Applications of the Geometry-Sensitive Ensemble Mean for Lake-Effect Snowbands and Other Weather Phenomena

JONATHAN J. SEIBERT,[a] STEVEN J. GREYBUSH,[a] JIA LI,[b] ZHOUMIN ZHANG,[c] AND FUQING ZHANG[a]

[a] *Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania*
[b] *Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania*
[c] *College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania*

(Manuscript received 9 August 2021, in final form 19 October 2021)

ABSTRACT: Ensembles of predictions are critical to modern weather forecasting. However, visualizing ensembles and their means in a useful way remains challenging. Existing methods of creating ensemble means do not recognize the physical structures that humans could identify within the ensemble members; therefore, visualizations for variables such as reflectivity lose important information and are difficult for human forecasters to interpret. In response, the authors create an improved ensemble mean that retains more structural information. The authors examine and expand upon the object-based Geometry-Sensitive Ensemble Mean (GEM) defined by Li and Zhang from a meteorological perspective. The authors apply low-intensity thresholding to WRF-simulated radar reflectivity images of lake-effect snowbands, tropical cyclones, and severe thunderstorms and then process them with the GEM system. Gaussian mixture model–based signatures retain the geometric structure of these phenomena and are used to compute a Wasserstein barycenter as the centroid for the ensemble; D2 clustering is employed to examine different scenarios among the ensemble members. Three types of ensemble mean image are created from the centroid of the ensemble or cluster, which each improve upon the traditional pixel-wise average in different ways, successfully capture aspects of the ensemble members' structure, and have potential applications for future forecasting efforts. The adjusted best member is a better representative member, the Bayesian posterior mean is an improved structure-based weighted average, and the mixture density mean is an outline of the key structures in the ensemble. Each is shown to improve upon a simple arithmetic mean via quantitative comparison with observations.

KEYWORDS: Lake effects; Bayesian methods; Statistical techniques; Ensembles; Clustering; Machine learning

1. Introduction

Modern weather forecasting relies heavily upon the use of ensembles of predictions- using multiple models or variations on a single model to generate several different versions of a forecast (Kalnay 2002). These methods provide meteorologists with a range of potential outcomes and uncertainties (i.e., probabilistic forecasts instead of deterministic), the consensus of which allows for more accurate predictions and warnings for the most probable event (Hirschberg et al. 2011; Karstens et al. 2015). While more resource-intensive, ensemble forecasting as a whole is a significant improvement over single-model forecasting. However, forecasters cannot incorporate all the ensemble members (often dozens to hundreds) into a useful consensus just by considering the results of each member separately, such as in a postage stamp plot. Doing so provides too much information for a human to easily reconcile, leading to information overload. Especially because of the limited time available to create forecasts, having too much available information obfuscates the forecasting process, rather than clarifying it (Sivillo et al. 1997).

To address the challenge of ensemble visualization, the scientific community has developed visual summary tools and several types of ensemble mean—an approximate consensus reached by averaging the ensemble members. These tools are very useful for certain applications, such as track predictions for tropical cyclones, which often display either each member's predicted path as well as a simple average [e.g., a plume diagram as in Figs. 4 and 11 of Wu et al. (2010)] or cones of probabilities based on the spread of the ensemble. Another way to display ensembles is a spaghetti-style plot, in which a selected contour value for each member is displayed on a map [e.g., cover of Kalnay (2002), discussed in Roberts et al. (2019)]. In this way, the consensus and uncertainty information of an ensemble, which is critical to modern forecasting (Joslyn et al. 2007; AMS 2008; Novak et al. 2008; Demuth et al. 2009), can be efficiently conveyed to forecasters making time-sensitive decisions. Spaghetti- and plume-style plots unfortunately do not adapt well to all fields. For example, although a spaghetti-style plot can outline each member in an ensemble of snowbands, doing so only coarsely depicts how the members concentrate in and generally agree on the central region and loses information on the intensity of the snowbands' interiors. Another widespread technique, the arithmetic mean, is fairly well suited to continuous meteorological variables commonly displayed in two dimensions, such as temperature. However, applying it to ensembles of a more physically structured variable, with sharp gradients, such as reflectivity (precipitation strength), also results in a greater loss of information (to be demonstrated later). This loss of information, which impacts some variables more than others,

F. Zhang: Deceased.

Corresponding author: Jonathan J. Seibert, jjs5895@psu.edu

DOI: 10.1175/MWR-D-21-0212.1

can be just as dangerous as information overload—either can result in increased subjectivity and inaccuracy in forecasts, potentially leading to increased casualties during severe weather events. Thus, how to best display ensembles' information is critical to modern weather forecasting. The visualization challenge has been examined and emphasized by Sivillo et al. (1997), Kalnay (2002), and the National Research Council (2006), yet remains insufficiently addressed.

Several approaches have been developed in recent years to combating the difficulties of ensemble visualization while best utilizing the advantages of probabilistic forecasting. For example, Lee et al. (2009) and Eipper et al. (2019) employed a best-member approach that treated the member with the smallest mean absolute error (to observations) as the ensemble consensus. Roberts et al. (2019) provide a style of plot combining paintball plot (all members plotted together, each in a different color; e.g., Greybush et al. 2017, their Fig. 16) of member threshold exceedance with neighborhood maximum ensemble probability contours that could provide useful probabilistic guidance. The performance-weighted averaging approach to ensemble means improves on the simple arithmetic mean, weighting each member's contribution to the mean by some measure of prior performance (Woodcock and Engel 2005; Greybush et al. 2008). The Bayesian model–averaging technique (Raftery et al. 2005) makes particular use of the weighted average of posterior probability distributions, as well as observed prediction skill, to predict an improved forecast.

These techniques reduce the information loss inherent to pixel-wise averaging, but remain insufficient for ensemble fields relating to clouds and precipitation. Effectively, any variation on the arithmetic mean treats each point in the region as separate. It does not recognize any physical structure in the ensemble members, resulting in an average that is not informed by the physical "objects" that a human would see. A truly "geometry sensitive" mean would retain much more of that information by identifying each member's components as distinct objects, and accounting for their structure and relative location in the calculation of the average. Figure 1 provides an illustrative example: suppose a small number of ensemble member thunderstorms with similar structure and intensity of precipitation, but varying locations. An equally weighted mean yields a smeared-out ellipse of weak precipitation (feature blurring). A more realistic mean would be a single, centralized cell that retains the major structures and intensity of each member.

This is especially important because of the value of those physical structures when forecasting, such as the shape of a bow echo thunderstorm or the intensity of a particular snowband. ("Intensity" will hereinafter refer to the brightness of an image for a particular point, region, or object; for this application, intensity represents composite reflectivity.) In essence, for a variable such as temperature, with much less variation over a short distance, an arithmetic mean will tend to produce a reasonable result. However, cloud and precipitation related variables have much stronger spatial gradients and variations and differences in location, intensity, and orientation of features. Thus, an arithmetic mean will tend to blur features and obscure the interpretation.

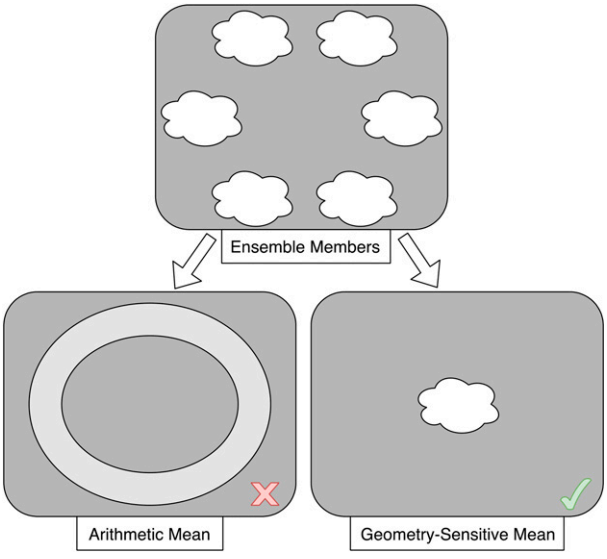

FIG. 1. (bottom left) Idealized arithmetic (pixel-wise) mean vs (bottom right) GEM for (top) an ensemble of storm cells. Whereas the arithmetic mean takes the strict average on a per-pixel basis, resulting in loss of intensity and structure information, the ideal GEM gives a more realistic mean that retains that information.

Object-based techniques bearing similarities to the authors' have been employed in meteorology before, but have generally been focused on forecast verification and not been adapted to ensemble consensus applications. The three systems most relevant to this approach are those of Davis et al. (2009, hereinafter D09), Nehrkorn et al. (2014, hereinafter N14), and Han and Szunyogh (2016, hereinafter HS16). N14 and HS16 both developed "feature alignment" techniques based on defining the difference, or error, between a forecast and its matching observation for the same time, thus defining vector difference fields. D09's Method for Object-Based Diagnostic Evaluation (MODE) attempts to identify corresponding features in a forecast and its verifying observation. It was originally developed to approximate the viewpoint of a human when identifying objects. D09 identifies individual objects by using a combination of radial smoothing and thresholding to mask the variable field. They use an "interest function" to roughly describe the likelihood of match between each object pair, based on weighted similarities in all described attributes. N14 developed a solution specifically for data assimilation. Using a nonlinear cost function minimization, they generate a field of displacement vectors, constrained to be smooth and nondivergent to get physically realistic results. These displacement fields are then used to adjust the forecasts to minimize error with respect to their corresponding observations, improving future forecasts down the data assimilation chain. This technique was also applied to an ensemble of forecasts, using the mean square error (MSE) as "distance" between the values of the members and a simple mean to choose a best member. HS16's "optical flow" method operates more iteratively, transitioning the forecast image to best match the observation by iteratively moving larger to

smaller sections of pixels. Thus, coarse structures are matched first, and then finer details, which "preserves global precipitation" and some finer structures. However, they had to work around the accompanying tendency to over-converge precipitation elements. A subsequent paper (Han and Szunyogh 2018) further developed HS16's algorithm by employing it and a rigid motion adjustment repeatedly, in sequence.

This research aims to address information loss and overload in ensemble forecasting by demonstrating a new method of creating improved ensemble means, using a different approach to object-based design. This method, the geometry-sensitive ensemble mean (GEM), considers the underlying geometry in each member, capturing the physical structure of the phenomena they depict to reduce feature blurring. Specifically, this paper will examine the principles laid out in the proof-of-concept (Li and Zhang 2018) from a meteorological perspective, and discuss direct applications for the meteorological community through the example of an ensemble of radar reflectivity imagery of lake-effect snowbands, with tropical cyclone and severe thunderstorm reflectivity ensembles as supporting cases. (Henceforth, the authors discuss ensemble members as images, rather than fields of values.) To achieve a geometry-sensitive mean, the authors have created three distinct new types of mean image: the adjusted best member (ABM), the Bayesian posterior mean (BPM), and the mixture density mean (MDM). These collectively compose GEM, originally defined in Li and Zhang (2018). The ABM and BPM utilize best-member and Bayesian posterior mean components similar to those of Lee et al. (2009) and Raftery et al. (2005). However, our approach differs in that the criterion for weighting ensemble members uses a nontraditional distance metric to a related ensemble centroid, rather than relying upon outside information (e.g., prior performance). Additionally, by using three distinct types of mean in conjunction, the authors hope to overcome the disadvantages of each individual type, such as a best-member alone losing information on the ensemble's spread. For the purposes of this paper, a centroid (also known as a barycenter) is defined as the arithmetic mean of the location of all points in a group. GEM's "ensemble centroid" relies upon this principle as well, but requires additional background that will be given in section 3. GEM is based on preserving the basic form of the structures in each ensemble member through object recognition via concepts from computer vision. MODE is conceptually similar, attempting to simulate human vision, and distinguishes objects by combining smoothing and thresholding. GEM employs thresholding (section 2a), but identifies objects using two distinct clustering algorithms, also addressing N14's concerns about manual object recognition. Being based on adjustment of forecasts to match observations, both N14 and HS16's methods share more similarities with the GEM best member (section 2d) than the geometry-based ensemble centroid at GEM's core. In general, GEM finds the best match or alignment for the entire ensemble, rather than matching one forecast to its corresponding observation. In addition, GEM employs the Wasserstein distance (WD) metric, which considers differences in both value and location (section 2c). The WD is similar in nature to D09's simpler attribute-based matching system, but is based in statistical methods and does not require explicit attributes beyond location, intensity, and shape.

The remainder of the paper will be organized as follows: section 2 will describe the data and techniques such as image processing, centroid calculation, and averaging methods (Fig. 2 shows the overall flow of the GEM system). Section 3 will discuss the results of applying this modified GEM system to the three reflectivity ensembles and interpretation for use in forecasting, as well as quantitatively comparing those results with observations. Section 4 will draw conclusions and outline areas for further development.

2. Data and methods

a. Data and preprocessing

The data for this research consists of three separate ensembles of simulated composite radar reflectivity images, generated by the Weather Research and Forecasting (WRF) Model. Each ensemble forms one of the three cases: lake-effect snowbands, a tropical cyclone, and severe thunderstorms.

The first dataset, forming the core of the study, contains 30 members (9 of which are displayed in Fig. 3) depicting composite reflectivity for simulated lake-effect snowbands. The raw reflectivity images were generated by WRF, version 3.7.1 (Skamarock et al. 2008), using the Advanced Research WRF dynamics core. The model was run over three one-way nested domains: a 27-km outer grid over the Great Lakes region, a 9-km intermediate grid, and a 3-km convection-allowing inner grid containing Lakes Huron, Erie, and Ontario. For the purposes of this paper, only the results from the innermost 3-km grid were used. The physics used for this run included Thompson et al. (2008) graupel microphysics, Rapid Radiative Transfer Model (RRTM) longwave radiation (Mlawer et al. 1997), Dudhia (1989) shortwave radiation, Mellor–Yamada–Janjić boundary layer (Janjić 1990), and the Noah Land Surface Model's surface physics (Ek et al. 2003). Initial and boundary conditions were provided by the Global Forecast System. Perturbations for each ensemble member were generated by NCEP's fixed background error covariance method for 3DVAR, "CV3" (Saslo and Greybush 2017 and references therein). The model was run from 1200 UTC 10 December 2013 to 1200 UTC 12 December 2013. The images chosen are valid for 1700 UTC 11 December 2013 (Saslo and Greybush 2017; Greybush and Saslo 2018).

The second dataset, from Minamide (2018), is a 60-member ensemble of simulated reflectivity images of the 2017 Hurricane Harvey. The model was run with a similar 27–9–3-km two-way nested grid, again with only the innermost being used here. Model physics used for this run were the single-moment 6-class mixed-phase microphysics scheme (Hong and Lim 2006), the Yonsei University planetary boundary scheme (Hong et al. 2006), and RRTM for both longwave and shortwave radiation schemes. These images are valid for 0000 UTC 24 August 2017. Figure 4 displays the first nine members of this ensemble (see below for a description of thresholding). Because the ensemble members for the lake-effect snowband case and the tropical cyclone case were not generated in

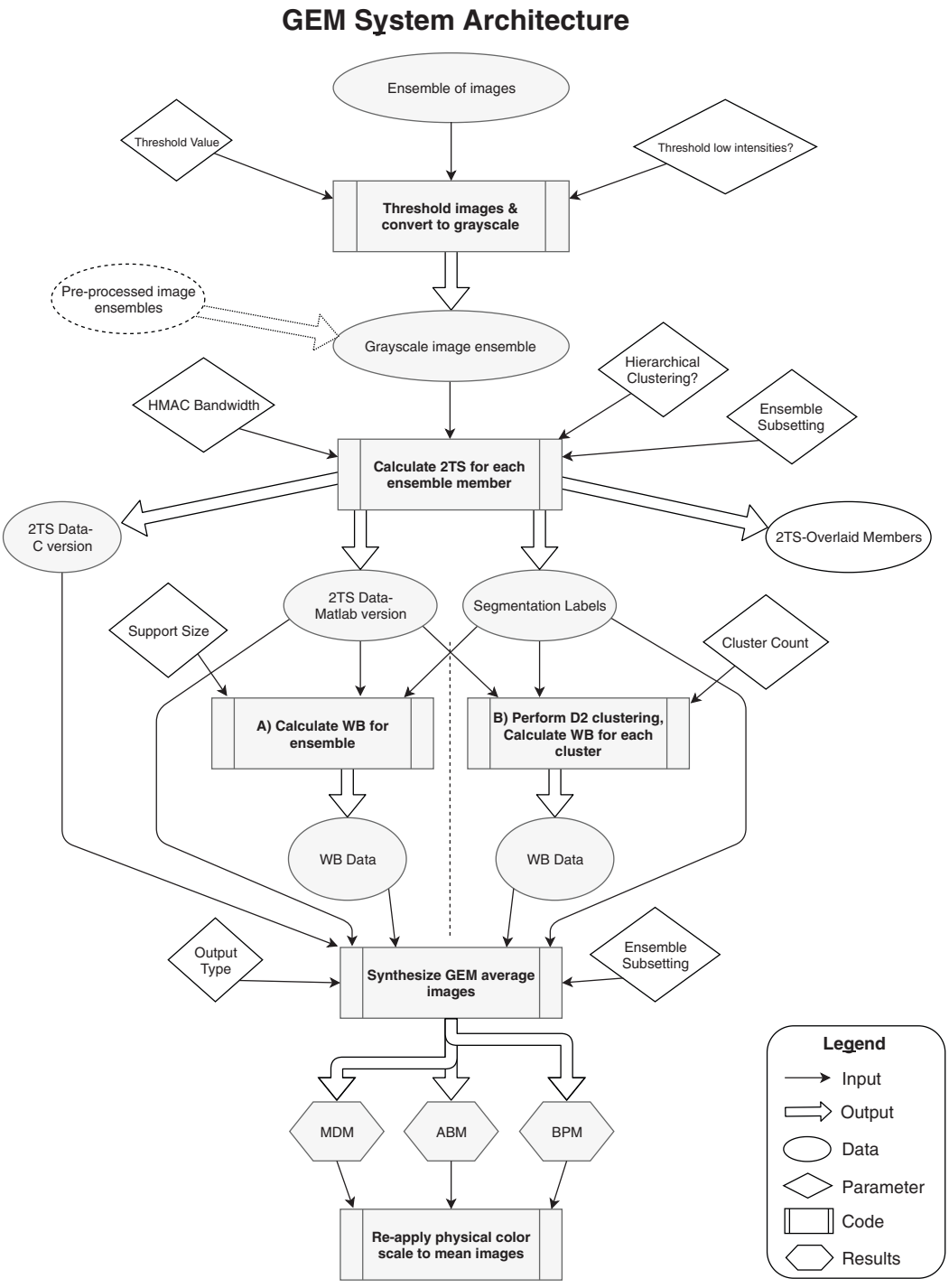

FIG. 2. Descriptive diagram of GEM system architecture. The central split denotes the choice of whether to use D2 clustering.

composite form, composite reflectivity images for each member were created by calculating the maximum reflectivities across all vertical levels.

The third dataset, from Hanson (2016), is a 30-member ensemble of composite reflectivity images of simulated severe thunderstorms over Pennsylvania and Maryland. This ensemble was also generated by WRF 3.7, using the same model physics and resolution as the first case. These are valid for 2300 UTC 20 April 2015. Figure 5 displays the first nine members of this ensemble.

In the initial stage of the GEM system for a given ensemble, each member image is put through low-intensity

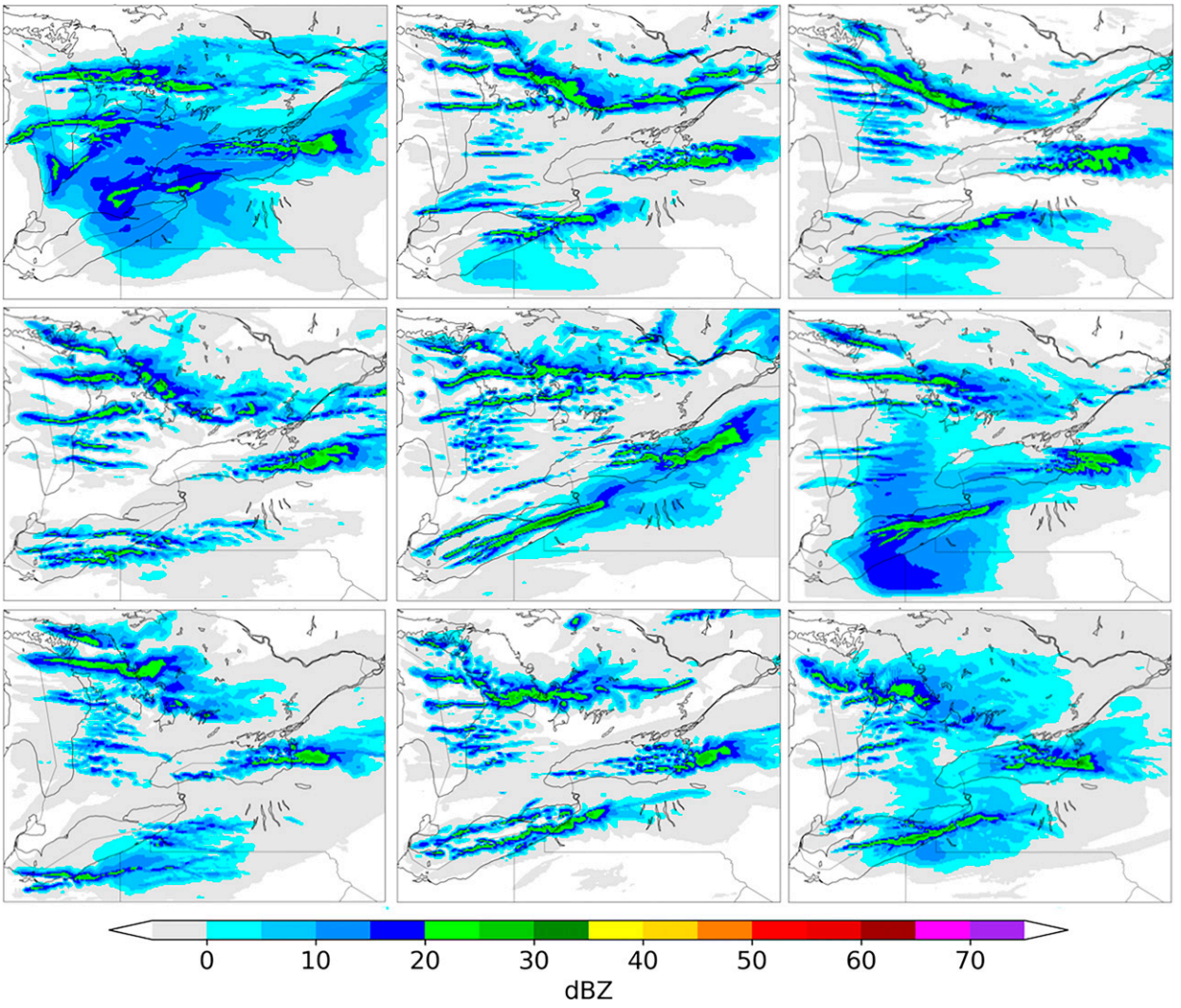

FIG. 3. Postage stamp plot of composite reflectivity (dBZ) from the first nine images of the snowband ensemble. All have a similar basic structure but significant variance in band orientation, intensity, and other properties. Note that dBZ is a logarithmic variable, having meaning at and below 0, but that values below 0 are here considered negligible.

thresholding, in which all regions of minimal reflectivity—below the specified threshold—are set to the threshold value. This focuses the analysis on high-precipitation regions of interest, and allows the algorithm—and human forecasters—to more easily distinguish discrete objects, such as individual bands within multiband snow systems coming off Lake Huron. (The GEM system is also run with the nonthresholded version of the ensembles for the sake of comparison.) The threshold used in this research is 0 dBZ, because the unit scale of reflectivity is logarithmic, and values below 0 generally indicate negligible amounts of precipitation. However, the GEM system is designed to allow any threshold value.

After thresholding, each image is then converted from reflectivity values (units of dBZ) to decimal grayscale (0–1) using the following equation:

$$\text{gray} = \frac{\text{dB}Z - e_{\min}}{e_{\max} - e_{\min}}, \tag{1}$$

where $e_{\max}$ and $e_{\min}$ are the ensemble maximum and ensemble minimum reflectivity, respectively. (Note that $e_{\min}$ becomes equal to the threshold value beforehand, if one is applied.) This shifts the range of dBZ values present in each image such that the minimum intensity becomes 0, and all values are normalized to fractions of 1 by the adjusted ensemble maximum reflectivity. This process makes it much easier to represent the reflectivity values in the main calculations and to allow for use of the Wasserstein distance (section 2c) while preserving the physical relationships between the intensities of the ensemble members. In addition, because this conversion occurs after the thresholding step, the low reflectivity values that were set to the threshold

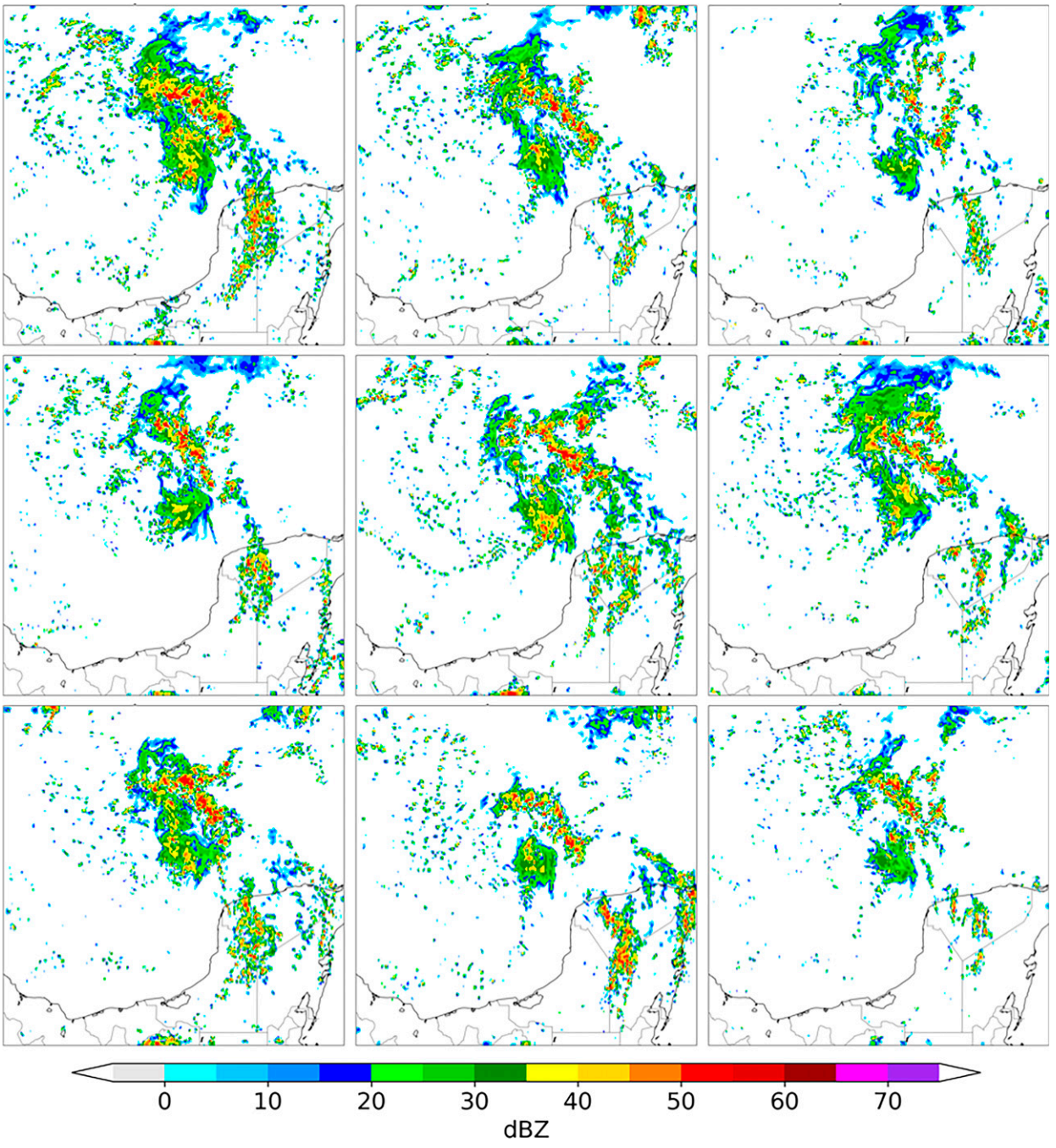

FIG. 4. As in Fig. 3, but for the first nine members of the tropical cyclone ensemble, thresholded.

value are mapped to 0 in grayscale. Thus, the minimal reflectivity values are still effectively removed from the calculation, while also allowing the remaining values to be defined more precisely across the entire grayscale range. This process is demonstrated in Fig. 6. The final results are mapped back from grayscale to the physically interpretable dBZ units of reflectivity using the inverse of Eq. (1), divided by 255 (to return to decimal grayscale before conversion), and with the mean of the original members' minimum and maximum reflectivity replacing the overall ensemble minimum and maximum.

It is possible to use a quantile of the above-threshold values (or of the members' maxima) instead of said means, but the authors chose to use said means over a quantile, as they offer the best consistent representation of the lake-effect snowband ensemble, which is the primary case for this research. Systems were added to the GEM code to allow for use of various quantile methods if desired. In addition, while it is possible

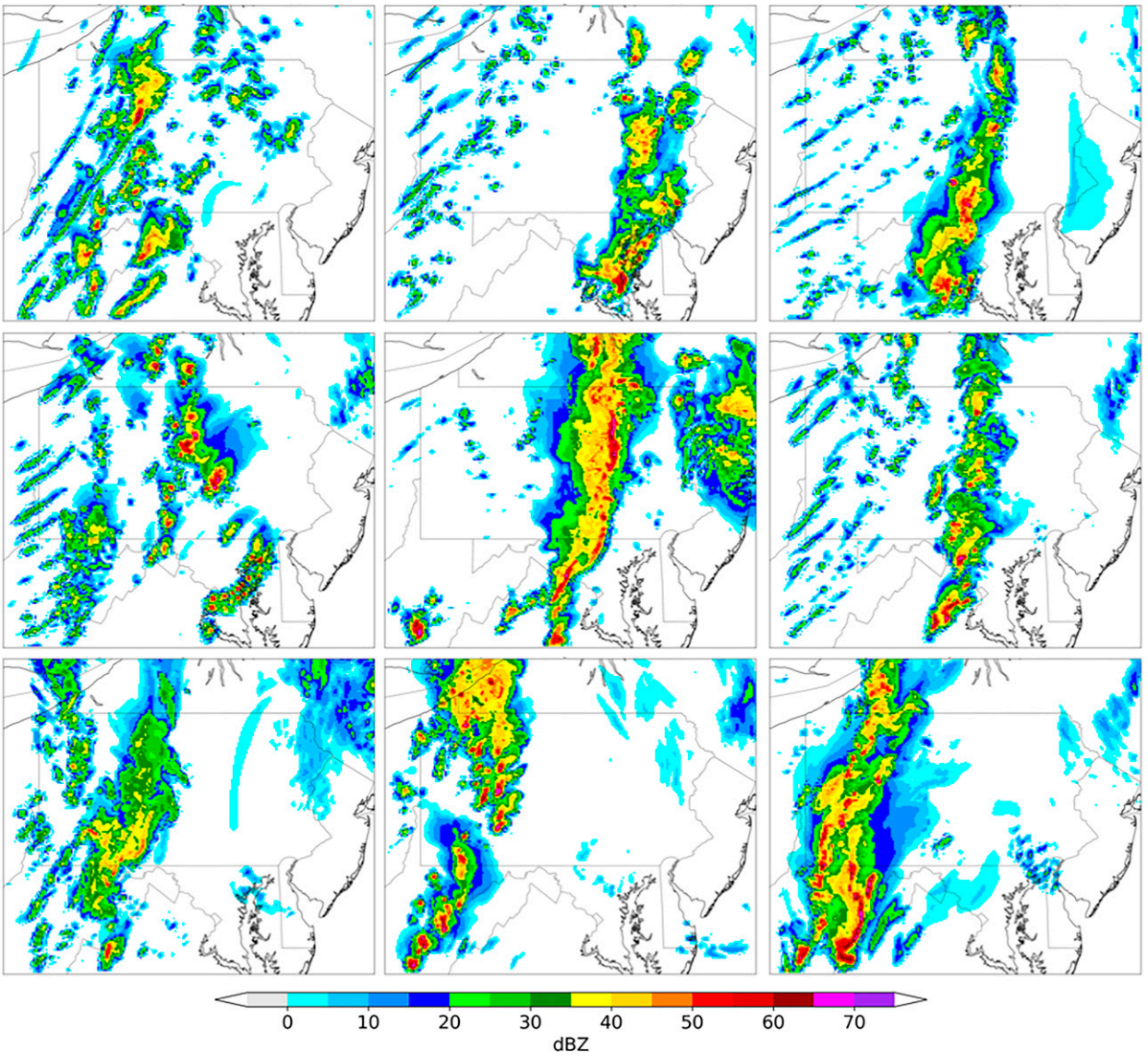

FIG. 5. As in Fig. 3, but for the first nine members of the severe thunderstorm ensemble, thresholded.

that some error is introduced by transforming the logarithmic dBZ values linearly, the main system is agnostic to the variable in use and is designed to create results with value ranges that are as faithful to the ensemble as possible.

b. Two-tiered signature

To calculate the ensemble centroid, GEM generates a two-tiered signature (2TS) that captures the geometry of physical structures depicted in each image by splitting them into smaller objects called "patches." Similarly to a pixel count, the number of patches used to define the structures in each image determines the granularity of the image and the scale of the features captured by individual patches. If too many patches are used, it effectively becomes a pixel representation, defeating the point of the method, and if too few are used, they cannot capture enough detail in the physical structures of the image (Fig. 7). Neither extreme retains significant geometric information, and so the optimal setting lies between the two.

The two tiers of the signature for a given image are then defined by the patches describing their contents. The first tier is the centroid of each patch, weighted by the total patch intensity. The second tier is defined by the parameters of a Gaussian distribution fitted to the intensity-weighted locations of the other points in the patch. Thus, the 2TS is effectively a Gaussian mixture model. Taken together, the 2TS captures the location, overall intensity, and shape of each patch, similarly to the SAL method (Wernli et al. 2008) discussed in D09 and HS16.

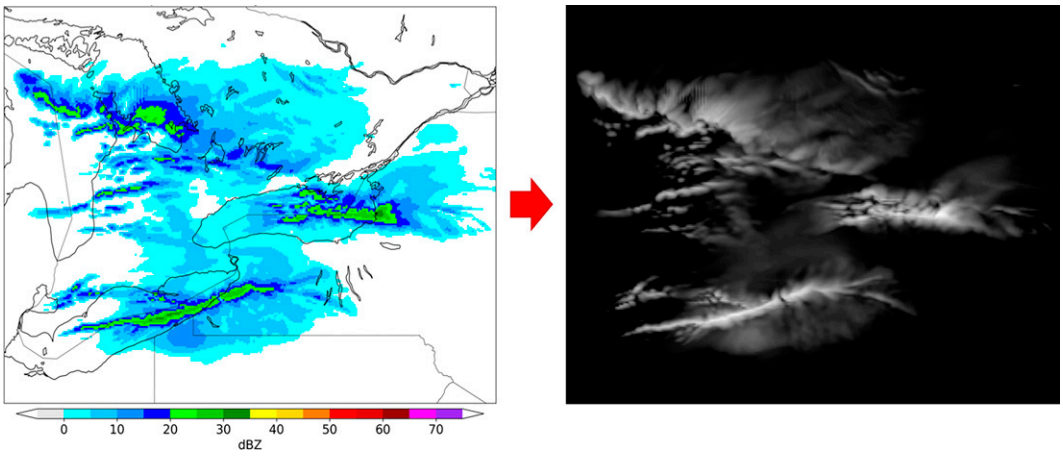

FIG. 6. The (left) thresholded and (right) grayscaled versions of snowband ensemble member 9.

Figure 8 demonstrates how an image is broken into patches and fitted with a signature. First, all pixels with nonzero intensity are clustered into "cells" using weighted *k*-means [an example can be seen in Fig. 2b of Li and Zhang (2018)]. Doing so provides a coarser-resolution version of the image that still retains intensity and structure—useful when the total number of pixels is large. (This implementation of *k*-means adds clusters until the average squared distance between cluster mean and members falls below 0.005.) Li et al.'s (2007) hierarchical mode association clustering (HMAC) algorithm is then used to merge these cells into patches. The total intensity, intensity-weighted centroid (red Xs), and covariance of each patch become the 2TS. The Gaussians defined by those values are shown as red ellipses. The centroids of those patches are known as "support points." The "support size" is how many there are, representing the granularity of the calculation. In general, support points are points of reference in each set, by which the differences between the sets are calculated. Each point has a set of coordinates (two dimensional, in this case) and an intensity, representing the mean location and total intensity of its patch. The shape (covariance) of the patch is tracked separately by the second tier of the signature. Collectively, these patches and their attributes define the geometry of the physical structures in the image. The only details of the HMAC algorithm that are affected by GEM are the input bandwidth values (see section 2e); additional details can be found in the references.

c. *Wasserstein distance and barycenter*

Prior work in object-based forecasting has used simple Euclidean distance between objects as one of the criteria—or the only criterion—for matching two corresponding objects in different images. However, Euclidean distances between points with differing intensities cannot independently represent meaningful differences between their images. The authors employ the Wasserstein distance between images to make up for this shortcoming.

The Wasserstein distance (Rachev 1985) is the counterpart of the Euclidean distance-based mean for Euclidean vectors.

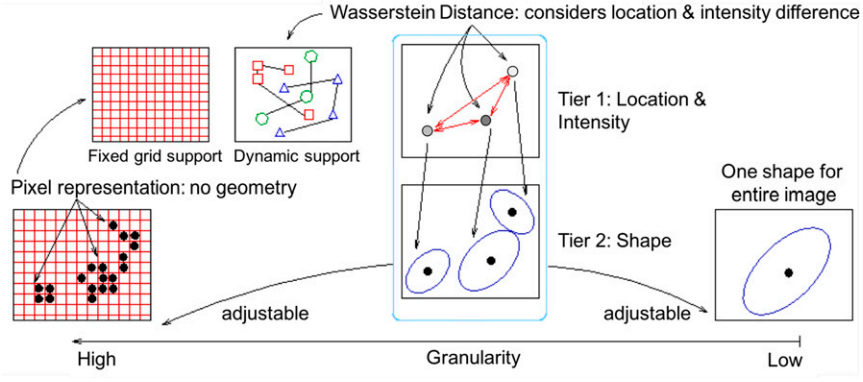

FIG. 7. Schematic diagram illustrating the two-tiered signature (2TS). The 2TS captures the structural information by describing the location, intensity (shown here in grayscale), and shape of each cloud patch in a dynamic fashion—the support points are not constrained to specific preset locations. Too many or too few patches both fail to describe useful details in the geometry. The figure is adapted from Li and Zhang (2018).

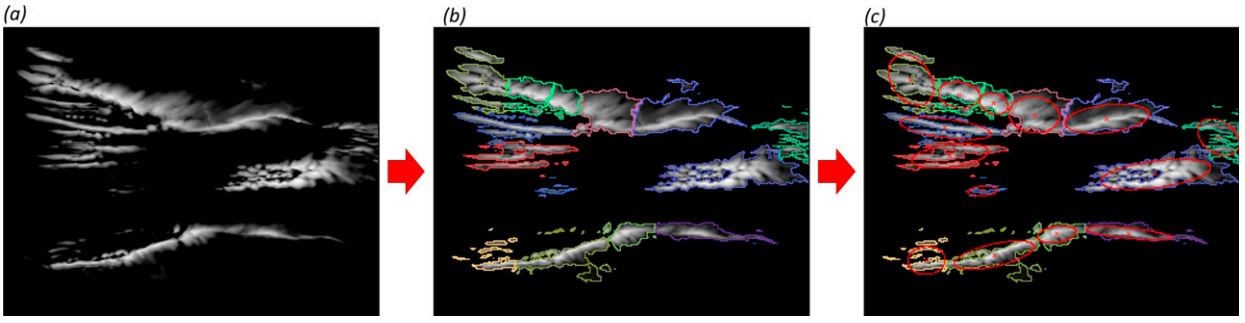

FIG. 8. Schematic illustrating the process of clustering pixels into cloud patches and fitting the Gaussian of their 2TS: (a) the starting reflectivity source image, in grayscale; (b) image broken into patches, outlined in different colors; and (c) intensity-weighted centroid (red star) and Gaussian distribution (red ellipse) calculated for each patch.

The WD was originally used to define distance between probability distributions, but can also be employed for geometrically dispersed intensity distributions (i.e., grayscale images) summarized by support points. (The authors now consider each member image to be defined by a set of support points as described above, and we will refer to them as such.) Applied to this case, the smallest possible WD between images matches each support point with its closest counterpart from the other image, both in Euclidean distance and value (or intensity). It is the smallest possible sum of the products of the matching support points' differences in location and their differences in value. As each of these is a vector, their product is a matrix, and the new distance is the sum of all the elements in that matrix. The WD is also known as the earth-mover's distance—this method seeks out the minimum possible work required to balance out the differences, as if manually shifting piles of dirt from one location to another to equalize the two sets. The matrix of "matching weights" minimized by the algorithm provides instructions on how to move the probabilities (or "piles of dirt") from one distribution to the other with the least effort. One advantage of this method is that the number of support points in the two distributions do not need to be the same. Part of the reason that existing methods of creating ensemble means are insufficient for certain meteorological purposes is that those methods do not account for the structural information that makes ensemble members physically meaningful. Using the WD to create a mean retains that critical information on the geometric structure of the distributions being averaged.

The WD is defined formally as follows: Let W be the Wasserstein distance between two distributions under the L2 norm ($\|\ \|$) in n-dimensional space. (The L2 norm is the length of a vector, defined for a vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$ as $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.) Let the superscripts a and b indicate the two sets of support points D^a and D^b, the subscripts i and j indicate the current support point in their corresponding sets (i for set a; j for set b), and m^a and m^b indicate the corresponding support size (total number of support points). (Note that m^a and m^b are not required to be equal between sets.) The $x_i^{(a)}$ and $x_j^{(b)}$ are then the support point coordinates in n-dimensional space with $w_i^{(a)}$ and $w_j^{(b)}$ being their corresponding intensities. Then

$$\left\{W\left[D^{(a)}, D^{(b)}\right]\right\}^2 := \min_{\{\pi_{i,j} \geq 0\}} \sum_{i=1}^{m^a} \sum_{j=1}^{m^b} \pi_{i,j} \left\| x_i^{(a)} - x_j^{(b)} \right\|^2 \quad \text{s.t}$$

$$\sum_{i=1}^{m_a} \pi_{i,j} = w_j^{(b)},\ j = 1, 2, \ldots, m^b - 1,\ m^b$$

$$\sum_{j=1}^{m_b} \pi_{i,j} = w_i^{(a)},\ i = 1, 2, \ldots, m^a - 1,\ m^a, \qquad (2)$$

where $\pi_{i,j}$ is the set of "matching weights" between the two sets. The matching weights are the values that are optimized by linear programming in order to minimize the total WD. The values themselves add up to the intensity of the support point they match to, and the sum of those weights (for each point) is multiplied by the distance between each corresponding point to produce the WD. As such, the matching weights are a function of both Euclidean distance and difference in value between each support point in a set and all support points in the other. The L2 norm is here used to assign a length to the vector of the Euclidean distance between each possible pair of support points.

The Wasserstein barycenter (WB) is then the centroid of all members' sets of support points, defined by minimizing the total WD between itself and each ensemble member. It thus acts as the centroid of the ensemble—a kind of skeletal mean. Specifically, the WB of the ensemble's first-tier signatures (intensity and location) is the first-tier signature of a theoretical, idealized ensemble "mean" and is used as the ensemble centroid. This study uses Ye et al. (2017)'s algorithm to calculate the WB.

The WD can also serve as a measure of ensemble variance. Specifically, the average of the squared WD between each ensemble member and one of the GEM images would then estimate both the position and intensity variance among the ensemble.

d. Ensemble mean images

The three types of ensemble mean image generated around the WB are the mixture density mean, the Bayesian posterior

TABLE 1. GEM parameters tested (all) and chosen as optimal for this case (boldface type). Nearly infinite combinations are possible with other values for each parameter, but full sets of results were tested only for combinations listed herein. WB-RM was disabled for D2 testing only, because the smaller ensemble sizes caused nonrepresentative rigid adjustments in the means.

Bandwidth values (HMAC)	Support sizes	Threshold values	WB-RM	Rescaling
A (5, 8)	**Automatic**	**0 dBZ**	**Rotation and translation**	Ensemble min/max
B (3, 5)	9	10 dBZ	Rotation only	**Member mean min/max**
C (2, 3.5)	18		Translation only	
D (2)	60		**None**	

mean, and the adjusted best member. These are later compared with a simple pixel-wise average (PWA) of the images—the traditional ensemble mean used in meteorology.

The MDM is a physical representation of the 2TS of the unobserved "true" mean: the first tier (patch centroids) of the signature is the WB of the ensemble's first tier. The second tier is calculated via covariance fusion—an average of the patches corresponding to each support point across the ensemble, weighted by the total intensity of the patches and normalized by the mean total intensity of the ensemble [see full equation in Li and Zhang (2018), section 3.2.2]. Effectively, the MDM is what the WB would look like when fully fleshed out by the Gaussian distribution of its support points' intensities. It serves to demonstrate the skeleton of the ensemble, highlighting structures common among the ensemble members.

The BPM treats the ensemble members as random samples to estimate the mean as the truth. Li and Zhang (2018) define the prior probability distribution of the true mean as a normal distribution with the MDM as its mean and the average WD between each ensemble member and the WB as the variance. From this definition, the BPM—as the posterior mean—can be expressed as the sum of each ensemble member, weighted proportionally to the WD between each of them and the WB, plus the MDM, weighted proportionally to the average WD, such that the sum of all the weights is 1. Effectively, the BPM morphs the MDM into a more realistic shape (closer to any individual ensemble member) by including both the MDM itself and the ensemble members in an average, and weighting the lowest-WD members more heavily to better reflect realistic shapes without distorting the mean. If all of the members are roughly equidistant (in WD), then the BPM becomes a roughly equally weighted average, but includes the MDM as an extra ensemble member, thus still better spatially informed than a simple arithmetic mean. By constructing the mean using the WD as the member weights, the BPM does not require outside information on member performance to operate. Also of note is that if an ensemble member is much closer to the centroid than the others, the BPM will very strongly resemble that member, as it will weight it accordingly.

The ABM is a variant on an in-sample mean. Each ensemble member is rotated and translated to best align with the barycenter by minimizing the WD between them. The ABM then selects the adjusted member with final lowest pixel-wise MSE to the MDM image. In doing so, the ABM provides a realistic example of what may occur, while also indicating the most representative location and orientation of the precipitation features. An option to select the adjusted best member based on solely the WD has been implemented, but the MSE is used here for the final step to remain consistent with Li and Zhang (2018).

The MDM and BPM are rescaled such that their maximum intensity is 255 (in grayscale) following its calculation, prior to conversion back to dBZ, to make the results visually easier to interpret. [Examples of each mean image can be found in Figs. 9, 10, 12, and 13, described in more detail below, and the full equations are in Li and Zhang (2018).]

e. Variations

The GEM system was run with varied sets of parameters, in order to test the sensitivity of the mean images to these choices. Results are generated for the specified ensemble at several levels of granularity, with and without thresholding, rotation or translation (when computing the ABM), and rescaling. By "granularity" the authors refer to the spatial resolution of the WB calculation, defined by the support size—the larger the support size, the more points of reference are available to define the member and WB signatures. The granularity is initially controlled by the Gaussian kernel bandwidths for the HMAC algorithm: larger bandwidths merge more components into one patch, reducing the support size, and therefore the granularity. Granularity can be finely adjusted by altering the support size after the point in the system execution at which the member 2TSs are calculated (see Fig. 2 for illustration). This study examines four specific bandwidths, with HMAC input values A (5, 8), B (3, 5), C (2, 3.5), and D (2). These parameters will require tuning on a per-application basis to generate the most useful results—the choice is subjective because of the structural differences between different phenomena. After testing the GEM system with the parameter variations described above, the authors determined that bandwidths B and C produced the most applicable results for the three ensembles under examination. Therefore, the bandwidth-B subset of results were chosen to be shown and discussed below. Bandwidth A, as the fastest setting for GEM, was used for D2 clustering tests. These variations are summarized in Table 1. Not all permutations of results have been completed for the tropical cyclone and thunderstorm ensembles at this time, as those cases were added to serve as proof-of-concept that GEM can function on other phenomena.

f. D2 clustering

The authors also apply the D2 clustering technique (Ye and Li 2014) to the GEM system. This technique can be used to section an ensemble into two or more different groups, which

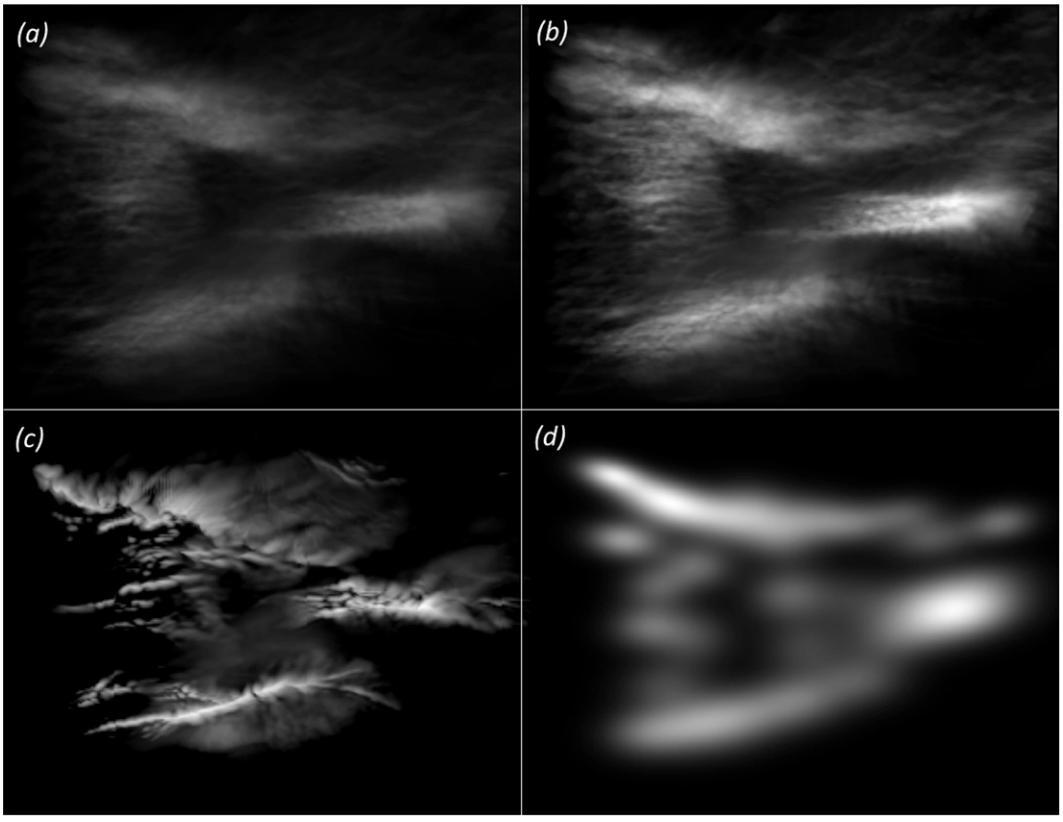

FIG. 9. Thresholded snowband ensemble results using GEM bandwidth B, shown as grayscale images: (a) PWA, (b) scaled BPM, (c) ABM, and (d) scaled MDM.

could present multiple forecast scenarios based on the major structural patterns present in the ensemble (e.g., Kowaleski and Evans 2016). Effectively, D2 splits the input ensemble into smaller groupings, based on their WD to multiple centroids. The algorithm functions iteratively, much like *k*-means clustering: it first creates *N* first-guess centroids and assigns all members to one of them by smallest WD. It then loops through, calculating the WD between each member and each centroid and adjusting the centroids and their member assignments to minimize total WD between each centroid and its members. D2 was chosen over further uses of k-means or similar algorithms to remain consistent with Li and Zhang (2018).

g. Comparison with observations

To quantify the improvement of these GEM images over the PWA, the authors obtained observation data for the specified time for the lake-effect case from the 11 NEXRAD stations with ranges in the model domain (product "NCR"), created a composite reflectivity version, and calculated a root-mean-square error (RMSE) value between each mean image and these composited observations. As the resolution of the observations was higher than that of the ensemble used to create the mean images, each was re-gridded to the resolution of the other. This was accomplished using bilinear interpolation (for upscaling the mean images) and a dynamic moving-box average (for downscaling the observations). Additionally, as it was noted that the northwestern snowband present in most ensemble members was not present in the observations, a second comparison was made with both observations and mean images masked to include only data points within 150 km of each NEXRAD station. This maximum "view range" of 147.7 km (rounded out to 150 km) was computed by accounting for the minimum possible height of the radar beam given distance, curvature of Earth, and minimum elevation angle, given that lake-effect snowbands are unlikely to appear at heights above 3 km.

3. Results and interpretation

Overall results for the three ensembles demonstrate that GEM is capable of capturing the geometric structure of weather phenomena, and shows agreement between all results for a given ensemble. Figure 9 shows results for the snowband ensemble at bandwidth B in grayscale, as it was the output format for the original GEM system, alongside a simple pixel-wise average (PWA; e.g., arithmetic mean) for comparison. In short, the ABM acts as a better representative member for the ensemble, the BPM as a better structure-based weighted average, and the MDM as an outline of the key structures in the ensemble. However, as these images are difficult to quantify and interpret in grayscale, the recolored versions (Fig. 10) were created and plotted against a map of

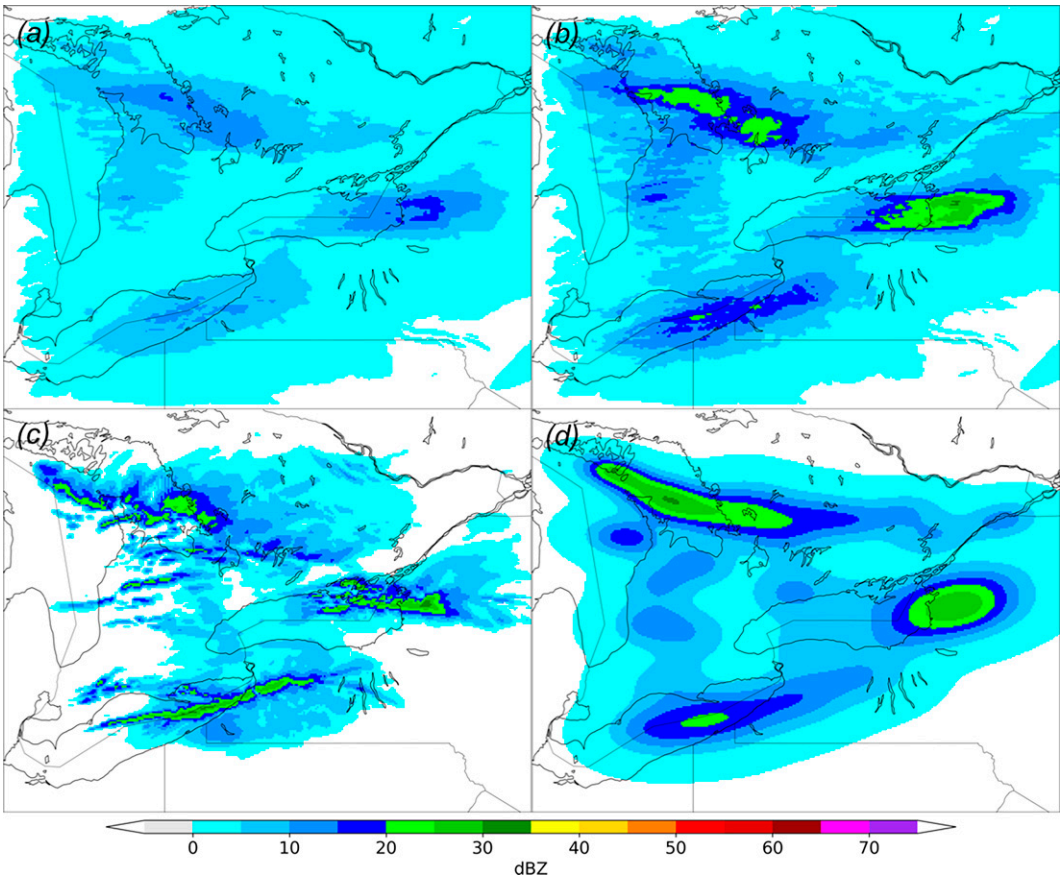

FIG. 10. Thresholded snowband ensemble results using GEM bandwidth B, shown as composite reflectivity images: (a) PWA, (b) scaled BPM, (c) ABM, and (d) scaled MDM.

their corresponding geographical locations. (For the purposes of plotting, thresholded results treat any dBZ values of 0 or less after conversion as "no data," shown as regions of white.)

Figures 9–12 display the unscaled PWA and ABM alongside the scaled BPM and MDM for their respective cases. This format demonstrates the improvements the GEM images make over the basic PWA, in both structure and intensity. (The ramifications of the rescaling process are discussed later in this section.) As an ensemble member modified only in position and orientation, rescaling the ABM is unnecessary—its reflectivity values are already as realistic as the initial model results.

Each of the three ensemble means can be interpreted as a form of probability or risk estimation. The MDM and BPM can both be read as the most likely locations (among the ensemble members) for common structures and points of peak intensity. The MDM is a generalized indicator of where a given proportional intensity of precipitation is likely to fall, whereas the BPM is a more representative version that follows the shape of the member reflectivity structures more closely. The BPM can potentially give a better idea of the current stage of development for the ensemble-depicted phenomenon and its relevant physical structures. Additionally, the wider, semi-realistic central structures of the BPM indicate the range of variation among the main structures of the ensemble, as each is formed by a weighted average of the overlapping structures from each member. The MDM gives a more theoretical idea of where the regions of highest precipitation are likely to occur in an area. In that way, it could be used similarly to a risk map, indicating regions of slight, moderate, and high risk of certain amounts of reflectivity (or other variable being examined). However, not much should be read into the stratification visible in Figs. 10–14—it is artificially created by assigning colors to 5-dBZ-wide swaths of the value range.

The primary result of note from Fig. 10 is that the three mean images successfully capture the basic structures of the ensemble, with some improvement over the PWA (which has significant feature blurring) and are in general agreement with each other in terms of location and intensity. The BPM and PWA are very similar in structure, although not identical. However, the BPM (and MDM) capture the intensity of the snowbands much better than the PWA, having the same maximum intensity as the ensemble average. The MDM shows a generalized map of where the maxima of reflectivity are likely to appear, mapping out the structures common to the entire ensemble. The BPM refines those patterns into a more representative shape that is marginally better at focusing the regions of high reflectivity than the PWA—although not as

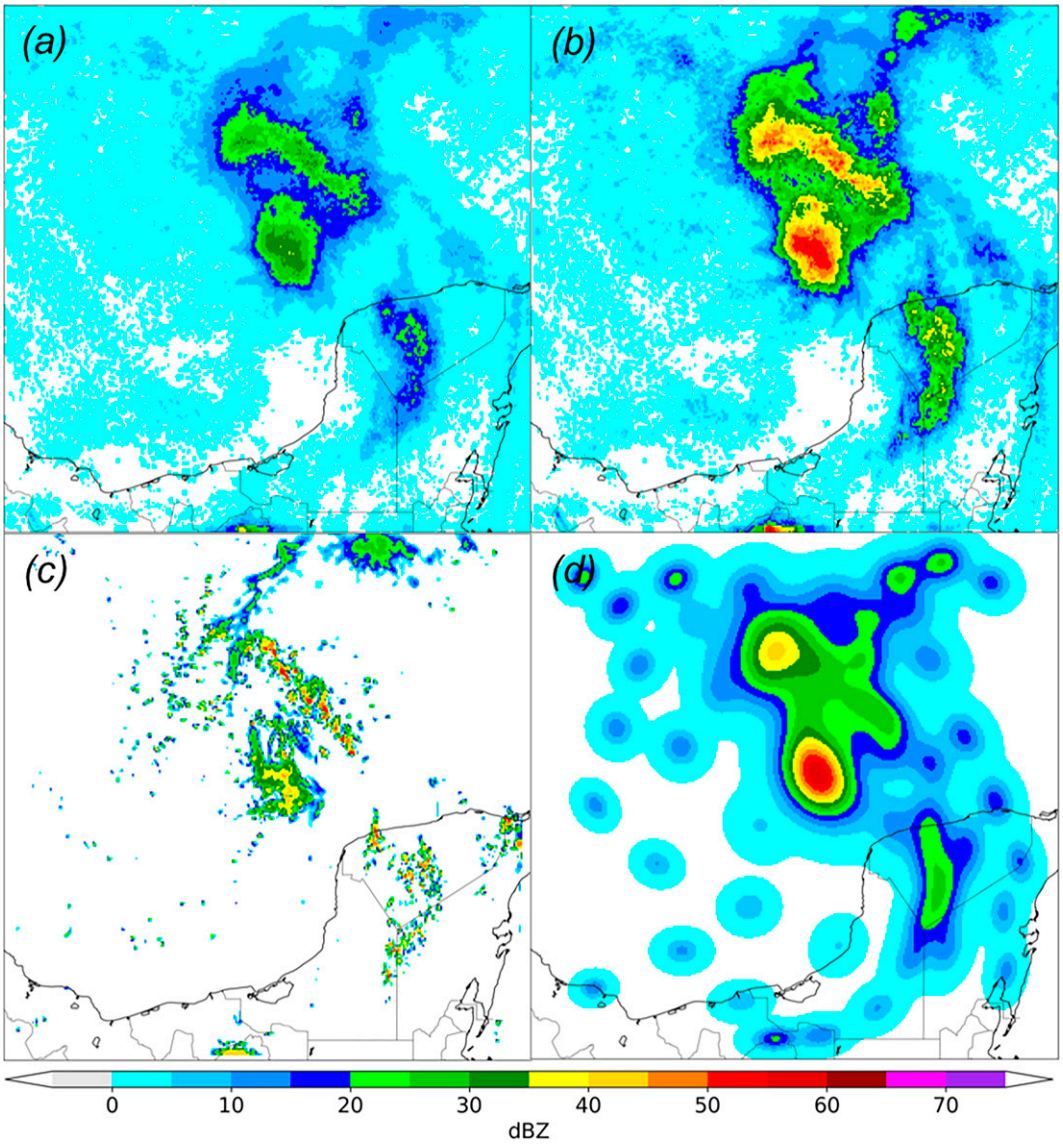

FIG. 11. Thresholded tropical cyclone ensemble results using GEM bandwidth B, shown as composite reflectivity images: (a) PWA, (b) scaled BPM, (c) ABM, and (d) scaled MDM.

focused as any individual member. Both the BPM and PWA still succeed in indicating the general areas where significant reflectivity might be expected, but are too dispersed to be taken as a literal representation of a likely atmospheric state. The northwestern snowband in the BPM (from Lake Huron) more accurately captures the narrowing and splitting partway along the band's length (considering 20 dBZ and above in the BPM) that is commonly found in the ensemble members. An uninterrupted band of consistent width is uncommon, although it does occur in the ensemble. The other noteworthy difference is that the PWA's southernmost snowband (off Lake Erie) displays a dual band, an artifact of the varied latitudes of the members' bands, whereas the BPM displays only a single band, as is much more common among the ensemble members.

In contrast to the other two, the ABM is essentially the most representative ensemble member. The ABM for this set of results differs from the unmodified member 9 (Fig. 3, bottom right) mostly in the translation of the ABM to the northeast. This is in accordance with its alignment to the WB, showing the most representative position and orientation while still retaining its shape. It is possible for this rigid motion to shift the ABM away from physical features on the ground it was tied to—such as this ABM's eastern snowband no longer being entirely over Lake Ontario—but this adjustment could also potentially serve to correct model biases of a similar nature. This result indicates that snowbands under these basic conditions are more likely to form or be oriented farther to the northeast, with structures similar to those displayed by the ABM. Additionally, GEM allows for this rigid

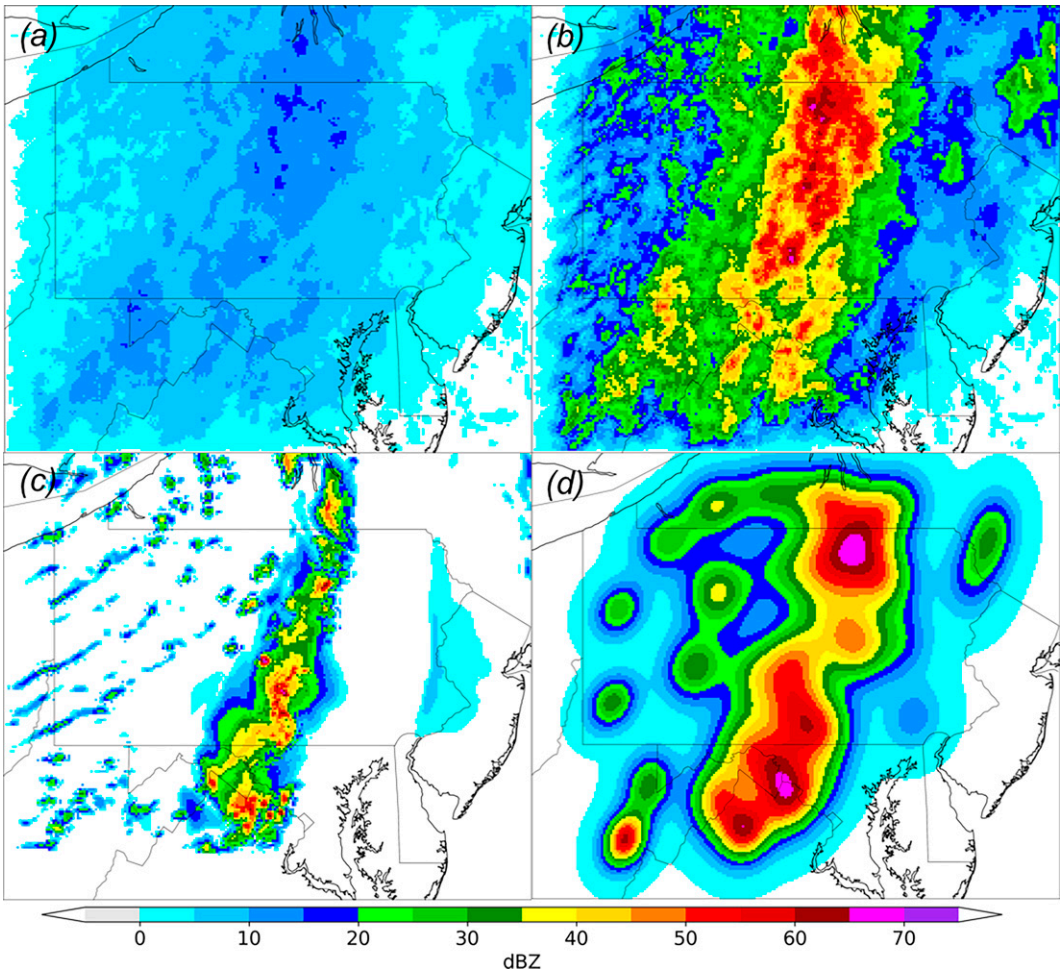

FIG. 12. Thresholded thunderstorm ensemble results using GEM bandwidth B, shown as composite reflectivity images: (a) PWA, (b) scaled BPM, (c) ABM, and (d) scaled MDM.

motion adjustment to be turned off, in the event that the ensemble it is applied to causes the adjustment to become nonrepresentative.

Table 2 displays the results of the RMSE comparison between observations (Fig. 13a) and the GEM images outlined in section 2g, quantifying their improvement over a standard PWA. The GEM images demonstrate noticeable reduction in error from the PWA in all tested comparison methods, with the ABM having the largest reduction in all cases. The significant advantage of the ABM over the other two mean types in an RMSE comparison was expected, as it is a rigidly transformed ensemble member with no structural alteration, and is much closer to the observations than a fully GEM-synthesized mean. However, these results do conclusively show that, for at least this set of results, the BPM and MDM images also constitute a significant quantitative improvement over a simple PWA—at least 17% less error than the PWA in all cases. Strictly as a function of comparison with a PWA, then, the GEM images are more "realistic" means. A visualization of the 150-km mask is shown in Fig. 13b.

Figure 11 shows the results for the tropical cyclone (TC) case. Their most notable feature is the accuracy with which the MDM traces out the spiral rainbands of Hurricane Harvey. The patches that described the original members were clearly able to capture the important structures in the cyclone, given sufficient granularity. Although GEM has not been tested on less sparse TC variables such as brightness temperature, it is evident that for a variable such as reflectivity, GEM is capable of describing ensemble structures well even given relatively wispy individual members. GEM thus accomplishes the authors' primary goal for this application—creating a descriptive mean for such a variable field without losing much structural information. The MDM even succeeds in describing the storm cells separate from the main body of the cyclone to the northeast and far south with 2–3 patches each. The TC case contains both large-scale and very small-scale features present in the image, and is an example where the use of a larger or smaller number of support points to capture the salient features could be explored.

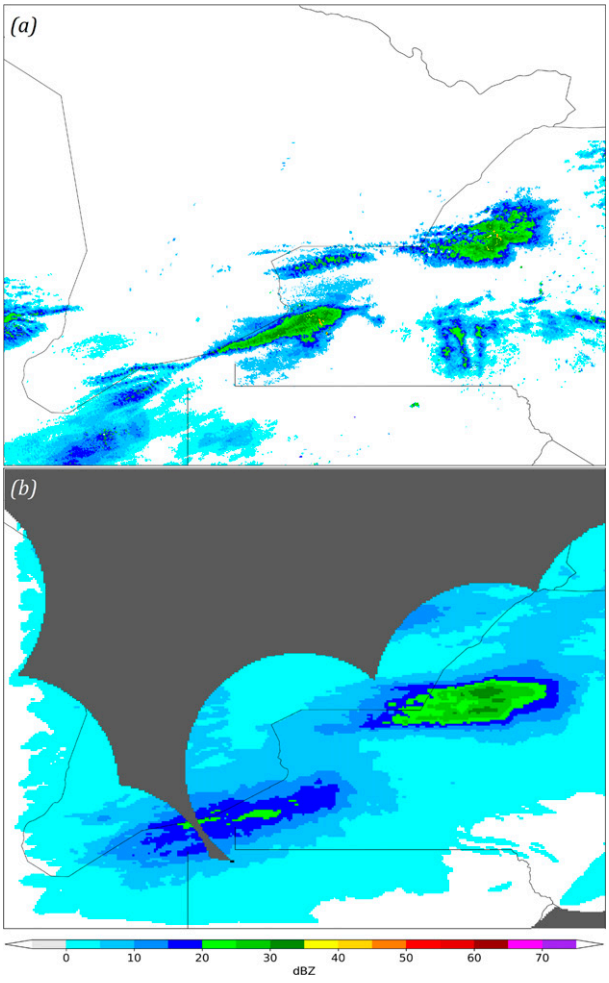

FIG. 13. (a) Observation data for the snowband case from NEXRAD, colored according to the dBZ scale used for results; (b) a sample BPM result, with the 150-km mask applied over top (dark gray) to illustrate the restricted range of the available NEXRAD data.

Figure 12 displays the corresponding results for the thunderstorm ensemble. GEM still arrives at a reasonable consensus for location and relative intensity, despite not capturing the thunderstorm ensemble as well as the snowband and TC ensembles, possibly due to the higher maximum intensities or larger location spread. In addition, comparison with Fig. 14 demonstrates that the thunderstorm case BPM is a significant improvement over the PWA, more successfully focusing the strongest reflectivity into the central line. While the rescaled BPM still has too broad of an area with high reflectivities to be truly representative, its shape is much closer to the individual members than the PWA, which has much stronger hotspots in the northeast and southwest from the members that are strongest in those locations. Effectively, WD-based weighting of the BPM here improves upon the PWA by recognizing spatial outliers and weighting them accordingly. The MDM here illustrates that spatial consensus, describing the common central storm line and frequent hotspots along it. It also indicates the aforementioned outliers and lower-intensity features to the west without being overly affected by them.

The other two cases show similar results when comparing their scaled (Figs. 10–12) and unscaled (Fig. 14) BPM and PWA images. The BPM is a marginal improvement over the basic PWA for the snowband and TC cases, as it is likely to be for most ensembles. The unscaled snowband BPM is more able to distinguish boundaries between bands than the PWA, as shown by the more distinct separation of the areas above 5 dBZ. The unscaled BPM has most of the same basic shape and structures of the PWA, but focuses the regions of higher intensity slightly more into the cores of each snowband. (This is also true of the TC case.) Both are too dispersed to be particularly readable without the application of rescaling. Similarly, while the unscaled snowband MDM is not as useful in identifying regions of distinctly higher risk, it still captures the locations and relative intensities more strongly than the corresponding PWA, and displays the consensus of shape more thoroughly, especially in the NW band.

In both the TC case and thunderstorm case (Figs. 11 and 12), it is apparent that the rescaling process can result in some or all regions of concentrated high reflectivity in the mean images being too strong, broad, or both to be representative. This is especially evident in Fig. 12, where the maximum reflectivities of the BPM and MDM are significantly higher and present in broader areas than some individual members. This effect has only appeared when reflectivities in excess of 50 dBZ are present, and is likely a side effect of having a few ensemble members with relatively high maximum reflectivities influencing the ensemble maximum. Further, this effect is mitigated by the use of the member-averaged minimum and maximum dBZ values in the average images' conversion back to dBZ, as these values are inevitably less extreme than the overall ensemble minimum and maximum, and capture the ensemble consensus more strongly. While this flaw in GEM is significant, GEM is intended to better summarize an ensemble, not to provide literal depictions of possible observations. The relative intensities and structures defined by the GEM images still serve to illustrate the areas of highest risk, and the rescaling and recoloring provides an easier way for forecasters to visually distinguish those areas.

This issue can also be partially addressed with the use of quantile-based rescaling (i.e., substituting a specific quantile of the ensemble's above-threshold values for the mean of the members' maxima). This method reduces the impact of outlier maximums, but also reduces the amount of visual distinction between lower intensities in ensembles with lower maxima, such as the snowband case. As such, this workaround is not suitable for all cases.

Figure 15 demonstrates the sensitivity testing employed, examining the response of each mean image type to varied support size. The MDM is highly sensitive to granularity, as shown by the strong response from changing support size alone. The support size 9 MDM is clearly composed of just those nine Gaussian shapes, tracing a general triangular shape, but the support size 18 and 60 MDM images begin to distinguish the eastern snowband from the southern, among other increasingly detailed features. This detail is largely a function of the support size, and the smoothness a result of

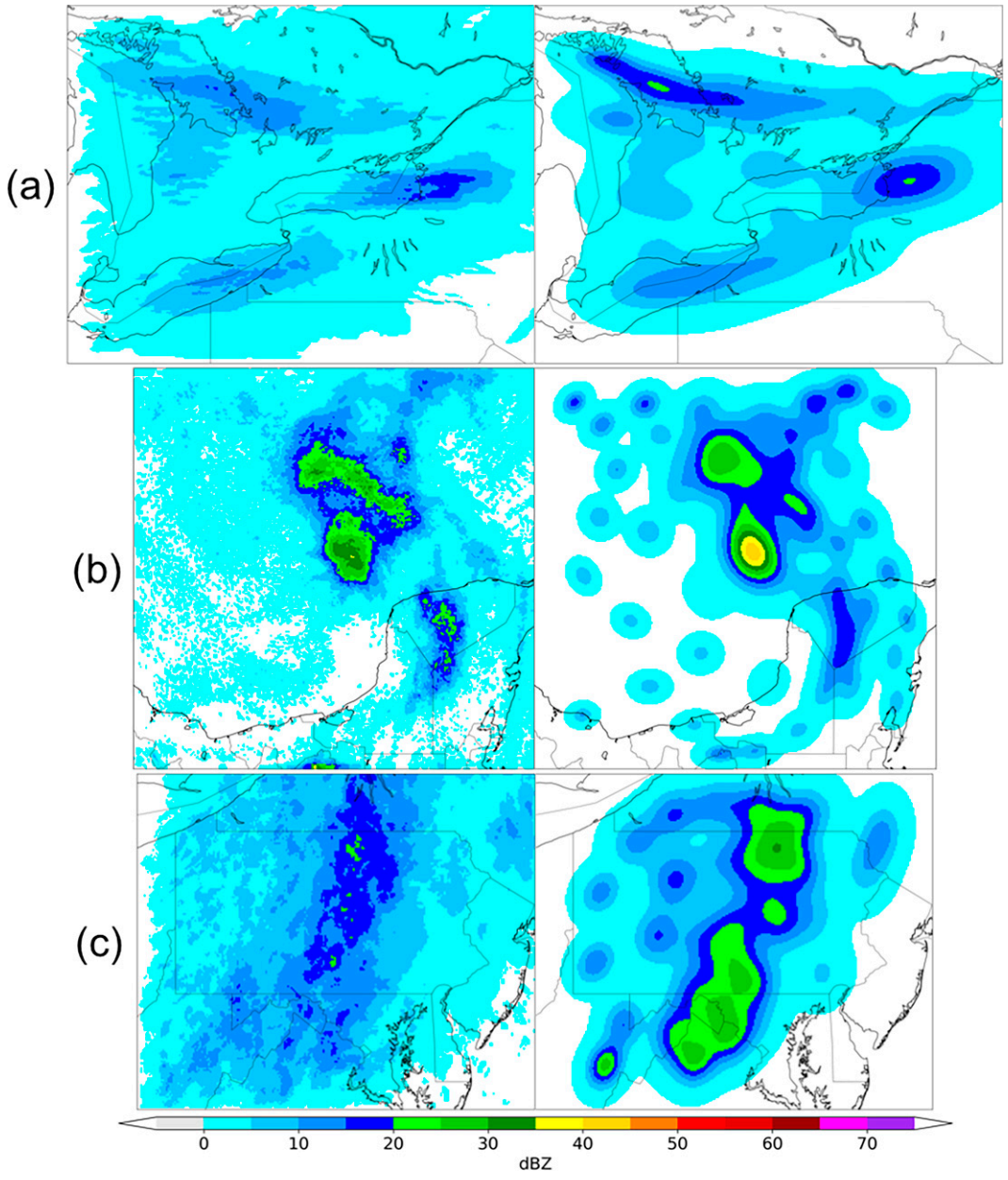

FIG. 14. Unscaled (left) BPM and (right) MDM using bandwidth B for each set of thresholded ensembles, shown as composite reflectivity images: (a) snowband case, (b) TC case, and (c) thunderstorm case.

Gaussian shapes for each support point overlapping. However, raising the support size much farther can diminish the intensity assigned to each to the point where there is no longer any overlap, return the MDM to an uninformative series of bright points. The other mean image types do not respond as strongly to changes in support size, but still display significant structural differences between bandwidth values (not shown). Additionally, the thresholded results are more representative of the primary snowband structures than the non-thresholded results, likely because of lower interference from regions of low precipitation.

The testing that has been conducted with D2 clustering has demonstrated that it is capable of concisely showing several different scenarios from among the ensemble, as determined by the members' primary structures (see the example in Figs. 16–18). This could be very useful for demonstrating major possibilities, using each cluster's mean, from among the ensemble's predictions without overloading the forecaster with information (e.g., Kowaleski and Evans 2016). D2 clustering can successfully group member images by rough scenario, and can do so on a finer level than a human eye alone when aided by the WD metric. There are clear similarities

TABLE 2. Breakdown of the RMSE values between the observations (obs) and GEM snowband mean images. Because the observations use a different resolution than the snowband ensemble, one or the other had to be regridded to match. To avoid bias, the results of both are shown. In addition, tests were conducted with the 150-km radar mask applied to both the GEM images and observations, to ensure fair comparison given the limited range of the available NEXRAD stations. The percent improvement of each RMSE value over that of the PWA for each category is given in parentheses. Each GEM image has at least 17% less error than the PWA. All values are rounded to one decimal place.

	Model regridded to obs resolution	Model regridded to obs resolution (150-km mask)	Obs regridded to model resolution	Obs regridded to model resolution (150-km mask)
ABM–obs	43.6 (35.5%)	30.6 (31.9%)	43.7 (35.4%)	31.0 (31.8%)
BPM–obs	55.0 (18.7%)	35.6 (20.8%)	55.0 (18.8%)	36.1 (20.5%)
MDM–obs	56.1 (17.1%)	35.6 (20.8%)	56.0 (17.2%)	36.1 (20.5%)
PWA–obs	67.6	44.9	67.7	45.4

between the dominant patterns within each cluster, such as the weaker southern snowband in Fig. 16, versus the more elongated bands in Figs. 17 and 18. Thus, these results show that D2 can employ the WD to good effect in numerically grouping ensemble members by structure and intensity. In addition to the ABM chosen by MSE to the MDM, each cluster member in the figures is also ranked by smallest WD to the cluster centroid, indicating other atmospheric states more or less representative of the ensemble subset for their specific scenario. Although the cluster members were chosen by minimizing WD, there is clear disagreement in all three clusters between the MSE-based best member (ABM) and

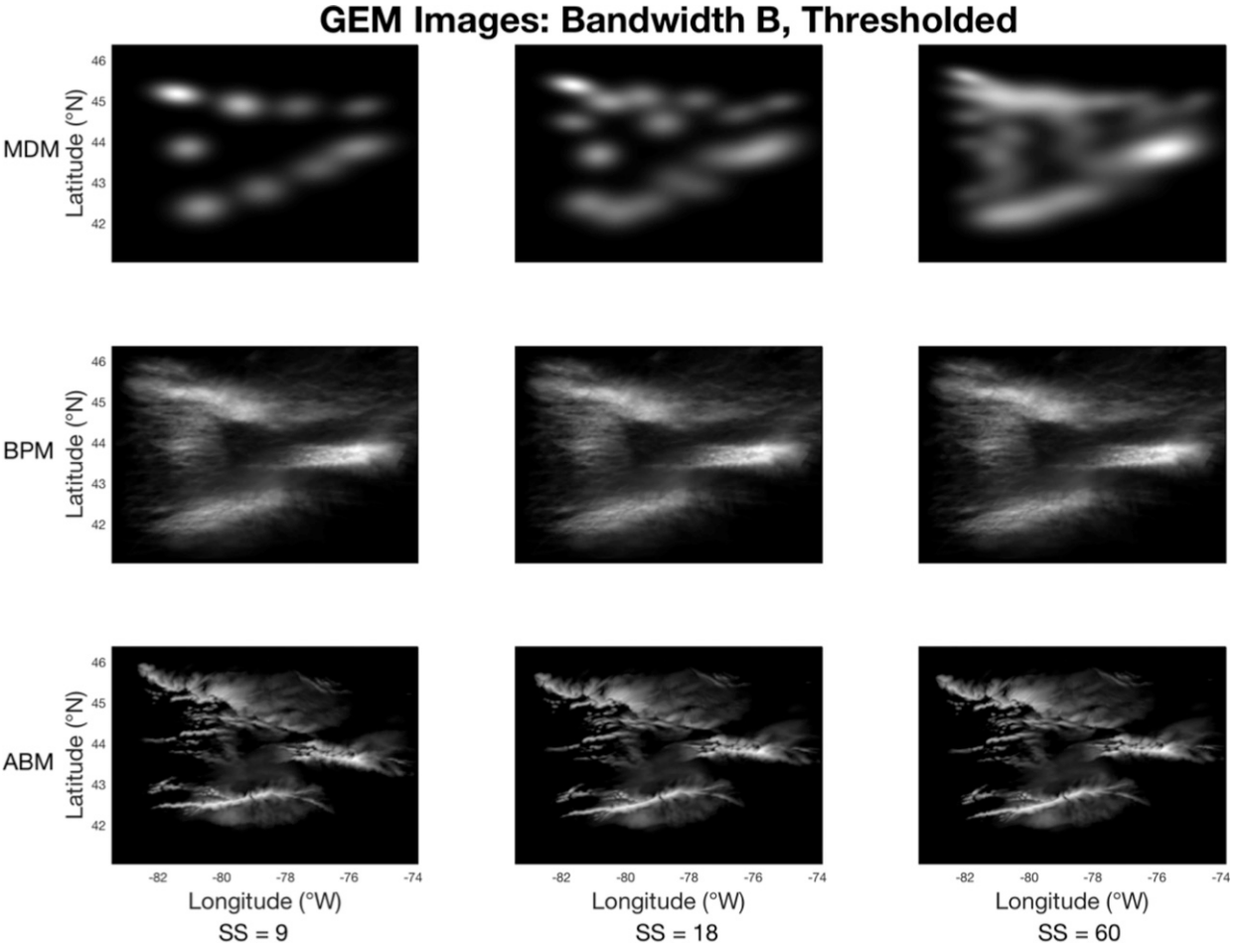

FIG. 15. Results for sensitivity testing the GEM for (top) MDM, (middle) BPM, and (bottom) ABM on bandwidth B using support sizes (left) 9, (center) 18, and (right) 60. The MDM shows the greatest sensitivity to support size, although all mean image types display recognizable variance between different bandwidths.

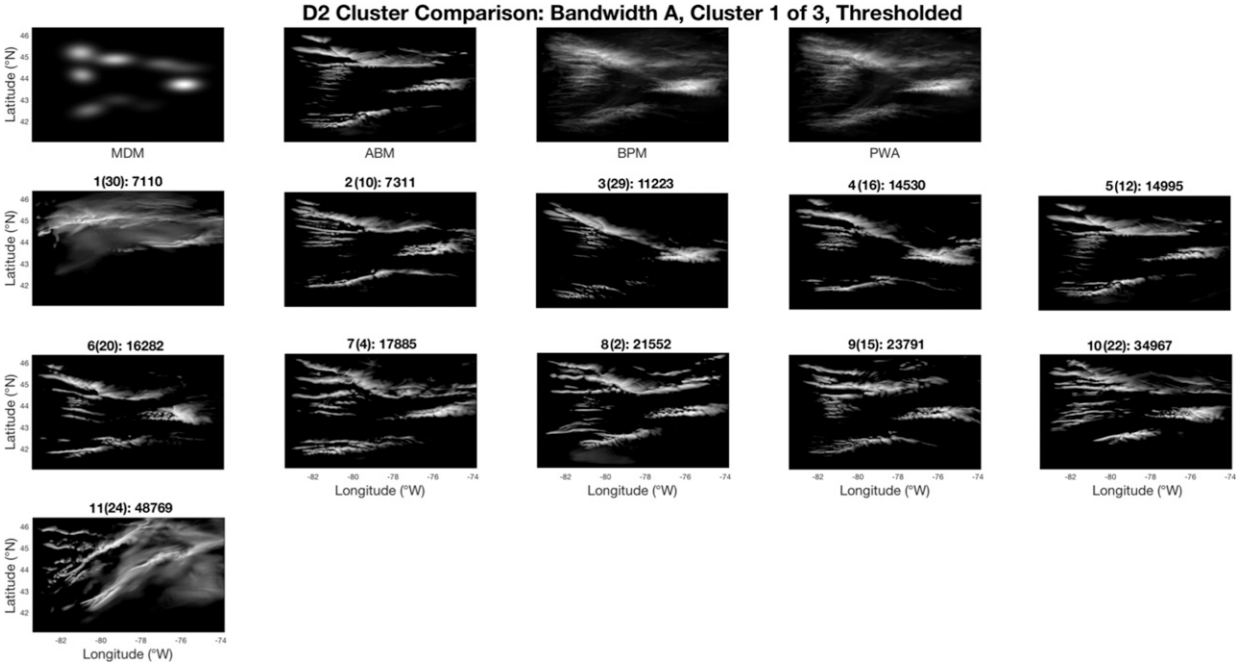

FIG. 16. Results for the first of three clusters generated through D2 clustering, using GEM bandwidth A, shown as grayscaled images: (top) GEM images for the cluster; (remaining rows) each ensemble member assigned to the cluster, sorted by WD to cluster centroid, labeled as "[Position in sorted list]([Member number]): [WD to cluster WB]."

the WD-based best member (cluster member 1), illustrating the difference between the priorities of each distance metric. Because the authors have not yet developed a method of quantifying which option is more useful, both options are displayed here.

The testing with D2 clustering has also highlighted a potential problem with these methods—when the ensemble is split into clusters of fewer than 30 members, the translation and rotation fitting process employed in the ABM calculation results in poor adjustment. Those ABM results tended to be

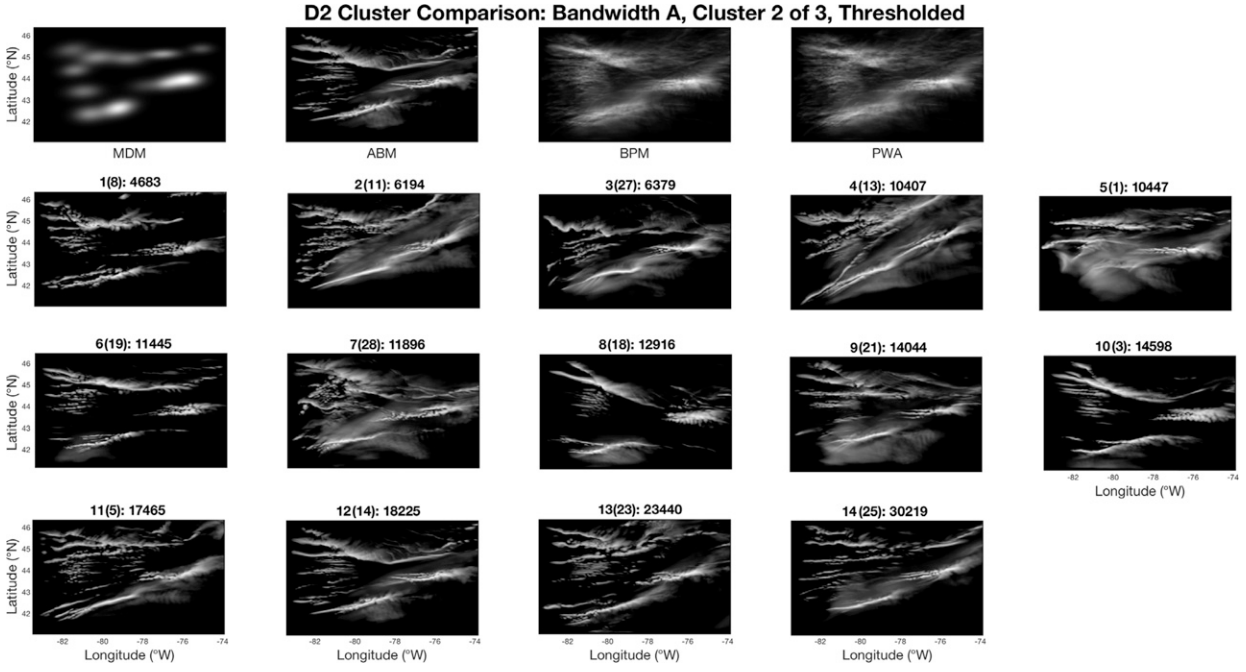

FIG. 17. As Fig. 16, but for the second of the three clusters.

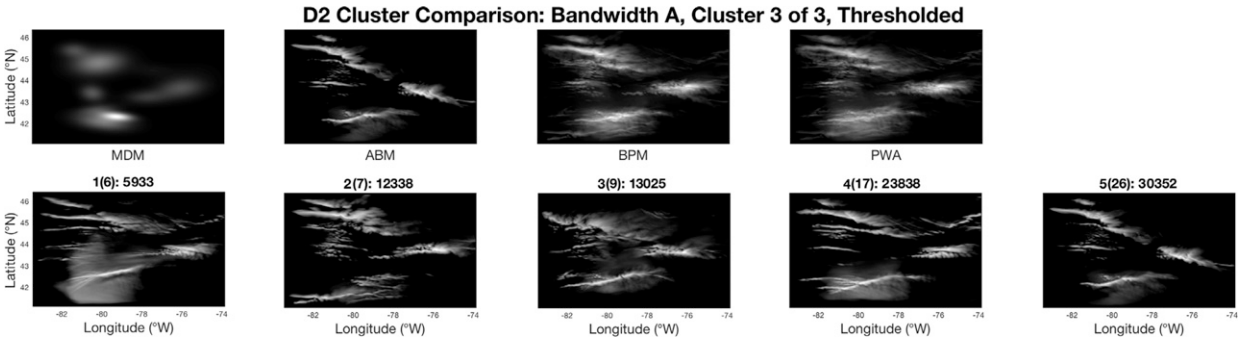

FIG. 18. As Fig. 16, but for the third of the three clusters.

shifted and rotated unrealistically far, often being partially out of the image domain. As a result, the ABM images shown in the D2 results here were calculated with that process disabled, and are simple best members without rotation or translation. It is likely that this effect is due indirectly to the small sample size, and that the reduced number of members in each cluster resulted in a barycenter that was not descriptive enough to match to the members.

4. Conclusions

The authors apply an expanded version of the GEM system to three ensembles of composite reflectivity images: a lake-effect snowband ensemble, a tropical cyclone ensemble, and a severe thunderstorm ensemble. For each ensemble, each member image is thresholded and grayscaled (to aid in image processing, clarity, and object recognition). Two-tiered signatures are then constructed to describe the geometry of each image, and used to compute the ensemble's Wasserstein Barycenter. That centroid is then used with the signatures to compute three types of ensemble mean, each of which has been shown to retain more useful information than a traditional pixel-wise average. The MDM is a visualization of the centroid, the BPM is an improved weighted average, and the ABM is an improved best member.

Each type of mean will have different applications for forecasting, and provides an ensemble mean with a different level of realism. The MDM outlines the key structures present among the ensemble, can be interpreted as a generalized risk map for the intensities it depicts, and visually demonstrates the geometric centroid of the ensemble. As a better weighted average, the BPM is semi-realistic, showing areas that can contain variations on the basic structures of the ensemble, and possibly indicating the current stage of development for the phenomenon depicted. It refines the outlines of the MDM by incorporating ensemble members weighted by WD to the centroid. Finally, ABM can act as a better representative member, indicating a likely position and orientation while maintaining the structure of the ensemble member closest to the centroid in WD space. Each of these has been shown to be closer to a "fully realistic" mean than the PWA, through RMSE comparison with observations.

The authors have improved upon the original GEM system by applying low-intensity thresholding to the source images before processing, and by restoring units and colors of dBZ to the output images and input ensembles for easier interpretation. The thresholding is designed to allow the algorithm and human forecasters to more easily distinguish individual objects among the reflectivity fields.

As demonstrated by the three cases examined above, GEM is capable of capturing the distinct structures present in multiple types of weather phenomena and creating useful means for each. Not all phenomena will be equally well suited to processing via GEM, but the authors have shown that GEM can describe improved ensemble means for at least lake-effect snowbands, tropical cyclones, and severe thunderstorms. The less representative results for the thunderstorm case are potentially influenced by two factors—1) the wider spread of ensemble member structures and locations and 2) the intensity rescaling. However, GEM's consensus is still reasonable in location and relative intensity despite that increase in variance, and the rescaling is an optional feature, intended to preserve maximum intensity and make results easier to read and interpret. These results potentially indicate that GEM as a whole, and the rescaling feature in particular, may be more suited to lower-intensity phenomena. Cases with higher maximum ensemble intensities also tend to have a greater variance in maximum member intensity, contributing to the above problem. In all cases, rescaling the MDM and BPM increases the visibility of the features they describe for the ensemble.

Current applications of D2 clustering to the GEM system indicate that WB-guided D2 is capable of identifying multiple rough scenarios, demonstrating several major possibilities for the ensemble without overloading the forecaster with excessive information. Each scenario has its own mean images, and offers the cluster members, sorted by WD to the cluster centroid, as more specific possibilities within each. The authors note that ensemble sizes of less than 30 members, such as when breaking a 30-member ensemble into two of more clusters, create nonrepresentative ABM images unless the rigid motion is disabled.

This work also has implications for potential innovations in ensemble data assimilation contexts. Respecting the underlying objects or features during data assimilation could result in improved analyses that are more representative of the

underlying phenomena. Potential applications include comparing observations with the prior or adjusting the posterior ensemble in WD space, rather than RMSE, or using a BPM in place of a pixel-wise ensemble mean.

To generate best results for specific phenomena and individual cases, GEM will require some tuning via deliberate choice of bandwidths, support sizes, and threshold values. The authors have not yet devised a system for assisting with this process, but have determined that a threshold value of 0 dBZ is a reasonable starting point for reflectivity, given the variable's logarithmic nature. Additionally, the default set of four bandwidth options included in GEM span a range wide enough to assist in identifying an appropriate granularity for a given application.

GEM can be of aid for forecasting purposes, but may also be useful for other purposes, such as verification or data assimilation. It is possible that GEM-guided analysis will prove useful in identifying atmospheric processes or improving models in ways that have not yet been explored, as a result of introducing object-based averaging to the pool of available techniques. As the GEM system is further developed, its applications should become even more significant for the meteorological community.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Award AGS-1745243. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank two anonymous reviewers and the *Monthly Weather Review* editor for their insights and suggestions.

Data availability statement. All NEXRAD observation data used in this research are openly available from NOAA NCEI (https://www.ncdc.noaa.gov/nexradinv/). All ensemble simulations used in this study, drawn from Saslo and Greybush (2017), Minamide (2018), and Hanson (2016), are stored on The Pennsylvania State University's ICDS server along with the observation data, and they can be made available upon request to the authors.

REFERENCES

AMS, 2008: Enhancing weather information with probability forecasts: An information statement of the American Meteorological Society. *Bull. Amer. Meteor. Soc.*, **89**, 1049–1053, https://doi.org/10.1175/1520-0477-89.7.1041.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Demuth, J. L., B. H. Morrow, and J. K. Lazo, 2009: Weather forecast uncertainty information: An exploratory study with broadcast meteorologists. *Bull. Amer. Meteor. Soc.*, **90**, 1614–1618, https://doi.org/10.1175/2009BAMS2787.1.

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2.

Eipper, D. T., S. J. Greybush, G. S. Young, S. Saslo, T. D. Sikora, and R. D. Clark, 2019: Lake-effect snowbands in baroclinic environments. *Wea. Forecasting*, **34**, 1657–1674, https://doi.org/10.1175/WAF-D-18-0191.1.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta Model. *Climate Dyn.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Greybush, S. J., and S. Saslo, 2018: Insights from a convective-allowing ensemble data assimilation and prediction system for lake-effect snow. *Eighth Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., https://ams.confex.com/ams/98Annual/webprogram/Paper331262.html.

——, S. E. Haupt, and G. S. Young, 2008: The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Wea. Forecasting*, **23**, 1146–1161, https://doi.org/10.1175/2008WAF2007078.1.

——, S. Saslo, and R. Grumm, 2017: Assessing the ensemble predictability of precipitation forecasts for the January 2015 and 2016 East Coast winter storms. *Wea. Forecasting*, **32**, 1057–1078, https://doi.org/10.1175/WAF-D-16-0153.1.

Han, F., and I. Szunyogh, 2016: A morphing-based technique for the verification of precipitation forecasts. *Mon. Wea. Rev.*, **144**, 295–313, https://doi.org/10.1175/MWR-D-15-0172.1.

——, and ——, 2018: A technique for the verification of precipitation forecasts and its application to a problem of predictability. *Mon. Wea. Rev.*, **146**, 1303–1318, https://doi.org/10.1175/MWR-D-17-0040.1.

Hanson, G. S., 2016: Impact of assimilating surface pressure observations from smartphones on a regional, high resolution ensemble forecast: Observing system simulation experiments. M.S. thesis, Dept. of Meteorology and Atmospheric Science, The Pennsylvania State University, 47 pp., https://etda.libraries.psu.edu/catalog/28709.

Hirschberg, P. A., and Coauthors, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.*, **92**, 1651–1666, https://doi.org/10.1175/BAMS-D-11-00073.1.

Hong, S.-Y., and J. Lim, 2006: The WRF Single-Moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Janjić, Z. I., 1990: The step-mountain coordinate: Physical package. *Mon. Wea. Rev.*, **118**, 1429–1443, https://doi.org/10.1175/1520-0493(1990)118<1429:TSMCPP>2.0.CO;2.

Joslyn, S., K. Pak, D. Jones, J. Pyles, and E. Hunt, 2007: The effect of probabilistic information on threshold forecasts. *Wea. Forecasting*, **22**, 804–812, https://doi.org/10.1175/WAF1020.1.

Kalnay, E., 2002: *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 341 pp.

Karstens, C. D., and Coauthors, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30**, 1551–1570, https://doi.org/10.1175/WAF-D-14-00163.1.

Kowaleski, A. M., and J. L. Evans, 2016: Regression mixture model clustering of multimodel ensemble forecasts of Hurricane Sandy: Partition characteristics. *Mon. Wea. Rev.*, **144**, 3825–3846, https://doi.org/10.1175/MWR-D-16-0099.1.

Lee, J. A., L. J. Peltier, S. E. Haupt, J. C. Wyngaard, D. R. Stauffer, and A. Deng, 2009: Improving SCIPUFF dispersion forecasts with NWP ensembles. *J. Appl. Meteor. Climatol.*, **48**, 2305–2319, https://doi.org/10.1175/2009JAMC2171.1.

Li, J., and F. Zhang, 2018: Geometry-sensitive ensemble mean based on Wasserstein barycenters: Proof-of-concept on cloud simulations. *J. Comput. Graph. Stat.*, **27**, 785–797, https://doi.org/10.1080/10618600.2018.1448831.

——, S. Ray, and B. G. Lindsay, 2007: A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, **8**, 1687–1723, https://doi.org/10.5555/1314498.1314555.

Minamide, M., 2018: On the predictability of tropical cyclones through all-sky infrared satellite radiance assimilation. Ph.D. dissertation, The Pennsylvania State University, 202 pp., https://etda.libraries.psu.edu/catalog/15449mum373.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

National Research Council, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts.* National Academies Press, 124 pp., https://doi.org/10.17226/11699.

Nehrkorn, T., B. Woods, T. Auligné, and R. N. Hoffman, 2014: Application of feature calibration and alignment to high-resolution analysis: Examples using observations sensitive to cloud and water vapor. *Mon. Wea. Rev.*, **142**, 686–702, https://doi.org/10.1175/MWR-D-13-00164.1.

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

Rachev, S.-T., 1985: The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.*, **29**, 647–676, https://doi.org/10.1137/1129093.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Saslo, S., and S. J. Greybush, 2017: Prediction of lake-effect snow using convection-allowing ensemble forecasts and regional data assimilation. *Wea. Forecasting*, **32**, 1727–1744, https://doi.org/10.1175/WAF-D-16-0206.1.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Wea. Forecasting*, **12**, 809–818, https://doi.org/10.1175/1520-0434(1997)012<0809:AEFP>2.0.CO;2.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487, https://doi.org/10.1175/2008MWR2415.1.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting*, **20**, 101–111, https://doi.org/10.1175/WAF-831.1.

Wu, C.-C., G.-Y. Lien, J.-H. Chen, and F. Zhang, 2010: Assimilation of tropical cyclone track and structure based on the ensemble Kalman filter (EnKF). *J. Atmos. Sci.*, **67**, 3806–3822, https://doi.org/10.1175/2010JAS3444.1.

Ye, J., and J. Li, 2014: Scaling up discrete distribution clustering using ADMM. *Proc. Int. Conf. Image Processing (ICIP)*, Paris, France, Institute of Electrical and Electronics Engineers, 5267–5271, https://doi.org/10.1109/ICIP.2014.7026066.

——, P. Wu, J. Z. Wang, and J. Li, 2017: Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Trans. Signal Process.*, **65**, 2317–2332, https://doi.org/10.1109/TSP.2017.2659647.