitation amount (every 4 mm) when using 2PC MSLP FZC from Fig. 16a, plotted as in Fig. 15b.

clustering are close to the 90-member ensemble mean; the result of a large first cluster.

Finally, the 12-h APCP weighted index distributions are compared for 2PC and IDISP FZC (Fig. 14). As with MSLP, the average weighted index distribution value of the smallest cluster (index 5) is <20% for short lead times, meaning that the smallest cluster is less likely to be the SNA than what is suggested by its membership count in this time frame. An overestimated probability of the smallest cluster at short lead times also occurs with 95% confidence in 2PC for KMC, in Euclidean space for AHC Ward's linkage and SOM, and in IDISP space for AHC average linkage and KMC. Likewise, for both medium and long lead times, the average weighted index distribution of the largest cluster (index 1) is <20%. The probability that the largest scenario for medium and long lead times is the SNA is less than what the number of

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC


a.) December 2012 case 6-day lead time FZC projected onto 3-day lead time 2PC space.

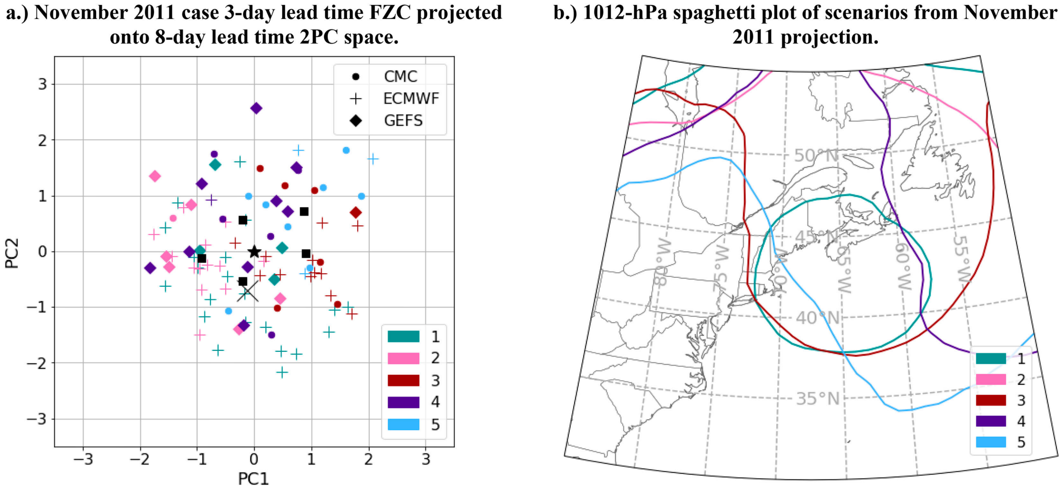
b.) 1012-hPa spaghetti plot of scenarios from December 2012 projection.

FIG. 19. (a) Cluster plot for the December 2012 case MSLP scenarios (3-day lead time) when using 2PC FZC from a 6-day lead time. (b) The 1012-hPa spaghetti plot from mean of scenarios generated from (a).

members suggest. At long lead times, the weighted index distribution is <20% for all clustering approaches for the largest cluster. For medium lead times, all algorithms in 2PC space, FZC of Euclidean, and AHC Ward's linkage, AHC full linkage, and KMC for IDISP are also <20% for the first index. Compared to MSLP, 12-h APCP is skewed more toward a lower probability of occurrence for the largest cluster and a higher probability of occurrence for the smallest cluster. Otherwise, like MSLP, the probability of the largest cluster being the SNA decreases with increasing lead time, and the probability of the smallest cluster being the SNA increases with increasing lead time. Overall, weighted index distribution results demonstrate that a forecaster should be particularly careful interpreting probabilities of 12-h APCP clusters as well, especially at long lead times, as the largest cluster is less likely to be the SNA and the smallest cluster is more likely to be the SNA than suggested by the cluster size.

c. *Cross-field clustering comparisons*

Using clusters obtained from MSLP to generate 12-h APCP clusters results in a reduction in forecast skill. Across the board, the magnitude of displacement and rate errors increase compared to the original 12-h APCP clusters (Fig. 15). In more than 50% of the cases, the displacement and magnitude errors were larger when using cross-field clustering compared to same-field clustering for 12-h APCP forecasts (Table 2), with the results consistent across clustering approaches and well within the 95% statistical significance thresholds of each other. The use of cross-field clustering results in less distinct precipitation scenarios than if precipitation itself were to be used to cluster. A second case is used to better understand the issues with cross-field clustering, with an IT of 0000 UTC 16 November 2011 and a VT of 0000 UTC 24 November 2011 (Fig. 16, 8-day lead time). The EOF1 and EOF2 of MSLP and 12-h APCP represent an intensity


a.) November 2011 case 3-day lead time FZC projected onto 8-day lead time 2PC space.

b.) 1012-hPa spaghetti plot of scenarios from November 2011 projection.

FIG. 20. (a) Cluster plot for November 2011 case MSLP scenarios (8-day lead time) when using 2PC FZC from a 3-day lead time. (b) The 1012-hPa spaghetti plot from mean of scenarios generated from (a).

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

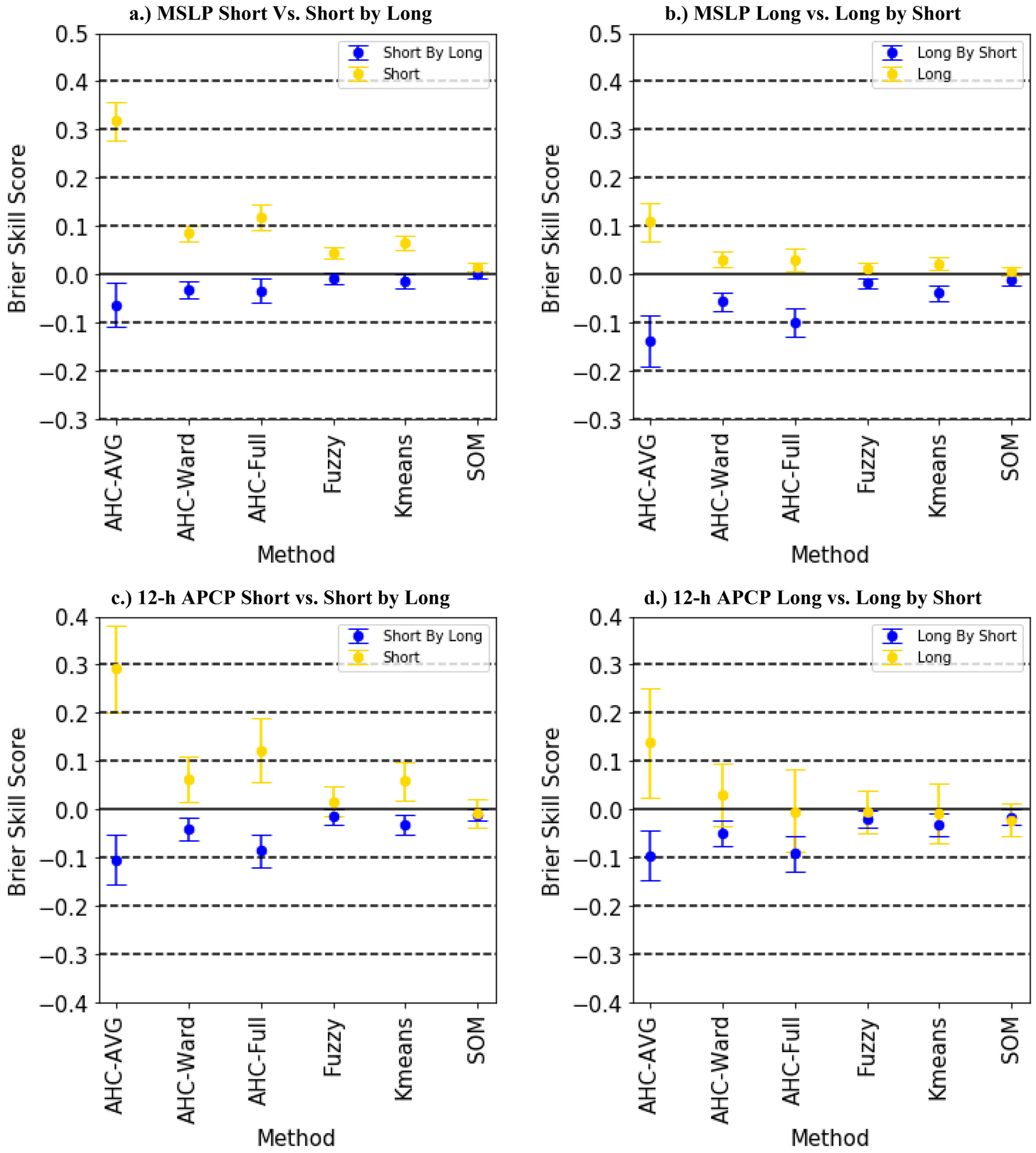


FIG. 21. Comparison of mean Brier skill score values when using alternate lead times (blue) as opposed to not using alternative lead times (yellow), plotted using 95% confidence intervals. (a) Long lead time clusters to cluster short lead times vs using the short lead time clusters themselves, for MSLP. (b) Short lead time clusters to cluster a long lead time vs using the long lead time clusters themselves, for MSLP. (c) As in (a), but for 12-h APCP. (d) As in (b), but for 12-h APCP.

shift and an east–west storm track shift, respectively. Finally, the scenarios generated from 2PC FZC of 12-h APCP (Fig. 17) are compared to the scenarios generated from cross-field clustering (Fig. 18). The largest 12-h APCP cluster contains nearly twice as many members (41/90, 46%) as the largest MSLP-based cluster (23/90, 26%). Both of these scenarios contain little precipitation. Fewer model members exist in the "no storm" scenario for MSLP than 12-h APCP, suggesting that other MSLP-based scenarios include ensemble members with no identifiable surface cyclone or no precipitation; clustering APCP using MSLP alone is insufficient to represent precipitation; any associated MSLP features must not have been sufficiently discrete for the clustering algorithms to pick them up. A diverse set of MSLP patterns, scattered across the MSLP 2PC space, can produce no precipitation and be near each other in the 12-h APCP space.

d. *Alternative lead time clustering comparisons*

To test the impact of using the same clusters at one lead time for alternative lead times, the clusters derived at later lead times

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

TABLE 3. The percent of times using a different lead time to cluster gives a SNA with a lower MSLP displacement error magnitude or magnitude of intensity error than using the original lead times to cluster, along with 95% bootstrapped confidence intervals. The bold font is used to highlight the different lead times.

	Magnitude of displacement error	Magnitude of intensity error
Using **medium** lead time to cluster short lead time	48.0% ± 4.5%	49.0% ± 4.7%
Using **long** lead time to cluster short lead time	44.2% ± 4.6%	46.8% ± 4.5%
Using **short** lead time to cluster medium lead time	34.8% ± 4.7%	42.8% ± 4.8%
Using **long** lead time to cluster medium lead time	35.4% ± 4.6%	41.9% ± 4.8%
Using **short** lead time to cluster long lead time	39.3% ± 4.5%	44.3% ± 4.6%
Using **medium** lead time to cluster long lead time	41.1% ± 4.7%	43.8% ± 4.7%

are used to cluster forecasts at earlier lead times (e.g., day 6 to cluster day 3) and vice versa (e.g., day 3 to cluster day 6). All possible pairs of lead time sets (short, medium, long) were tested.

An example from the December 2012 case visualizes using the day-6 (later) lead time clusters to cluster the day-3 forecast for 2PC FZC (Fig. 19). An example from the November 2011 case uses the day-3 (earlier) lead time clusters to cluster the day-8 lead time clusters (Fig. 20). The phase space clustering plots of alternative lead time clusters projected onto the lead time being tested for are noisy and demonstrate little coherence (Figs. 19a and 20a). When applying long lead times on short lead times, only scenario three (maroon) suggests there may be a small factor beyond noise, given the slight difference in positioning of the 1012-hPa isobar for this scenario (Fig. 19b). Using the short lead time on long lead time shows more variation between clusters, but it is difficult to know if this variation is attributable to a factor other than model ensemble spread (Fig. 20b). Averaged Brier skill scores across cluster spaces consistently show that not using an alternative lead time provides greater skill than using the alternative lead times to cluster (Fig. 21). In addition, using the original clusters more likely gives a SNA with a lower displacement and lower intensity error for MSLP than using alternative lead time clusters (Table 3). The results for 12-h APCP are quantitatively similar (not shown).

6. Summary and future work

This study evaluated several different clustering strategies for a 90-member ensemble by providing results from two case studies and evaluating the statistical differences for 180 cold season extratropical cyclones along the U.S. East Coast. This study demonstrates how each clustering strategy performed and if significant differences exist in the resulting scenarios. Key results are presented in Table 4.

For MSLP, the IDISP clustering space on average gives a more skillful scenario nearest to the analysis than any of the other clustering approaches. The SNA generated from IDISP gives lower displacement errors than the other clustering spaces. However, there are two caveats to clustering in IDISP space. First, when the cyclone center is located at the edge of a region, by definition, variance information is lost in the direction perpendicular to the region edge. Second, if multiple cyclones are present, IDISP may "jump" between pressure minimums, and clustering is sensitive to this behavior. In general, IDISP is a strong candidate for use in clustering when a single cyclone is completely enclosed within the tested region. The 2PC space clustering presents a viable alternative to IDISP when IDISP's caveats are problematic. The SNA still performs better than the 90-member ensemble mean, as shown in previous studies (e.g., Zheng et al. 2019).

For 12-h APCP, there is no clustering approach which clearly outperformed the rest. The version of IDISP tested for 12-h APCP did not produce as strong of a result as for MSLP, with the SNA performing marginally better than the 90-member ensemble mean and similarly to the 2PC space.

Some clustering algorithms in Euclidean space are not recommended to be used. Euclidean space is prone to producing a single large cluster. The inherent challenges related to its high dimensionality, along with the dataset being too noisy, inhibit the creation of evenly sized clusters. Dimensionality reduction should be performed (e.g., 2PC, IDISP), before applying a clustering algorithm to MSLP or APCP.

Forecasters must be careful when assuming that the number of members within each cluster is equivalent to the probability of the scenario being the SNA. There are situations where the

TABLE 4. A summary of key results.

MSLP clusters	Intensity displacement space gave the lowest magnitude of displacement errors For some approaches the smallest cluster at short lead times, and the largest cluster at long lead times, were less likely to be the SNA than the number of members suggested
12-h APCP clusters	The low discreteness of Euclidean space led to high sensitivity to outliers, particularly for average/full linkage, making these approaches not recommended for use with 12-h APCP The largest cluster at long lead times was much less likely to be the SNA than what was suggested by the number of members
12-h APCP by MSLP	A diversity of MSLP patterns can all lead to a little/no precipitation pattern, limiting the usefulness of using MSLP to cluster 12-h APCP
Alternative lead times	Cluster membership is inconsistent from lead time to lead time and less skillful than not using alternative lead times

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

number of members within a cluster was not representative of that cluster's probability of being the most accurate. The smallest cluster at short lead times and the largest cluster at long lead times are both less likely to represent the SNA than the number of members would suggest. The 12-h APCP especially shows difficulty representing the suggested probability of its largest cluster at long lead times.

As for clustering algorithms, only AHC average or full linkage show differences in behavior from other clustering algorithms. AHC average and full linkages tend to leave outliers by themselves, sometimes considering just a single ensemble member as a "cluster." The weighted index distribution results are helpful in illuminating when outlier sensitive algorithms may be of interest to a forecaster. At short lead times, scenarios consisting only of outliers might not be useful to the forecaster, as the smallest scenarios are less likely to be the SNA than what the number of members would suggest. At long lead times, when larger clusters underperform and smaller clusters overperform, the outlier-sensitive AHC average and full linkage algorithms may be more useful.

Using cross-field and alternative lead time clustering does not improve results over original clusters. There may be situations where the use of MSLP to cluster 12-h APCP could be useful to a forecaster. However, a forecaster must be aware that ensemble members with and without precipitation will average with each other and reduce scenario variability. The strategy might be most useful when only a few null precipitation model runs exist in the ensemble, but further testing is needed to verify if this is the case. The alternative lead time strategy appears to be unviable due to inconsistency in cluster membership from lead time to lead time; however, there may be some synoptic patterns which may be more conducive to alternative lead time clustering, which was beyond the scope of this analysis and could be the focus of future studies. Finally, additional work is needed to see if a mass field may be clustered using a cross-field method (e.g., 12-h APCP clusters may be used to cluster MSLP). Further work is needed to understand whether there are better clustering approaches that exist. Use of a supervised machine learning technique to attempt the identification of the SNA would indicate if any predictive ability of scenarios exists beyond the number of members within each cluster. Sensitivity to changes in the horizontal resolution and the domain size are important factors but were not considered in this study so as to isolate the variables of focus. Other important fields, such as wind, potential temperature, and 500-hPa heights, remain untested. The clustering strategies used in this study remain far from exhaustive. An example of a viable clustering space may be to use the first PC of several atmospheric parameters, such as clustering PC1 of MSLP versus PC1 of APCP to cluster both MSLP and APCP. This study only tested clustering algorithms where 100% of ensemble members contributed to a cluster, including outliers. A method where clusters could be formed without including outliers and the associated inevitable skew may give better results (e.g., density-based spatial clustering with noise). While clustering generally shows promise as a method to simplify analysis of model ensembles, forecasters should remain aware that clustering is unlikely to resolve or improve a poorly distributed set of ensembles, such as those at long lead times that appear to give a chaotic set of solutions. Effectiveness of clusters likely varies substantially from scenario to scenario.

Acknowledgments. This research was supported by grants from NOAA CSTAR (NA17NWS4680004) and NOAA JTTI (NA20OAR4590361). The authors thank Minghua Zheng for her guidance, assistance, prior work, and for providing the MSLP ensemble data. The authors would also like to thank Dr. Edmund K. M. Chang and Dr. Sultan Hameed for their ideas, insight, and assistance with this work. We thank the three reviewers who helped improve previous versions of this manuscript.

Data availability statement. All the raw ensemble data used in this study may be retrieved from either the TIGGE archive or the Copernicus Climate Data Store. All MSLP and 12-h APCP data for all cyclone events are stored online (https://drive.google.com/drive/folders/1PEm3_tvwW83jNvO_q28icELx4GVWsR_J?usp=sharing) and a hard disk drive at Stony Brook University. Please contact the corresponding author for access.

APPENDIX

Selected Equations

a. Principal component projection

The von Storch (1999) projection equation is defined as

$$a = \frac{\mathrm{cov}(\mathbf{A}'\mathbf{E}')}{\mathrm{var}(\mathbf{E}')}, \qquad \text{(A1)}$$

where a is the projection of the analysis onto the PC space, $\mathbf{A}'$ is the standardized analysis relative to the 90-member ensemble mean, and E_i is the ith leading EOF corresponding to the PC desired.

b. Silhouette score

The silhouette score is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max(a_i, b_i)}, \qquad \text{(A2)}$$

where n is the number of clusters, a_i is the mean intracluster distance between all phase points in the cluster, and b_i is the mean distance to the nearest cluster centroid for each phase point, for the ith cluster.

c. Adjusted Rand index

Before formulating the ARI, it is better to first understand the Rand index (RI) (Rand 1971). Following Rand's original formula, let S = a set of elements [1, 2, …, i], clustered by method A into groups [A_1, A_2, …, A_m] and by method B into groups [B_1, B_2, …, B_n]. For example, S might be [1, 2, 3, 4, 5], group A_1 may be [1, 2], group A_2 may be [3, 4, 5], group B_1 may be [4, 5], and group B_2 may be [1, 2, 3]. Next, define

1) w: The number of pairs of elements in S that are in the same group in A and in the same group in B.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

2) x: The number of pairs of elements in S that are in different groups in A and in different groups in B.
3) y: The number of pairs of elements in S that are in the same group in A and in different groups in B.
4) z: The number of pairs of elements in S that are in different groups in A in in the same group in B.

For example, the pair (1, 2) counts as an instance of "w," since both algorithms agree that the elements should belong to the same group. The sum $w + x$ is the number of agreements, and $y + z$ is the number of disagreements. The RI is (number of agreements)/(number of agreements + number of disagreements) or

$$\mathrm{RI} = \frac{w + x}{w + x + y + z}. \tag{A3}$$

The ARI adjusts the RI to account for random chance (Hubert and Arabie 1985). Following with Vinh et al. (2009), the process of calculating ARI index is best understood by construction of a contingency table:

$$\begin{matrix} r_{11} & \cdots & r_{1n} & x_1 \\ \vdots & \ddots & \vdots & \vdots \\ r_{m1} & \cdots & r_{mn} & x_m \\ y_1 & \cdots & y_n & \end{matrix} ,$$

where

$$r_{ij} = |A_i \cap B_j|, \tag{A4}$$

$$y_1 \ldots y_n = \sum_{i=0}^{m} r_{i1} \ldots \sum_{i=0}^{m} r_{in}, \tag{A5}$$

$$x_1 \ldots x_n = \sum_{j=0}^{n} r_{1j} \ldots \sum_{j=0}^{n} r_{mj}. \tag{A6}$$

Defining a generic term p, where the function $f(p) = 0.5 \times p(p - 1)$, the ARI is defined as

$$\mathrm{ARI} = \frac{\sum_{ij} f(r_{ij}) - \left[\sum_i f(x_i) \sum_j f(y_j)\right] \Big/ f(n)}{0.5 \times \left[\sum_i f(x_i) + \sum_j f(y_j)\right] - \left[\sum_i f(x_i) \sum_j f(y_j)\right] \Big/ f(n)}, \tag{A7}$$

where n is the total number of r objects.

REFERENCES

Arthur, D., and S. Vassilvitskii, 2007: K-means++: The advantages of careful seeding. *Proc. 18th Annual ACM-SIAM Symp. on Discrete Algorithms*, New Orleans, LA, SIAM, 1027–1035.

Blake, E. S., T. B. Kimberlain, R. J. Berg, J. P. Cangialosi, and J. L. Bevin II, 2013: Tropical cyclone report: Hurricane Sandy (AL182012). NHC Tech. Rep., 157 pp., https://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf.

Booth, J. F., H. E. Hieder, 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.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bytheway, J. L., and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long lived convective precipitation in the central U.S. *J. Adv. Model. Earth Sci.*, **7**, 1248–1264, https://doi.org/10.1002/2015MS000497.

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

Crossett, C. C., A. K. Betts, L.-A. L. Dupigny-Giroux, and A. Bomblies, 2020: Evaluation of daily precipitation from the ERA5 global reanalysis against GHCN observations in the northeastern United States. *Climate*, **8**, 148, https://doi.org/10.3390/cli8120148.

Defays, D., 1977: An efficient algorithm for a complete link method. *Comput. J.*, **20**, 364–366, https://doi.org/10.1093/comjnl/20.4.364.

Dolan, R., and R. E. Davis, 1992: An intensity scale for Atlantic coast northeast storms. *J. Coastal Res.*, **8**, 840–853.

Duda, J. D., and W. A. Gallus Jr., 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, https://doi.org/10.1175/WAF-D-13-00005.1.

Dunn, J. C., 1973: A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. *J. Cybern.*, **3**, 32–57, https://doi.org/10.1080/01969727308546046.

Elkan, C., 2003: Using the triangle inequality to accelerate *k*-means. *Proc. 20th Int. Conf. on Machine Learning (ICML-2003)*, Washington, D.C., AAAI Press, 147–153, https://dl.acm.org/doi/10.5555/3041838.3041857.

Hart, N. C. G., S. L. Gray, and P. A. Clark, 2015: Detection of coherent airstreams using cluster analysis: Application to an extratropical cyclone. *Mon. Wea. Rev.*, **143**, 3518–3531, https://doi.org/10.1175/MWR-D-14-00382.1.

Hersbach, H., and Coauthors, 2019: ERA5 monthly averaged data on single levels from 1979 to present. Copernicus Climate Change Service (C3S), accessed 21 January 2021, https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?tab=overview.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Hubert, L., and P. Arabie, 1985: Comparing partitions. *J. Classification*, **2**, 193–218, https://doi.org/10.1007/BF01908075.

Kiel, B. M., 2021: Comparison of clustering approaches in a multi-model ensemble for U.S. East Coast winter storms. M.S. thesis, Dept. of Marine and Atmospheric Sciences, State University of New York at Stony Brook, 197 pp.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC

Kohonen, T., 1982: Self-organized formation of topically correct feature maps. *Biol. Cybern.*, **43**, 59–69, https://doi.org/10.1007/BF00337288.

Korfe, N. G., and B. A. Colle, 2018: 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.

Lamberson, W. S., M. J. Bodner, J. A. Nelson, and S. A. Sienkiewicz, 2023: The use of ensemble clustering on a multimodel ensemble for medium-range forecasting at the Weather Prediction Center. *Wea. Forecasting*, **38**, 539–554, https://doi.org/10.1175/WAF-D-22-0154.1.

Lopes, A. M., and J. A. T. Machado, 2015: Dynamical analysis and visualization of tornadoes time series. *PLOS ONE*, **10**, e0120260, https://doi.org/10.1371/journal.pone.0120260.

Ma, C.-M., and E. K. M. 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.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, CA, University of California Press, 281–297, https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992?tab=ChapterArticleLink.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroligas, 1996: The ECMWF ensemble prediction system: Methodology and results. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Neal, R., D. Fereday, R. Cocker, and R. E. Corner, 2016: A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteor. Appl.*, **23**, 389–400, https://doi.org/10.1002/met.1563.

North, G. R., T. L. Bell, R. F. Calahan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110**, 699–706, https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2.

Ohba, M., and S. Sugimoto, 2019: Differences in climate change impacts between weather patterns: Possible effects on spatial heterogeneous changes in future extreme rainfall. *Climate Dyn.*, **52**, 4177–4191, https://doi.org/10.1007/s00382-018-4374-1.

Rand, W. M., 1971: Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.*, **66**, 846–850, https://doi.org/10.1080/01621459.1971.10482356.

Reusch, D. B., R. B. Alley, and B. C. Hewitson, 2007: North Atlantic climate variability from a self-organizing map perspective. *J. Geophys. Res.*, **112**, D02104, https://doi.org/10.1029/2006JD007460.

Ross, T. J., 2010: *Fuzzy Logic with Engineering Applications*. 3rd. ed. Wiley, 606 pp.

Rousi, E., C. Anagnostopoulou, K. Tolika, and P. Maheras, 2015: Representing teleconnection patterns over Europe: A comparison of SOM and PCA methods. *Atmos. Res.*, **152**, 123–137, https://doi.org/10.1016/j.atmosres.2013.11.010.

Rousseeuw, P. J., 1987: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

Sokal, R. R., and C. D. Michener, 1958: A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, **38**, 1409–1438.

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 generations of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Uccellini, L. W., and J. E. T. Hoeve, 2019: Evolving the national weather service to build a Weather-Ready Nation: Connecting observations, forecasts, and warnings to decision-makers through impact-based decision support services. *Bull. Amer. Meteor. Soc.*, **100**, 1923–1942, https://doi.org/10.1175/BAMS-D-18-0159.1.

Vinh, N. X., J. Epps, and J. Bailey, 2009: Information theoretic measures for clusterings comparison: Is a correction for chance necessary? *Proc. 26th Annual Int. Conf. on Machine Learning*, Montreal, QC, Canada, Association for Computing Machinery, 1073–1080, https://doi.org/10.1145/1553374.1553511.

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

Ward, J. H., Jr., 1963: Hierarchal grouping to optimize an objective function. *J. Stat. Assoc.*, **58**, 236–244, https://doi.org/10.1080/01621459.1963.10500845.

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

——, ——, and ——, 2019: Evaluating U.S. East Coast winter storms in a multimodel ensemble using EOF and clustering approaches. *Mon. Wea. Rev.*, **147**, 1967–1987, https://doi.org/10.1175/MWR-D-18-0052.1.

Brought to you by NOAA Central Library | Unauthenticated | Downloaded 08/28/24 03:55 PM UTC