2011 2nd International Conference on Environmental Science and Technology
IPCBEE vol.6 (2011) © (2011) IACSIT Press, Singapore

Calibrating surface temperature forecasts using BMA method over Iran

Iman Soltanzadeh
Institute of Geophysics
University of Tehran
Tehran, Iran
imasol@gmail.com

Majid Azadi and G. Ali Vakili
Atmospheric Science & Meteorological Research Center
(ASMERC)
Tehran, Iran
Azadi68@hotmail.com & vakili7@yahoo.com

***Abstract*—Using Bayesian Model Averaging (BMA), an attempt was made to obtain calibrated probabilistic numerical forecasts of 2-meter temperature over Iran. The ensemble makes use of three limited area models (WRF, MM5 and HRM), with WRF used five times with different configurations. The resulting ensemble of eight members was run for a period of 6 months (from December 2008 to May 2009) over Iran. The 48-h raw ensemble outputs were calibrated using BMA technique for 120 days using a 40 days training sample of forecasts and corresponding verification data.**

The calibrated probabilistic forecasts were assessed using flatness of rank histogram and attribute diagrams. Results showed that application of BMA improved the calibration of the raw ensemble. Using the weighted ensemble mean forecast as a deterministic forecast it was found that the deterministic-style BMA forecasts performed usually better than the best member's deterministic forecast.

Keywords-Bayesian Model Averaging (BMA); NWP; post-processing

I. INTRODUCTION

Ensemble forecasting is a numerical prediction method that samples the uncertainties in initial conditions and model formulation. Thus, rather than producing a single deterministic forecast, multiple forecasts are produced by making small alterations to either the initial conditions or the configuration of the forecast model, or both. Ensemble forecasts have been operationally implemented on the synoptic scale (Toth and Kalnay 1993; Houtekamer et al. 1996; Molteni et al. 1996) and on the mesoscale (Stensrud et al. 1999; Wandishin et al. 2001; Grimit and Mass 2002; Eckel and Mass 2005). Despite their relatively high skill, they tend to be under-dispersed and thus uncalibrated, especially for weather quantities at the surface.

In the last couple of years various statistical methods such as logistic regression (Wilks 2006), Bayesian Model Averaging (Raftery et al. 2005), non-homogeneous Gaussian regression (Gneiting et al. 2005) and Gaussian ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005), among others, have been developed for calibrating the raw ensemble forecasts.

In this study the Bayesian Model Averaging (BMA) technique, proposed by Raftery et al. (2005), has been used to calibrate the raw outputs of a multimodel multi-analysis ensemble for 2-m temperature at 299 meteorological stations over Iran. In BMA, parameters (weights and variances) for a mixture of distributions (e.g. Gaussians) are estimated over a sliding-window training period of forecasts and observational data. Parameter estimation is accomplished by maximizing the log-likelihood or minimizing the CRPS.

BMA was originally developed for weather quantities in which PDFs could be approximated by normal distributions, such as temperature and sea level pressure. Bao et al. (2010) extended and applied the BMA to 48-h forecasts of directional variable of the wind direction using von Mises densities as the component distributions centered at the individually bias-corrected ensemble surface wind direction and could get consistent improvements in forecasts.

The present study aims at producing calibrated surface temperature forecasts at 299 meteorological stations scattered across Iran using a multimodel multi-analysis for the period of 15 December 2008 to 11 June 2009. Predictive forecast PDFs' skills are evaluated using reliability. Point or deterministic-style BMA forecasts are compared with deterministic forecast of individual members using standard verification scores.

The paper is organized as follows: the BMA procedure is described briefly in Section II, while the implementation details are presented in section III. Verification results are discussed in Section IV and finally, conclusions and proposal for further works are drawn in Section V.

II. CALIBRATION METHOD

Bayesian Model Averaging (BMA) was proposed by Raftery et al. (2005) as a statistical post-processing approach for combining different model forecasts and producing full predictive PDFs from ensembles, subject to calibration and sharpness. In the following a brief description of the method is presented, but the reader is referred to Raftery et al. (2005) for full details. The BMA predictive PDF of weather quantity to be forecast is a weighted average of PDFs defined around each individual bias-corrected ensemble member. The individual PDFs for the ensemble members need not to be Gaussian or even the same. In this study, as in Raftery et al. (2005), the weather quantity to be forecast, y, is temperature whose behavior can be estimated by normal distribution. Hence a Gaussian distribution, $h_k(y|f_k)$, is defined around each individual forecast, f_k, conditional on f_k being the best forecast in the ensemble. The BMA predictive PDF is then a conditional probability for forecast quantity y given K model forecasts $f_1, \dots, f_k$, and is given by:

$$p(y|f_1, \dots, f_k) = \sum_{k=1}^{K} w_k g_k(y|f_k) \qquad (1)$$

Where w_k is the posterior probability of forecast k being the best one, and is based on forecast k's relative performance over a training period. The w_k's are nonnegative probabilities and their sum is equal to 1, that is, $\sum_{k=1}^{K} w_k$. Here K, the number of ensemble members, is equal to 7. $g_k(y|k)$ is a univariate normal PDF with mean $f_k=k + b_k f_k$, (bias-corrected forecast) that is a linear function of forecast f_k, and standard deviation σ^2 assumed to be constant across ensemble members. This situation is denoted by:

$$y|f_k \sim N(a_k+b_k f_k, \sigma^2 \quad (2)$$

Coefficients a_k and b_k in the mean of the individual PDFs vary with time and location and are estimated by a linear regression of observed temperature, y, on model k forecasts, f_k, in the training period, for each time and location separately. This regression can be considered as a preliminary debiasing of the deterministic forecasts in the ensemble. The K weighs or posterior probabilities w_k and variance $\sigma^2$are estimated using maximum likelihood (Fisher et al. 1992). For a fixed set of training data and underlying normal probability model, the method of maximum likelihood selects values of the model parameters that maximize the likelihood function, that is, the value of the parameter vector under which the observed data were most likely to have been observed. For both mathematical simplicity and numerical stability, usually the log-likelihood function is used for maximization rather the likelihood function itself. Usually, the maximum value of the log-likelihood function is evaluated using the expectation-maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997). The BMA deterministic forecast also can be calculated by weighted averaging of the k deterministic forecasts using w_k as weights, that is:

$$\sum_{k=1}^{K} w_k(a_k f_k + b_k) \quad (3)$$

III. The ensemble system and data

Forecasts of the Weather Research and Forecasting (WRF: Skamarock et al. 2001, 2008) model with five different configurations, the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5: Dudhia 1993; Grell et al. 1995) and the High Resolution Model (HRM: Majewski 1991, Majewski and Schrodin 1994) of Deutcher Wetter Dienst (DWD) both with one configuration for 2-m temperature, 48-hour in advanceare used in this study to build a seven-member ensemble. The main differences between different model setups pertain to convective and boundary layer parameterization schemes. WRF and MM5 are used with non-hydrostatic option whereas HRM is hydrostatic. The initial and boundary conditions come from the operational 12Z runs of global forecasting system (GFS) of NCEP (National Center for Environmental Prediction) for MM5 and WRF and of DWD's global model (GME) for HRM models respectively. The integration period goes from 15 December 2008 to 11 June 2009. MM5 and WRF were run with two nested domains, with the larger domain covering south-west middle east from 10° to 51°north and from 20° to 80° east and the smaller domain covers Iran from 23° to 41° north and from 42° to 65° east. The spatial resolutions are 45 and 15-Km for the coarser and finer domains respectively. The inner domain in HRM, covers an area from 25° to 40° north and from 43° to 63° east with spatial resolution of 14-km. Forecasts out to 48 h ahead for the inner domains were used to form the ensemble of forecasts.

The data used in this study consists of 12Z observations of 2-m temperature at 299 irregularly spaced synoptic meteorological stations scattered all over the country from 15 December 2008 to the 11 June 2009 and corresponding 48-h forecasts from the above mentioned eight members of the ensemble bilinearly interpolated to the observation sites. Using N days as training period, the BMA predictive PDF, Eq. (1), and BMA deterministic forecast, Eq. (3), for the 2-m temperature were calculated and evaluated for each station site and the remaining days.

IV. Training period

The sample of past days, N, used as training period, in estimating the unknown parameters (a_k, b_k, w_k and σ^2) in Eq. (2) is a sliding training window, such that new coefficients are estimated for each day using the most recent N days as training period. In principal, length of the training period must be long enough that it does not lead to over fitting. However, longer training periods increase statistical variability and shorter training periods makes the forecast system to respond more quickly to changes in the model's error patterns due to changes in the weather regime. It is clear that using short training periods have the advantage of data availability and ease of computations. A balance should be made in determining the length of the training period. In this study, as in Raftery et al. (2005), several experiments were conducted with training period changing from 10 to 65-days. Fig. 1 shows the mean absolute error (MAE), root mean squared error (RMSE) and the continuous rank probability score (CRPS) of the BMA deterministic forecasts versus training days. As is seen from the Figure, a minimum error occurs around 20 and 40 days of training period. But increasing the training days beyond 40 does not change the results significantly and the error remains almost constant or even increases. So, a 40 days window was selected as the training period.

V. Results and conclusions

A. Raw ensemble

Fig. 2 presents an example of the conditional quintile plot (CQP) for 48-h temperature forecast of one member (member 6: MM5) in the ensemble for the period mentioned in section 2. It is clear that deterministic

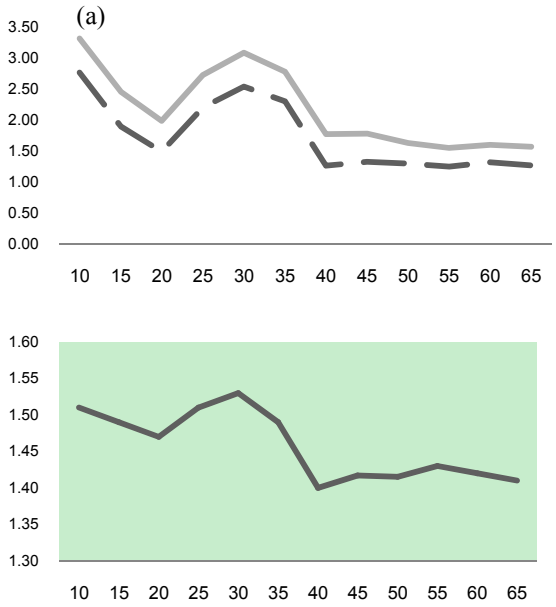

Figure 1. Comparison of training period length for surface temperature: diagram of (a) MAE (dashed line), RMSE (solid line) and (b) CRPS for from 10 to 65 days.

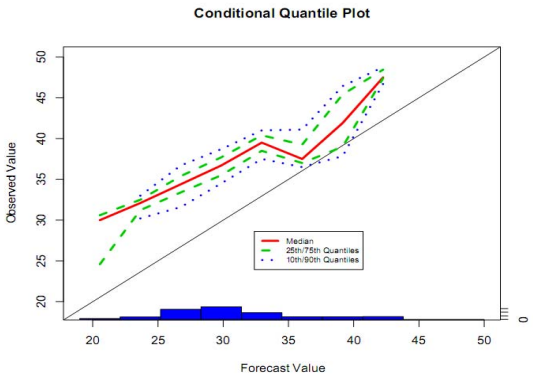

Figure 2. Conditional Quantile Plot (CQP) diagram for MM5 2-m temperature forecasts.

forecasts, corresponding to the 7th ensemble member show a cold bias and thus an debiasing of the forecasts are needed. The CQPs related to other ensemble members (not shown here) show similar results, i.e. a cold bias.

Rank histograms are very useful tools for evaluating an ensemble forecast system performance (e.g., Hamill and Colucci 1997, 1998; Hou et al. 2001; Stensrud and Yussouf 2003). A rank histogram is a histogram of the observation ranks when pooled in the sorted forecasts of the ensemble members. Hamill (2001) shows how to use rank histograms for evaluating ensemble forecasts appropriately. Rank histogram of the raw ensemble is presented in Fig. 3(a). Under-dispersion of the ensemble is reflected in the non-uniform shape of its rank histogram. As it is seen from the figure the rank histogram for the raw ensemble is a sloped one showing a consistent bias in the ensemble forecast. The ensemble members have under-

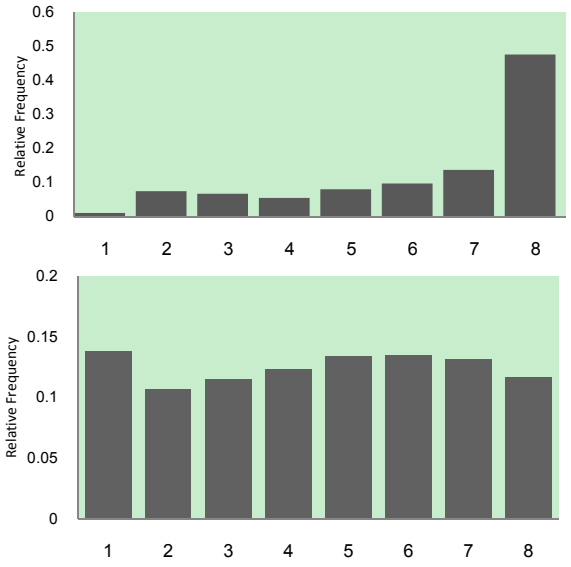

Figure 3. (a) Rank histogram of the raw ensemble for 48-h surface temperature forecasts and (b) Probability Integral Transform (PIT) histogram of calibrated BMA outputs.

forecast or cold bias, such that around 50% of the times the observed temperature was greater than all the ensemble member values. This result is consistent with the results presented in Fig. 2. The under-dispersion for the raw ensemble has been reported in many other studies. The goal of post-processing is to correct for such known forecast errors, i.e. to construct a calibrated ensemble with statistical properties similar to observations.

B. Deterministic BMA forecast

For comparing the deterministic BMA forecast (Eq. 3) with the deterministic forecast corresponding to the best member of the ensemble, the mean absolute error (MAE) and percentage of successful forecasts (forecasts with less that 2 °C difference from the verifying observation) are considered. Results show that MAE of the deterministic forecasts of the seven members of the ensemble system are between 2.2 to 6.7 °C, while that of the BMA deterministic forecast is lower and around 2 °C (Figures are not presented here).

Percentage of the successful forecasts for seven members of the ensemble, shown in Fig. 4, ranged from 3.6% to 56.5% for the first and fifth members respectively. This score for the BMA deterministic forecast is close to 62% which is again better that those of all ensemble members. It is thus seen that the BMA deterministic forecast outperforms all other seven deterministic forecasts corresponding to the ensemble members.

C. Calibrated ensemble

One main aim of the ensemble forecasting is to account for various uncertainties in the ensemble system inherent in the probabilistic forecasts. Fig. 5 (a and b) show two

TABLE I. WEIGHTS OF TWO SYNOPTIC STATIONS OF PIRANSHAHR AND NOSHAHR IN INITIALIZED AT 1200 UTC ON 1 FEB. AND 1 MAY 2009, RESPECTIVELY.

Station	Date	WRF1	WRF2	WRF3	WRF4	WRF5	MM5	HRM
Noshahr	20090401	0.0112	7.1×10^{-6}	3.02×10^{-3}	0.216	0.246	0.175	0.348
Piranshahr	20090501	6.8×10^{-3}	2.98×10^{-6}	0.0115	0.212	0.399	0.170	0.20

Figure 4. Histogram of percentage of successful forecast of raw data and BMA outputs for ±2 oC error.

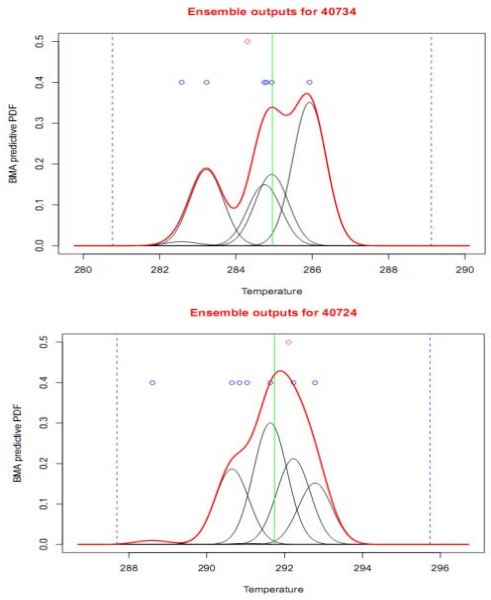

Figure 5. BMA predictive PDF (thick red curve) and its seven components (thin black curves) for the 48-h surface temperature for two synoptic stations of (a) Piranshahr and (b) Noshahr in itialized at 1200 UTC on first February and first May 2009, respectively. Also shown are the ensemble member forecasts and range (solid horizontal line and black bullets), observation (red bullets), the BMA 90% prediction interval (dotted lines) and the median of the PDF (solid green vertical line).

examples of the final predictive BMA for 48-h forecast of 2-m temperature valid at first February and May 2009, along withtheir seven normal PDF components issued for two synoptic stations of Piranshahr and Noshahr located in the west and north of the country. As is seen, the calibrated BMA PDF that is a weighted sum of its seven components, is a non-normal distribution. Table I shows the calculated weights given to each member of the ensemble. It is seen that for Noshahr, the weights in descending order are given to HRM, WRF5, WRF2 and MM5 members respectively, while for Piranshahr the weights in descenig order are given to WRF5, WRF4, HRM and MM5. As mentioned above, a higher weight given to a member means that the member is more useful. But, as mentioned by Gneiting et al. (2005), low weight for a member does not mean necessarily a lower performance of that member. If there are colinearities between two (or more) ensemble members, the weights given to one of the members might be low, though it might be still more skillful.

One important aim of applying the BMA technique is to obtain a well calibrated ensemble with reduced under-dispersion. The fact that the predictive PDFs are calibrated is reflected in the uniformity of the post-processed ensemble rank histogram for 48-h forecasts, presented in Fig. 3(b). As the figure shows, the BMA has been very successful in calibrating the raw ensemble forecasts.

A more detailed comparison of the BMA calibrated forecasts with the raw ensemble can be obtained from the attribute diagram for probability forecast of particular quintiles. Attribute diagrams for two temperature categories "equal to 0 °C" and "20 to 25 °C" is presented in Fig. 6. As is seen, the BMA technique provides reliable and skillful probabilistic forecasts of freezing temperature when calculated for all 299 station locations over Iran (Fig. 9a) and shows significant improvement over the raw ensemble. Hypothesis testing showed that the improvement was significant at the 95% confidence level. Similar results could be obtained for other quintiles e.g. 20 to 25 °C interval (Fig. 6b).

VI. CONCLUSION

This paper describes the results of 48-h probabilistic surface temperature forecasts over Iran for the period of 15 December 2008 to 11 June 2009 using Bayesian Model Averaging for calibration the ensemble outputs. The ensemble system consists of the WRF with five different configurations, MM5 and HRM both with one configuration. The initial and boundary conditions come from the operational 12Z runs of GFS for MM5 and WRF and of GME for HRM models respectively.

The probabilistic forecasts were accomplished for 299 synoptic station locations scattered across Iran. The experiment was set up in such a way that it could be run in real time operations; the BMA was trained on recent realizations of the forecast errors, and then applied to the subsequent forecasts in the 3 months test period. Based on the results of several experiments with different training sample sizes from 10 to 65-day, a 40-day window was selected as the training period.

Overall results showed that the BMA technique is successful at removing most, but not all of the under-dispersion exhibited by the raw ensemble and thus attaining higher reliability in the probabilistic forecasts that could be used in an operational framework.Using the weighted

ensemble mean forecast as a deterministic forecast it was found that the deterministic-style BMA

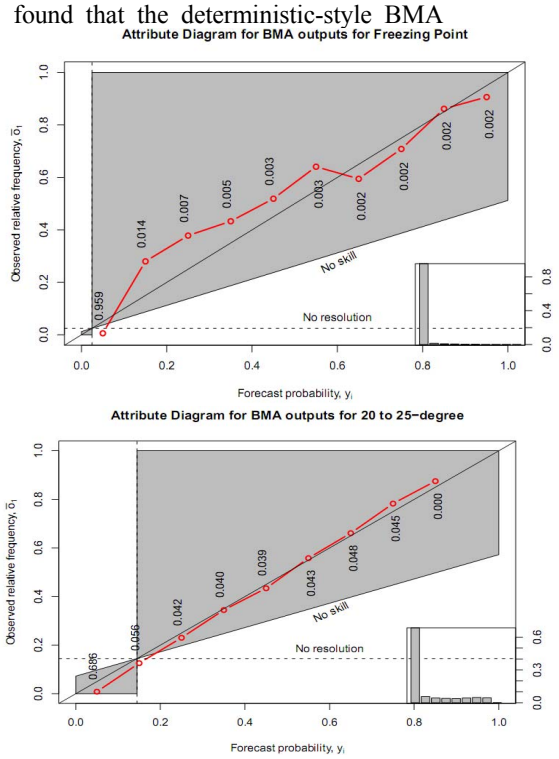

Figure 6. Attribute diagrams for two temperature categories "equal to 0 °C" and "20 to 25 °C" with 5 °C.

forecasts performed almost always better than the best member's deterministic forecast in the ensemble.

REFERENCES

[1] Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P. and Raftery, A. E.: Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction. Mon. Wea. Rev., 138, 1811–1821, 2010.

[2] Dudhia, J.: A nonhydrostatic version of the Penn State/NCAR mesoscale model: Validation tests and the simulation of an Atlantic cyclone and cold front, Mon. Wea. Rev., 121, 1493–1513, 1993.

[3] Dempster, A., Laird, N. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.

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

[5] Fisher, R. A., Thornton, H. G. and Mackenzie, W. A.: The accuracy of the plating method of estimating the density of bacterial populations, Annals of Applied Biology, 9, 325-359, 1922. [CP22 in Bennett 1971 , vol. 1.]

[6] Grell, G. A., Dudhia, J., and Stauffer, D. R.: A description of the fifth-generation Penn State/NCAR mesoscalemodel (MM5), NCAR Tech. NoteTN-398+STR, 122 pp, 1995.

[7] Grimit E., and Mass, C.: Initial results of a mesoscale short-rangeensemble forecasting system over the Pacific Northwest, Weather and Forecasting, 17, 192–205, 2002.

[8] Gneiting, T., Raftery, A. E., Westveld , A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation, Mon. Wea. Rev., 133, 1098-1118, 2005.

[9] Hamill, T. M.. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Wea. Rev. 129:550–560, 2001.

[10] Hamill, T. M., and Colucci, S. J.: Verification of Eta–RSM shortrange ensemble forecasts. Mon. Wea. Rev., 125, 1312–1327, 1997.

[11] Hamill, T. M., and Colucci, S. J.: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. Mon. Wea. Rev., 126, 711–724, 1998.

[12] Hou, D., E. Kalnay, and Droegemeier, K. K.: Objective verification of the SAMEX'98 ensemble forecasts. Mon. Wea. Rev., 129:73–91, 2001.

[13] Houtekamer, P. L., and Derome, J.: The RPN ensemble predictionsystem. Proceedings, ECMWF Seminar on Predictability. Vol. II. ECMWF, 121–146, 1996. [Available from ECMWF, Shinfield Park,Reading, Berkshire RG2 9AX, United Kingdom.]

[14] Majewski, D.: The Europa-Model of the DWD, ECMWF Seminar on numerical methods in Atmospheric Science, 2, 147-191, 1991.

[15] Majewski, D. and Schrodin, R.: Short description of Europa-Model (EM) and Deutschland Model (DM) of the DVD. Q. Bull, 1994.

[16] McLachlan, G. and Krishnan, T.: The EM algorithm and extensions. Wiley series in probability and statistics. John Wiley & Sons, 1997.

[17] Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T.: The ECMWF ensemble prediction system: Methodol- ogy and validation. Quart. J. Roy. Meteor Soc., 122, 73–119, 1996.

[18] Raftery, A. E., Balabdaoui, F., Gneiting, T., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Technical Report no. 440, Department of Statistics, University of Washington, 2003.

[19] Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weath. Rev., Vol. 133, pp. 1155-1174, 2005.

[20] Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles. Tellus, 55A, 16–30, 2003.

[21] Skamarock, W. C., Klemp, J. B. and Dudhia, J.: Prototypes for the WRF (Weather Research and Forecast) model. Preprints, Ninth Conf. on Mesoscale Processes, Fort Lauderdale, FL, Amer. Meteor.Soc., CD-ROM, J1.5, 2001.

[22] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D., Barker, D., Wang, W. and Powers, J.G.: A description of the Advanced Research WRF Version 3, NCAR Tech. Note NCAR/TN-475+STR, 2008.

[23] Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E.: Using ensembles for short-range forecasting, Mon. Wea. Rev., 127, 433-446, 1999.

[24] Stensrud, D. J. and Yussouf, N.: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. Mon. Wea. Rev., 131:2510–2524, 2003.

[25] Toth, Z. and Kalnay, E.: Ensemble forecasting at NMC: Thegeneration of perturbations. Bull. Amer. Meteor. Soc., 74, 2317–2330, 1993.

[26] Wandishin, M. S., Mullen, S. L., Stensrud, D. J. and Brooks, H.E.: Evaluation of a short-range multimodel ensemble system, Mon.Wea. Rev., 129, 729-747, 2001.

[27] Wang, X. and Bishop, C. H.: Improvement of ensemble reliability with a new dressing kernel, Quart. J. Roy. Meteor. Soc., 131, 965-986, 2005.

[28] Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd Edition, Academic Press, 627 pp, 2006.

[29] Mason, I. B.: A model for assessment of weather forecasts, Austral. Met. Mag., 30, 291-303, 1982.

[30] Wilson, L. J., Beauregard, S., Raftery, A. E. and Verret, R.: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian modelaveraging. Mon. Wea. Rev., 135, 1364-1385, 2007.