UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MULTI-SCALE OBJECT-BASED PROBABILISTIC FORECAST EVALUATION OF WRF-BASED CAM ENSEMBLE CONFIGURATIONS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

ANDREW WILKINS

Norman, Oklahoma
2020

MULTI-SCALE OBJECT-BASED PROBABILISTIC FORECAST EVALUATION OF WRF-BASED CAM ENSEMBLE CONFIGURATIONS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Aaron Johnson, Chair

Dr. Xuguang Wang, Co-Chair

Dr. Jason C. Furtado

TABLE OF CONTENTS

Abstract

Recent developments in numerical weather prediction have included increased usage of ensemble forecasts in contrast to single, deterministic forecasts. In particular, convection-allowing model (CAM) ensembles have been utilized as they contain a distinctive ability to predict convective initiation location, mode and morphology. Such information can be extremely useful to forecasters predicting severe weather threats associated with particular storm modes and their morphological transitions spatially and temporally. Numerous studies verifying CAM ensemble forecast performance have been conducted. However, the primary focus from these studies have been on neighborhood-based verification of spatial coverage rather than convective mode and morphology. A limitation of neighborhood-based verification methods is their inability to adequately represent storm scale features of greatest subjective interest to forecasters. In contrast, as an alternative to neighborhood-based methods, recent object-based probabilistic framework has been introduced to assess forecasts, providing a unique setting to re-evaluate aspects of optimal CAM ensemble design with a focus on mode and morphology prediction.

Herein, we adopt an optimized object-based probabilistic (OBPROB) forecasting method in conjunction with a traditional neighborhood-based (NMEP) method to evaluate forecasts of four separately designed 10-member ensembles. The ensemble configurations evaluated include single-model, single-physics (SMSP) configurations, single vs. multi-model, single vs. multi-physics, and multi-model vs multi-physics. Due to the implementation of multiple verification techniques and separate verification of objects on different spatial scales, comparisons of forecast performance in terms of mesoscale precipitation locations and forecasted storm mode and morphology can be made explicitly, leading to insight on optimal CAM ensemble design for specific storm modes and morphologies. Both OBPROB and NMEP frameworks show ensembles with a more diverse member-to-member design improve probabilistic forecasts over single-model, single-physics designs through greater sampling of forecast uncertainties. However, individual case studies suggest each methods' systematic results are reliant on separate forecast features. For example, subjective analysis shows neighborhood-based verification, even at high reflectivity thresholds, are impacted by both convective and stratiform precipitation whereas the OBPROB method explicitly focuses on convective precipitation. OBPROB verification of storm morphology forecasts also display the need for probabilistic calibration to improve ensemble reliability.

1. Introduction

Numerical weather prediction (NWP) since the early 2000s has taken advantage of advances in computational resources through analysis of high-resolution, convection-allowing models (CAMs; Done et al. 2004; Kain et al. 2006, 2008; Weisman et al. 2008; Schwartz et al. 2009). More recently, specific emphasis has been placed on improving convection-allowing ensemble forecasts (CAEs; Clark et al. 2010; Schwartz et al. 2010, 2015; Xue et al. 2010; Johnson and Wang 2012, 2017; Duda et al. 2014, 2016, 2017; Romine et al. 2014; Johnson et al. 2017; Gasperoni et al. 2020; Johnson et al. 2020, hereafter J20). For example, Schwartz et al. (2010) demonstrate that high precipitation and severe weather forecasts of CAEs have potential to be enhanced by post-processing methods while providing forecasters with simple, accessible products. Duda et al. (2017) showed underdispersion of CAE forecasts can be improved through incorporation of land surface model (LSM) perturbations and indicated the need for inclusion in CAE designs. Furthermore, Gasperoni et al. (2020) found multi-model and multi-physics ensembles are superior ensemble designs with respect to single-model, single-physics designs in the context of mesoscale precipitation location forecasts. Such CAE studies suggest certain ensemble designs are optimal depending on the forecast aspect of interest.

The benefit of using convection-allowing models and ensembles is that they contain a distinctive ability to predict convective initiation location, mode and morphology. This ability is exceptionally useful to operational forecasting in association with predicted storm mode and morphology. Accurately predicting convective mode and morphology can have significant impacts on anticipated severe weather, as discrete, isolated supercells are more prone to tornadoes and large hail, whereas organized linear systems present a much greater threat for severe straight-line winds and flash flooding. Therefore, adequately depicting optimal CAE design with respect to storm mode and morphology can translate to impacts related to operational forecast improvements.

Experiments comparing different CAE designs in convective events have mainly placed emphasis on spatial coverage through neighborhood-based verification methods or subjective evaluations (e.g., Clark et al. 2010; Clark 2017; Gasperoni et al. 2020; Schwartz et al. 2010; Schwartz and Sobash 2017; Carlberg et al. 2018). Neighborhood-based methods provide an improved framework to traditional gridpoint-based methods as high amplitude features are considerably less sensitive to spatial displacements (Ebert 2008). A limitation associated with neighborhood-based methods are their limited ability to quantify the fidelity of model simulated convective scale features through the same smoothing process that results in less sensitivity to spatial displacements (Gilleland et. al 2013; J20). To alleviate these limitations, object-based frameworks have been documented to retain convective scale details while providing objective information about forecast aspects of interest (Davis et al. 2006a,b,2009; Johnson et al. 2011a,b,2013,2020; Wolff et al. 2014; Clark et al. 2014; Stratman and Brewster 2017).

Comparison of storm morphology forecasts with different CAE configurations have yet to be evaluated directly through objective verification metrics. Application of object-based frameworks, have primarily been utilized in deterministic rather than ensemble forecast settings (e.g., Davis et al. 2006a,b,2009; Johnson et al. 2011a,b,2013; Wolff et al. 2014; Clark et al. 2014; Stratman and Brewster 2017). Of select studies that have focused on forecasted storm mode (e.g., Carlberg et al. 2018), verification has typically been evaluated subjectively. An appropriate verification framework well suited to provide probabilistic guidance on ensemble forecasted storm mode and morphology was recently developed and applied in J20. Denoted as an object-based probabilistic (OBPROB) verification tool, the J20 OBPROB framework provides an opportunity to evaluate CAE design in terms of probabilistic forecasts of storm mode and morphology.

Explicitly forecasted in CAEs, storm mode and morphology require uncertainty of the physical processes that are closely related to their fruition (Schumacher and Clark 2014). Select studies have relied on data assimilated lateral boundary condition (LBC) and initial condition (IC) perturbations within a single-model, single-physics ensemble to sample forecast uncertainty and achieve member-to-member spread (e.g., Schwartz et al. 2015,2019; Gowan et al. 2018). Motivation behind the usage of such simple design ensembles is rooted in their ease of maintenance, postprocessing and statistical interpretations (Schwartz et al. 2019). However, many studies verifying CAE forecasts have found single-model, single-physics ensembles still lack forecast spread needed to more accurately represent forecast uncertainty (e.g., Duc et al. 2013; Schumacher and Clark 2014; Schwartz et al. 2014; Johnson and Wang 2017). To improve ensemble spread, incorporation of within ensemble physics parameterization diversity has been found to more adequately distribute latent heating profiles (e.g., Schumacher and Clark 2014) and associated cold pool evolution (e.g., Johnson and Wang 2017). Therefore, such studies have objectively analyzed forecast variables related to forecasted storm mode and morphology, however have not explicitly evaluated CAE storm morphology forecast performance. Nonetheless, it is hypothesized that larger variable variance provided by a multi-physics ensemble design will result in improved storm morphology spread and lead to greater sampling of forecast uncertainty.

A second method in designing CAEs is related to dynamical core diversity. In a multi-model ensemble, members are comprised of two or more dynamical cores (e.g., Ebert 2001; Wandishin et al. 2001; Eckel and Mass 2005; Candille 2009; Johnson and Wang 2012, hereafter JW12; Melhauser et al. 2017; Gasperoni et al. 2020). Much like a multi-physics design, the use of multiple dynamical cores in a single ensemble forecasting system has been found to be advantageous with respect to single-model ensembles due to better sampling of flow uncertainty (e.g., Candille 2009). Although evaluated at coarser grid-spacing, Wandishin et al. (2001) found even members from a less skillful model can increase forecast skill in a multi-model design. Such features are desirable in a forecasting context, however multi-model ensembles also have downfalls. Gasperoni et al. (2020) notes that in addition to potential clustering of forecasts in a multi-model ensemble, costly increases in computational resources are needed to maintain multiple dynamical cores. Ideally though, apart from greater resource demand, multi-model ensembles have been described as a "worthwhile compromise" between reliability and resolution pitfalls found in their single-model counterparts (e.g., Candille 2009). Translating these findings to a convective scale forecast, it is hypothesized that a multi-model ensemble will improve upon constituent model forecasts of storm morphology due to increased sampling of uncertainty and better representation of error growth.

The primary purpose of this study is to objectively evaluate the impacts of ensemble design choices on convective mode and morphology using a further developed and optimized OBPROB technique. In addition, a secondary goal of the study includes supplementary analysis of OBPROB forecasts and how they compare to a traditional neighborhood-based method will be completed to demonstrate the usefulness of the OBPROB method and exploit distinct verification method sensitivities. Through explicit analysis of ensemble storm morphology forecasts in accordance with comparison to conventional neighborhood-based techniques, a better understanding of ensemble design choices and their impacts of forecasts of storm morphology will be gleaned. A third goal of this study is to further develop the OBPROB products to improve their usefulness for operational convective forecasting settings.

The rest of the paper is organized as follows. Section 2 describes the OBPROB method in full, including object definition, matching and probabilities, in addition to specific optimizations made to the method for this study. Section 3 describes the ensembles used, ten retrospective case studies, and separate verification metrics to be shown in the results. Section 4 describes the results, which are split into multiple subsections detailing the overall objective verification, subjective interpretation and comparison to traditional neighborhood-based results. The main conclusions are then summarized and discussed in section 5.

2. OBPROB methodology

To assess forecast performance, many studies focusing on convective forecasts have utilized the annual NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs) hosted by the Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL; Clark et al. 2012b;, 2018; Gallo et al. 2018, 2018). Studies such as J20, exploit the opportunity to dissect forecaster input from the HWT, further molding and tuning forecasting methods to bridge the gap between forecaster desires and model development constraints. For example, J20 used forecaster specifications from the 2017 and 2018 HWT experiments to finely tune forecast parameters used in the object-based probabilistic (OBPROB) forecasting technique that was introduced in JW12. The OBPROB technique is advantageous in that convective mode of modeled storms are retained for verification, which is essential to severe weather forecasting (e.g., Gallus et al. 2008; Duda and Gallus 2010; Smith et al 2012; Pettet and Johnson 2003). In addition, plots generated from the OBPROB method concisely display probabilities of explicitly resolved convective scale features, providing a verification metric that can be evaluated while greatly reducing time consuming subjective interpretation. Applicable to ensemble settings, the tuned OBPROB method in J20 creates a unique framework to verify storm mode and morphology forecasts of convection-allowing ensembles. Figure 1 visualizes the OBPROB technique's advantages through a cartoonish eight-member ensemble. Assigning a probability of being matched to each object in Figure 1a, member 1's forecast contains a red object of 12.5% probability and a blue object of 87.5% probability as seven out of eight members contain a similar object. Therefore, it can be seen more explicitly that the OBPROB method can concisely convey both the explicitly resolved convective scale structure and forecast uncertainty in Figure 1a, while also providing a verification metric that can evaluate the ensemble probability distribution of convective storm mode.

The OBPROB method procedure starts with first defining objects, then moves to matching objects, and finishes with object probability and verification. Objects are defined in each members' reflectivity forecast as described in section 2a. The object matching process takes each individual members' forecast objects and compares distinctive physical object attributes to other member forecasts to define object similarity (section 2b). Finally, object probabilities are assigned based on the total number of members with a matching object and are plotted onto a single, easy to interpret plot (section 2c). The resultant plot and corresponding object probabilities are utilized in objective verification of forecasted storm morphology through observed objects from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) system (section 2d).

a. Object definition

Object definition in this study is similar to the method outlined in J20. First, a 6-km (2-grid point) Gaussian smoother is applied to ensemble reflectivity fields to reduce grid-scale noise while retaining convective scale features of interest to the storm morphology forecast. Objects are then defined using a predefined dBZ threshold above which, all closed contoured values are outlined. For each object defined, attributes are then calculated, including object area, which is the total number of grid points within the object, object longest axis, which is the length of grid points of the largest axis, object aspect ratio, which is the length of the object divided by the width, and object centroid location, which is a simple x,y coordinate. Objects that contain an area less than 42 grid points are then omitted to remove any aliased objects whose diameter falls below the effective resolution of the ensemble, 7Δx (J20; Skamarock 2004).

b. Object matching

As in J20, the object matching process follows a much more simplified version compared to previous methods (e.g., MODE from Davis et al. 2009). The MODE utility incorporates a piecewise linear similarity function for each object attribute, basing additive total interest on each individual interest value through associated confidence and weight values determined by the user (e.g., Davis et al. 2009). Comparatively, the OBPROB simplification involves removal of the piecewise similarity function and corresponding confidence and weight of interest values while integrating the Gaspari and Cohn (1999) function in its place. The Gaspari and Cohn function used in this study is an approximated form which contains *e*-folding values and a value of zero at a specified location, giving the function an approximate Gaussian shape (J20). Figure 2 details the explicit shape of the functions used for each of the three matching parameters used to define total interest between two objects. Specific *e*-folding values are 200-km for object centroid distance (Figure 2a), 0.33 for object area (shown as 1.0-0.33 since larger ratios indicate larger interest; Figure 2b) and 0.5 for object aspect ratio (Figure 2c). In accordance with subjective interpretation by the author and participants in the HWT, these parameters have been tuned to maximize realistic object matching, including a change in the aspect ratio *e*-folding value from 0.2 (e.g., used in J20), to 0.5 (Figure 2c). Based on two objects individual object attribute interest values, a total interest, *I*, is calculated:

$$I_{total} = f_{a1}\ x\ f_{a2}\ x\ f_{a3} \tag{1}$$

From Eq. 1, as in J20, total interest can be defined as the product of each individual attribute interest value, f_a. Due to the multiplicative nature of the total interest calculation, confidence values and associated weights are not needed as the total interest *I* will be automatically weighted by the interest values themselves through changes in the *e*-folding interest function (i.e., if an attribute interest value is near-zero, the resultant total interest will also be near-zero, J20). If a pair of objects total interest exceeds that of a predefined threshold, the two objects will match, increasing the objects' probability. Here we use a matching threshold of 0.35. This value is adjusted from J20, which used a matching threshold of 0.2. The adjustment is due to a shift in aspect ratio interest calculations to a "ratio-of-ratios" calculation, similar to what is already used for object area, which takes the compliment of the ratio of aspect ratios of two matched objects divided by the complement of the *e*-folding ratio. The matching threshold of 0.35 was then tested through subjective analysis and was found to be in agreement realistic interpretations.

c. Object probabilities

Previous studies have suggested multiple methods regarding optimal post-processing of object-based, probabilistic ensemble forecasts. One option would be to choose a control member from which object probabilities are calculated and displayed (JW12). While control members may have less error on average, this method may limit the distribution of object possibilities, through misrepresentation of the "center" of the forecast distribution. Furthermore, for this study, each member forecast is equally likely, making a control member forecast potentially suboptimal. Other proposed methods include choosing a representative member prior to examination, however, Schwartz et al. (2014) suggests that such a method has yet to be shown to systematically perform superiorly to a randomly selected member. Potential reasons for representative members failing to be consistently more skillful could be resultant from elimination of lower probability possibilities (J20). In comparison to previous methods, the OBPROB method used in this study constructs a plot that concisely displays representative object probability distributions, while retaining lower probability possibilities. The step-by-step method, which is also outlined in J20, is as follows:

1. Compile all ensemble forecasted objects into a single array with corresponding probabilities calculated from the number of ensembles with a matching object (i.e., 8 out of 10 members with a matching object equals a probability of eighty percent).
2. Sort object probabilities in descending order, with ties in probability going towards the object with highest average total interest from objects in matching members.
3. Plot the highest probability object.
4. Remove the highest probability object, in addition to all associated matching objects from the total array of objects, giving a new, shortened array.
5. Repeat steps 2-4 until no objects remain in the array.

Figure 3 outlines a real-world example of this process, showing explicitly how individual member analysis (Figure 3a-j) can be extremely time consuming. Furthermore, Figure 3l indicates the prime issue in neighborhood-based forecasts smoothing out convective scale details as one singular high probability contour extends from Minnesota through Kansas, while an object-based paintball plot (Figure 3k), which plots a simple overlay of each individual member forecasts onto a single plot, shows somewhat more explicitly two separate linear squall lines are forecasted. The OBPROB plot (Figure 3m) can be interpreted similarly to the paintball plot, however, a key difference is the simplicity of the OBPROB plot. The simplistic nature arises due to redundant objects being plotted only once (i.e., the red object in Nebraska had 8 out of 10 members with a matching object, leaving 8 objects to be plotted once; represented by the red Nebraska object). Additionally, Figure 3m shows the OBPROB plot retains lower probability objects, characterized by the blue object in western Nebraska. Thus, the OBPROB plot, while accelerating subjective interpretation of explicitly resolved storm morphology forecasts through simplified plots, also provides a unique verification metric through a distribution of object probabilities, which can be used to evaluate ensemble performance in relation to forecasted storm mode and morphology.

d. Object verification

The verification procedure of forecast objects to observation objects is comparable to the object matching process described above in section 2b. Every object represented in the final OBPROB plot, is compared to observation objects through verbatim interest functions and total matching thresholds for consistency. The main differences arise due to theoretical interpretation of what an object represents. Similar to J20 and JW12, final OBPROB plot objects are interpreted to represent a discrete set of "events" forecasted by the ensemble. As a consequence, it can be expected that a single forecast object is only allowed to match to one observation object. However, another result of this interpretation is that multiple forecast objects can potentially match to a single observed event. In other words, because each individual object represented in the OBPROB plot is representative of a separately forecasted event, each object is given an opportunity to match to any observed event. In practice, for the vast majority of scenarios, the OBPROB method keeps this from occurring since similar objects are merged into a single, higher probability object. The probabilities associated with each respective object then, determines how skillful the forecasted event is. For example, if a ten percent probability object matches to an observed object, the subsequent skill of that forecast is deemed poor as only one out of ten members forecasted a similar object. On the contrary, if a ten percent probability object does not match to an observed object, the ensuing skill is deemed good since only one out of ten members forecasted such an event.

e. OBPROB optimization

The OBPROB method described herein and in J20, was also further optimized with respect to ensemble storm mode and morphology forecasts. The optimizations can be split into three main sections: Model bias adjustment through reflectivity percentile analysis developed by Skinner et al. (2018), stratiform observation object filtering via a similar percentile analysis method, and scale separation of objects into single-cell, multi-cell, and mesoscale organized (denoted by meso-gamma, meso-beta, and meso-alpha, respectively).

The first optimization in the OBPROB methodology for this study was accounting for model bias in reflectivity fields. Subjective analysis revealed the need to adjust for model reflectivity bias, as certain members resulted in forecast objects of unrealistic size and shape due to precipitation amplification. Table 1 details a hypothetical example of the technique used which is the same method developed and used in Skinner et al. (2018). Here, we analyze the observation reflectivity distribution, locating the percentile at which 40 dBZ occurs, which is the value that is used to define observation objects in this study. We then take the observation reflectivity percentile and compare it to the forecast reflectivity distribution, finding the corresponding forecasted dBZ value at the same percentile value. For example, in Table 1, the 40 dBZ contour in the observations corresponds to the 85th percentile of the observation distribution. Therefore, the respective forecasted dBZ value used to define forecast objects must correspond to the 85th percentile of the forecast distribution. In Table 1, the 85th percentile correlates to the 45 dBZ contour, adjusting for a five dBZ model bias. Relative forecast dBZ values used to define objects ranged from 55 dBZ to 32 dBZ depending on the ensemble and ensemble member and time of day, indicating bias adjustment was necessary for consistency between ensembles, within ensemble members, and diurnal variations. Bias correction was performed singularly for single-model, single-physics ensembles, however, was implemented separately for each member for the multi-physics ensemble. The end result led to more continuous object matching from member to member, in addition to directly justifiable comparisons to observation objects and their attributes.

Since this study has a specific focus on verification of storm mode and morphology, the second optimization to the OBPROB method included stratiform observation object filtering. Illustrated in Figure 4, radar bright band effect often resulted in observation objects coinciding with stratiform precipitation not associated with the actual observed storm mode of interest (Figure 4a). Such objects resulted in subsequent matching of forecast objects, leading to somewhat ambiguous verification conclusions. In order to account for such stratiform objects impacting storm morphology verification, a similar reflectivity percentile analysis used to adjust reflectivity bias was applied. In this context, a 46 dBZ threshold criterion for the 95th percentile of within-object reflectivity was implemented, effectively reducing trailing stratiform objects while retaining convective objects, allowing for a greater focus on verification of convective storm mode and morphology (Figure 4b).

The final optimization in the OBPROB method included separation of forecast objects based on convective organization scale (i.e, single-cell, multi-cell, and mesoscale organized). Prior works on convective organization (e.g., Houze 1993), suggest there are different dynamical complexities at different convective organization scales. Houze (1993) defines these dynamical discontinuities to be different for systems of longest horizontal axis greater than 100-km. For this study's purpose, the 100-km (33-grid point) horizontal length is used as a starting point from which we adjust to evenly distribute object sample sizes into each convective organization scale. Category bounds are 45-km (15 grid points) and 75-km (25 grid points), where objects containing a longest horizontal axis less than 45-km are considered single-cell, between 45 and 75-km are multi-cell, and greater than 75-km are mesoscale organized.

Due to different dynamical complexities at each organizational scale, distinctive predictability of convective system attributes is expected. It is widely accepted that this is true as attributes such as spatial location, size, and shape of largely organized convective systems are much more predictable than attributes for discrete, loosely organized convection (e.g., Zhang et al. 2006; Trentmann et al., 2009; Barthlott et al. 2011; Keil et al. 2014). To ensure the verification process is aligned with the predictability at each respective scale, objects of mesoscale organized convective systems, multi-cell systems, and single-cell systems are verified separately with varying verification techniques. Figure 5 illustrates the differing verification methods for each scale separation described. For mesoscale organized systems (Figure 5a), the novel OBPROB method is used solely as some predictability of storm mode and morphology is expected. For smaller, less organized convection (i.e., multi-cellular and single-cellular), object probabilities are returned to grid point space through a Gaussian-smoothed contour plot (Figure 5b,c). Through the return to a grid point space, relative uncertainty about specific object locations are accounted for, matching the predictability we expect at these scales. Both meso-beta and meso-gamma scales use a Gaussian radius of 10 grid points (30-km), which was subjectively shown to best represent realistic probabilities.

3. Experimental design

a. Experiment description

To address the impacts ensemble design has on storm mode and morphology forecasts, four ensemble-to-ensemble comparisons from four different ensembles were conducted, as outlined in Tables 1 & 2. Each ensemble consisted of ten members (a control member and 9 re-centered EnKF perturbations) initialized from the final EnVar analysis described in Gasperoni et al. (2020). Of the four ensembles listed in Table 2, NMMB and ARW-SP were single-model, single-physics designs. NMMB consisted of Ferrier-Aligo microphysics (Aligo et al. 2018), Mellor-Yamada-Janjic (MYJ) boundary layer physics (Janjic 1994), and NOAH land surface model scheme (Tewari et al. 2004), whereas ARW-SP included Thompson microphysics (Thompson et al. 2008; Thompson and Eidhammer 2014), Mellor-Yamada-Nakanishi-Niino (MYNN) boundary layer physics (Nakanishi and Niino 2009), and the RUC land surface model scheme (Benjamin et al. 2004). ARW-MP was a multi-physics design containing four different microphysics schemes, three planetary boundary layer (PBL) schemes and two land-surface model (LSM) schemes. The four microphysics schemes included in ARW-MP were the aforementioned Thompson scheme, National Severe Storms Laboratory (NSSL) bulk two-moment scheme (Mansell et al. 2010), Morrison two-moment scheme (Morrison et al. 2009), and recently developed P3 scheme (Morrison and Milbrandt 2015). ARW-MP PBL parameterizations consisted of MYJ, MYNN and the Yonsei University Scheme (YSU; Hong et al. 2006). Finally, LSM schemes utilized in designing ARW-MP were NOAH and RUC. The fourth and final ensemble, MM, was composed of five members from NMMB and ARW-SP; the only multi-model ensemble in this study. Since MM consisted of NMMB and ARW-SP members, corresponding physics parameterizations were also kept.

Each ensemble-to-ensemble comparison, outlined in Table 3, is devised to address specific impacts certain ensemble design choices have on storm morphology forecasts. By doing so, information ranging from how specific scheme choices affect the forecast, to how different design diversities affect the forecast, can be gleaned. In other words, organizing the experiment through ensemble-to-ensemble comparisons allows evidence to surface regarding which environments and/or storm morphologies specific ensemble designs are more skillful in forecasting.

As in Gasperoni et al. (2020), all ensembles were verified over ten retrospective case studies from 2015 to 2016 (Table 4). The cases selected make up a diverse set of synoptic scale forcing, geographical location, time of day, and observed storm mode and morphologies, enhancing the ability to address specific ensemble design sensitivities.

All object-based forecasts were then compared to a traditional, neighborhood-based forecast. Due to this study's focus on extreme events (i.e., strong convective precipitation), neighborhood maximum ensemble probability (NMEP) was the selected method as is recommended by Schwartz and Sobash (2017). Previous studies suggest that optimally addressing model error may rely on combinations of techniques to address uncertainties (Berner at el. 2015; Duda et al. 2016; Gasperoni et al. 2020; Jankov et al. 2019). Here we composed a method comparison to assess more explicitly how ensemble forecasts of approximate mesoscale precipitation locations coincide with respective forecasts of storm mode and morphology and if they are sensitive to similar initial forecast uncertainties. Additionally, the utilization of multiple verification methods allowed examination of the OBPROB method's ability to effectively resolve convective scale details and provide unique probabilistic storm morphology information; something the OBPROB method is uniquely suited to evaluate.

b. Verification methods

Verification of object-based and neighborhood-based forecasts were performed in terms of forecasted composite reflectivity fields. To ensure consistency between the separately defined NMMB and ARW domains, the MM ensemble used ARW member forecasts that were bilinearly interpolated to the NMMB grid. Observation verification fields were radar mosaics obtained from the MRMS. Due to discrepancies between the ARW and NMMB domains, observations also required consistency through bilinear interpolation. Here, observation objects and their attributes from the ARW domain were kept, with respective object locations interpolated to match corresponding NMMB locations. Doing so ensured observation objects themselves, in addition to object attribute interest values, were consistent between verification domains. Ensuing forecast verification was then performed based on relative ensemble domain, with ARW ensembles (ARW-SP & ARW-MP) verified using ARW observations, and NMMB ensembles (NMMB & MM) verified using ARW observations interpolated to the NMMB grid.

Objective verification metrics of ensemble forecasts were separated by method. Object-based verification was further separated by storm organizational scale described above (meso-alpha, meso-beta, meso-gamma): Meso-alpha object verification included the Brier score (BS; Brier 1950), while meso-beta and meso-gamma verification utilized the fractions Brier score (FBS; Roberts and Lean 2008; Schwartz et al. 2010). Each ensemble BS was then used in ensemble-to-ensemble comparisons, effectively making reported values a BSS relative to the ensemble in comparison. Other object-based forecast metrics included resolution, reliability and sharpness, calculated to analyze ensemble spread. In addition, for the object-based verification, ensemble object attributes were verified. Normalized attribute distributions were completed for three attributes of interest: object area, object longest axis and object aspect ratio. Forecasted distributions were then compared to corresponding observation object attribute distributions and subjectively evaluated.

Neighborhood-based verification uniformly used the Brier skill score (BSS; Wilks 2006) for all three corresponding radii. Specifically, the BSS equation used is as follows:

$$BSS = 1 - FBS/FBS_{ref} \tag{2}$$

In Eq. 2, the Fractions Brier score (FBS) is calculated as a domain-wide mean squared difference of NMEP fields and observed neighborhood probability fields. In order to avoid ambiguity from the BSS calculations, the reference forecast (FBS_{ref}) was calculated as a climatological probability of event occurrence averaged over every grid point within the domain (Hamill 2005). The overall ensemble BSS is then used in ensemble-to-ensemble comparison scores, creating a new verification value relative to the compared ensemble.

Statistical significance of probabilistic verification for both techniques included a one-sided permutation resampling method used in JW12. To ensure independent samples, separate procedures were taken based on the verification technique. For the object-based approach, subsampling procedures were separated based on convective organization scale. For mesoscale organized objects, since objects are not confined to grid points in a traditional sense, we treat them as forecasts for separate events and thus as independent samples themselves. Additionally, because each ensemble is forecasting a slightly different set of "events", statistical significance at this scale does not include paired samples between ensembles. Instead, each forecasted object in an ensemble-to-ensemble comparison is randomly reassigned to an ensemble in the permutation resampling test.

For multi-cell and single-cell objects, by returning to a grid point based space, a different methodology is used. Here, for uniformity between ensembles, we define subsample subdomains based on observation object locations. Depending on the organizational scale category the observation object falls into, a subdomain is defined by distance from the object centroid location: 200-km for meso-beta, and 150-km for meso-gamma. By implementing the previously used e-folding centroid location interest distance for object matching, if two observation object subdomains overlap, the associated objects can be deemed as correlated and co-joined into one singular subsample. Independent samples are then separated, with all forecasted values at grid points within an observation object subdomain counting toward one sample and all forecasted values outside of the observation object subdomains counting towards a separate sample. To test the credibility that each subsample taken in this method is independent, scatterplots are constructed and are displayed in Figure 6. With Pearson correlation coefficients plotted in the top left of each panel, the relationship between different subsamples is small, as correlation coefficients often reside at or below 0.3 (i.e., the typical accepted threshold for weak correlation; Wilks 2006). Therefore, since subdomains generated contain a weak relationship, they are deemed to be independent samples. In a much simpler manner, for the neighborhood-based approach, a single daily contingency table sample from each case was used (e.g., Gasperoni et. al 2020; Hamill 1999).

4. Results

a. Objective Verification

1) OBJECT ATTRIBUTE DISTRIBUTIONS

i) ARW-SP VS NMMB (SMSP)

Normalized distributions of each object attribute and every ensemble are calculated in Figure 7. Objects are aggregated over every forecast hour and case study giving systematic objective verification. Object attribute distributions are then compared to respective observation object attribute distributions subjectively and objectively through resampling tests for statistical significance. Resampling tests were conducted by first calculating overall distribution differences as seen subjectively in Figure 7. Cumulative differences for all objects greater than and less than 300 km for object area are separately compared to 1000 resampled differences where each of the 10 cases were utilized with replacement to simulate a full variety of possible storm mode cases. Significance is then tested at the 95% confidence interval (i.e., if the overall distribution difference is more than 95% of resampled differences, the actual distribution difference is deemed significant). Resampling tests revealed ensemble forecast distributions produce statistically significant differences when compared to observations, indicating the distribution discrepancies in Figure 7 are larger than sampling variability and are not case dependent. Contrarily, object attribute distribution differences from ensemble to ensemble did not result in statistically significant differences. Figure 7a,d show overall results for object area for NMMB and ARW-SP, respectively. Compared to observations, NMMB shows generally higher forecast frequencies for smaller scale objects below 300 grid points, and generally lower frequencies for medium to large scale objects above 300 grid points. Similarly, compared to observations, ARW-SP also generates more relative frequencies of objects with small area and less frequencies for objects with larger area. Although the transition zone where both ensembles cross the observation distribution is slightly different, overall object area distributions between the SMSP ensembles are largely similar. For object longest axis, NMMB (Figure 7b) results display comparable distributions for objects of short longest axis with lower frequencies of longer objects compared to observations. Conversely, ARW-SP longest axis frequencies (Figure 7e) show larger relative frequencies for objects of short longest axis. With ultimately both distributions eventually falling below observation frequencies as longest axis increases, the main discrepancies between NMMB and ARW-SP are confined to short object frequencies. The largest differences between SMSP ensemble forecasts are the object aspect ratio frequency distributions. NMMB (Figure 7c) exhibits analogous results to longest axis results with nearly equivalent forecast frequencies for large aspect ratios above ~0.45 and lower frequencies for small aspect ratios above ~0.45. In comparison, ARW-SP results (Figure 7f) show an increasingly obvious over-forecasting of objects with considerably larger aspect ratio frequencies above ~0.6 and under-forecasting of smaller aspect ratio frequencies below ~0.6.

All SMSP object attribute plots suggest, in varying levels of clarity, that ARW-SP forecasts additional small, short, circular objects compared to NMMB. This notion is supported mostly by increased frequencies in large aspect ratios, but can also be seen for small longest axes. While the increase in forecasts of small, short, circular objects may slightly skew ARW-SP forecast distributions, the lack thereof for NMMB forecasts indicates a different process is responsible for distribution differences at large area, length, linear objects. It is hypothesized from these distribution differences that NMMB struggles to effectively grow upscale small, circular objects into larger linear systems. Supplemental subjective analysis will provide further detail regarding this hypothesis in section 4b.

ii) ARW-SP VS ARW-MP (SPMP)

The second and third rows of Figure 7 highlight object attribute results for ARW-SP and ARW-MP, respectively. Results for ARW-MP object area (Figure 7g), show similar features to ARW-SP (Figure 7d) with higher forecast frequencies of small area objects and lower frequencies of large area objects compared to observations. In contrast, ARW-MP object longest axis forecasts (Figure 7h) reside closely with observations for objects of short longest axis length. This is particularly noticeable for single-cell object forecasts where the ARW-SP distribution clearly shows a spike in frequencies. ARW-MP longest axis forecasts, much like ARW-SP, eventually drop below the observed frequency distribution for larger object longest axes, specifically for mesoscale organized objects. In similar fashion, ARW-MP object aspect ratio forecasts (Figure 7i) closely resemble observed frequencies for large aspect ratios, with a drop-off at small ratios. These results are much improved to ARW-SP forecasts which showed clear differences for large object aspect ratios.

As a whole, the addition of multi-physics into the ARW ensemble design shows greatest objective impact on object attribute forecasts of single-cell objects. Although minor, ARW-MP object attribute distributions are superior to ARW-SP distributions, most noticeably for object longest axis and object aspect ratio. Thus, it should be noted that since the main contrasting qualities between forecasted object attributes are confined to smaller, shorter objects, differences in probabilistic verification of larger, mesoscale organized objects must be resultant from other aspects of the forecasts.

iii) MM VS NMMB (MMSM)

Object attribute distributions for NMMB and MM are summarized in Figure 7a-c, and Figure 7j-l, respectively. Frequency distributions for object area show MM (Figure 7j) forecasts contain a more definitive separation at 200 grid points, with higher forecast frequencies below 200 grid points and lower frequencies above 200 grid points. This result is slightly different than NMMB which is shown to comprise a more loose relationship about the 300 grid point mark. Plots for MM longest axis (Figure 7k) and NMMB (Figure 7b) show MM frequencies are influenced by ARW members as is indicated by the spike in frequencies for single-cell objects. However, as longest axis increases, both ensemble distributions converge, dropping slightly below observed frequencies. Comparable trends are seen in forecasted object aspect ratio. The associated MM verification plot (Figure 7l) illustrates that contrary to the NMMB (Figure 7c), forecasted frequencies are clearly higher than observations for large aspect ratios, with both ensembles falling below observations at small aspect ratios below ~0.45-0.5.

Contrary to expectations, when comparing results from MM and NMMB, object attribute verification suggests the addition of multiple dynamical cores into the ensemble design does not clearly improve forecasts. In fact, for more circular objects of short horizontal length, MM distributions are displaced further from observations in comparison to NMMB. Additionally, as in the SPMP experiment, MMSM ensemble object attribute distributions for larger, mesoscale organized objects are altogether similar. This indicates once again, that probabilistic verification of forecasted storm mode and morphology at large scales is largely independent from systematic object attribute distribution biases.

iv) ARW-MP VS MM (MPMM)

Evaluation of the ARW-MP and MM ensemble object attribute distributions can also be made through analysis of panels g—i and j—l of Figure 7, respectively. Object area distributions for ARW-MP and MM (Figure 7g,j) both result in higher frequencies for small objects and, although bin-to-bin variance occurs, lower frequencies for larger objects. Likewise, respective object aspect ratio plots (Figure 7i,l) show nearly identical distribution values in comparison to observations. Characterized by the peak seen in small object longest axis lengths (Figure 7k), MM forecasts diverge from ARW-MP forecasts (Figure 7h), however, this is the only subjectively noticeable location where differences between the MPMM ensemble longest axis distributions clearly arise.

With exception to short longest axis lengths, ARW-MP and MM contain two of the most alike object attribute distributions. Considering all object attribute distributions together, differences between convective object attributes for a multi-model design versus a multi-physics design mostly occur for objects in short horizontal length, with the multi-physics design more closely resembling observations. Consequentially, much like previous ensemble-to-ensemble object attribute comparisons (i.e., SPMP & MMSM), probabilistic verification of ensemble large scale forecasted storm mode and morphology must be dependent on other forecast aspects.

2) PROBABILISTIC VERIFICATION

i) ARW-SP VS NMMB (SMSP)

Objective evaluation of ensemble storm mode and morphology forecast performance through the unique probabilistic information provided by the OBPROB method is summarized in Figure 8. Scores are plotted via a “heatmap” where each box represents a bin of three forecast hours for a certain convective organization scale. Three hour bins are used to decrease sampling noise as sample sizes per forecast hour are significantly less in the OBPROB object space compared to typical gridpoint space. Values shown are in essence a BSS, with the first ensemble listed as the forecast and the second ensemble listed as the reference forecast. Thus, colors of red indicate positive skill whereas colors of blue indicate negative respective skill for the first ensemble listed. Overlaid quantities are p-values associated with permutation resampling tests for statistical significance. P-values in green are deemed statistically significant and p-values in black are not statistically significant at the 80% confidence level.

Figure 8a summarizes systematic object-based probabilistic verification results for the SMSP ensembles. For mesoscale organized objects (top row), ARW-SP contains more skillful storm morphology forecasts for the first two forecast bins. After forecast hour six, NMMB contains the superior probabilistic storm mode forecast for the remainder of the forecast period. Overlaid p-values indicate the probabilistic improvements given by ARW-SP at early lead times are statistically significant, while times where NMMB contains better verification are not. This would suggest that while late forecast bins show NMMB to be more skillful, this result could be due to sampling uncertainty. At the multi-cell and single-cell scales (second and third rows) a more uniform result is reached showing NMMB to have unanimously superior probabilistic storm morphology forecasts. In addition, many forecast bins show the probabilistic improvements NMMB forecasts provide are statistically significant for five out of six bins at the meso-beta (multi-cell) scale and three out of six bins at the meso-gamma (single-cell) scale. Furthermore, two of these forecast bins, specifically forecast bin three at the meso-beta scale and forecast bin six at the meso-gamma scale, are statistically significant above the 99% confidence level. Therefore, we conclude that the while the ARW-SP ensemble may contain clearly superior probabilistic storm morphology forecasts at early lead times for large scales, there is clear benefit provided at smaller scales for less organized convection from NMMB.

Comparing the object-based verification of storm mode and morphology to the neighborhood-based verification of approximate mesoscale locations of precipitation for the SMSP ensemble comparison (Figure 8b) yields interesting results. Focusing on the neighborhood radius of 16 (loosely corresponding to the meso-alpha scale object verification) similar trends surface for both verification methods with ARW-SP more skillful at early lead times and NMMB more skillful at late lead times. The portion of the forecast period at which NMMB forecasts become superior are different between the methods, however, the trends still remain similar. At smaller convective organization scales (meso-beta and meso-gamma), verification results are largely contrasting as ARW-SP contains superior forecast BS for mesoscale precipitation locations up until late lead times. These results are suggestive of object-based and neighborhood-based methods being sensitive to separate aspects of the forecast. Reasons for the different sensitivity will be further explored in section 4c.

ii) ARW-SP VS ARW-MP (SPMP)

The second ensemble-to-ensemble comparison (ARW-SP versus ARW-MP) probabilistic verification results are located in Figure 8c. At the meso-alpha scale, ARW-MP generally becomes more skillful as lead time increases with statistically significant differences after the 6-8 hour forecast bin. The presence of statistical significance at only times where ARW-MP was more skillful suggests the verification improvements provided are systematic, while early lead times where ARW-SP provided positive skill, may be caused by sampling uncertainty. In contrast to the meso-alpha scale, meso-beta and meso-gamma scale verification shows the multi-physics ensemble design provides uniform storm morphology forecast improvements. Moreover, all bins at both scales show statistically significant improvements of ARW-MP over ARW-SP. Overall, object based probabilistic verification of storm mode suggests ARW-MP introduces systematic improvements for all convective organization scales with more pronounced differences at larger scales.

Corresponding neighborhood-based verification of SPMP ensembles (Figure 8d) show forecasts of mesoscale precipitation locations coincide well with storm morphology forecasts.

Specifically, at a radius of 16 (corresponding to the meso-alpha scale) both methods show ARW-MP increasing in relative skill as lead time increases. NMEP plots do not display the same results for portions of the large scale forecast where ARW-SP object-based results were more skillful, however both methods lack statistically significant differences at the first forecast bin. Extremely similar conclusions are found for meso-beta and meso-gamma scales between verification methods as both showcase uniformly, ARW-MP forecasts of precipitation locations and storm morphology are superior. One highlighted difference between the techniques is associated with changing of convective organization scale. In the object-based results, ARW-MP shows the most benefit as scale increases, while neighborhood-based results show ARW-MP most skillful as scale decreases. The differing verification trends are suggestive of the neighborhood-based method smoothing over convective scale features pertinent to the object-based verification, particularly at large radii. Visual representation of neighborhood-based smoothing effects are described in greater detail in section 4c.

iii) MM VS NMMB (MMSM)

Verification results for how a multi-model ensemble design impacts object-based storm morphology forecasts is outlined in Figure 8e. Outcomes at the meso-alpha scale show probabilistic improvements at all forecast bins, accentuated by large increases in skill at early lead times seen for the first two forecast bins. Four out of six bins are statistically significant, including the first two containing significance above the 99% confidence interval. Moving down to smaller convective organization scales, multi-cell and single-cell results show very little probabilistic differences in ensemble performance, with statistical significance found only for the first bin at the meso-beta scale. The stark difference between scale verification results could be caused by two different factors: Either the OBPROB method is effectively filtering out which objects should be contained in the mesoscale organized scale and which should not, or the probabilistic benefits to a multi-model ensemble are mostly encompassed in forecast objects of mesoscale organized convective systems. Subjective analysis of multi-cell and single-cell probabilities (not shown) suggests the OBPROB method indeed effectively filters out objects of smaller scale. However, increased sample sizes from additional cases beyond the scope of this study would also likely lead to more pronounced differences for meso-beta and meso-gamma verification.

Figure 8f shows associated NMEP verification results for the MMSM ensemble comparison. In terms of probabilistic improvements, MM is seen to not only provide increased skill at all lead times and radii, but also do so above the 99% confidence level. This result is contrary to the object-based results which showed at multi-cell and single-cell scales, probabilistic improvements given were minimal and far from being statistically significant. On the other hand, at large radii and convective organization, MM is seen to significantly improve both storm morphology and mesoscale precipitation location forecasts. Punctuated by the largest objective skill increase at early lead times, neighborhood-based and object-based methods agree that a multi-model ensemble design has greatest impact to forecasts for large scale convective features at early lead times.

iv) ARW-MP VS MM (MPMM)

OBPROB verification of the MP vs MM ensemble-to-ensemble comparison (MPMM) is displayed in Figure 8g. At the meso-alpha scale, MM is mostly superior in probabilistic storm morphology forecasts. ARW-MP shows improvements for the 15-17 hour forecast bin, however an associated p-value of 0.603 indicates that respective skill differences are likely due to sampling uncertainty. Among the forecast lead times that showed BS improvements from MM, only one, the 3-5 hour forecast bin, is statistically significant above the 80% confidence level. Verification of multi-cell objects also resulted in MM skill increases, with all forecast bins displaying additive skill from MM. Four of these bins produced significant differences. Single-cell object verification shows multi-physics and multi-model ensemble designs produce little skill differences in terms of forecasted storm morphology. These probabilistic differences are rather trivial as only the 15-17 hour forecast bin contained statistical significance from MM. Therefore, object-based calculations show that while MM provides a more consistent skill improvement, aside from multi-cell object forecasts, only select forecast periods (i.e., forecast hours 3-5 for meso-alpha, 15-17 for meso-gamma) are representative of systematically improved storm mode forecasts. The select nature of systematic verification improvements at meso-alpha and meso-gamma scales indicates MPMM storm morphology forecast differences occur at specific stages of storm development and morphology. Such physical differences will be further discussed in section 4b.

Neighborhood-based verification of forecasted precipitation locations for MPMM are displayed in Figure 8h. Aside from the 0-2 hour forecast bin at a radius of 16, all bins for all radii resulted in statistically significant improvements of mesoscale precipitation locations from MM. This is considerably different from the object-based results of forecasted storm mode, particularly at small and large convective organization scales (i.e., meso-gamma and meso-alpha). Furthermore, NMEP verification shows MM successively improving upon ARW-MP forecasts as lead time increases, with the final forecast bin at a radius of 16 showing greatest improvement. Object-based results are contradictory to this result at the final forecast lead time, which showcased ARW-MP providing positive respective skill.

v) SUMMARY

Probabilistic verification of ensembles using OBPROB led to four key differences between ensemble designs and forecasted convective storm morphology at large convective scales. First, represented by superior forecasts from ARW-SP, for single-model, single-physics ensembles, choice of model core and physics parameterizations was found to have greatest impact during early lead times. Second, compared to a single-physics design, the addition of multiple physics parameterizations created increasing benefit as lead time increases with significant improvements after forecast hour five. Third, implementing a multi-model ensemble with respect to a single-model ensemble provided clear benefit that was pronounced at early lead times. Fourth, respective improvements from a multi-model and multi-physics ensemble led to superior probabilistic verification from the multi-model forecasts for the majority of forecast lead times with forecast hours 3-5 containing significant improvements.

Comparison of OBPROB verification to a neighborhood-based (NMEP) verification engendered important method differences. Highlighted by MPMM verification, superior ensemble forecasts of approximate mesoscale precipitation locations does not necessarily imply the forecasts will also contain the more optimal storm morphology forecast. Subsequently, a second conclusion is that forecasts of convective precipitation locations and convective storm mode are sensitive to differing aspects of the forecast. This notion is found to be true for multiple ensemble-to-ensemble comparisons (i.e., SMSP, MPMM) suggesting there may be specific physical quantities responsible for such verification discrepancies. These method comparison conclusions are in agreement with prior findings that suggest multiple verification methods are needed to fully quantify CAE error uncertainty (e.g., Berner at el. 2015; Duda et al. 2016; Jankov et al. 2019; Gasperoni et al. 2019; J20). Furthermore, these conclusions indicate differences between object and neighborhood-based methods are connected to more than just smoothing over of convective scale details from neighborhood-based techniques.

3) QUANTIFYING FORECAST SPREAD

i) ARW-SP VS NMMB (SMSP)

Understanding of forecast spread from ensemble convective storm morphology forecasts and how it contributes to skill differences is completed through reliability diagrams (Wilks 2006). Reliability diagrams not only plot ensemble reliability, which is the deviation from the diagonal, but also indirectly reflect ensemble resolution, which is the variation about the horizontal climatological base rate. If an ensemble over-forecasts probabilities, points will occur in the space below the diagonal and vice versa for under-forecasting. Thus, perfect reliability occurs when forecast probabilities match the observed frequency. A "no-skill" line halfway between perfect reliability and the climatological base rate is also plotted for reference. For simplicity and ease of interpretation, resolution and reliability values for each ensemble are plotted in the top of each reliability diagram. Specifically, reliability indicates the degree to which a forecast can be taken at face value (i.e., for high reliability, forecast objects of 50% probability should verify 50% of the time) and resolution specifies the ability of the forecast to differentiate separate events of varying frequencies of occurrence (Sanders 1963; Murphy 1971, 1973, 1986; Stephenson et al. 2008). Associated ensemble sharpness plots are displayed below reliability diagrams as an additional indicator of the effects of forecast spread in terms of storm morphology (object) and approximate location/intensity (neighborhood). Sharpness plots show corresponding probability frequencies (counts) that fall in each probability bin in the reliability diagram. For meso-alpha scales 10 forecast bins are used, equally spaced from 10% to 100%. For meso-beta and gamma-scales due to probabilities failing to reach high values, reliability diagrams are incomplete with probabilities reaching 50% and 30%, respectively.

Object-based reliability diagrams are shown in Figure 9. Characterized by blue (NMMB) and green (ARW-SP) contours, meso-alpha scale reliability and corresponding sharpness (Figure

9a,d) show ARW-SP contains more reliable forecasts of high probability bins (i.e., 90% and 100%), with NMMB more reliable for 10% probability objects. For intermediate probabilities from 50% to 80%, NMMB contains more reliable forecasts, however these probability bins contain significantly smaller sample sizes. Due to a sizable contribution to forecast skill from large sample sizes, the increase in reliability for high probability objects from ARW-SP could be one reason for the verification improvements seen when highest probability objects are most expected (i.e., early lead times). Additionally, calculated resolution values show ARW-SP to have higher resolution for mesoscale organized objects compared to NMMB. ARW-SP resolution remains superior for meso-beta forecasts (Figure 9b), however drops for meso-gamma forecasts (Figure 9c) where NMMB has greater resolution. With the vast majority of samples for meso-beta and meso-gamma scales (Figure 9e,f) coming from the 0-10% probability bin, both SMSP ensembles justifiably have comparable reliability at these scales. While some discrepancies occur between the ensembles at higher probabilities around 30%, a main point of emphasis is the noticeable over-forecasting bias both ensembles have for object-based contour plot probabilities.

ii) ARW-SP VS ARW-MP (SPMP)

Also shown in Figure 9, ARW-SP (green) and ARW-MP (black) reliability, resolution and sharpness are displayed for each convective scale. For the meso-alpha scale, in addition to improved resolution, ARW-SP bin-to-bin reliability is largely comparable to ARW-MP, with various probability bins producing superior reliability. For example, at the smallest probability bin (i.e., 10%) ARW-SP is slightly more reliable, whereas for the 100% probability bin, ARW-MP produces more reliable forecasts. Although ARW-SP contains as many reliable probability bins as ARW-MP, ARW-MP still results in superior overall reliability. With ensemble sharpness (Figure 9d) showing a much greater occurrence of 100% probability objects from ARW-SP, lower

ARW-SP reliability indicates that overall SPMP verification may hinge greatly upon verification of high, 100% probability objects. Furthermore, ensemble sharpness shows more evenly distributed object probabilities for the ARW-MP ensemble, indicative of forecast diversity affecting the respective verification of the SPMP ensembles. Results for meso-beta and meso-gamma scales (Figure 9b,c) show ARW-MP with higher resolution and reliability. Arrival at this conclusion is consistent with probabilistic verification which showed ARW-MP producing unanimously better skill compared to ARW-SP.

iii) MM VS NMMB (MMSM)

The addition of multiple dynamical cores into ensemble design is shown to improve large scale storm morphology forecast reliability and resolution. Summarized in Figure 9a, MM (red) is shown to result in greater forecast reliability when compared to NMMB (blue) for all forecast probability bins except 40 and 50%. The associated sharpness diagram (Figure 9d) gives more insight as to why MM struggles with these intermediate probabilities. Indicated by the peak seen at 50%, MM forecasts clearly contain clustering problems as ARW-SP and NMMB members agree upon separate respective convective events. Due to higher occurrences of 50% probabilities forecasted by MM, a large over-forecasting bias seen in the reliability diagram develops. Despite undesirable clustering issues, MM still contains the most diverse distribution of forecasts, depicted by a more evenly distributed ensemble sharpness plot. For meso-beta scale probabilities (Figure 9b), MM still results in greater resolution and reliability, however at the meso-gamma scale (Figure 9c), NMMB is shown to have superior reliability and resolution.

iv) ARW-MP VS MM (MPMM)

Superior convective storm morphology reliability and resolution are found from a multi-model design with respect to a multi-physics design. Highlighted in Figure 9a, MM not only contains higher resolution, but also superior reliability for eight out of ten probability bins (i.e., all except 40 and 50% probability bins). In other words, MM is not only able to better separate differently observed events, but also generate respective object probabilities that are much closer to observed frequencies. At meso-beta and meso-gamma scales (Figure 9b,c) MM and ARW-MP resolutions switch, with MM containing better resolution for multi-cell forecasts, and ARW-MP having superior resolution for single-cell forecasts. It is important to note however, that sample sizes at points largely responsible for resolution calculation values drop significantly (Figure 9e,f). Thus, the main point of emphasis from the smaller two convective organization scales remains the fact that object probabilities are considerably over-forecasted for both MM and ARW-MP ensembles.

v) SUMMARY

Evaluation of ensemble spread through reliability and sharpness diagrams provide further insight into the probabilistic verification of ensemble-to-ensemble comparisons at large convective scales. SMSP ensemble results revealed that while NMMB contained better overall reliability, ARW-SP forecasts of 100% probability objects, which contributed greatly to overall verification due to a large sample size, were more reliable. Ensemble spread also proved pertinent to SPMP probabilistic verification. Although ARW-SP contained superior resolution, ARW-MP resulted in better overall reliability, particularly at the 100% probability bin, in addition to more evenly distributed object probabilities shown through ensemble sharpness. The more evenly distributed object probabilities can be linked to greater sampling of forecast uncertainty from ARW-MP, which is important to probabilistic verification. Of all ensembles, MM provided the most reliable forecasts with greatest resolution. True for both MMSM and MPMM comparisons, MM was able to better distinguish observed events while better representing observed frequencies of probabilities. The only pitfall to MM forecasts was a tendency for constituent model members to separately agree on their own respective forecasts (i.e., clustering). Nonetheless, with the most even spread of object probabilities shown via ensemble sharpness, it is prevalent that reliability, resolution, and forecast diversity all contribute to the overall superior verification of MM to NMMB and ARW-MP.

b. Subjective evaluation

1) SINGLE-MODEL, SINGLE-PHYSICS (NMMB VS ARW-SP)

Subjective analysis of object-based results is performed to quantify specific physical differences responsible for objective verification discrepancies. In turn, it can be seen more explicitly why one ensemble is preferable over another at specific lead time and convective organization scales. For this study, examination of the physical differences between ensemble storm morphology forecasts is confined to the meso-alpha scale. Confinement to mesoscale organized objects is done for two reasons: To place emphasis on objects largely responsible for severe weather events of greatest subjective interest to forecasters and to directly evaluate the unique OBPROB method in terms of depicting ensemble storm morphology forecasts. For each individual ensemble-to-ensemble comparison, focus is directed to portions of the forecast which generate statistically significant probabilistic verification differences. However, for ease of interpretation, a singular case study, May 16, 2015 (hereafter M16), representative of individual ensemble-to-ensemble physical differences, is chosen.

Described in Table 4, the M16 case study is composed of many desirable features related to ensemble storm morphology forecast performance. Due to strong synoptic scale forcing, initial supercellular dryline convection in the southern Plains quickly grew upscale into a north-south oriented squall line. Multiple broken convective segments associated with the squall line encompassed a rather expansive region, extending from southern Texas through Iowa and Minnesota, allowing for increased variability among ensemble forecasts. Furthermore, much of the convection was of severe caliber. Official filtered SPC storm reports resulted in 50 tornado reports, 140 wind reports, and 42 hail reports, for a total of 232 severe weather reports. Consequently, during its evolution, this system was of great subjective interest to forecasters, making it a prime case for evaluation of ensemble storm morphology forecast performance.

OBPROB plots describing the forecasted storm morphology for M16 are located in Figure 10. Valid at 11 Z May 17, 2015, observation objects (Figure 10i) depict multiple broken linear segments in the overarching system. With two meso-alpha scale (maroon) objects in southern Texas, another near the Dallas, Fort Worth area, and a fourth in southwest Arkansas, much of the stronger storms at this analysis time are located in southern portions of the synoptic scale system, while northern segments into Missouri and Iowa have begun to, if not already deteriorate. Comparisons of the observed storm morphology structure to SMSP ensemble forecasts can be made through Figure 10a (ARW-SP) and Figure 10b (NMMB). Recall, systematic verification between SMSP ensembles showed ARW-SP producing statistically significant improvements respective to NMMB forecasts at early lead times. Figure 11, which shows probabilistic verification specifically for the May 16, 2015 case study, shows that at the displayed time in Figure 10, object-based results are representative of systematic results as ARW-SP contains superior storm mode forecasts compared to NMMB (Figure 10a). From Figure 10a,b, some of the physical reasons behind ARW-SP improvements at early forecast periods are clearly represented. With respect to NMMB, ARW-SP has a much larger object probability distribution indicative of greater member-to-member forecast spread. NMMB forecasts are too certain of convective organization in southwest Arkansas and central Texas, leading to poor verification of high probability objects.

In fact, the one member whose forecast deviated enough from other ensemble members in Arkansas matched to observations. This conclusion is consistent with other case studies, as NMMB forecasts tend to lack member-to-member spread, leading to poor verification of high probability objects.

While containing larger object probability distributions in comparison to NMMB forecasts, ARW-SP still commonly result in overconfident object probabilities. As seen in Figure 10a, ARW-SP forecasts extend strong convective objects into Missouri, well after these segments decayed in observations (Figure 10i). The overconfidence found in northern extensions of north-south oriented squall lines, is seen consistently in other case studies, indicative of a northern edge bias in the ensemble. Other subjective differences seen from representative cases for SMSP probabilistic results (not shown) suggest NMMB forecasts, with respect to ARW-SP forecasts, tend to underdo upscale growth of convective systems, either lacking in number of mesoscale organized morphologies (i.e., M16), or horizontal scale of linear systems. Both conclusions are consistent with the object attribute results which hinted at NMMB failing to grow small scale objects into mesoscale organized systems consistently.

2) SINGLE-PHYSICS, MULTI-PHYSICS (ARW-SP VS ARW-MP)

Corresponding M16 OBPROB plots for ARW-SP and ARW-MP are located in Figure 10c,d, respectively. Justified by the M16 SPMP verification (Figure 11b)[1], physical differences seen in these plots are representative of systematic probabilistic verification results which showed ARW-MP containing superior storm morphology forecasts at middle and late lead times. Emphasized by a single, matching 100% probability object associated with observed convection in AR, ARW-MP effectively distinguishes which forecast environments are of high certainty and

[1] Non-binned results (not shown) show explicitly for forecast hour 12, ARW-MP outperforms ARW-SP.

which are not compared to ARW-SP forecasts, which show just one out of three high probability objects matching to observations. In other words, ARW-MP accurately decreases forecast object probabilities in regions of larger potential variability, while retaining high probabilities in settings where forecasted convection verifies. This result is depicted by decreased probabilities for objects associated with the northern extension of the Missouri squall line, and southern extension of Texas convection from 100% in ARW-SP to 10% and 70% in ARW-MP. Separate outcomes from this feature in ARW-MP forecasts can be seen in other case studies (not shown) where ARW-MP's member-to-member diversity allows for better post-convection environment modification to where new convection has a higher likelihood to verify. In contrast, a central downfall to a multi-physics design with respect to storm morphology forecasts is the increased member-to-member diversity has potential to persist systems well after observed decay in select members, leading to poor performance for cases where convection fully decays. Overall, in conjunction with improved probabilistic skill as lead time increases, subjective evaluation of member-to-member diversity having an advantageous affect in object verification for ARW-MP, is consistent with what would be originally expected from a multi-physics, single-physics ensemble comparison.

3) SINGLE-MODEL, MULTI-MODEL (NMMB VS MM)

MMSM ensemble comparison OBPROB plots for M16 are also shown in Figure 10. Recall, probabilistic verification of MMSM ensembles detailed large improvements from a multi-model design, particularly at early forecast lead times. M16 verification for MMSM ensembles (Figure 11c) are consistent with systematic results at forecast hour 12, showing MM provides large skill increases. Initial impressions from MM forecasts (Figure 10e) are clearly related to the obvious increase in object probability distributions and ensuing forecast diversity when compared to NMMB forecasts (Figure 10f). However, further inspection suggests MM not only increases spread from NMMB, but also decreases unlikely high probabilities produced by constituent ARW members in northern Missouri and southcentral Texas. Therefore, MM not only immediately identifies uncertainties in the forecast even at early lead times, but also effectively adjusts for individual model biases through decreases in probabilities of northern bias objects from ARW, while increasing forecast spread is not found in NMMB. Simply put, MM is able to separate out and take the best portions of the forecast from each constituent model, resulting in superior storm morphology verification. Subjective analysis conclusions are supportive of objective results which showed MM having the higher resolution and reliability compared to NMMB.

4) MULTI-MODEL, MULTI-PHYSICS (MM VS MP)

Physical qualities representative of statistically significant probabilistic improvements from MM for forecast hours 3-5 of the MPMM ensemble comparison are also found through subjective analysis of MM (Figure 10g) and ARW-MP (Figure 10h) OBPROB plots for M16. Individual verification of M16 show forecast hours 12-14 are characteristic of systematic improvements found at forecast hours 3-5 with MM largely improving upon the skill from ARW-MP. The M16 case as a whole, shows MM better samples forecast uncertainty than ARW-MP. For example, while ARW-MP forecasts a matching 100% probability object associated with observed Arkansas convection, the ensemble also forecasts a matching 30% probability object. The resultant verification of both these objects is slightly worse than the MM forecast which forecasts matching 80% and 40% objects. Additionally, in southcentral Texas convection MM generates much more forecast diversity with three objects of 70, 20, and 10% probability, compared to ARW-MP's single, non-matched 70% object. While the 10% MM object matches to observations instead of, say, the 70% object, the increased member-to-member diversity still results in a BS improvement to the ARW-MP forecast of southcentral Texas convection.

Subjective analysis of other case studies representative of MM improvements over ARW-MP (not shown) indicate MM also better samples forecast uncertainty through the reduction of non-matching very high probability objects produced in ARW-MP forecasts. However, the decrease in object probabilities associated with an increase in member-to-member forecast spread for MM does not necessarily imply a lack thereof for high probability ARW-MP objects. Further scrutiny of individual member forecasts from ARW-MP reveals forecast spread amongst members, although less than MM, is still generated by ARW-MP at early lead times. The difference resides in the fact that member-to-member forecasts do not generate enough spread to cause objects to be considered forecasts for separate events. It is hypothesized then, that MM better represents how forecast uncertainty and diversity are expected to grow as lead time increases through generation of meaningful spread, with very high probability objects confined mostly to early forecast lead times.

c. Comparison to NMEP

Besides analyzing how ensemble design affects storm morphology forecasts, one of the main questions this study aims to answer is how the novel OBPROB verification method compares to a traditional neighborhood-based method. More specifically, some of the questions are whether good forecasts of approximate mesoscale precipitation locations imply good forecasts of storm morphology, and if differences in verification of scale separated convective objects (i.e., single-cell, multi-cell, and mesoscale organized) correspond well to a simple change in neighborhood radius. Probabilistic verification results already discussed hint that both verification techniques are potentially sensitive to different forecast features meaning overall results are not necessarily connected. To further investigate why this is the case, subjective analysis of ensemble NMEP plots from M16 are generated and compared to existing OBPROB plots.

NMEP plots used in subjective analysis of neighborhood-based verification results of M16 (Figure 11) are shown in Figure 12. Comparing verification of SMSP ensembles, OBPROB results show ARW-SP to be more skillful than NMMB (Figure 11a) whereas NMEP results (Figure 11e) show NMMB to contain additive still. Corresponding NMEP plots for ARW-SP and NMMB (Figure 12a,b) show the verification differences are most likely due to exceptional extension of northern bias probabilities in ARW-SP well into Iowa, leading to poorer verification. In comparison, northern bias object probabilities, while also upwards of 100%, only extend into northern Missouri. In addition, lack of forecast spread seen in NMMB object-based forecasts is much less apparent in NMEP results as probabilities similarly spread with respect to ARW-SP probabilities. As a result, it is clearly displayed that superior forecasts of mesoscale precipitation location do not imply storm morphology forecasts will also be more skillful.

A prime advantage of object-based techniques are their ability to quantify convective scale features that neighborhood-based methods typically smooth out. A principal example of this can be found in NMEP forecasts of ARW-SP and ARW-MP ensembles for M16 (Figure 12c,d). For ARW-SP, a high probability contour reaches from central Iowa through to northern Texas continuously, indicating no sections of lower probabilities along the way. Contrarily, object-based plots (Figure 10a) of the same forecast indicate multiple objects of varying probability through this region. Consequently, not only can NMEP probabilities differ considerably from OBPROB probabilities, but various independent storm morphologies are smoothed over.

Analysis of NMEP plots associated with the fourth ensemble-to-ensemble comparison (MPMM) indicate another difference between neighborhood and object-based methods. Figure 12g,h show MM and ARW-MP ensemble verification can be affected by non-convective observation precipitation (i.e., stratiform precipitation) as observation contours are found to extend into Missouri and Illinois. As demonstrated in object-based observations (Figure 10i) precipitation in these regions at the 11 Z analysis time are not convective due to system decay. Since NMEP observation contours exist at this time, it is justifiable to conclude that neighborhood-based verification is influenced by both convective and stratiform precipitation. However, this is an expected conclusion as the neighborhood-based forecast probabilities are dependent upon precipitation fields simply exceeding a predefined threshold. Even so, the potential influence stratiform precipitation has on NMEP verification bolsters the argument that verification techniques and their results can be complementary.

Finally, the impacts of separating objects based on convective organization scale versus simply increasing by neighborhood radius can be seen in Figure 12e,f. Verification results from both techniques indicate MM contains superior storm morphology and mesoscale precipitation location forecasts (Figure 11c,g), however the agreement in verification is impacted through separate aspects of the forecast. Aside from being hindered from higher probabilities located in Iowa associated with constituent ARW members, MM precipitation location forecasts (Figure 12e) are largely comparable from large scale convective systems in Arkansas and Texas compared to NMMB (Figure 12f). However, after consideration of probabilities associated with smaller scale convective/stratiform precipitation in Nebraska and South Dakota, it is apparent that high probability NMMB forecasts in this region negatively affected overall verification. Although it may not be a sole reason for final ensemble performance, the presence of single and multi-cell convection still impacts neighborhood-based forecast verification at large radii. In turn, comparisons between OBPROB and NMEP forecasts show that simple increases in neighborhood radii do not necessarily correlate to an increase in focus on larger convective organization.

5. Discussion and conclusions

Convection-allowing ensembles contain a distinct ability to forecast convective initiation, location and evolution. Commonly used by forecasters for guidance on storm morphology, CAEs can provide probabilistic information about forecasted storm mode and morphology. In limited studies, CAE design impacts on skill and probabilistic forecasts have primarily focused on spatial coverage of precipitation through neighborhood-based methods rather than explicit verification of convective storm morphology. Of select experiments that have directly evaluated modeled convection, none have objectively verified CAE configuration impacts on storm morphology forecasts, as greater emphasis has been placed on deterministic forecast settings. The usefulness of defining optimal ensemble design in terms of predicted storm mode, is rooted in the fact that certain storm modes are associated with distinct severe weather threats. Therefore, through better understanding of optimal CAE design, CAEs can become far more valuable in convective forecasting settings.

Instead of answering which ensemble design is "best", four ensemble-to-ensemble comparisons were designed to answer specific questions related to ensemble design choices and storm morphology forecasts. For how certain model and physics parameterization choices impact storm morphologies two single-model, single-physics ensembles (ARW-SP and NMMB) were compared. Impacts related to how ensemble diversity through varying physics schemes impacts storm morphology forecasts, were analyzed through comparisons of a single-model, single-physics ensemble (ARW-SP) to a single-model, multi-physics ensemble (ARW-MP). Evaluations of forecasts from a single-model ensemble (NMMB) with respect to a multi-model ensemble (MM) gave insight to how varying model dynamical cores affects probabilistic storm morphologies. Finally, relative impacts from a multi-model design (MM) compared to a multi-physics design

(ARW-MP) were compared to determine which design choice is more optimal related to forecasts of storm morphology.

To address the impacts ensemble design has on forecasted storm mode and morphology, a newly optimized OBPROB technique is applied. Optimizations of the OBPROB method from previous installations (i.e., J20) are designed to help create a framework that is specific to verification of ensemble storm morphology forecasts. The first optimization was model bias adjustment through reflectivity percentile analysis (Skinner et al. 2018). Model reflectivity bias adjustment allowed for more continuous object matching from member-to-member, reducing objects of unrealistic size and shape in addition to normalizing ensemble forecasts to where differences are specifically related to morphological distinctions rather than model bias. The second optimization was filtering of stratiform observation objects. Caused by radar bright-band effect, reduction of stratiform objects permitted focus on verification of storm mode and morphology. The final optimization of the OBPROB method was scale separation of objects based on convective organization (i.e., single-cell, multi-cell and mesoscale organized). Classification of objects ensured the verification process was aligned with the predictability expected at each respective scale. The end result was an OBPROB plot composed solely of mesoscale organized objects, allowing for interpretation similar to an operational convective outlook (J20).

For the first ensemble-to-ensemble comparison (SMSP), objective and subjective verification suggest model and physics scheme choices affect storm morphology forecasts. Evaluated first, object attributes revealed largest differences between SMSP ensembles were punctuated by ARW-SP forecasting an increased amount of small, short, circular objects. With similar distributions for large, long, and linear objects, object attribute distributions suggested NMMB struggled to grow small scale objects upscale. Probabilistic verification supported this notion as ARW-SP was found to produce statistically significant improvements with respect to NMMB forecasts of mesoscale organized objects at early lead times when upscale growth was most pronounced. Otherwise, NMMB forecasts were more skillful. Reliability and resolution results showed ARW-SP to have higher resolution, and NMMB with superior overall reliability. However, analysis of largest individual probability bin sample sizes displayed high, 100% probability objects to be more reliable from ARW-SP, which provides a level of consistency with early lead time probabilistic results where highest probabilities are most expected. Subjective analysis of OBPROB plots show the statistically significant probabilistic differences at early lead times were related to both probabilistic distributions of objects and actual storm morphology forecasts. Compared to ARW-SP, subjective analysis revealed NMMB forecasts greatly lacked spread, leading to unmatching, very high probability objects and poorer verification. As far as storm morphology, NMMB struggled to grow single-cell and multi-cell objects upscale, resulting in a lack in horizontal scale of linear systems needed to match observations. Therefore, individual model and scheme choices can impact how ensembles grow convective systems upscale.

In relation to a single-physics versus multi-physics ensemble design (SPMP), objective and subjective verification are consistent in that as lead time increases, a multi-physics ensemble design becomes increasingly beneficial. Predominantly at middle and late lead times, object-based verification of mesoscale organized objects resulted in statistically significant skill increases from ARW-MP. With the main discrepancies between object attribute distributions confined to small, short, circular objects, objective verification clearly relied on different features of SPMP forecasts. Analysis of ensemble spread through reliability, resolution and sharpness revealed such features to include more reliable forecasts from ARW-MP, particularly for high probability objects, and larger forecast spread through more evenly distributed object probabilities. Subjectively, objective results were manifested in ARW-MP member-to-member diversity. With larger spread from individual member forecasts, ARW-MP better distinguished which environments should be of high probability and which contained larger uncertainty, ultimately leading to superior verification. Member diversity includes better post-convection environment modification to where new convection has a higher likelihood to verify; a concept important to operational forecasters. It is concluded then, that benefits of a multi-physics ensemble with respect to a single-physics ensemble are contained in individual member forecast diversity, allowing for greater discernment of environmental uncertainty as lead time increases.

Concluding evaluations of single-model versus multi-model ensemble storm morphology forecasts suggest greatest benefit is realized at the mesoscale organized scale (meso-alpha). Of the probabilistic improvements at the meso-alpha scale, statistically significant differences above the 99% confidence level were discovered at early lead times. Although object attribute distributions showed discernable differences mainly for short length objects, quantification of forecast spread unveiled objective reasons for superior verification from MM. The addition of multiple-dynamical cores greatly improved ensemble reliability and resolution. Furthermore, more evenly distributed object probabilities from ensemble sharpness diagrams imply forecast diversity may also impact probabilistic verification. Subjective analysis of the M16 case study, demonstrates increased skill is resultant from relative bias reduction from each constituent model as MM increases forecast spread lacking in NMMB forecasts while reducing the northern bias associated with ARW members. Due to subsequent diversity, MM effectively takes skillful portions from each respective model forecast, leading to increased sampling of initial forecast uncertainties.

Analysis of probabilistic verification of large scale storm morphology forecasts generated by a multi-model (MM) ensemble versus a multi-physics (ARW-MP) ensemble largely led to sampling uncertainty. With exception to the 3-5 hour forecast bin improvement from MM, probabilistic differences were not statistically significant. Considering object attribute distributions, very small probabilistic differences are supported as attribute distributions were markedly similar. Incorporating ensemble spread analysis, MM resulted in greatly improved reliability and resolution compared to ARW-MP. However, a downfall to MM forecasts were issues with clustering. Indicated by a peak in 50% probability occurrence, each constituent model members separately agreed on individual forecasts. Even though overall object probabilities were most evenly distributed from MM, clustering of forecasts was one undesirable quality. Subjectively, physical differences characteristic of the probabilistic skill improvements seen from forecast hours 3-5 in MM were from a decrease in non-matching, very high probability objects. The reduction of non-matching, very high probability objects showed MM better samples early lead time forecast uncertainty, but also unearthed a separate quality about ARW-MP forecasts. Further scrutiny on individual member forecasts revealed ARW-MP members, while producing a 100% probability object, still contained member-to-member spread. Consequently, it is concluded that skill improvements from a multi-model ensemble are from a greater sampling of initial condition uncertainty, resulting in meaningful spread not found from early lead times in a multi-physics ensemble. Therefore, MM forecasts better represent how forecast uncertainty and diversity are expected to grow, with very high probability objects confined mostly to early in the forecast period.

Although focus in this study was placed on largely organized convective objects due to their largest contribution to forecasted storm mode and morphology and greatest predictability, objective verification of multi-cell and single-cell objects was still performed. For SMSP ensembles, NMMB forecasts were more skillful at multi-cell and single-cell scales where some improvements were statistically significant. Reliability and resolution results are consistent with superior NMMB forecasts as NMMB produced better reliability for both multi-cell and single-cell scales. These results suggest while NMMB may struggle to grow small scale objects upscale, their approximate forecasted locations and intensities are superior to ARW-SP. SPMP ensembles produced even more pronounced probabilistic differences at multi-cell and single-cell scales where significant improvements from ARW-MP were found unanimously. Reliability diagrams support probabilistic results as ARW-MP contained better reliability and resolution for both scales. Objectively then, the addition of multiple physics parameterizations not only improves large scale storm morphology forecasts, but produce benefit for approximate locations and intensities of multi-cell and single-cell objects. In contrast, benefits of MM to NMMB were greatly dampened at smaller convective organization scales. With insignificant probabilistic differences found at multi-cell and single-cell, it is hypothesized either the OBPROB method effectively separates out objects of differing convective organization, or multi-model ensemble improvements are mainly found for large scale forecasts. Evaluation of MMSM reliability and resolution showed MM with greater reliability, however improved quantities are quite small. Additional case studies may reveal more prominent differences for MM improvements upon NMMB. Conversely, in comparison to ARW-MP, MM produced a majority of probabilistic skill improvements at multi-cell and single-cell scales, especially at the meso-beta scale where four out of six forecast bins were statistically significant differences. While superior reliability and resolution belonged to MM forecasts at the multi-cell scale, results reversed for single-cell objects as ARW-MP produced forecasts of greater reliability and resolution. Objective verification of respective CAE design impacts suggest then that MM improvements of approximate location and intensity of multi-cell objects are not as profound at the single-cell scale. Therefore, additional case studies to enhance probabilistic differences are needed to effectively discern meaningful differences from MM vs ARW-MP at the meso-gamma scale.

Previous studies suggest that optimally addressing model error may rely on combinations of techniques to address uncertainties (Berner at el. 2015, Duda et al. 2016, Jankov et al. 2019, Gasperoni et al. 2019). To assess if verification techniques are indeed sensitive to separate aspects of the forecast, and display the usefulness object-based methods provide relative to explicitly resolving convective scale features, a neighborhood-based technique (NMEP) was compared to the OBPROB method. Objective comparisons through probabilistic verification showed superior forecasts of approximate mesoscale precipitation locations do not necessarily imply superior forecasts of storm morphology. Influenced by non-convective precipitation, NMEP verification was shown to be dependent on forecast aspects separate from storm morphology. Different verification findings support the conclusion that individual verification methods are sensitive to separate uncertainties of the forecast. Subjective analysis also bolstered previous studies' conclusions (e.g., Gilleland et al. 2009, J20) as many NMEP probability contours were shown to smooth out convective scale details pertinent to the storm morphology forecast. Finally, analysis of NMEP plots revealed verification at larger radii were still swayed by forecasts of small scale convective precipitation. This conclusion supports the claim that the OBPROB method is able to separate objects of larger convective organization, and simple increases in neighborhood radii do not necessarily correlate to an increase in focus on larger convective organization.

The main focus of this study was on verification of convective objects through forecasted reflectivity fields. Other atmospheric variables closely related to forecasted storm mode severity such as updraft helicity (UH) and maximum estimated size of hail (MESH) were not analyzed. With additional cases to account for decreased sample size, future work should focus on partitioning ensemble storm morphology forecasts based on more specific classification of storms (i.e., strongly rotating) and how ensembles depict not only storm mode and morphology but storm severity. While this study revealed certain differences of storm morphology resulting from different CAE designs, consideration of additional cases in a long-term CAE may enhance and reveal additional differences. Particularly at smaller convective scales, information regarding ensemble forecasts of multi-cell and single-cell objects could be further analyzed to expose other key impacts ensemble design has on forecasted storm mode and morphology.

Acknowledgments

This study was supported by NOAA Award NA17OAR4590187. The author would like to acknowledge and thank Nicholas Gasperoni and Yongming Wang for generating ensemble forecast and observation data and associated discussions of neighborhood-based conclusions. All computational post processing and verification utilized resources provided by the University of Oklahoma (OU) Supercomputing Center for Education and Research (OSCER).

References

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. Mon. Wea. Rev., 146, 4115-4153.

Barthlott C., and Coauthors, 2011. Initiation of deep convection at marginal instability in an ensemble of mesoscale models: A case-study from COPS. *Q. J. R. Meteorol. Soc.* ***137***(S1): 118-136.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model. Mon. Wea. Rev., 132, 473-494.

Berner, J., S. Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev*., 139, 1972-1995.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665.

Carlberg, B. R., W. A. Gallus, and K. J. Franz, 2018: A Preliminary Examination of WRF Ensemble Prediction of Convective Mode Evolution. *Wea. Forecasting*, **33**, 783–798.

Clark, A. J., W. A. Gallus, Jr., M. Xue, and F. Kong, 2010: Growth of spread in convection-allowing and convection-parameterizing ensembles Wea. Forecasting, 25, 594-612.

Clark, A. J., W. A. Gallus, and M. L. Weisman, 2010: Neighborhood-Based Verification of Precipitation Forecasts from Convection-Allowing NCAR WRF Model Simulations and the Operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

Clark, A. J., and Coauthors, 2012b: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74.

Clark, A. J., R. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29**, 517–542.

Clark, A. J., 2017: Generation of Ensemble Mean Precipitation Forecasts from Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1569–1583.

Clark, A. J., and Coauthors, 2018: The 2018 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. 29th Conf. on Severe Local Storms, Stowe, VT, Amer. Meteor. Soc., 14B.8.

Davis, C. A., B. G. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784.

Davis, C. A., B. G. Brown, and R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795.

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.

Done, J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5,** 110–117. Doi:10.1002/asl.72.

Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171.

Duda, J. D., and W. A. Gallus Jr., 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Wea. Rev*., 142, 2198-2219.

Duda, J. D., X. Wang, F. Kong, M. Xue, and J. Berner, 2016: 742 Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Mon. Wea. Rev*., 144, 1887-1908.

Duda, J. D., X. Wang, and M. Xue, 2017: Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design. *Mon. Wea. Rev*., 145, 2001-2025.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568.

Gallo, B. T., and Coauthors, 2018: Spring Forecasting Experiment 2018 conducted by the Experimental Forecast Progam of the NOAA Hazardous Weather Testbed: Program overview and operations plan. Internal Tech. Doc., 46 pp.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757.

Gasperoni, N. A., X. Wang, and Y. Wang, 2020: A Comparison of Methods to Sample Model Errors for Convection-Allowing Ensemble Forecasts in the Setting of Multiscale Initial Conditions Produced by the GSI-Based EnVar Assimilation System. *Mon. Wea. Rev.*, **148**, 1177–1203.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430.

Gilleland, E., 2013: Testing Competing Precipitation Forecasts Accurately and Efficiently: The Spatial Prediction Comparison Test. *Mon. Wea. Rev.*, **141**, 340–355.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of Mountain Precipitation Forecasts from the Convection-Permitting NCAR Ensemble and Operational Forecast Systems over the Western United States. *Wea. Forecasting*, **33**, 739–765.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, 14, 155-167.

Hamill, T. M., and J. Juras, 2005: Brier Skill Scores, ROCs and Economic Value Diagrams Can Report False Skill.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. Mon. Wea. Rev., 134, 2318-2341.

Houze, R. A., Jr., 1993: *Cloud Dynamics*. Academic Press, San Diego, 573 pp.

Janjic, Z. I., 1994: The step-mountain ETA coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. Mon. Wea. Rev., 122, 927-945.

Jankov, I., and Coauthors, 2019: Stochastically perturbed parameterizations in an HRRR based ensemble. *Mon. Wea. Rev*., 147, 153-173.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693.

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

Johnson, A., and X.Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425.

Johnson, A., and X. Wang, 2017: Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Wea. Forecasting*, 32, 289-315.

Johnson, A., X. Wang, and S. Degelia, 2017: Design and implementation of a GSI-based convection-allowing ensemble-based data assimilation and forecast system for the PECAN field experiment. Part II: Overview and evaluation of a real-time system. *Wea. Forecasting*, 32, 1227-1251.

Johnson, A., X. Wang, Y. Wang, A. Reinhart, A. J. Clark, and I. L. Jirak, 2020: Neighborhood- and Object-Based Probabilistic Verification of the OU MAP Ensemble Forecasts during 2017 and 2018 Hazardous Weather Testbeds. *Wea. Forecasting*, **35**, 169–191.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21,** 167–181.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting,* **23,** 931–952.

Keil C, Heinlein FA, Craig G. 2014. The convective adjustment time-scale as indicator of predictability of convective precipitation. *Q. J. R. Meteorol. Soc.* **140**: 480–490.

Mansell, E. R., 2010: On sedimentation and advection in multimoment bulk microphysics. *J. Atmos. Sci.*, **67**, 3084–3094.

Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688.

Morrison, H., G. Thompson, and V. Tatarskii, 2009: Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One- and Two-Moment Schemes. *Mon. Wea. Rev.*, **137**, 991–1007.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties. Part I: Scheme Description and Idealized Tests. *J. Atmos. Sci.*, **72**, 287–311.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10,** 155–156.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor*., **12,** 595–600.

Murphy, A. H., 1986: A new decomposition of the Brier score: Formulation and interpretation. *Mon. Wea. Rev.*, **114,** 2671–2673.

Nakanishi, M., and H. Niino, 2009: Development of an 828 improved turbulent closure model for the atmospheric boundary layer. J. Meteor. Soc. Japan, 87, 895-912.

Pettet, C. R., and R. H. Johnson, 2003: Airflow and precipitation structure of two leading stratiform mesoscale convective systems determined from operational datasets. *Wea. Forecasting*, **18**, 685–699.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon. Wea. Rev, 136, 78-97.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev*., 142, 4519-4541.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2,** 191–201.

Schumacher, R. S., and A. J. Clark, 2014: Evaluation of Ensemble Configurations for the Analysis and Prediction of Heavy-Rain-Producing Mesoscale Convective Systems. *Mon. Wea. Rev.*, **142**, 4108–4138.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137,** 3351–3372.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather Forecasting*, 25, 262-280.

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, 29, 1295-1318.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Wea. Forecasting*, **30**, 1645–1654.

Schwartz, C. S., G. S. Romine, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015: A Real-Time Convection-Allowing Ensemble Prediction System Initialized by Mesoscale Ensemble Kalman Filter Analyses. *Wea. Forecasting*, **30**, 1158–1181.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev*., 145, 3397-3418.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's Real-Time Convection-Allowing Ensemble Project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (MRMS) severe weather and aviation products: Initial operating capabilities. Bull. Amer. Meteor. Soc, 97, 1617-1630.

Stephenson, D. B., C. A. S. Coelho, and I. T. Jolliffe, 2008: Two Extra Components in the Brier Score Decomposition. *Wea. Forecasting*, **23**, 752–757.

Stratman, D. R., and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 May 2011 tornadic supercells to microphysics parameterizations. *Mon. Wea. Rev.*, **145**, 2697–2721.

Tewari, M., and Coauthors, 2004: Implementation and verification 872 of the unified NOAH land surface model in the WRF model. 20th Conf. on Wea. Analysis and Forecasting/16th Conf on NWP, Seattle, WA, Amer. Met. Soc., 11-15

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.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. J. Atmos. Sci., 71, 3636-3658.

Trentmann J., and Coauthors, 2009. Multi-model simulations of a convective situation in low-mountain terrain in central Europe. *Meteorol. Atmos. Phys.* **103**: 95-103.

Wandishin, M. S., S. L. Mullen, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23,** 407–437.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.

Wolff, J. K., M. Harrold, T. Fowler, J. Halley-Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472.

Xue, M., Y. Jung, and G. Zhang, 2010: State estimation of convective storms with a two moment microphysics scheme and an ensemble Kalman filter: Experiments with simulated radar data *Quart. J. Roy. Meteor. Soc*, 136, 685-700.

Zhang FQ, Odins AM, Nielsen-Gammon JW. 2006. Mesoscale predictability of an extreme warm-season precipitation event. *Weather and Forecasting* **21**: 149–166.

Tables

Table 1. Visual representation of reflectivity bias adjustment process developed by Skinner et al. 2018.

Observations		**Forecasts**	
%	**dBZ**	**%**	**dBZ**
100	**60**	**100**	**70**
95	**50**	**95**	**60**
90	**45**	**90**	**50**
85	**40**	**85**	**45**
80	**35**	**80**	**40**
75	**30**	**75**	**35**

Table 2. Ensembles and their design.

Ensemble	**Dynamical Core**	**Member Number**	**Microphysics**	**LSM**	**PBL**
NMMB	NMMB	0-9	Ferrier-Aligo	MYJ	NOAH
ARW-SP	ARW	0-9	Thompson	MYNN	RUC
MM	NMMB	0-4	Ferrier-Aligo	MYJ	NOAH
	ARW	5-9	Thompson	MYNN	RUC
ARW-MP	ARW	0	Thompson	MYNN	RUC
		1	Thompson	MYJ	NOAH
		2	NSSL	YSU	NOAH
		3	NSSL	MYNN	NOAH
		4	Morrison	MYJ	NOAH
		5	P3	YSU	NOAH
		6	NSSL	MYJ	NOAH
		7	Morrison	YSU	NOAH
		8	P3	MYNN	NOAH
		9	Thompson	MYNN	NOAH

Table 3. Ensemble-to-ensemble comparisons.

Ensemble Comparison	Description
ARW-SP vs NMMB (SMSP)	Comparing how model and scheme choices impact storm mode and morphology forecasts.
ARW-SP vs ARW-MP (SPMP)	Analyzing the effects of physic scheme diversity on storm mode and morphology forecasts.
MM vs NMMB (MMSM)	Investigating the impacts of model dynamical core diversity on storm mode and morphology forecasts.
ARW-MP vs MM (MPMM)	Examining the relative effects model core and physic scheme diversity has on storm mode and morphology forecasts.

Table 4. Similar to Gasperoni (2020), ten retrospective case studies from 2015 to 2016 and their storm morphology description.

Case Date	Initialization Time	Synoptic Forcing	Case Description
May 16, 2015	2300 UTC	Strong	Single-cell dryline convection growing upscale into long lived squall line from TX to MO
May 25, 2015	1300 UTC	Strong	Multi-cell convection with large upscale growth into bowing squall line in southeast TX

June 26, 2015	0400 UTC	Weak	Nocturnal, bowing MCS, KS to MO;
			Nocturnal MCS Ohio Valley;
			Ensuing daytime convective initiation

July 14, 2015	1900 UTC	Strong	Southward advancing QLCS with associated cold front through decay, MS and OH valley
Sept 11, 2015	0100 UTC	Moderate	Supercellular convection growing upscale into squall line with advancing cold front
May 22, 2016	2300 UTC	Moderate	Isolated convection becoming outflow dominant QLCS, western TX
June 17, 2016	2000 UTC	Weak	Southward advancing squall line with bowing segment, southeastern US
July 06, 2016	0100 UTC	Weak	Southward propagating squall line growing in horizontal scale, MN to IL; convective clusters in KS and NE
July 07, 2016	0000 UTC	Weak	Supercellular convection growing upscale into bowing MCS, SD to MO
July 10, 2016	0400 UTC	Weak	Single and multi-cellular convection growing upscale into nocturnal MCS, Dakotas to IA

Figures

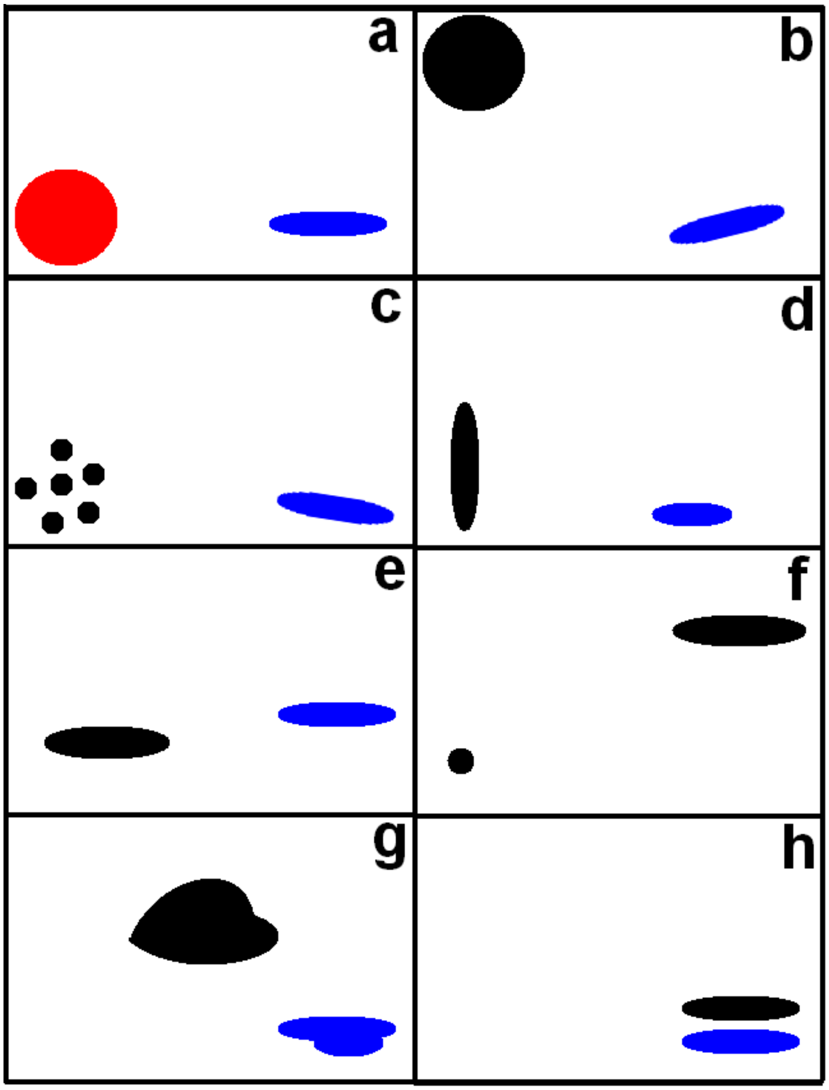


Figure 1. From JW12, visual demonstration of the basic concept of OBPROB. Each panel (a-h) represents an individual member forecast in which, respective forecast objects are defined.

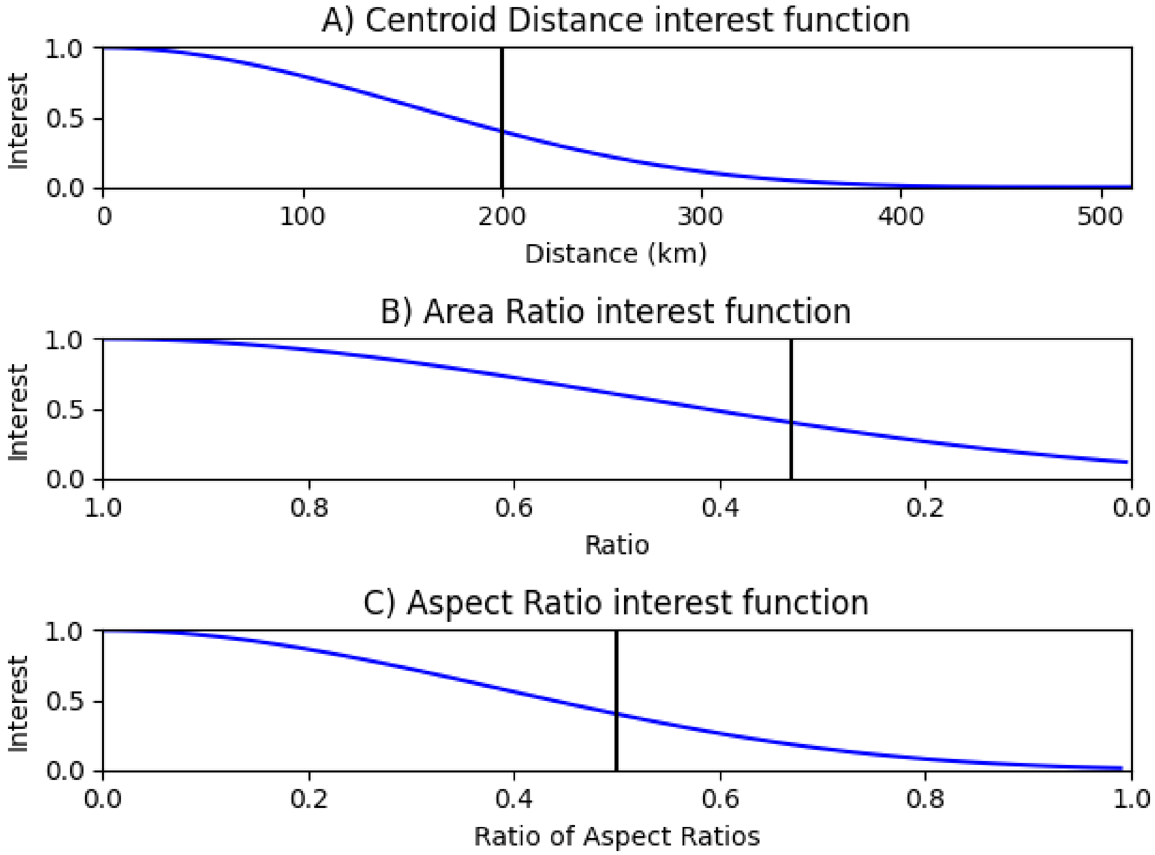


Figure 2. Interest functions to define similarity (i.e., interest) of the centroid location, area and aspect ratio attributes between two objects. E-folding values are marked by a black line.

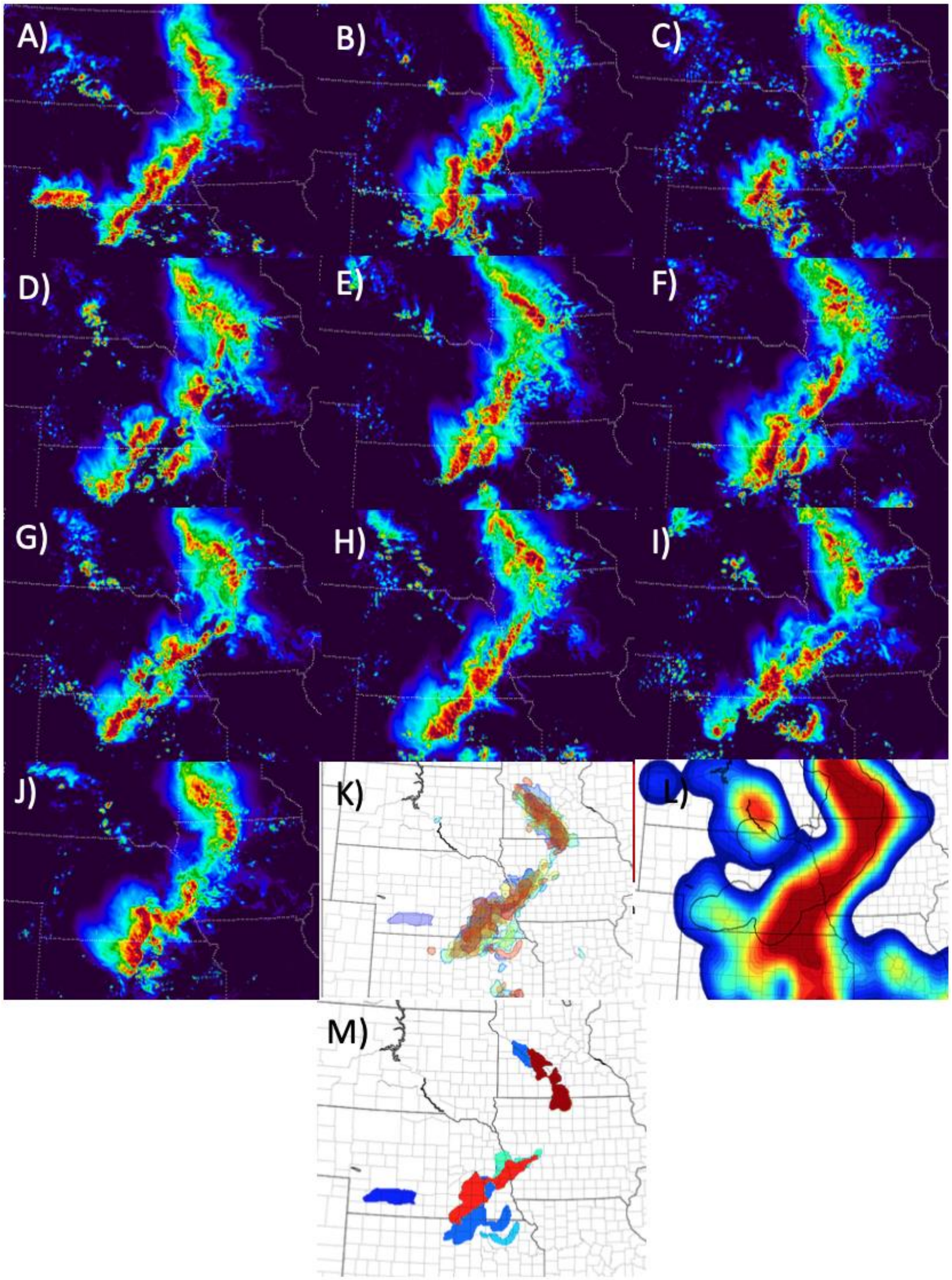


Figure 3. (A-J) Ten member Nonhydrostatic Multiscale Model on the B grid (NMMB) ensemble forecast initialized 00 Z, 7 July, 2016, valid at 06 Z July 07, 2016, (B) corresponding paintball, (C) neighborhood maximum ensemble probability (NMEP), and (M) OBPROB plots.

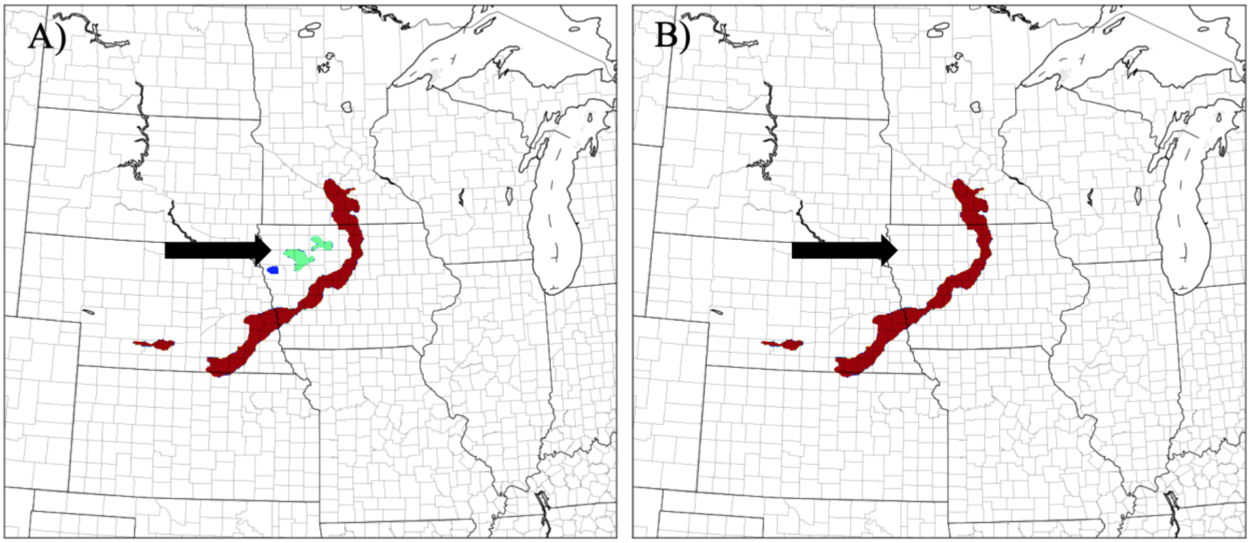


Figure 4. OBPROB observation object plots (A) before stratiform object filtering and (B) after stratiform object filtering. Observation object colors indicate convective scale category. Maroon objects correspond to meso-alpha or mesoscale organized convection, cyan objects represent meso-beta or multi-cell convection and blue objects signify meso-gamma or single-cell convection.

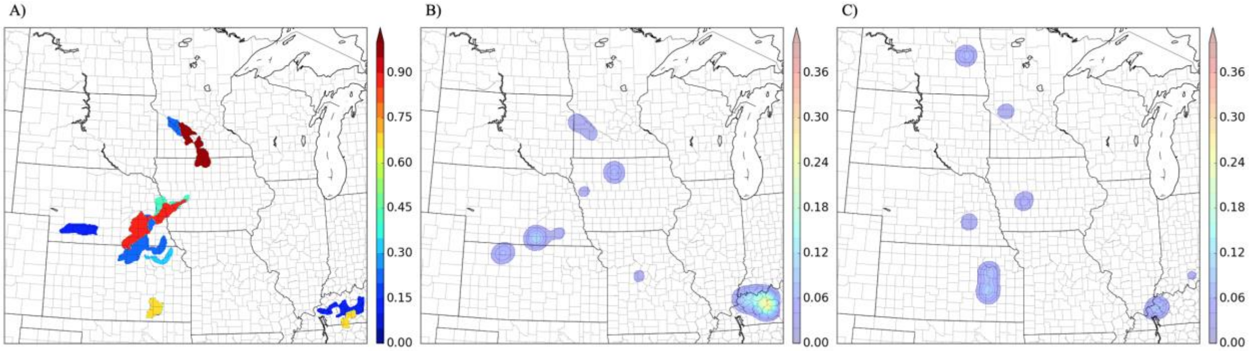
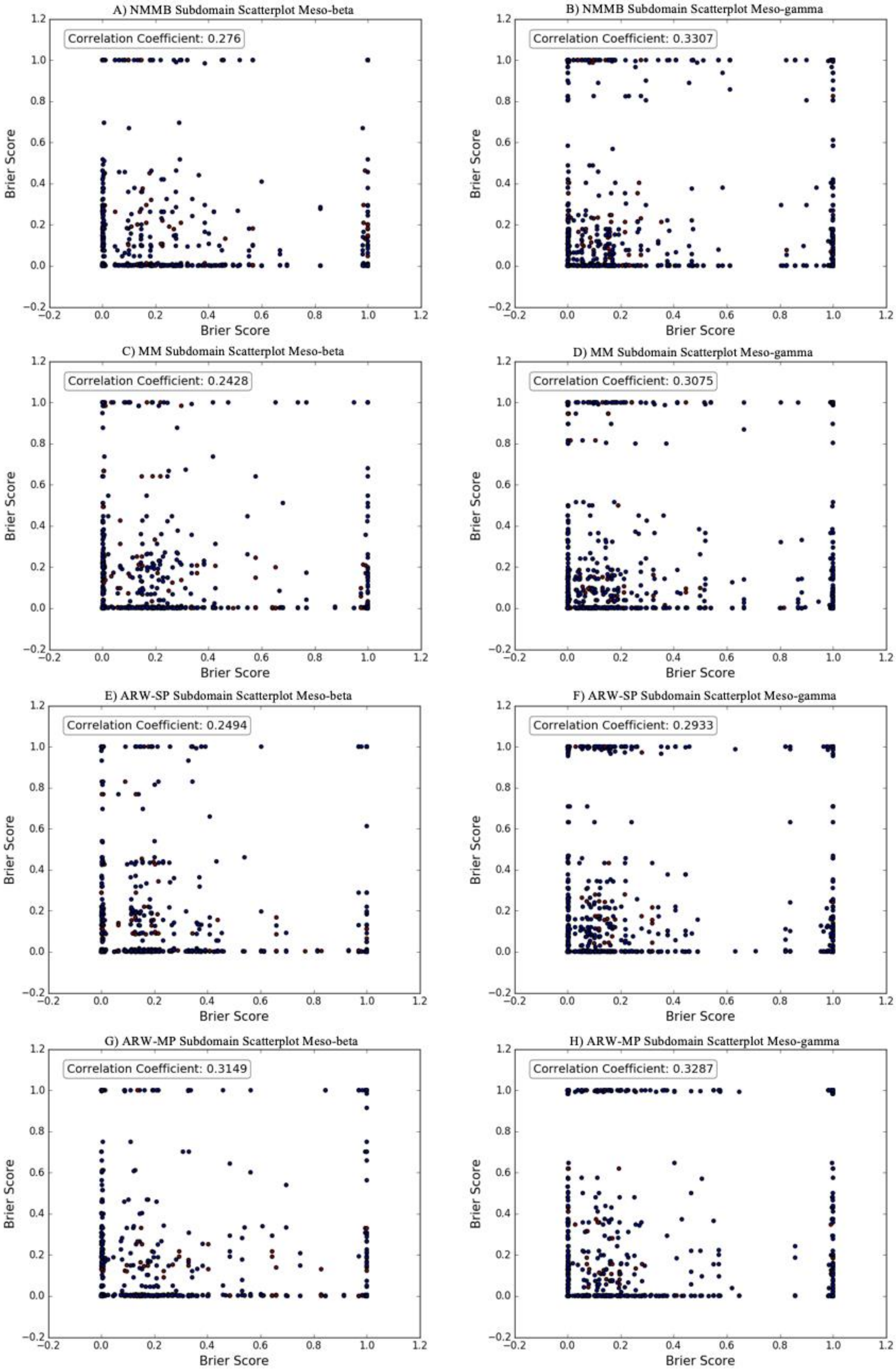


Figure 5. Plotting examples of A) meso-alpha (mesoscale organized) objects, B) meso-beta (multi-cell) objects, and C) meso-gamma (single-cell) objects. Meso-alpha objects use the OBPROB method, while meso-beta and meso-gamma use a Gaussian smoothed contour plot of object probabilities.



A) NMMB Subdomain Scatterplot Meso-beta

B) NMMB Subdomain Scatterplot Meso-gamma

C) MM Subdomain Scatterplot Meso-beta

D) MM Subdomain Scatterplot Meso-gamma

E) ARW-SP Subdomain Scatterplot Meso-beta

F) ARW-SP Subdomain Scatterplot Meso-gamma

G) ARW-MP Subdomain Scatterplot Meso-beta

H) ARW-MP Subdomain Scatterplot Meso-gamma

Figure 6. Ensemble subsampling scatterplot for meso-beta and meso-gamma scales. Each point represents a single subdomain sample. Respective Pearson correlation coefficients are plotted in the top left of each panel.

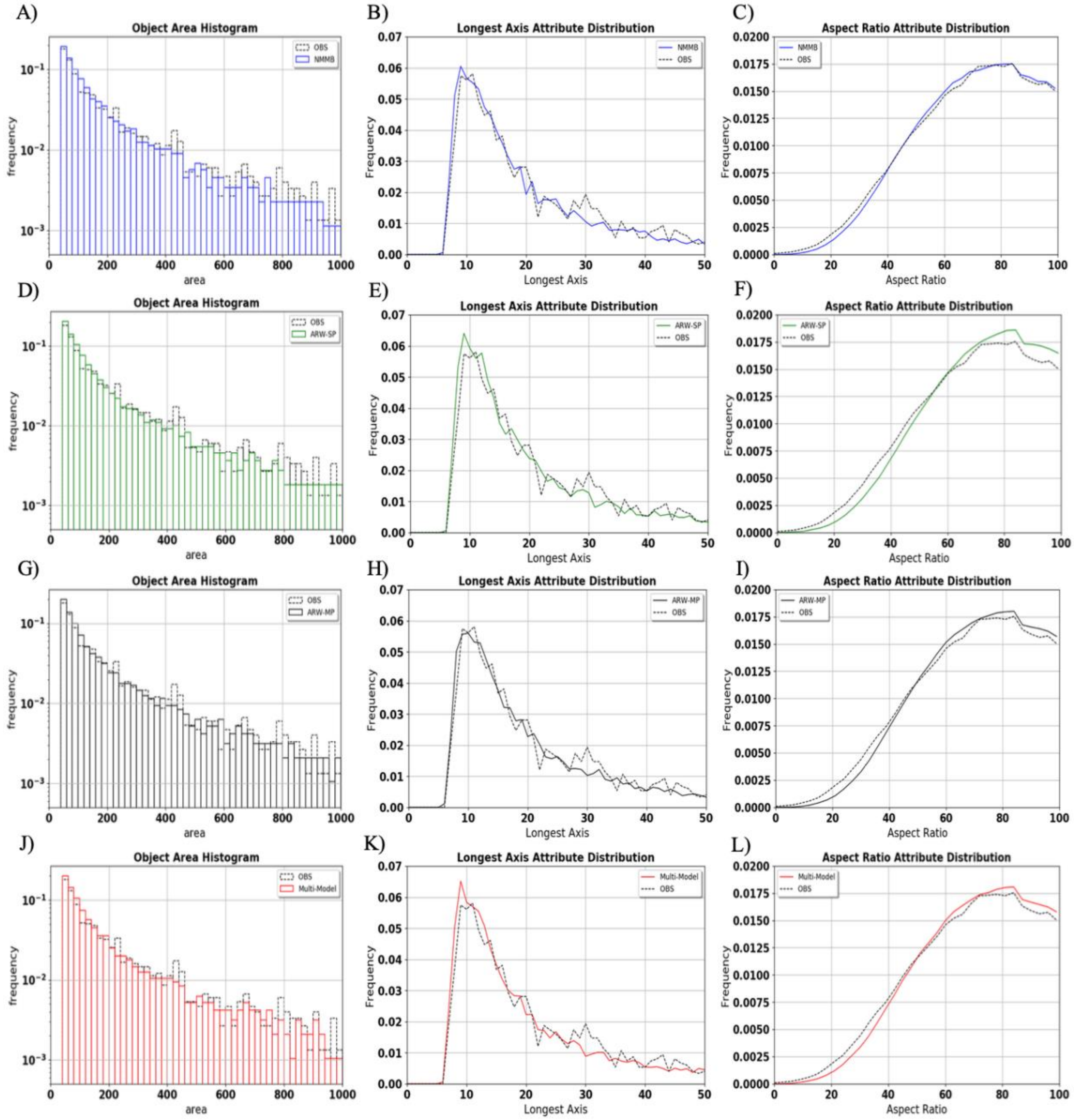


Figure 7. Normalized ensemble object attribute distributions for each ensemble (rows). Left column: Ensemble object area frequency distribution (solid) and observations (black, dashed) binned every 20 gpts. Center column: Ensemble object longest axis frequency distribution rounded to the nearest grid point (solid) and observations (black, dashed). Right column: Ensemble object aspect ratio distribution where each plotted point is averaged with +/- 2 nearest points.

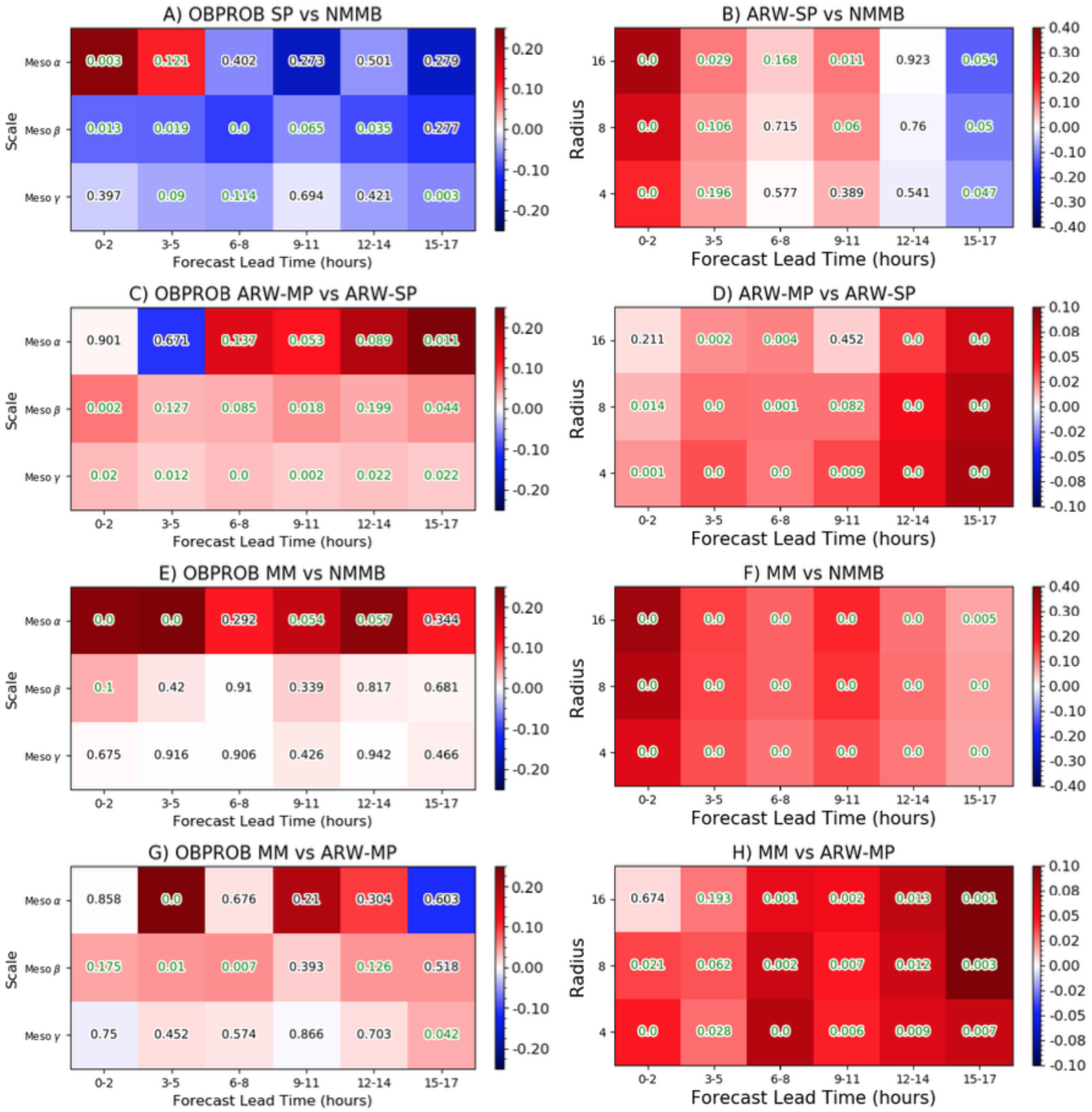


Figure 8. Left column: Object-based Brier score averaged difference over all cases based on lead time and binned every three forecast hours. Values shown are subtracted where red indicates added skill and blue indicates negative respective skill for the first ensemble listed in the title. Overlay p-value colors indicate presence of statistical significance (green) at the 80% confidence level. Right column: Same as left column except for neighborhood-based results. Meso-alpha, beta, and gamma correspond to a radius of 16, 8, and 4, respectively.

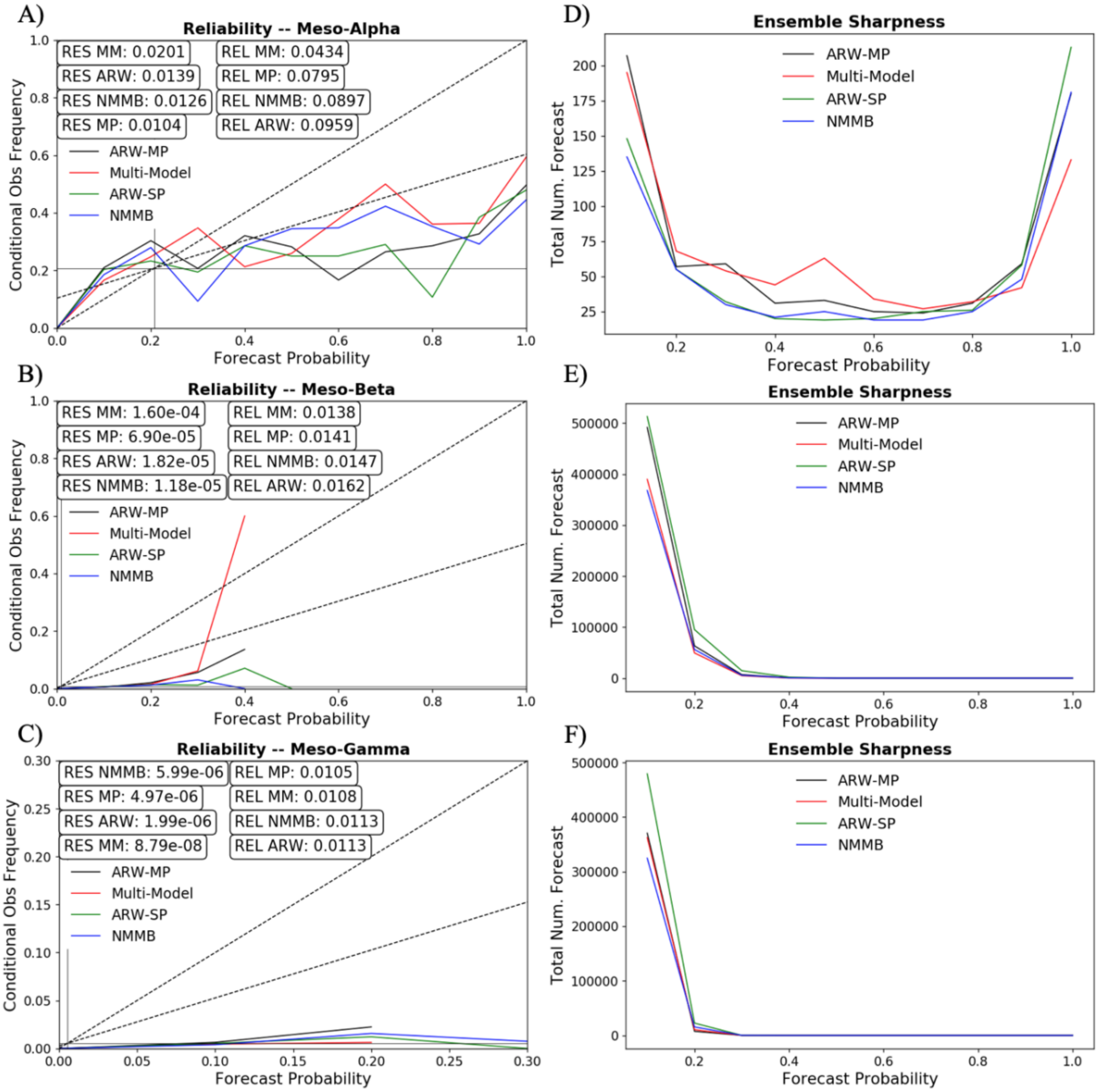


Figure 9. Meso-alpha, meso-beta, and meso-gamma reliability diagrams and corresponding sharpness plots. Reliability diagrams for meso-beta, and meso-gamma are not full diagrams given the low probability nature of the contour plots at these scales.

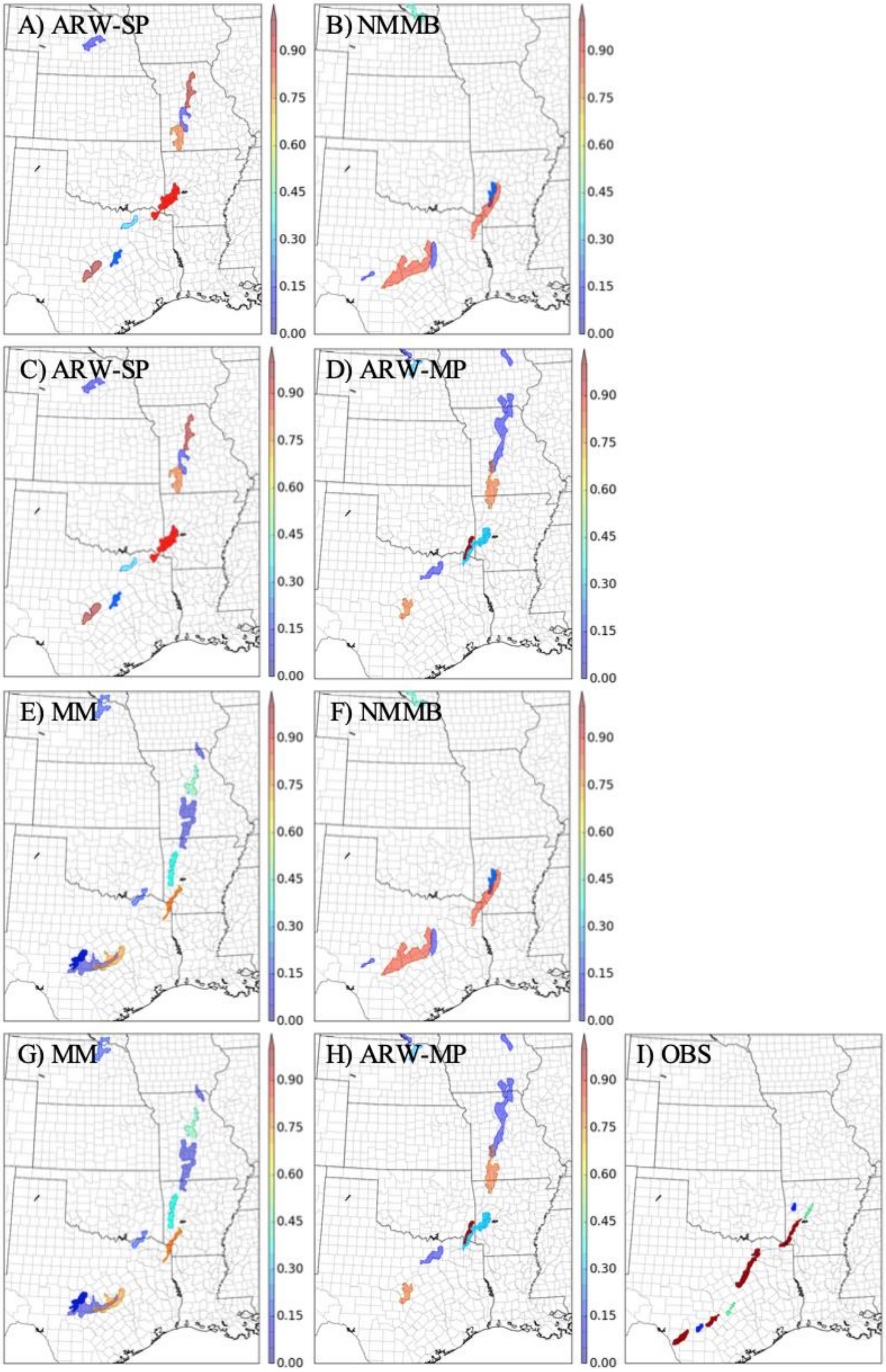


Figure 10. OBPROB plots for ensemble-to-ensemble comparisons from May 16, 2015 case study, valid at 01 Z. Each ensemble-to-ensemble comparison is organized by column: SMSP (A-B), SPMP (C-D), MMSM (E-F), and MPMM (G-H). Objects with transparency of 1.0 are matched to observations and objects with 0.5 transparency are not. As in Figure 4, observation plot (I) object colors indicate convective scale category.

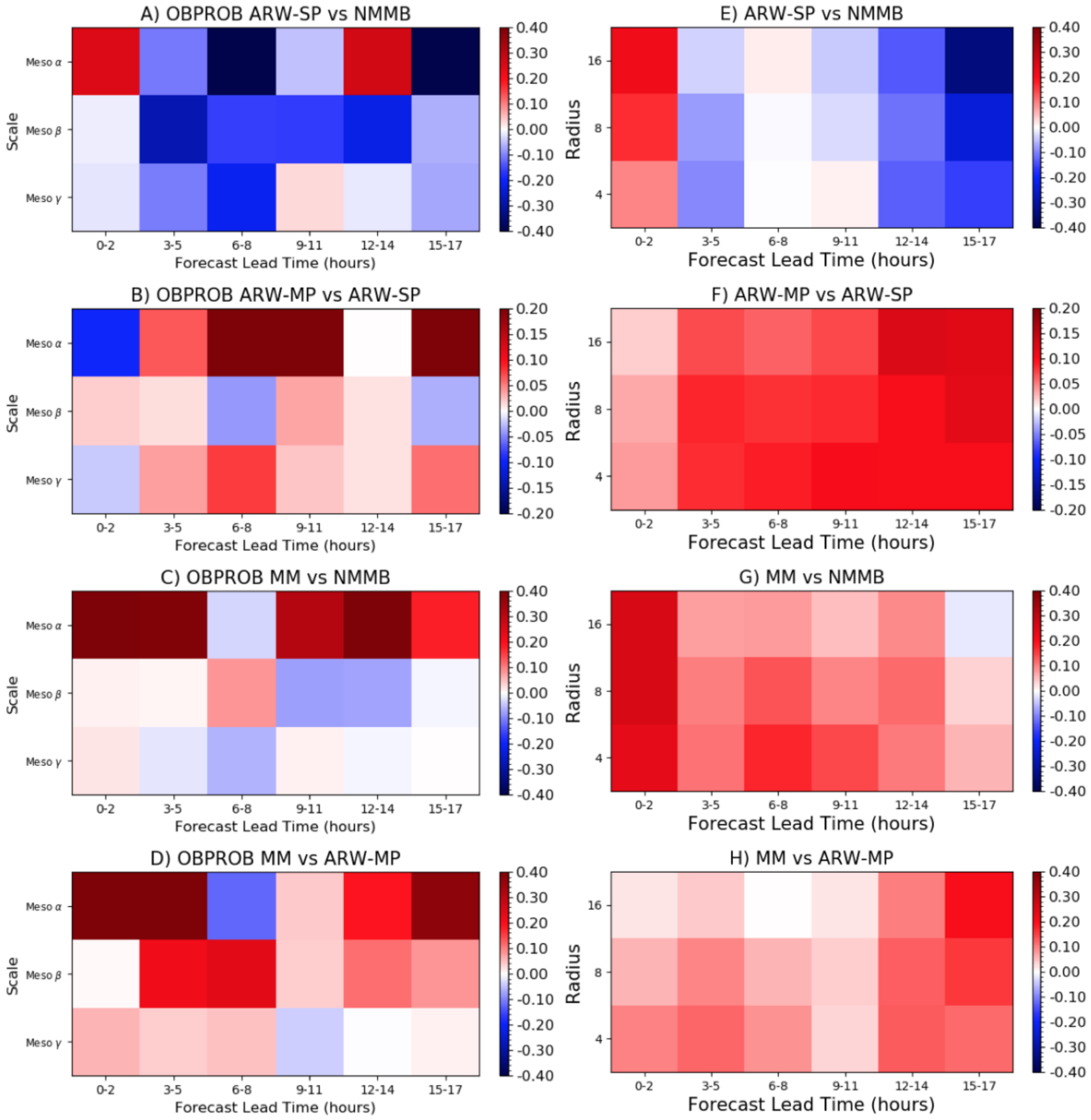


Figure 11. As in Figure 8, but for the May 16, 2015 case study. Statistical significance is not plotted due to data being representative of one case.

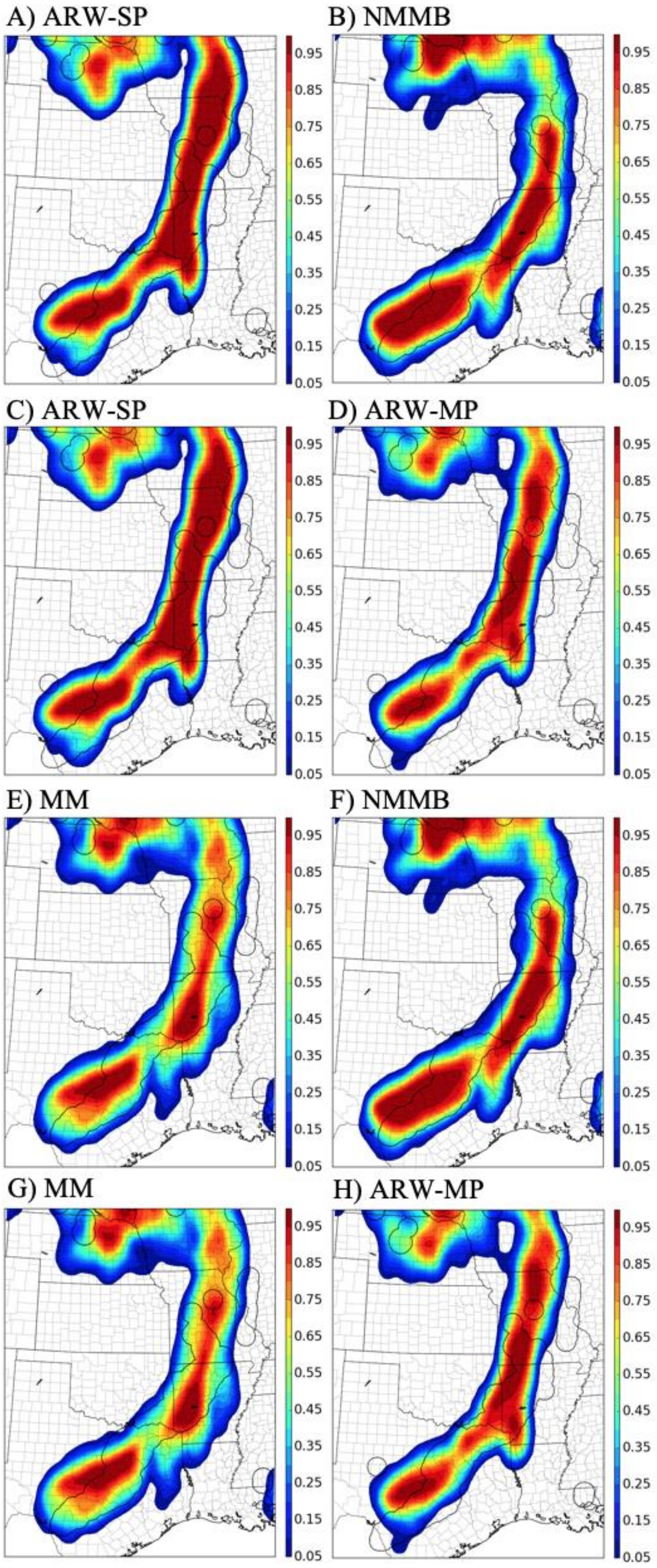


Figure 12. As in Figure 10, but for NMEP plots with a radius of 16 grid points.