3.1 VERIFICATION AND CALIBRATION OF PROBABILISTIC PRECIPITATION FORECASTS DERIVED FROM NEIGHBORHOOD AND OBJECT BASED METHODS FOR A MULTI-MODEL CONVECTION-ALLOWING ENSEMBLE

Aaron Johnson* and Xuguang Wang

University of Oklahoma School of Meteorology and
Center for Analysis and Prediction of Storms (CAPS), Norman, OK

1. INTRODUCTION

The convection-allowing ensembles produced at the National Oceanographic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Experiments (e.g., Kong et al. 2009) provided an opportunity to study ensemble calibration at a convection-allowing resolution. Schaffer et al. (2011; referred to hereafter as SCH11) examined variations on a two-parameter reliability based method of ensemble precipitation forecast calibration using these data. One goal of the present study is to compare the effectiveness of several calibration methods for improving the skill of convection-allowing probabilistic precipitation forecasts during the 2009 HWT Spring Experiment.

Traditional grid point based probabilistic forecasts can have similar limitations as traditional grid point based deterministic verification methods at convection-allowing resolution (e.g., Gilleland et al. 2009). Neighborhood based methods have been proposed to derive probabilistic precipitation forecasts using nearby grid points to reduce such limitations. Object based methods have also been proposed to identify and compare *deterministic* forecast and observed weather systems at potentially different grid points. Another goal of the present study is to propose a way of deriving object based *probabilistic* forecasts which, along with the neighborhood method, will be used to verify the convection-allowing ensemble during the 2009 HWT Spring Experiment.

An object based hierarchical cluster analysis found that the precipitation forecasts systematically clustered based on model dynamic cores (Johnson et al. 2011). The last goal of the present study is to evaluate the impacts of model diversity on the skill of the neighborhood and object based probabilistic precipitation forecasts and the dependence of such impacts on calibration.

2. DATA AND METHODS

Since 2007 CAPS has generated experimental daily real-time, convection-allowing (4 km grid spacing), ensemble forecasts for the NOAA HWT Spring Experiments. In this study, the 2009 ensemble forecasts are verified and calibrated. The 2009 ensemble contained 20 members; 10 with Weather Research and Forecasting-Advanced Research WRF (WRF-ARW; Skamarock et al. 2005), 8 with WRF-Nonhydrostatic Mesoscale Model (WRF-NMM; Janjic 2003) and 2 with Advanced Regional Prediction System (ARPS; Xue et al. 2003). Initial and Lateral Boundary Condition (IC/LBC) perturbations were obtained from NCEP Short Range Ensemble Forecasts (SREF; Du et al. 2006) and perturbations to multiple physics schemes were included as detailed in Table 1. Quantitative Precipitation Estimates (QPEs) from the National Severe Storm Laboratory Q2 product are used as the verification data, referred to as observations. Twenty-six days of forecast and observation data between 30 April 2009 and 6 June 2009 are used

To explore the impact of the model diversity on the probabilistic precipitation forecast skill, three 8-member sub-ensembles denoted as “ARW”, “NMM”, and “MODEL” are defined as follows. “NMM” contains all 8 NMM members. “ARW” contains the 8 ARW members with the same IC/LBC perturbations used in the NMM sub-ensemble to emphasize model rather than IC/LBC differences (Table 1). A Multi-model sub-ensemble, “MODEL”, is defined by randomly choosing 4 members from each of the ARW and NMM subgroups on each day while still using the same 8 IC/LBC perturbations.

2.1 Neighborhood based probabilistic forecasts

Neighborhood Ensemble Probability (NEP; Schwartz et al. 2010) is the percentage of grid points from all ensemble member forecasts within a search radius that exceed a threshold. Thresholds of 2.54 mm, 6.5 mm, and 12.7 mm for both hourly and 6-hourly accumulations are used with a search radius of 48 km.

* *Corresponding Author Address*: Aaron Johnson, School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Suite 5900, Norman OK 73072; email: ajohns14@ou.edu

2.2 Object based probabilistic forecasts

The Method for Object based Diagnostic Evaluation (MODE; Davis et al. 2006) is used to define objects as contiguous areas exceeding a threshold of 6.5 mm after smoothing with a 4 grid point averaging radius. Attributes describing each object, such as area, aspect ratio and centroid location, are then calculated. The similarity of objects is quantified using Total Interest (please see Davis et al. 2006 and Johnson and Wang 2012 for further details) which is a function of the similarity of the attribute values.

An object based probabilistic forecasting method is proposed as follows. An object is determined to have occurred if the Total Interest between the forecast and observed objects exceeds a matching threshold. The forecast probability of occurrence is generated for each object in a control member forecast as the fraction of ensemble members with a matching object.

2.3 Verification method

Forecasts are verified using Brier Skill Score (BSS) where the reference forecast is obtained from the observed frequency of the event during the 2007-2010 HWT Spring Experiment periods, excluding the day of the forecast. The verification event is observed precipitation exceeding a threshold for grid point based forecasts and an observed object matching the forecast object for the object based forecasts.

Grid point based forecasts are verified every hour for the first 6 hours and every 3 hours thereafter for hourly accumulations and every 6 hours for 6-hourly accumulations. Object based forecasts are verified at the 1, 3, 6, 12, 18, 24 and 30 hour (6, 12, 18, 24 and 30 hour) lead times for hourly (6-hourly) accumulations. Statistical significance is determined using a one-sided Wilcoxan Signed Rank test for grid point based forecasts and permutation resampling for object based forecasts. To reduce the sensitivity of the object based probabilistic verification to the choice of the control member, BSS is aggregated over 50 repetitions with different randomly chosen control members for each forecast day.

2.4 Calibration methods

To test the sensitivity of the calibrations to the length of the training period, both 10-day and 25-day training periods are used. Unless otherwise stated, results are shown using the 10-day training period.

2.4.1 Reliability based method

Reliability based calibration is applied by placing the forecast probability into a discrete bin, determining the observed frequency for forecasts in that bin during a training period, and using that observed frequency as the calibrated forecast probability. Five hundred bins of variable width are used such that there are an approximately equal number of samples in each bin.

2.4.2 Two-parameter reliability based method

A two-parameter reliability based calibration of NEP based probabilistic forecasts has been proposed by SCH11. This method is similar to the reliability based method described above, with two key differences. First, the un-calibrated forecasts are divided into as many evenly spaced bins as the number of ensemble members, instead of 500 unevenly spaced bins. Second, each of the bins is further divided into 7 smaller bins based on the average forecast accumulation at all grid points used to compute the NEP.

An analogous two-parameter reliability based method is used for the object based probabilistic forecasts. The verification event for object based forecasts is the observation of a matching object, rather than the exceedance of a precipitation threshold. Therefore, the forecast probability bins are divided into 3 smaller bins based on the area of the objects instead of the average accumulation.

2.4.3 Logistic Regression

Logistic Regression (LR; Hamill et. al 2004) consists of fitting the following equation to a period of training data.

$$P = 1 - \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^{N} \beta_i x_i)} \qquad \text{(Eq. 1)}$$

In Eq. 1, P is the forecast probability, x_i are the predictors, N is the number of predictors, and β_i are the fitted coefficients. For neighborhood based forecasts the mean and standard deviation of each member's NEP, raised to the ¼ power, are used as the two predictors. For the object based forecasts the number of members with a matching object and the natural logarithm of the forecast object area are used as the two predictors.

2.4.4 Member-by-member bias adjustment

Calibration is also accomplished by adjusting for the bias of each member using the Cumulative Distribution Function (CDF) of forecasts and observations during training. For grid point based forecasts, each forecast accumulated precipitation value is replaced with the observed value that had the same cumulative probability as the un-calibrated forecast value during training. For the object based forecasts, the CDF of Total Interests between the control forecast objects and objects in other members is adjusted to be consistent with the CDF of Total Interests between the control forecast objects and observed objects.

3. NEIGHBORHOOD RESULTS

3.1 Uncalibrated verification

NEP probabilistic forecasts are more skillful than the traditional ensemble probability (TRAD), derived from the percentage of members forecasting the event at a single grid point (Fig. 1). TRAD has negative skill at many lead times while NEP has negative skill only for the least skillful lead times and thresholds.

Un-calibrated forecasts have skill maxima in the overnight and early morning hours (approximately 9-15 and 27-30 hour lead times, valid at 09-15 UTC and 03-06 UTC, respectively) and at the 1 hour lead time for hourly accumulation. There are skill minima during the afternoon for both accumulation periods (approximately 18-24 hour lead times, valid 18-00 UTC), at the 6 hour lead time for 6-hourly accumulation and at the 2-4 hour lead times for hourly accumulation (Fig. 1). Both the afternoon skill minimum and the minimum at 2-4 hours of forecast time correspond to the maxima in accumulated precipitation forecast bias of most members (not shown). The more pronounced 2-4 hour compared to the18-24 hour minimum corresponds with greater under-dispersion at 2-4 hour lead times. Such under-dispersion may be a result of assimilating the same radar data into all members without convective scale IC perturbations.

3.2 Calibrated verification

After calibration, the probabilistic forecasts for all accumulation thresholds have positive skill (Fig. 1). Most of the skill increase occurs during the un-calibrated skill minima. As such, the diurnal cycle of the forecast skill is less pronounced after calibration. During the un-calibrated skill minima, the skills of the probabilistic forecasts calibrated by various calibration methods are qualitatively similar in that differences in skill among calibration methods are smaller than differences in skill between calibrated and un-calibrated forecasts (Fig. 1). Although qualitatively similar, there are some significant differences among the calibration methods. In general, LR tends to be best for hourly accumulations and NEPrel tends to best for 6-hourly accumulations. NEPba tends to be less skillful than LR and NEPrel for the under-dispersive 2-4 hour lead times. Thus, in addition to the intended application (e.g., lead time, time of day and threshold) the types of errors to be corrected should be considered when choosing a calibration method. In general, SCH11 (NEPba) is the most (least) sensitive to the length of the training period based on the magnitude and significance of the differences in skill for 10 and 25 days of training (Fig. 2).

3.3 Sub-ensemble verification

Un-calibrated NEP forecasts from the ARW and NMM sub-ensembles have significantly different skill, with ARW being the most skillful, for most lead times and thresholds after the first couple of forecast hours (Fig. 3). There are many lead times when NMM has negative skill and ARW has positive skill. MODEL is only significantly more skillful than ARW for the 6.5 mm/hr threshold at the 30 hour lead time. The fact that ARW is similarly or significantly more skillful than MODEL suggests that there is little advantage of the multi-model ensemble for un-calibrated probabilistic precipitation forecasts using the neighborhood method.

The calibration decreases the differences in the skill among the ARW, NMM and MODEL sub-ensembles (Fig. 4). In contrast to the un-calibrated probabilistic forecast skill, the calibrated probabilistic forecast skill of the ARW sub-ensemble is generally only significantly greater than the NMM sub-ensemble skill for the smaller accumulations at longer accumulation periods (e.g., Fig. 4d,e). Unlike the un-calibrated probabilistic forecast skill, the calibrated multi-model sub-ensemble, MODEL, is more skillful than the calibrated single-model ARW and NMM sub-ensembles beyond the 24 hour lead time for all thresholds except 12.7 mm/hr. the larger skill of the MODEL sub-ensemble at longer lead times suggests that the inclusion of models with different convective scale attractors may allow for a more complete sampling of the true forecast probability distribution for the next-day convective scale precipitation forecasts.

4. OBJECT BASED RESULTS

4.1 Uncalibrated verification

The un-calibrated object based probabilistic forecasts have negative skill except for the 1-hour accumulation at 1 hour lead time and the 6-hour accumulation at 18 and 24 hour lead times (Fig. 5). The un-calibrated object based skill is negative at more lead times than the un-calibrated NEP and TRAD forecasts for the 6.5 mm thresholds corresponding to the 6.5 mm threshold used to define objects. Such result is consistent with the expectation that forecasting the probability that a specific object will occur is more difficult than forecasting whether a precipitation threshold will be exceeded.

Like the neighborhood based un-calibrated probabilistic forecasts (e.g., Fig. 1b), the un-calibrated hourly object based probabilistic forecasts have a diurnal cycle with skill minima at 3 and 18 hour lead times and skill maxima at 1 and 12 hour lead times (Fig. 5a). The more pronounced skill minimum at the 3 hour lead time, compared to the 18 hour lead time, is associated with greater bias in forecast Total Interest at the 3 hour lead time. Thus the ensemble members are systematically more similar to each other than to the observations. This can also be interpreted as a sign of ensemble under-dispersion.

4.2 Calibrated verification

All calibration methods improve on the skill of un-calibrated object based probabilistic forecasts (Fig. 5). Unlike the neighborhood based forecasts, the skills of bias adjusted forecasts are significantly less than the skills of forecasts calibrated with LR, SCH11 and NEPrel at most lead times. Only LR, SCH11 and NEPrel result in skillful forecasts at all lead times for both accumulation periods (Fig. 5). At most lead times LR is also significantly more skillful than SCH11 and NEPrel. Like the neighborhood based forecasts, calibration results in the greatest skill increases during periods of un-calibrated skill minima. Like the neighborhood based forecasts, the SCH11 calibration is the most sensitive calibration to the length of the training period with significantly higher skill for the longer training period at many lead times (Fig. 6).

4.3 Sub-ensemble verification

Like the neighborhood based forecasts, the un-calibrated object based probabilistic forecasts from the single-model ARW sub-ensemble are significantly more skillful than those from the single-model NMM sub-ensemble (Fig. 7a,b). Unlike the neighborhood based forecasts, there is no advantage of the un-calibrated multi-model sub-ensemble over the better ARW single-model sub-ensemble at any lead time (Fig. 7a,b). Like the neighborhood based probabilistic forecasts, the skill minimum at the 3-6 hour lead times is relatively more pronounced for the NMM sub-ensemble than the ARW and MODEL sub-ensembles.

The calibrated object-based probabilistic forecasts from the 8 member sub-ensembles have positive skill at all lead times (Fig. 7c,d). Like the neighborhood based probabilistic forecasts, the calibration reduces the differences in skill among the sub-ensembles (Fig. 7c,d). Unlike the neighborhood based probabilistic forecasts, there is no advantage of the multi-model sub-ensemble for the object-based probabilistic forecasts before or after calibration, even at the later lead times.

5.SUMMARY

The un-calibrated neighborhood based probabilistic forecasts show diurnal cycles of skill, with minima often below the level of no skill. Calibration primarily improves the neighborhood based probabilistic forecast skill during the un-calibrated skill minima. After calibration the neighborhood based probabilistic forecasts are skillful for all lead times and thresholds. During the un-calibrated skill minima the differences among calibration methods are smaller than the differences between calibrated and un-calibrated forecasts. No calibration method is superior for all lead times thresholds. The differences in skill between 10-day and 25-day training periods are also smaller than the differences between calibrated and un-calibrated forecasts. SCH11 is the most sensitive to training period length while NEPba is the least sensitive. Significant differences in skill between single-model ensembles are reduced by calibration. After calibration, the neighborhood based sub-ensemble forecasts have significantly higher skill after the 24 hour lead time with the multi-model ensemble than the single-model ensembles for some forecast thresholds.

The un-calibrated object based probabilistic forecasts have less skill than the un-calibrated neighborhood based forecasts. Calibration increases the skill of the object-based forecasts at all lead times especially, but not only, during the un-calibrated skill minima. Skillful forecasts are obtained at all lead times after calibration with LR, SCH11 and reliability based calibration but not after bias adjustment of Total Interest. LR is the most skillful calibration for the object-based probabilistic forecasts. The effect of training period length on the object-based probabilistic forecasts is also greatest for SCH11 and least for bias adjustment. Object based calibration also reduces the significant differences in skill between the single-model sub-ensembles. The calibrated object based sub-ensemble forecasts show no advantage of using multiple models.

Both neighborhood and object based forecasts have the most pronounced skill minima at 2-4 hour lead times which correspond to enhanced ensemble under-dispersion. Future work on how to include mesoscale and convective scale initial condition perturbations is suggested to reduce under-dispersion at early forecast lead times.

ACKNOWLEDGEMENTS

The authors are grateful to NSSL for the QPE verification data and NCAR for making MODE source code available. This research was supported by Science Applications International Corporation (SAIC) as a sponsor of the AMS Graduate Fellowship program and University of Oklahoma faculty start-up award 122-792100. The CAPS real-time forecasts were produced at the Pittsburgh Supercomputing Center (PSC), and the National Institute of Computational Science (NICS) at the University of Tennessee, and were mainly supported by the NOAA CSTAR program (NA17RJ1227). Some computing was also performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU). Fanyou Kong, Ming Xue, Kevin Thomas, Yunheng Wang, Keith Brewster and Jidong Gao of CAPS are thanked for the production of the ensemble forecasts.

REFERENCES

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

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members, *Preprint,WMO Expert Team Meeting on Ensemble Prediction System, Exeter*, UK, Feb. 6-10, 2006 [available online http://wwwt.emc.ncep.noaa.gov/mmb/SREF/reference.html].

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

Janjic´, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.,* **82,** 271–285.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Season-long ensemble clustering and implication for optimal ensemble design. *Mon. Wea. Rev.,* **139**, 3694-3710.

Johnson, A., X. Wang, 2012: Verification and calibration of neighborhood and object based probabilistic precipitation forecasts from a multi-model convection allowing ensemble. In Preparation.

Kong, F., M. Xue, K.W. Thomas, J. Gao, Y. Wang, K. Brewster, K.K. Droegemeier, J. Kain, S. Weiss, D. Bright, M. Coniglio, and J. Du, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment, *10th WRF Users' Workshop, NCAR Center Green Campus*, Boulder, CO, June 23-26, 2009, Paper 3B.7.

Schaffer, C. J., W. A. Gallus, M. Segal, 2011: Improving probabilistic ensemble forecasts of convection through the application of QPF-POP relationships. *Wea. Forecasting*, **26,** 319-336.

Schwartz, C.S., J.S. Kain, S.J. Weiss, M. Xue, D.R. Bright, F. Kong, K.W. Thomas, J.J. Levit, M.C. Coniglio, and M.S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. NCAR Tech Note NCAR/TN-468_STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.].

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Physics*, **82,** 139-170.

TABLES AND FIGURES

Member	IC	LBC	R	MP	PBL	SW	LSM
ARW CN	CN	NAMf	Y	Thom.	MYJ	Goddard	Noah
ARW C0	NAMa	NAMf	N	Thom.	MYJ	Goddard	Noah
ARW N1	CN – em	em N1	Y	Ferr.	YSU	Goddard	Noah
ARW N2	CN – nmm	nmm N1	Y	Thom.	MYJ	Dudhia	RUC
ARW N3	CN - etaKF	etaKF N1	Y	Thom.	YSU	Dudhia	Noah
ARW N4	CN- etaBMJ	etaBMJ N1	Y	WSM6	MYJ	Goddard	Noah
ARW P1	CN + em	em N1	Y	WSM6	MYJ	Dudhia	Noah
ARW P2	CN + nmm	nmm N1	Y	WSM6	YSU	Dudhia	Noah
ARW P3	CN + etaKF	etaKF N1	Y	Ferr.	MYJ	Dudhia	Noah
ARW P4	CN + etaBMJ	etaBMJ N1	Y	Thom.	YSU	Goddard	RUC
NMM CN	CN	NAMf	Y	Ferr.	MYJ	GFDL	Noah
NMM C0	NAMa	NAMf	N	Ferr.	MYJ	GFDL	Noah
NMM N2	CN - nmm	nmm N1	Y	Ferr.	YSU	Dudhia	Noah
NMM N3	CN - etaKF	etaKF N1	Y	WSM6	YSU	Dudhia	Noah
NMM N4	CN - etaBMJ	etaBMJ N1	Y	WSM6	MYJ	Dudhia	RUC
NMM P1	CN + em	em N1	Y	WSM6	MYJ	GFDL	RUC
NMM P2	CN + nmm	nmm N1	Y	Thom.	YSU	GFDL	RUC
NMM P4	CN + etaBMJ	etaBMJ N1	Y	Ferr.	YSU	Dudhia	RUC
ARPS CN	CN	NAMf	Y	Lin	TKE	2-layer	Noah
ARPS C0	NAMa	NAMf	N	Lin	TKE	2-layer	Noah

TABLE 1. Details of ensemble configuration with columns showing the members, Initial Conditions (ICs), Lateral Boundary Conditions (LBCs), whether radar data is assimilated (R), and which Microphysics scheme (MP; Thompson, Ferrier, WRF Single Moment 6-class, or Lin microphysics), Planetary Boundary Layer scheme (PBL; Mellor-Yamada-Janjic, Yonsei University or Turbulent Kinetic Energy-based scheme), Shortwave radiation scheme (SW; Goddard, Dudhia or Geophysical Fluid Dynamics Laboratory scheme), and Land Surface Model (LSM; Rapid Update Cycle or NOAH) was used with each member. Please see Johnson et al. 2011 for physics scheme references. NAMa and NAMf are the direct NCEP-NAM analysis and forecast, respectively, while the CN IC has additional radar and mesoscale observations assimilated into the NAMa. Perturbations added to CN members to generate the ensemble of ICs, and LBCs for the SSEF forecasts are from NCEP SREF (Du et al 2006). SREF members are labeled according to model dynamics: nmm members use WRF-NMM, em members use WRF-ARW (i.e., Eulerian Mass core), etaKF members use Eta model with Kain-Fritsch cumulus parameterization, and etaBMJ use Eta model with Betts-Miller-Janjic cumulus parameterization.

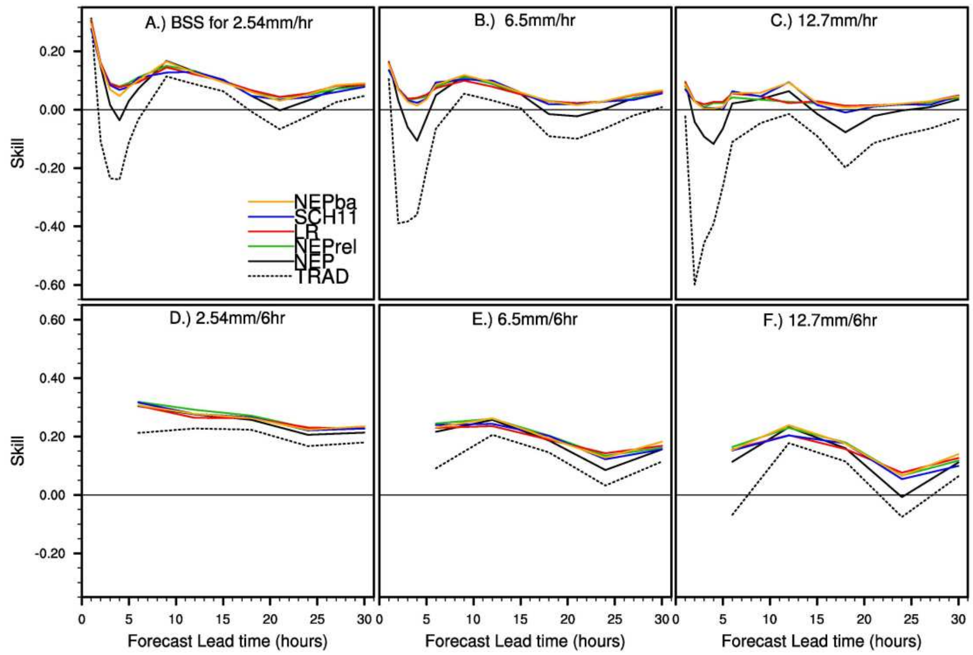

Figure 1. Brier Skill Score of traditional ensemble probability (TRAD), NEP without calibration (NEP), NEP calibrated using the reliability based method (NEPrel), Logistic Regression (LR), SCH11, and NEP from bias adjusted members (baNEP) for (a) 2.54mm/hr, (b) 6.5mm/hr, (c) 12.7mm/hr, (d) 2.54mm/6hr, (e) 6.5mm/6hr, and (f) 12.7mm/6hr.

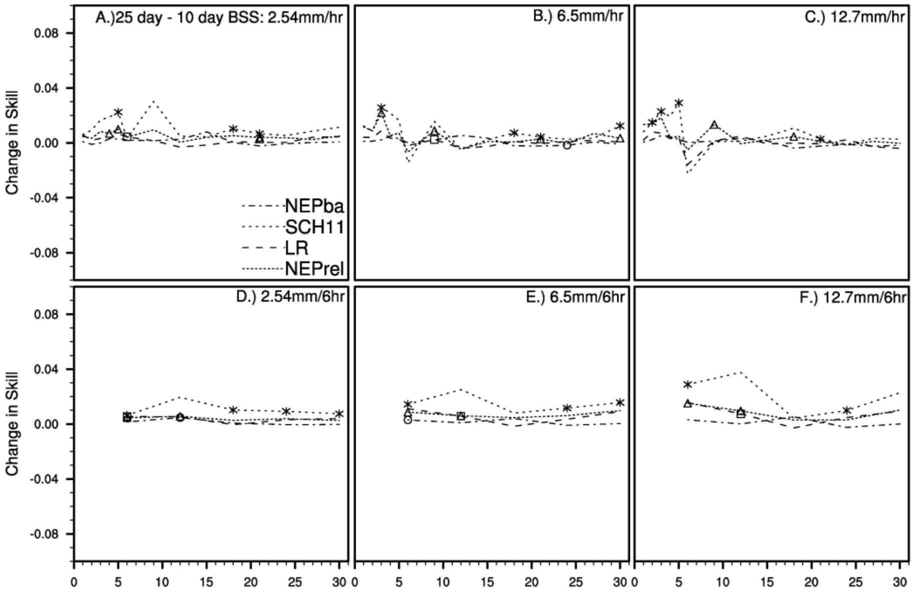

Figure 2. Difference in Brier Skill Score of neighborhood based probabilistic forecasts between 25 days and 10 days of training for ensemble calibration methods for (a) 2.54mm/hr, (b) 6.5mm/hr, (c) 12.7mm/hr, (d) 2.54mm/6hr, (e) 6.5mm/6hr, and (f) 12.7mm/6hr. Markers indicate statistically significant difference between 10 and 25 days of training at the 95% confidence level.

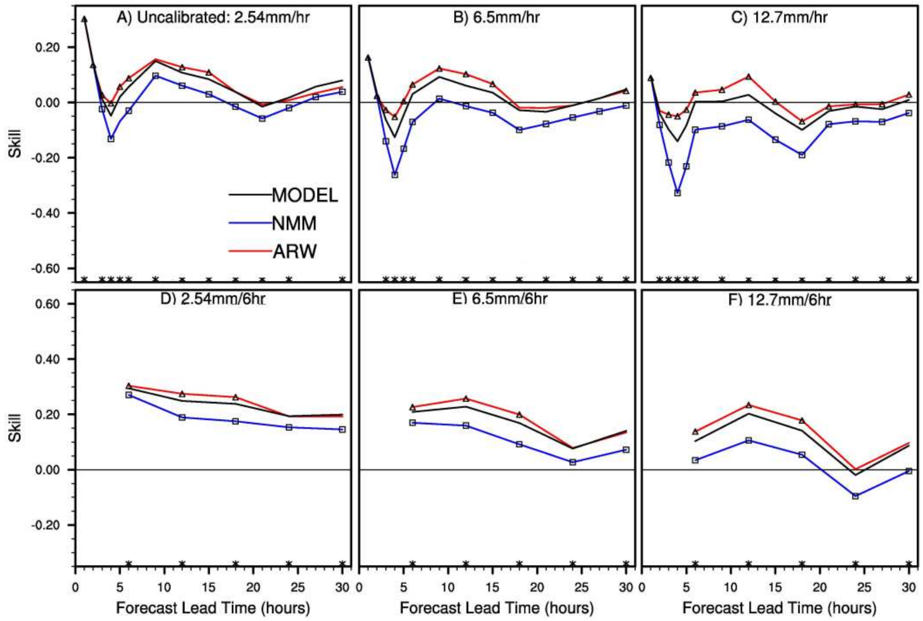

Figure 3. Brier Skill Score of uncalibrated NEP from single model (ARW and NMM) and multi-model (MODEL) sub-ensembles for thresholds of (a) 2.54mm/hr, (b) 6.5 mm/hr, (c) 12.7 mm/hr, (d) 2.54 mm/6hr, (e) 6.5 mm/6hr and (f) 12.7mm/6hr. Statistically significant difference from MODEL is indicated by a square or triangle, for NMM or ARW respectively, and significant difference between ARW and NMM is indicated by an asterisk along the horizontal axis.

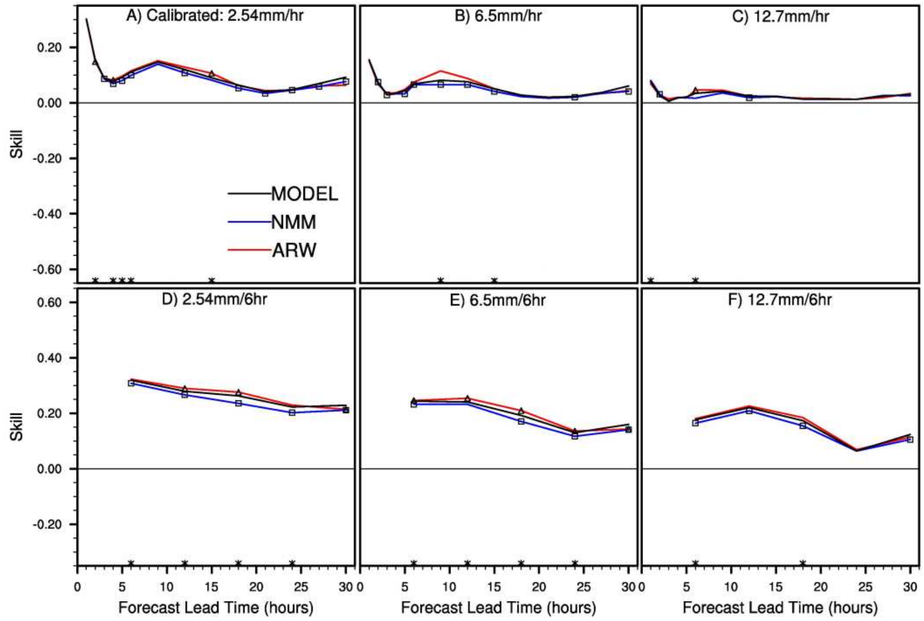

Figure 4. As in Figure 3, except for NEP forecasts calibrated using the reliability based method.

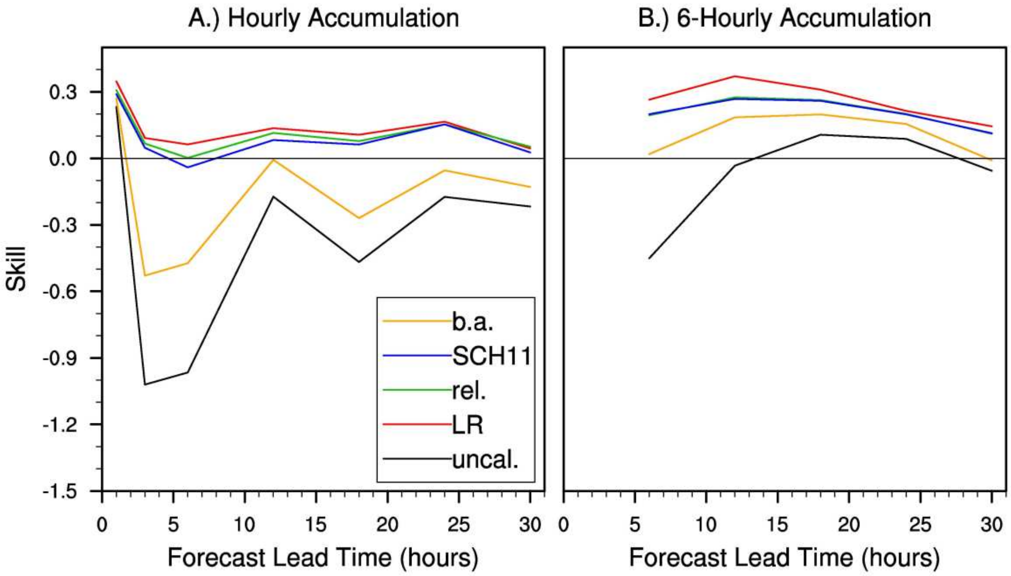

Figure 5. Brier Skill Score of object based probabilistic forecasts without calibration (uncal.), calibrated with logistic regression (LR), the reliability based method (rel.), the two-parameter reliability based method of SCH11 (SCH11), and the individual member bias (ba), for (a) hourly and (b) 6 hourly accumulation periods.

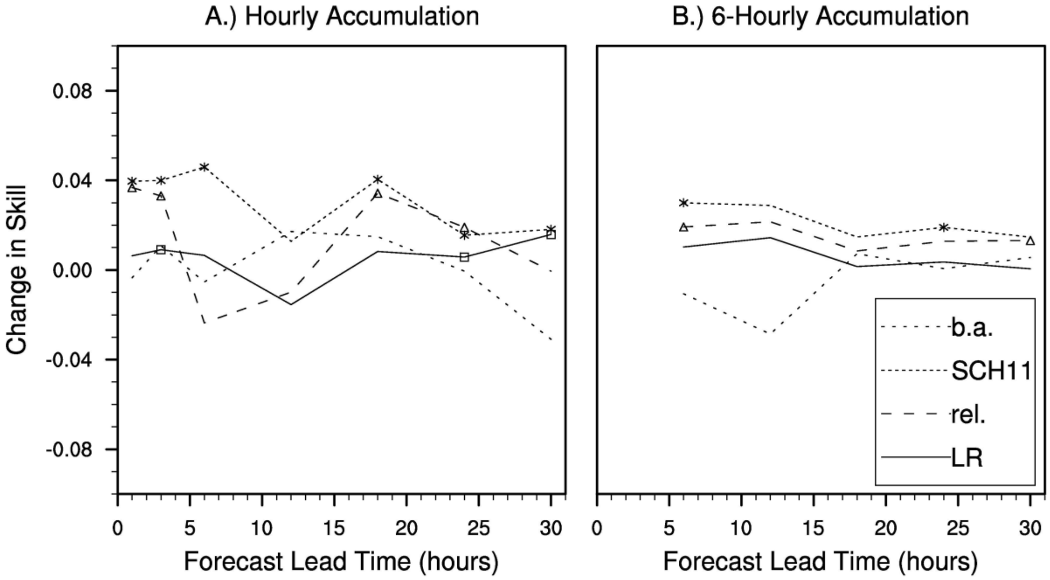

Figure 6. Difference in Brier Skill Score of object based probabilistic forecasts between 25 days and 10 days of training for ensemble calibration methods for (a) hourly accumulation and (b) 6-hourly accumulation. Markers indicate statistically significant difference between 10 and 25 days of training at the 95% confidence level.

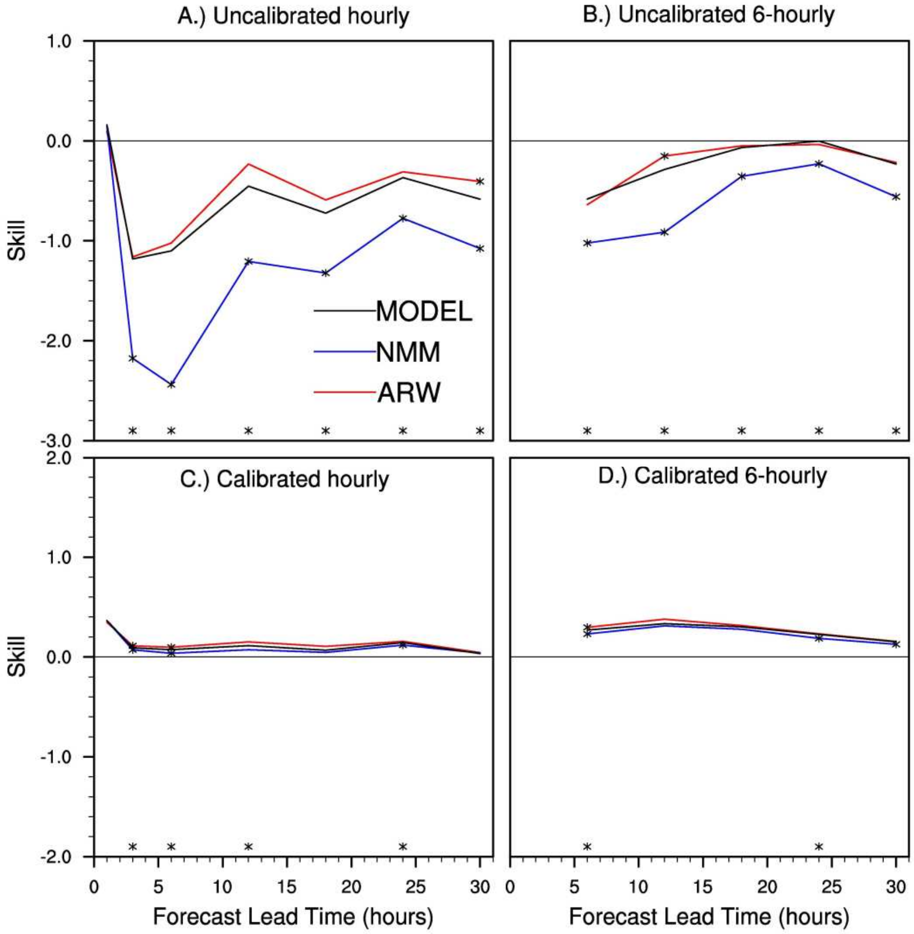

Figure 7. Brier Skill Score of (a,b) un-calibrated and (c,d) logistic regression calibrated object based forecasts from single model (ARW and NMM) and multi-model (MODEL) sub-ensembles for (a,c) hourly and (b,d) 6-hourly accumulation periods. Statistical significance is indicated as in Figure 3.