Regime Dependent Verification and Calibration of a 10-Member Convection-Permitting Ensemble during the 2019 HWT SFE

Soleil Cotterell[*][†]

Georgetown University
Washington, D.C.

Aaron Johnson

School of Meteorology, University of Oklahoma
Norman, OK

Xuguang Wang

School of Meteorology, University of Oklahoma
Norman, OK

ABSTRACT

3km grid spaced forecasts generated during the 2019 NOAA Hazardous Weather Testbed Spring Forecast Experiment (HWT SFE) by the Multiscale Data Assimilation and Predictability group at the University Oklahoma are verified using the Neighborhood Maximum Ensemble Probability (NMEP). The verification was first performed on variable HWT defined domains and later extended to a large fixed CONUS domain. 24-hour forecasts of hourly maximum composite radar reflectivity initialized at 0000 UTC on 26 days during the spring of 2019 were evaluated. Forecasts on the HWT domains were generally skilled albeit with a notable over-forecasting bias, while forecasts on the fixed domain were generally poor. Cumulative Distribution Function (CDF) bias correction was performed in each domain and forecasts were reverified. Further, the fixed domain was segmented into sub-domains and a novel regional CDF (RCDF) bias correction approach was undertaken. CDF corrected forecasts on the fixed domain were still poorer than climatology, but significantly more skilled than without calibration. RCDF corrected forecasts on the fixed domain were significantly more skilled than CDF forecasts and were the only forecasts to exceed climatological skill. Synoptic pattern classification using Self Organizing Maps (SOMs) identified physically realistic synoptic patterns occurring over a ten-year climatology. naïvely using the SOM-derived synoptic classification to remove bias from meteorologically similar synoptic flow regimes separately did not generally improve forecast skill compared to regime-blind bias correction, though an interesting exception is noted. Suggestions are made for improving the robustness of the regime-dependent calibration scheme.

1. Introduction

The use of convection allowing model (CAM) ensembles is becoming more common in operational settings owing to their ability to explicitly resolve convective scale processes. Recent efforts have utilized 3km horizontal grid spacing to explicitly resolve convective scale processes in the sense that grid scale deep convection is permitted (Weisman et al. 2008; Kain et al. 2008), though the convection itself is not fully resolved (Bryan et al. 2003). Grid spacing is not the only consideration for effective convective scale modeling. Data assimilation schemes are an active area of research and attempt to improve synoptic, mesoscale and convective scale atmospheric representation in the initialization phase. The Multiscale Data Assimilation and Predictability (MAP) group at the University of Oklahoma is particularly interested in these assimilation schemes and has fielded experimental convective scale models in the 2017, 2018 and 2019 stagings of the NOAA Hazardous Weather Testbed Spring Forecast Experiment.

Verification of these CAM ensembles poses a unique challenge particularly due to the double penalty issue, whereby small spatial displacement of small scale, high amplitude features (convective storms in this case) are penalized as both a flase alarm and a miss despite still consituting a useful forecast. This means that typical met-

[*] Current Affiliation: National Weather Center Research Experiences for Undergraduates Program, Norman, Oklahoma

[†] *Corresponding author address:* Soleil Cotterell, Georgetown University, 3700 O Street NW, Washington, D.C. 20057
E-mail: sac261@georgetown.edu

rics for verifying lower resolution forecasts are not readily applicable to higher resolution forecasts (Nurmi 2003). Resultantly, unique forecast verification techniques must be employed to accurately gauge the spatio-temporal usefulness of a CAM forecast. Probabilistic neighborhood based verification techniques are among the most common high resolution verification methods. The Neighborhood Maximum Ensemble Probability (NMEP) (Schwartz and Sobash 2017) scheme is particularly attractive as it accounts for small spatial displacements, while also replicating the interpretation of probabilistic convective hazard outlooks for nearby significant severe weather events produced by the Storm Prediction Center (SPC).

Using the aforementioned methods, CAM forecast output variables can be verified as probabilistic fields (Theis et al. 2005; Johnson and Wang 2012; Bouallegue and Theis 2014; Clark et al. 2018). Furthermore, statistical post-processing techniques have been developed to correct systematic biases in the derived probabilistic fields. For example, Johnson and Wang (2012) showed that standard methods such as logistic regression and cumulative distribution function (CDF) bias correction are effective in increasing CAM skill when used as statistical post-processing. Furthermore, Kolczynski Jr. and Hacker (2014) showed that differing atmospheric regimes can influence data assimilation analysis increments, while Wang et al. (2019) explored regime dependent characteristics of precipitation forecasts generated under differing precipitating regimes across the Great Plains.

We hypothesize accordingly that it should be possible to not only classify the regime dependent error characteristics of probabilistic CAM forecasts, but also develop methods to correct for this regime dependent bias. Unsupervised machine learning provides the tools needed to facilitate such large scale classification projects. In particular, techniques such as K-Means clustering and Self-Organizing Maps (SOMs) have been employed in objective classification schemes (Philippopoulos et al. 2014; Wang et al. 2019). In particular, Kolczynski Jr. and Hacker (2014) highlighted the potential for SOMs to be used in the identification of patterns that lead to significant model error.

The SOM is a particularly attractive option for atmospheric classification due to its ability to reduce the dimensionality of complex, non-linear datasets. While other regime classification studies have relied on linear dimensonality reduction methods such as principal component analysis (PCA) and clustering methods such as k-means clustering (Mo and Ghil 1988; Michelangeli et al. 1998; Robertson and Ghil 1999; Michailidou et al. 2012; Mote 1998). In contrast, the SOM is less dependent on the relationship among input data in the sense that it is effectively a non-linear generalization of PCA and thus can more effectively parse non-linear correlations among variables. This is particularly important for studies involving a highly dimensional input space in which effective data visualization is effectively impossible.

Table 1. MAP CAM Configuration (Clark et al. 2019)

Member	Model	ICs	Micro-Physics	PBL	Radiation	LSM	LBC
1	WRF-ARW	hybrid EnKF-Var[c]	Thompson[d]	MYNN[b]	RRTMG[e]	RUC	GFS-Control
2	WRF-ARW	rEnKf[a]	Thompson	MYNN	RRTMG	RUC	GEFS
3	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
4	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
5	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
6	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
7	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
8	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
9	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS
10	WRF-ARW	rEnKf	Thompson	MYNN	RRTMG	RUC	GEFS

[a] Re-centered Ensemble Kalman Filter
[b] Mellor-Yamada-Nakanishi-Niino scheme (Mellor and Yamada 1982; Nakanishi 2001; Nakanishi and Niino 2004)
[c] Wang and Wang (2017)
[d] Thompson and Eidhammer (2014)
[e] Iacono et al. (2008)

In this paper we intend to examine the feasibility of implementing flow regime dependent bias correction techniques. In Section 2, the neighborhood verification methodology is defined and techniques for clustering atmospheric flow patterns are described. The verification and calibration of CAM forecasts generated during the 2019 Spring Experiment is performed in section 3. Section 4 outlines the results of the SOM implementation. An appendix containing technical details of the SOM implementation is also included .

2. Methods

a. Convection Permitting Ensemble Configuratrion

For the past three years, the Multiscale Data Assimilation and Predictability (MAP) group at the University of Oklahoma has run a 3km resolution WRF-ARW architecture CAM ensemble in the annual NOAA Spring Forecast Experiment. In 2019, the model was initialized at 0000z every weekeday during the SFE and ran through the 36 hour lead time. Though all forecast lead times were considered, emphasis was placed on next-day time-scale forecasts (f21-f27 hour lead times). Synoptic and mesoscale observations were assimilated with a similar model configuration as the NCEP High Resolution Rapid Refresh Ensemble (HRRRE), but with the addition of a 3d Ensemble-Variational (EnVar) hydbrid data assimilation system developed in Wang and Wang (2017) and evaluated in Duda et al. (2019). This scheme assimilated hourly High Resolution Rapid Refresh (HRRR) forecasts between 1800-0000z and NEXRAD reflectivity data every 20 minutes from 2300-0000z on the day preceding initialization (Clark et al. 2019). Specific ensemble member configurations and parameterization schemes are listed in Table 1.

b. Neighborhood Verification Framework

For each forecast case, a domain of interest was defined based on expected convective trends. Within the domain, the neighborhood maximum ensemble probability (NMEP) method was used to generate probabilistic forecasts (Schwartz and Sobash 2017). Furthermore, corresponding observed NEXRAD data was interpolated bilinearly to the WRF grid. These observations were thresholded (Table 2) and used to assign a deterministic event/no event rating to each gridpoint over the times of interest. Using this data, the Brier Score (Brier 1950), BS_{for} was computed for each lead time over each forecast case.

Using observational data assimilated from the National Severe Storms Laboratory's (NSSL) Multi-Radar/Multi-Sensor System (MRMS), a reference climatology was assembled for all forecast variables listed in table 2. This climatology was used to compute the reference Brier Score, BS_{ref} against which variable specific model skill is evaluated. To facilitate this comparison, the Brier Skill Score (Brier 1950), BSS, was computed for each case according to the standard formula,

$$BSS = 1 - \frac{BS_{for}}{BS_{ref}} \qquad (1)$$

In order to gauge forecast bias and potential skill, a reliability and resolution analysis was performed. Standard reliability diagrams (RDs) and receiver operating characteristic (ROC) curves were produced for each forecast lead time. The area under the curve (AUC) of each ROC curve was calculated and used for objective classification of model resolution.

These analyses are carried out in two stages. In the first stage, model output is verified in localized regions where convective activity is most likely to occur on each day. By definition, these areas of interest (AOIs) change from day to day and thus the results of this stage may not generalize to larger scales. In the second stage, we expand our scope and verify model performance over a fixed verification domain (Figure 1).

c. Regime Clustering: SOMs

We choose to examine atmospheric conditions at 00z as it is near the typical peak of next-day severe convective activity. Specifically, 500mb heights, u and v components of 850mb-500mb bulk shear and surface based CAPE are chosen as input variables. To characterize atmospheric states, we utilize 00z initialized 0.5 degree grid spaced Global Forecast System (GFS) analyses generated during the month of May from 2008-2018. 2019 data is excluded from the clustering to avoid bias as we intend to focus our efforts on verifying 2019 HWT forecasts.

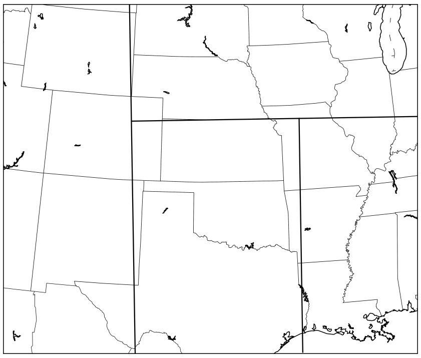


FIG. 1. The outer boundary represents the edge of the verification and assimilation domain for the SOM regime clustering. Interior lines represent domains within which regional bias correction will be performed.

As a compromise between efficiency and resolution, the GFS data is upscaled to 1x1 degree grids over a CONUS domain of interest (Figure 1). Upscaling is achieved by simple averaging over the grid with the resulting value being assigned to the grid centroid. For each predictor, the value at the centroid is appended to a vector hereafter referred to as a feature vector. Hence, each feature vector represents a forecast case and will have length $no.of\,grids \times 4$. These feature vectors may then be clustered into distinct atmospheric regimes using the SOM apparatus (Appendix I).

The number of patterns yielded depends on the size of the SOM. Though we would ideally like a small number of synoptic patterns given the small CAM forecast dataset available, we initially utilize a large 7×4 SOM, yielding 28 distinct synoptic regimes. This is done to facilitate an attempt at harvesting the emergent properties of large SOMs (Ultsch and Morchen 2009). K-means clustering is performed on the trained SOM to generate 2 consensus synoptic regimes. This number was chosen to ensure the robustness of our analysis given the small number of available forecasts.

Forecast cases are then classified using the Euclidean distance as an objective similarity metric. The distance between each feature vector and the vectors representing the two synoptic regimes are calculated in the usual way. The forecast case is then assigned to the synoptic regime corresponding to the smallest distance.

d. CDF Bias Correction

To correct for bias, the cumulative distribution function (CDF) correction method of Johnson and Wang (2012) was employed as a statistical post-processing technique. Given a forecast case, each forecast variable and its associated threshold of detection was considered. The observation percentile corresponding to the threshold was determined from a leave-one-out cross-validated distribution

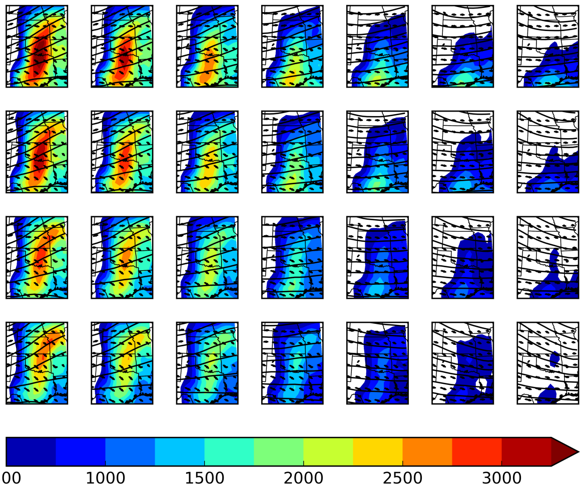
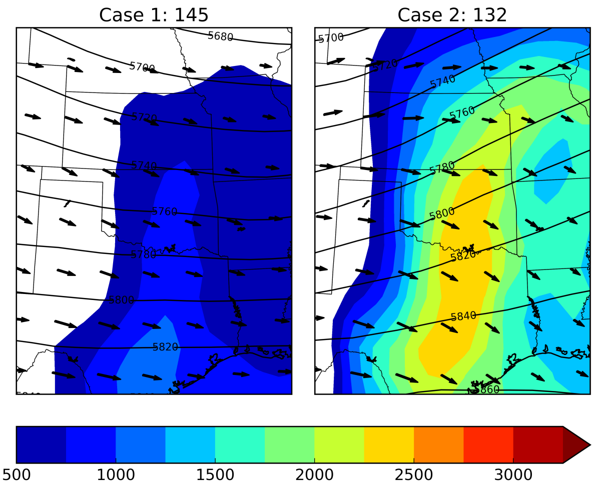


FIG. 2. 28 class SOM trained on 500mb heights (black contours), surface based CAPE (shaded) and 850-500mb bulk wind shear (arrows). Neighboring classes are objectively more similar than those father away.


Case 1: 145

Case 2: 132

FIG. 3. A k-means aided reduction of the 28-class SOM to 8 primary patterns. Shaded portions represent surface based CAPE, black contours denote 500mb heights and wind arrows denote 850-500mb bulk shear flow. The numbers above each diagram indicate the number of climatological days contained within each class. Case 1 denotes a marginally supportive severe environment, while case 2 suggests a more robust pattern.

of observations (Wilks and Hamill 2007). The computed percentile was then used to determine the forecast value at the corresponding percentile position. This value was taken to be the CDF corrected forecast threshold. For example, if the uncorrected forecast threshold for radar reflectivity was the 95th percentile of observations, then the corrected forecast threshold would be replaced by the 95th percentile value of the forecast distribution.

Calibration is also performed in two stages. In the first stage, distribution functions were assembled from forecasts and observations within each AOI, following the cross-validation method described above. All three variables within each forecast case were corrected in this fashion. Statistical significance testing using the bootstrapping technique of Hamill (1999) was then performed to determine the usefulness of the applied calibration.

The second stage of calibration is performed over the fixed domain. To explore the effects of regime and regional bias, we utilize a multifaceted approach. Firstly, each forecast is corrected using a cross-validated distribution assembled over all forecast cases. Secondly, forecasts cases are sorted and classified by synoptic regime. Forecasts within each regime are corrected using cross-validated distributions assembled from similarly classified forecast cases over the entire fixed domain. In both cases, we also partition the fixed domain into 4 regions based on the expected locations of regional bias (Figure 1). Each region is CDF corrected using forecasts and observations taken only from within that region.

3. Results and Discussion

a. Regime Clustering

The 28 atmospheric regimes classified by the SOM method are presented in a. Distinct regimes are readily apparent. The upper left of the map is consistent with midlevel ridging east of the plains mediating the advection of moist Gulf air into the Plains, resulting in a swath of high CAPE stretching from south Texas through western Missouri. Notably, the dominant shear pattern associated with strong advection of CAPE confirms that this pattern is a product of onshore surface flow. The lower right of the map suggests a strong zonal shear pattern and weak CAPE over the Plains concomitant with ridging to the west. This pattern precludes the advection of moisture from the south yielding a relatively stable environment.

The results of the k-Means reduction is presented in Figure 3. The 26 forecast cases were split equally between both regimes (Table 2). Case 1 demonstrates a marginal severe pattern, with largely zonal shear and marginal CAPE advection from the south. This case is representative of the nodes in the lower right of the SOM. On the other hand, case 2 presents an active severe regime with moderate to high CAPE present at most points east of the OK panhandle and directional bulk shear evident. This case is representative of the upper left nodes of the SOM.

b. *Verification on HWT Domains*

1) No Bias Correction

Within the areas of convective interest (AOI), uncalibrated composite radar reflectivity forecasts are more skillful than climatology at 0.05 significance at all considered lead times. Relative Operating Characteristic (ROC) analysis yielded areas under the curve (AUC) in excess of 0.8 for all considered lead times. This exceeds the typical 0.7 threshold suggested as a discriminator of forecast resolution. Attributes diagrams were produced and the associated reliability curves suggested that model performance was generally skillful, though negatively impacted by an over-forecasting bias.

Hail size forecasts generated using the maximum estimated size of hail (MESH) retrieval method were generally unskilled at all considered lead times. Significance testing confirmed this finding at and above the 0.05 significance level. Attributes diagrams indicated poor reliability, with a strong systematic over-forecasting bias (Figure), while ROC curves suggest good discrimination with AUC values close to 0.8.

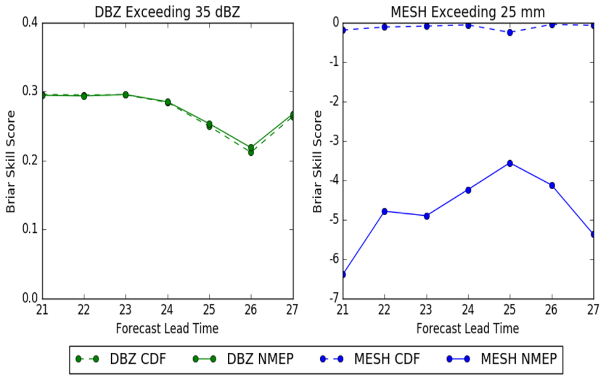


FIG. 4. Subplot showing the hourly averaged Brier Skill Score for radar reflectivity (left) and maximum estimated size of hail (right) taken over all forecast cases over all POIs

2) Bias Correction

CDF bias correction did little to correct the radar reflectivity over-forecast bias. In fact, bias correction slightly degraded forecast skill, though this degradation was not statistically significant. We hypothesize that bias averaging occurs due to the fact that the cumulative distribution function is sampled from different bias regimes owing to the lack of fixed domain in this stage of analysis.

Application of CDF correction to the MESH retrievals leads to vastly improved model skill, indicating consistently high bias over the entire WRF domain (Figure 4). However, the corrected forecasts are still less skillful than climatology at the 0.05 significance level. A limited exploration was conducted utilizing $0-3$ km updraft helicity as a proxy for hail size forecasting. Using 125 $m^2 s^{-2}$ as the event threshold for 1" hail yields a large increase in skill from uncalibrated hail size forecasts at all considered lead times.

TABLE 2. Absolute Brier Scores for each forecast case under different calibration schemes. The first column denotes the synoptic cluster (Figure 3) to which the forecast case belongs. The second column contains the forecast case's unique identifier. The remaining columns contain the uncalibrated Brier Score, fixed domain CDF corrected Brier score and regionally-corrected Brier score, respectively.

Class	Case	Reference Brier Score	Raw Brier Score	CDF Brier Score	RCDF Brier Score
1	1	0.0506	0.1372	0.1106	0.0622
1	2	0.1502	0.1619	0.1370	0.1178
1	3	0.0674	0.0918	0.0736	0.0691
1	5	0.1449	0.1687	0.1599	0.1552
1	9	0.1080	0.1565	0.1171	0.0761
1	10	0.1029	0.1637	0.1444	0.0985
1	11	0.0926	0.1290	0.1084	0.0883
1	12	0.0923	0.1309	0.1142	0.0933
1	13	0.0389	0.1036	0.0743	0.0444
1	17	0.1263	0.1813	0.1466	0.0908
1	19	0.0708	0.1041	0.0875	0.0571
1	25	0.1065	0.1580	0.1369	0.1165
1	26	0.1214	0.1595	0.1456	0.1231
2	4	0.0694	0.0942	0.0800	0.0691
2	6	0.0930	0.1054	0.0951	0.0844
2	7	0.1668	0.1923	0.1697	0.1400
2	8	0.1623	0.2592	0.2100	0.1546
2	14	0.0715	0.1089	0.0899	0.0751
2	15	0.1153	0.1287	0.1130	0.0920
2	16	0.1619	0.1707	0.1409	0.1160
2	18	0.1475	0.1879	0.1511	0.1012
2	20	0.1037	0.1517	0.1372	0.0972
2	21	0.0913	0.0963	0.0872	0.0673
2	22	0.0913	0.1382	0.1258	0.0900
2	23	0.1330	0.1462	0.1276	0.0940
2	24	0.1294	0.1287	0.1125	0.0857

The preliminary data presented in the following sections is drawn solely from the 24-hour lead time and is taken to be representative of the next-day lead time.

c. *Verification on Fixed Domain*

1) No Bias Correction

Reflectivity forecasts generated on the fixed domain were generally of lower skill than those on HWT domains. Though, this is to be expected given the comparatively larger area covered by the fixed domain. Evaluation of domain averaged model bias shows strong regional bias with positive bias noted west of the Rockies and over the northern Great Plains. These areas of regional bias are roughly spatially correlated with our fixed domain subdivisions.

Overall, reliability analysis indicates that forecasts suffer from significant over-forecasting bias and are generally unskilled (Figure 5). Though, a ROC AUC of 0.80 suggests that the model discriminates well between events and

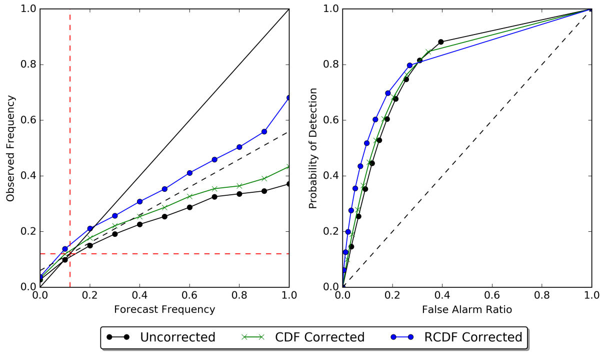


FIG. 5. (left) A reliability diagram performed in the fixed domain over all calibration regimes. The colored lines are reliability curves under different calibration regimes. The solid diagonal line is the line of perfect reliability (no forecast bias). The dashed diagonal line is the line of no skill - reliability curves above (below) this line represent forecasts that are skilled (unskilled). (Right) Receiver Operating Characteristic plots taken over each calibration routine in the fixed domain. Note that the probability of detection (POD) remains the same in each case, while the false alarm ratio progressively decreases. The RCDF routine provides the lowest false alarm ratio.

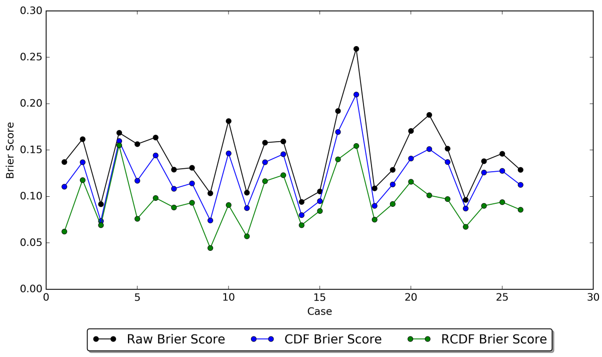


FIG. 6. A line plot of absolute Brier scores for each forecast case, in each calibration paradigm. Brier scores in the two calibration paradigm were significantly lower than in the uncalibrated paradigm. Brier scores in the RCDF calibrated paradigm were significantly lower than in the CDF calibrated paradgm.

TABLE 3. This table presents the results of calibration performed on forecast cases segregated by prevailing synoptic regime. The first half of the table presents results of the calibration performed on cases resident within the first regime, while the second presents the same for cases within the second regime.

Class	Case	CDF Brier Score	RCDF Brier Score
1	1	0.1115	0.0602
1	2	0.1349	0.1177
1	3	0.0731	0.0685
1	5	0.1598	0.1540
1	9	0.1126	0.0737
1	10	0.1441	0.0952
1	11	0.1079	0.0868
1	12	0.1144	0.0931
1	13	0.0733	0.0432
1	17	0.1425	0.0872
1	19	0.0872	0.0557
1	25	0.1366	0.1161
1	26	0.1450	0.1210
2	4	0.0804	0.0733
2	6	0.0954	0.0846
2	7	0.1697	0.1409
2	8	0.2126	0.1555
2	14	0.0911	0.0758
2	15	0.1136	0.0928
2	16	0.1406	0.1169
2	18	0.1552	0.1042
2	20	0.1381	0.1012
2	21	0.0879	0.0689
2	22	0.1262	0.0931
2	23	0.1276	0.0949
2	24	0.1123	0.0868

non-events. This suggests that poor model skill is mainly an artifact of poor calibration rather than a lack of predictive ability.

2) CDF Bias Correction

CDF bias correction performed over the fixed domain significantly improved skill in all forecast cases (Table 2; Figure 6), indicating the presence of systematic bias in each forecast case. However, average forecast skill remains negative (Table 4), suggesting the inability of CDF bias correction to sufficiently alleviate variable large-scale bias. Reliability analysis concurs with this finding, showing a marginal, yet perceptible improvement in the reliability curve compared to the calibrated case (Figure 5).

Given the strong performance of the model in the localized HWT domains and decent ROC AUC metrics in the fixed domain, the aforementioned results suggests that the CDF method is not fully accounting for extant sources of model bias.

3) Regime Dependent CDF Bias Correction

Considering all 26 cases together, it is apparent that regime dependent CDF corrected forecasts did not differ significantly from forecasts calibrated using the naïve CDF scheme. Performance of a regime-wise comparison similarly showed no significant intra-regime improvement in skill.

4) RCDF Bias Correction

RCDF forecasts were significantly more skillful and less biased that CDF forecasts (Figure 5). Furthermore,

TABLE 4. Case averaged Brier Scores and Brier Skill Scores under different calibration regimes. The first column corresponds to calibration applied to all cases without consideration for prevailing synoptic regime, while the second column presents the results of regime segregated calibration efforts. For the second column, scores from each synoptic regime are summed and then averaged together to facilitate accurate comparison with the first column.

Separate Calibration	No	Yes
Average Raw Brier Score	0.1444	0.1444
Average CDF Brier Score	0.1230	0.1228
Average RCDF Brier Score	0.0947	0.0947
Average Raw BSS	-0.3364	-0.3364
Average CDF BSS	-0.1365	-0.1367
Average RCDF BSS	0.1245	0.1238

the RCDF paradigm produce the only instance of positive case-averaged model skill (Table). This confirms our hypothesis that on large domains, region-blind CDF approach averages biases from dissimilar regions, thereby leading to ineffective calibration. The RCDF approach rectifies this deficiency.

Thus, it is apparent that these errors are not uniformly distributed over the domain, but are instead regionally concentrated. Moreover, the biases within these regions are likely persistent over most (if not all) forecast cases and thus constitute a regionally constrained systematic bias. Given this persistence, it is likely that these errors arise out of some deficiency in the model's physics, architecture or parameterization schemes.

5) REGIONAL AND REGIME DEPENDENT CDF BIAS CORRECTION

Similar to the regime dependent CDF bias correction scheme, there was no significant improvement in skill under the regional and regime dependent CDF bias correction scheme when all cases were considered together. However, a regime-wise comparison showed that there was a statistically significant improvement in skill within regime 1 under this calibration scheme, while there was no significant improvement within regime 2.

Given this result, we theorize that the first regime, which corresponds to a marginal severe environment, drives regionally dependent bias more strongly than the second more favorable synoptic regime. We further hypothesize that the lack of a similar result within regime 2 could be a result of regional characteristic averaging due to the reduction of 28 synoptic cases into 2. That is, regime 2 may not be fully representative of all cases assigned to it, in a synoptic sense. Though, given sample size constraints we cannot definitely comment on this disparity.

4. Conclusion

In this study, 2019 NOAA HWT SFE 24-hour lead time forecasts are verified and calibrated using a naïve CDF and newly proposed RCDF approach. The RCDF approach enables the exploration of the impact of regional forecast biases on model skill.

On localized HWT domains, the model generates highly skilled forecasts for hourly maximum composite radar reflectivity. CDF bias correction does not significantly improve model skill on this scale. On the expanded fixed domain, uncalibrated forecasts show no skill with respect to reference forecast. CDF bias correction induces a statistically significant improvement in skill, but forecasts still show no skill with respect to the reference forecast. RCDF forecasts increase skill significantly compared to both CDF and uncalibrated forecasts and are the only forecasts to demonstrate positive skill with respect to the reference forecast.

Using the SOM technique, we effectively facilitated the identification and clustering of physically realistic synoptic patterns occurring during the May severe weather season over a 9-year period. Using a k-means clustering routine, these patterns were combined to produce a reasonable representation of the two most dominant May severe weather patterns.

Naïve use of these SOM-derived synoptic classifications to remove bias from meteorologically similar synoptic flow regimes separately did not generally improve skill of the May 2019 CAE forecasts compared to regime-blind bias correction. The one exception to this observation was noted in regime 1. Within this region, the regime-dependent RCDF scheme induced a statistically significant improvement in forecast skill compared to the application of regime-blind RCDF correction within the same regime. A similar result was not observed in regime 2.

We hypothesize that the lack of a similar result within regime 2 could be a result of regional characteristic averaging due to the reduction of 28 synoptic cases into 2. That is, regime 2 may not be fully representative of all cases assigned to it, in a synoptic sense. Given sample size constraints we cannot definitely comment on this disparity, though we plan further study to investigate this dependence. In particular, we plan to work toward the generalization and application of these methods to experimental HRRRE forecasts, given the sizable repository of such forecasts. This will allow for a more robust representation and examination of the dominant synoptic states.

Acknowledgments. This study was supported by the NSF grant AGS-1560419. Thanks are extended to REU Director Daphne LaDue for facilitating this work in accordance with the NSF grant. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). We thank Amy McGovern, PhD and Ryan

Lagerquist for the invaluable background provided on the SOM technique.

APPENDIX

SOM Methodology

The SOM is a method of unsupervised competitive machine learning belonging to the class of artificial neural networks. In its simplest form, the SOM is a 2-dimensional grid of connected neurons (grid points). To each neuron we assign a vector of the same length as the feature vector hereafter referred to as weights. These weights are intilized either randomly or by sampling a 2-dimension principal component analysis subspace. This is discussed in some detail later on in the appendix.

At the most basic level, the training of the SOM grid is a competitive auction in the sense that neurons compete to ”win” each feature vector. The neuron that wins a feature vector auction will be closest to that feature vector in the Euclidean distance sense. This neuron is termed the best matching unit (BMU) and its weight is shifted closer to the feature vector that was won. In contrast to K-Means clustering, the SOM is endowed with a neighborhood function, hence any update of the BMU will also induce a shift in the weights of its neighbors (Wang et al. 2019). Furthermore, a learning rate parameter is defined which controls the degree to which a single update moves the winning neuron.

The training process is repeated for all 277 feature vectors over a user defined training length. As the training progresses, the neighborhood radius and learning rate are slowly reduced according to the neighborhood function chosen. Depending on the grid initialization chosen, the training processes may be either single staged or two staged. If the grid is initialized using random values, a coarse training must first be performed to loosely fit the map to the geometry of input data, followed by a fine tuning training to adjust the fit. On the other hand, use of Component Analysis (PCA) initialization negates the need for a two-step approach and can speed convergence of the technique in some circumstances (Kohonen et al. 1996). Given positive results from classification studies using PCA, we choose this initialization method and use a single fine-tuning training phase adopting parameters from Wang et al. (2019) and Kohonen et al. (1996).

Recent developments in SOM theory suggest that large maps can be used to facilitate emergent clustering, that is the identification of small data clusters not discernible using smaller scale maps (Ultsch and Morchen 2009). This is a good consideration as it should enable the identification and classification of rare atmospheric patterns. Though, given the small forecast sample size in this study, a study of these emergent properties is not feasible.

TABLE A1. Som configuration

Parameter	Value
Init.	PCA
No. Inputs	277
Dimensions (x,y)	7×4
Classes	28
Training length	277×100
Neighborhood Radius	3
Neighborhood Function	Gaussian
Learning Rate	0.02

Given this shortage of CAM forecast case, we must further reduce the 28 case in order to ensure a large enough sample size for comparison among regimes. K-means clustering is applied to post-training weights to facilitate an averaging of the 28 cases into 2 most probable cases. This configuration offered the most robust statistical framework, but also increased the chance of a forecast case being classified with a synoptic pattern not accurately/fully depicting the conditions in that case.

References

Bouallegue, B., and S. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Meteorological Applications*.

Brier, G., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**.

Bryan, G., J. Wyngaard, and J. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Monthly Weather Review*.

Clark, A., and Coauthors, 2018: The community leveraged unified ensemble (clue) in the 2016 noaa/hazardous weather testbed spring forecasting experiment. *Bulletin of the American Meteorological Society*.

Clark, A., and Coauthors, 2019: Spring forecasting experiment 2019: Program overview and operations plan. Tech. rep.

Duda, J., Y. Wang, X. Wang, and J. Carley, 2019: Comparing the assimilation of radar reflectivity using the direct gsi-based ensemble–variational (envar) and indirect cloud analysis methods in convection-allowing forecasts over the continental united states. *Monthly Weather Review*.

Hamill, T., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*.

Iacono, M., J. Delamere, E. Mlawer, W. Shepard, S. Clough, and W. Collins, 2008: Radiative forcing by long-lived greenhouse gases: calculations with the aer radiative transfer models. *Journal of the Atmospheric Sciences*.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Monthly Weather Review*.

Kain, J., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing nwp. *Weather and Forecasting*.

Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen, 1996: Som pak: The self-organizing map program package. Tech. rep.

Kolczynski Jr., W., and J. Hacker, 2014: The potential for self-organizing maps to identify model error structures. *Monthly Weather Review*.

Mellor, G., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Reviews of Geophysics and Space Physics*, **20**, 851–875.

Michailidou, C., P. Maheras, C. Anagnostopoulou, and I. Tegoulias, 2012: An objective classification of synoptic types over europe. *Advances in Meteorology, Climatology and Atmospheric Physics*.

Michelangeli, P.-A., R. Vautard, and B. Legras, 1998: Weather regimes: Recurrence and quasi stationarity. *Journal of the Atmospheric Sciences*, **52**, 1237–1256.

Mo, K., and M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. *Journal of Geophysical Research: Atmospheres*.

Mote, T., 1998: Mid-tropospheric circulation and surface melt on the greenland ice sheet. part ii: synoptic climatology. *International Journal of Climatology*, **18**, 131–145.

Nakanishi, M., 2001: Improvement of the mellor–yamada turbulence closure model based on large-eddy simulation data. *Boundary-Layer Meteorology*.

Nakanishi, M., and H. Niino, 2004: An improved mellor–yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*.

Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. Tech. rep.

Philippopoulos, K., D. Deligiorgi, and G. Kouroupetroglou, 2014: Performance comparison of self-organizing maps and k-means clustering techniques for atmospheric circulation classification. *International Journal of Energy and Environment*, **8**.

Robertson, A., and M. Ghil, 1999: Large-scale weather regimes and local climate over the western united states. *Journal of Geophysical Research: Atmospheres*, **12**, 1796–1813.

Schwartz, C., and R. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Monthly Weather Review*.

Theis, S., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, **12**, 257–268.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *Journal of the Atmospheric Sciences*.

Ultsch, A., and F. Morchen, 2009: Esom-maps: tools for clustering, visualization, and classification with emergent som. Tech. rep., University of Marburg, Germany.

Wang, J., X. Dong, A. Kennedy, B. Hagenhoff, and B. Xi, 2019: A regime based evaluation of southern and northern great plains warm-season precipitation events in wrf. *Weather and Forecasting*.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the gsi-based envar system: Methodology and experiment with the 8 may 2003 oklahoma city tornadic supercell. *Monthly Weather Review*.

Weisman, M., C. Davis, W. Wang, K. Manning, and J. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the wrf-arw model. *Weather and Forecasting*.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, **135**.