Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging

LAURENCE J. WILSON

Meteorological Research Division, Environment Canada, Dorval, Quebec, Canada

STEPHANE BEAUREGARD

Canadian Meteorological Centre, Meteorological Service of Canada, Dorval, Quebec, Canada

ADRIAN E. RAFTERY

Department of Statistics, University of Washington, Seattle, Washington

RICHARD VERRET

Canadian Meteorological Centre, Meteorological Service of Canada, Dorval, Quebec, Canada

(Manuscript received 27 May 2005, in final form 22 June 2006)

ABSTRACT

Bayesian model averaging (BMA) has recently been proposed as a way of correcting underdispersion in ensemble forecasts. BMA is a standard statistical procedure for combining predictive distributions from different sources. The output of BMA is a probability density function (pdf), which is a weighted average of pdfs centered on the bias-corrected forecasts. The BMA weights reflect the relative contributions of the component models to the predictive skill over a training sample. The variance of the BMA pdf is made up of two components, the between-model variance, and the within-model error variance, both estimated from the training sample. This paper describes the results of experiments with BMA to calibrate surface temperature forecasts from the 16-member Canadian ensemble system. Using one year of ensemble forecasts, BMA was applied for different training periods ranging from 25 to 80 days. The method was trained on the most recent forecast period, then applied to the next day's forecasts as an independent sample. This process was repeated through the year, and forecast quality was evaluated using rank histograms, the continuous rank probability score, and the continuous rank probability skill score. An examination of the BMA weights provided a useful comparative evaluation of the component models, both for the ensemble itself and for the ensemble augmented with the unperturbed control forecast and the higher-resolution deterministic forecast. Training periods around 40 days provided a good calibration of the ensemble dispersion. Both full regression and simple bias-correction methods worked well to correct the bias, except that the full regression failed to completely remove seasonal trend biases in spring and fall. Simple correction of the bias was sufficient to produce positive forecast skill out to 10 days with respect to climatology, which was improved by the BMA. The addition of the control forecast and the full-resolution model forecast to the ensemble produced modest improvement in the forecasts for ranges out to about 7 days. Finally, BMA produced significantly narrower 90% prediction intervals compared to a simple Gaussian bias correction, while achieving similar overall accuracy.

1. Introduction

For several decades, statistical methods have been used to postprocess the output of numerical prediction models into forecasts of sensible weather elements such as temperature and precipitation. The main goal of statistical postprocessing has been to interpret the model output variables to produce forecasts that are more accurate than could be produced directly from the model. The statistical postprocessing improves the accuracy by removing biases, and/or by increasing the correlation between forecast and observation. A prime example is model output statistics (MOS; Glahn and Lowry 1972),

Corresponding author address: Laurence J. Wilson, Environment Canada, 2121 Transcanada Highway, 5th Floor, Dorval, QC H9P 1J3, Canada.
E-mail: lawrence.wilson@ec.gc.ca

DOI: 10.1175/MWR3347.1

which for over 30 yr has been used in many centers to improve the output of deterministic operational forecasts. MOS typically uses linear statistical predictive techniques such as regression to relate a historical set of model forecast variables ("predictors") to surface observations. If the predictor set includes the model estimate of the predictand variable, then MOS explicitly accounts for the bias in the model forecast, or calibrates it to the training sample.

More recently, as changes in operational models became more frequent, it became desirable to develop adaptive statistical methods to postprocess model output. Adaptive procedures use shorter (smaller) samples of recent realizations of the forecast system to estimate and update the parameters of the statistical model. While small sample sizes are sometimes a problem, adaptive procedures respond quickly to changing statistical properties of the training sample, and bring the benefits of improvements to numerical weather prediction (NWP) models into the post processed products more quickly. Two examples are an updateable form of MOS (Wilson and Vallée 2002) and the Kalman filter (Simonsen 1991). The latter technique has been tested not only for the correction of model forecasts of surface variables, but also as a way of combining forecasts from different sources (Vallée et al. 1996).

All of the above techniques have been applied to the output of "deterministic" models, which produce one forecast of the surface variables of interest at each location. Since the early 1990s, ensemble systems have been developed and increasingly used for weather prediction at many centers. Ensemble systems take many forms, but they all provide multiple predictions of each forecast variable at each valid time, and the forecasts are generated using one or more deterministic models, in one or more versions, starting from different analyses. Examples include so-called poor man's systems involving a combination of existing models and analyses from different centers (Ziehmann 2000), systems based on different models and perturbed initial conditions (Pellerin et al. 2003), and single-model systems using perturbed initial conditions (Molteni et al. 1996; Toth and Kalnay 1997). The set of alternative forecast values obtained is usually interpreted as a sample from a probability density function (pdf), which is intended to represent the uncertainty in the forecast for each valid time and location.

Evaluations of ensemble pdfs that have been carried out often reveal that they are not calibrated. They may be biased in the sense that the pdf is not centered on the observation, and/or they are often found to be underdispersive, in the sense that the observations exhibit greater variance than the ensemble, on average (Toth et al. 2001; Pellerin et al. 2003; Buizza 1997; Hamill and Colucci 1997). Furthermore, gridded forecasts from ensembles are valid over large areas rather than at points. Statistical calibration of ensemble forecasts with respect to point observations can thus add some downscaling information to the forecasts.

Calibration of ensembles is also important if they are to be combined with other forecasts or other ensembles. This is because biases of individual ensembles will lead to artificially large ensemble spread when the ensembles are combined, and systematic errors of each ensemble will decrease the accuracy of the combined ensemble. In the recently planned North American Ensemble Forecast System (NAEFS; Toth et al. 2005), the importance of ensemble calibration has been recognized as part of the joint ensemble project. Both the Meteorological Service of Canada and the U.S. National Centers for Environmental Prediction are pursuing and comparatively evaluating ensemble calibration methods.

One rather promising method for calibrating ensemble forecasts is Bayesian model averaging (BMA). Widely applied to the combination of statistical models in the social and health sciences, BMA has recently been applied to the combination of NWP model forecasts in an ensemble context by Raftery et al. (2005). BMA is adaptive in the sense that recent realizations of the forecast system are used as a training sample to carry out the calibration. BMA is also a method of combining forecasts from different sources into a consensus pdf, an ensemble analog to consensus forecasting methods applied to deterministic forecasts from different sources (Vallée et al. 1996; Vislocky and Fritsch 1995; Krishnamurti et al. 1999). BMA naturally applies to ensemble systems made up of sets of discrete models such as the Canadian ensemble system.

Raftery et al. (2005) applied BMA to the University of Washington short-range ensemble, which is a five-member multianalysis, single-model ensemble (Grimit and Mass 2002). BMA was applied to surface temperature and mean sea level pressure forecasts, and the coefficients were fitted using forecasts and observations over a spatial region. They found that BMA was able to correct the underdispersion in both the temperature and pressure forecasts.

In our study we apply BMA to the Canadian ensemble system (Pellerin et al. 2003). This is a 16-member ensemble using perturbed initial conditions and eight different versions of each of two different NWP models. As in Raftery et al. (2005), we apply the BMA to surface temperature forecasts, but we test the method on single-station data, rather than over a spatial area. In this way, we attempt to calibrate the en-

semble forecasts with respect to local effects that may not be resolved by the forecast models. The BMA procedure is described in section 2; section 3 gives a brief summary of the Canadian ensemble system; the available data and experimental setup are discussed in section 4; section 5 describes the results; and we conclude with a discussion in section 6.

2. Bayesian model averaging

BMA is a way of combining statistical models and at the same time calibrating them using a training dataset. BMA is not a bias-correction procedure; models that are to be combined using BMA should be individually corrected for any bias errors before the BMA is applied. The BMA pdf is a weighted sum of the bias-corrected pdfs of the component models, where all the pdfs are estimated from the forecast error distribution over the training dataset.

Following the notation in Raftery et al. (2005), the combined forecast pdf of a variable y is

$$p(y|y^T) = \sum_{k=1}^{K} p(y|M_k, y^T)p(M_k|y^T), \quad (1)$$

where $p(y|M_k, y^T)$ is the forecast pdf based on model M_k alone, estimated from the training data; K is the number of models being combined, 16 or 18 in the present study; and $p(M_k|y^T)$ is the posterior probability of model M_k being correct given the training data. This term is computed with the aid of Bayes's theorem:

$$p(M_k|y^T) = \frac{p(y^T|M_k)p(M_k)}{\sum_{l=1}^{K} p(y^T|M_l)p(M_l)}. \quad (2)$$

The BMA procedure itself is applied to bias-corrected forecasts only. In this study, we consider two types of bias correction. The first is a simple adjustment of forecasts from each model for the average error over the training set:

$$f_k = a_k + y_k, \quad (3)$$

where f_k is the bias-corrected forecast for model k, y_k is the forecast value of the variable from model k, and a_k is the mean error for model k over the training dataset. This is referred to in the results section as "b1." The other bias-correction procedure considered is a simple linear regression fit of the training data using the corresponding model-predicted variable as the single predictor:

$$f_k = a_k + b_k y_k. \quad (4)$$

This is referred to in the results section as "FR."

Considering now the application of BMA to bias-corrected forecasts from the K models, Eq. (1) can be rewritten as

$$p(y|f_1 \ldots\ldots f_K, y^T) = \sum_{k=1}^{K} \omega_k p_k(y|f_k, y^T), \quad (5)$$

where $\omega_k = p(M_k|y^T)$ is the BMA weight for model k, computed from the training dataset, and reflects the relative performance of model k on the training period. The weights ω_k add up to 1. The conditional probabilities $p_k[y|(f_k, y^T)]$ may be interpreted as the conditional pdf of y given f_k, given that model k has been chosen (or is the "best" model or member), based on the training data y^T. These conditional pdfs are assumed to be normally distributed,

$$y|(f_k, y^T) \sim N(a_k + b_k y_k, \sigma^2), \quad (6)$$

where the coefficients a_k and b_k are estimated from the bias-correction procedures described above. This means that the BMA predictive distribution becomes a weighted sum of normal distributions, with equal variances, each one centered on the bias-corrected forecast from an ensemble member. A deterministic forecast can also be obtained from the BMA distribution, using the conditional expectation of y given the forecasts:

$$E[y|(f_1, \ldots, f_K, y^T)] = \sum_{k=1}^{K} \omega_k(a_k + b_k f_k). \quad (7)$$

This forecast would be expected to be more skilful than either the ensemble mean or any one member, since it has been determined from an ensemble distribution that has had its first and second moments debiased, using recent verification data for all the ensemble members. It is essentially an "intelligent" consensus forecast, weighted by the recent performance results for the component models.

The BMA weights and the variance σ^2 are estimated using maximum likelihood (Fisher 1922). The likelihood function is the probability of the training data given the parameters to be estimated, viewed as a function of the parameters, that is, the term $p(y^T|M_k)$ in Eq. (2). The K weights and variance are chosen so as to maximize this function (i.e., the parameter values for which the observed data were most likely to have been observed). The algorithm used to calculate the BMA weights and variance is called the expectation maximization (EM) algorithm (Dempster et al. 1977). The method is iterative, and normally converges to a local maximum of the likelihood. For a summary of the method as applied to BMA, the reader is referred to Raftery et al. (2005), and more complete details of the EM algorithm are given in McLachlan and Krishnan (1997). The value of σ^2 is related to the total pooled

error variance over all the models in the training dataset. (A free software package to estimate the BMA parameters called ensemble BMA, written in the freely available statistical language R, is available online at http://cran.us.r-project.org/src/contrib/Descriptions/ensembleBMA.html.)

It is possible to relax the assumption that the conditional distributions for the component models all have constant variance. We carried out tests where the variance was allowed to vary, and was estimated separately for each component model. This meant that, in addition to the 16 or 18 of each of a_k, b_k, and ω_k, 16 or 18 different values of σ_k^2 needed to be estimated, instead of a single σ^2 value. This significantly increases the number of independent BMA parameters that must be supported by the training sample.

3. The Canadian Ensemble System

Operational ensemble forecasts have been issued from the Canadian Meteorological Centre (CMC) since 1996. The ensemble strategy used at CMC has been described in Houtekamer et al. (1996) and Lefaivre et al. (1997). The version of the system that generated the data used in this study is described in Pellerin et al. (2003). The original ensemble consisted of eight members, which were eight variants of a global spectral model (SEF) that was used for operational deterministic forecasts in the late 1980s and early 1990s (Ritchie 1991). In 1999, eight additional members were added, consisting of eight variants of the Global Environmental Multiscale (GEM) model, a high-resolution version of which provides operational deterministic forecasts. The use of 16 different model versions is intended to account for uncertainties in the forecast due to limitations in model formulations. The 16 model versions are initialized with perturbed analyses; the perturbations are random and are produced through the data assimilation cycle in the model. In this way, errors due to the uncertainty in the initial conditions are accounted for. The SEF models have a horizontal resolution of T149 (equivalent to a grid length of about 140 km) and the GEM models have a resolution of 1.2°, about 130 km. The vertical resolution is 23 or 41 levels for the SEF members and 28 levels for the GEM members. The models differ from each other in several other ways, mostly related to the physical parameterization (details are contained in Pellerin et al. 2003).

One of the advantages of the BMA method is that it does not really matter what models are included. As long as sufficiently large training samples are available that are simultaneous in space and time for all models, the BMA can be applied to calibrate the ensemble of forecasts. The BMA weighting procedure will also ensure that the models containing the most accurate predictive information and unique predictive information in light of the full ensemble of models will be assigned high weights in the combined pdf. To test the behavior of the BMA when models from different sources were included in the ensemble, we conducted some experiments with an 18-member ensemble consisting of the original 16 members plus the control forecast and the operational full-resolution global model. The control forecast model is a version of the SEF model where some of the physical parameters that are varied for the eight SEF members are replaced by their averages. The control run uses an unperturbed analysis.

The full-resolution model is the version of the GEM global model that was operational at the time of the BMA experiments. This model has a uniform horizontal grid at 100-km resolution (0.9°) and 28 levels in the vertical, which means that the resolution is only slightly higher than the versions of GEM used in the ensemble. The GEM global is described fully by Coté et al. (1998).

4. Data and experiment

Data used in the experiment came from 1 yr (28 September 2003–27 September 2004) of surface temperature forecasts from the 16-member ensemble, supplemented with forecasts from the unperturbed control model and the full-resolution global model. All forecasts were interpolated to 21 Canadian stations for which a set of reliable observations was available. All observations are of 2-m temperature, taken at manned observing sites. The stations used in the study are distributed across Canada, and are representative of a variety of climatological regimes (Fig. 1). Forecasts were available at 24-hourly intervals out to 10 days, initialized at 0000 UTC each day of the 1-yr period. The full sample comprises 366 separate BMA analyses for each of 21 stations, which are used to produce a maximum of 7686 forecasts, based on the 7686 updated BMA analyses. In practice, the total sample on which the results are based is a little smaller than that due to missing forecasts or verifying observations.

The experiments were conducted in a recursive mode: the BMA was retrained each day through the 1-yr period, using a training sample period of the N previous days, where $N = 25$, 30, 35, 40, 50, 60, or 80. The training was carried out separately for each station and each 24-h forecast projection. Then the BMA coefficients were applied to the next day's forecast, as an independent case. In this way, one year of independent test data was accumulated to verify the performance of the technique. This kind of recursive tuning is similar in

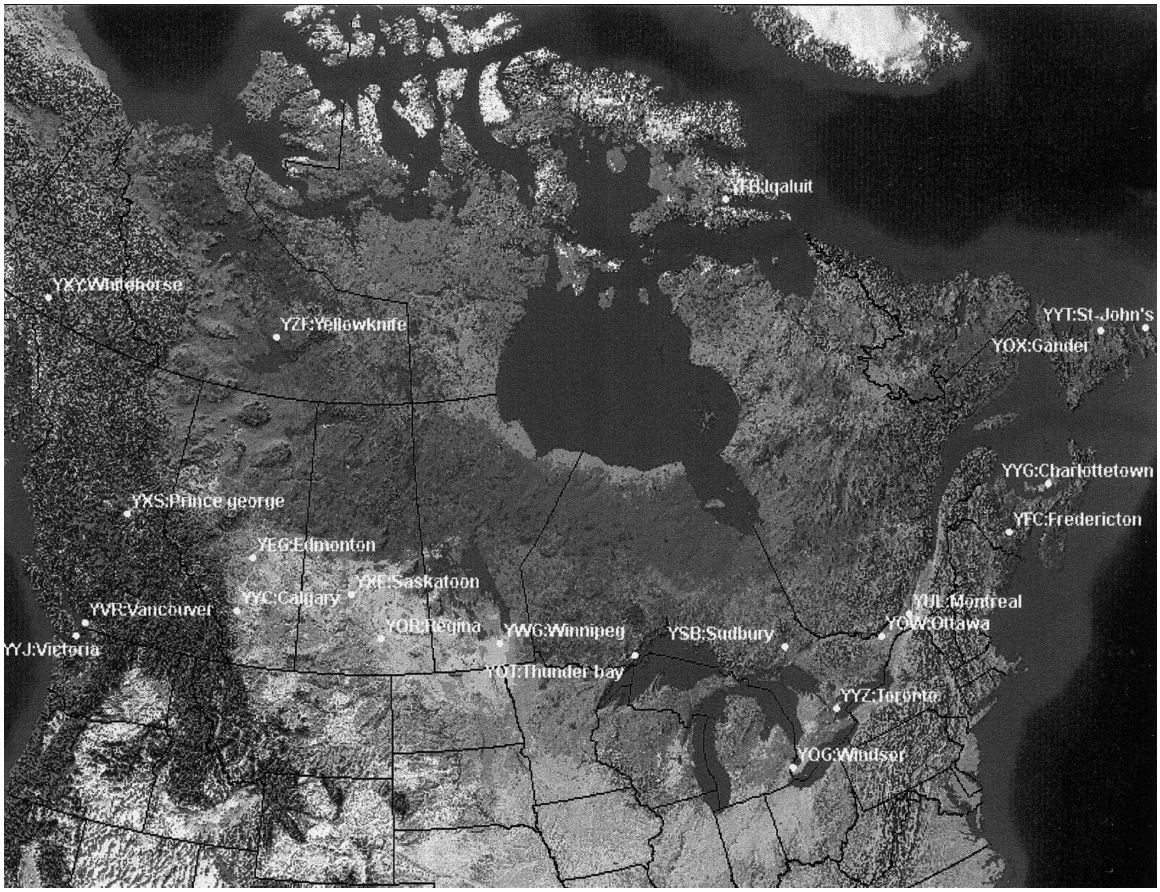

FIG. 1. Map showing the 21 stations used in the analysis.

some ways to the Canadian application of updateable model output statistics (UMOS; Wilson and Vallée 2003), except that BMA is retuned on a daily rather than weekly basis, and the training sample is of fixed size in each application. As with UMOS, there is also no direct dependence of any particular run on the previous runs; BMA coefficients were recalculated separately each time. With the addition of one day of new forecast data and the deletion of the oldest case from the training sample, there is a large overlap in the training dataset from one day to the next, especially for the longer training periods. One would expect, therefore, that the coefficients should not change rapidly from day to day.

Four different comparative experiments were carried out to try to optimize the parameters of BMA:

1) The length of the training period, N.
2) Constant variance over the ensemble members based on the total pooled error variance versus variable variance, estimated from the error distribution for each member. These experiments are referred to as "BMA" and "BMAvar," respectively.
3) Using a full regression to remove the bias versus additive bias removal only. The full regression involved application of Eq. (4) to each member of the ensemble over the training period, resulting in a set of ensemble forecasts for which the average error of all members is 0. The simple bias removal was accomplished by subtracting the mean error over the training sample from all the ensemble member forecasts [Eq. (3)]. These experiments are referred to as FR and bl, respectively.
4) The effect of adding the control forecast and the full-resolution forecast to the BMA (16 versus 18 members).

In addition to the original unprocessed set of ensemble forecasts, the various permutations lead to a total of 12 sets of processed ensemble forecasts, 6 for each of the 16- and 18-member ensembles: FR, bl, BMA after FR, BMA after b1, BMAvar after FR, and BMAvar after b1. All runs were tested for training periods ranging from 30 to 80 days. The shortest training period, 25 days, was dropped from consideration as soon as it became clear that it was too short.

The various experiments are assessed using three verification measures, the rank histogram (Anderson 1996; Hamill 2001), the continuous rank probability score (CRPS; Hersbach 2000), and the continuous rank probability skill score (CRPSS), a skill score with respect to climatology based on the CRPS. The first is used to check whether the ensemble spread is representative of the spread in the observations, and thus can give a good indication whether the second moment of the ensemble distribution has been calibrated by BMA. The CRPS measures the distance between the forecast continuous distribution function (cdf) and the observation expressed as a cdf, and measures the accuracy of the ensemble distribution. The CRPS is equal to the mean absolute error for deterministic forecasts. The skill score measures the percentage improvement of the ensemble distribution compared to a forecast of the long-term climatological distribution for the station and day of the year. These three measures are described more fully in the appendix.

5. Results

a. BMA weights

Figure 2 shows the time series of weights for one station (Montreal) for 24-, 120-, and 216-h projections, using a training period of 40 days. All 16 ensemble models are shown. This figure illustrates how the coefficient selection changes from day to day. There are several aspects to note here. First, the day-to-day changes in the coefficients can be quite large, especially for the shortest-range forecasts. This is perhaps surprising, given that the training sample changes by only 1 case in 40 from one day to the next. Second, different models are favored at different projections at the same time. This was true even for adjacent 24-h projections. Third, there is a tendency for the individual models to go in and out of "favor" for periods of as much as a month or more. Fourth, coefficients sometimes but rarely are close to 1, which means essentially that only one model was considered useful by BMA in that case. And finally, especially at the early projections, some models are rarely selected at all, for example SEF1 and 8 at 24 h and SEF7 at 120 h. We found that the length of the training period had relatively little effect on the coefficients. One can also note that sometimes model weights change pairwise. For example, at 216 h, SEF6 takes over from SEF5 for a couple of weeks in the winter. And, at 24 h, GEM 13 hands over to GEM 12 over a period of only a few days in the summer.

Although some models were practically never chosen through the whole year for some stations and projections, it turns out that all models were approximately equally important when the coefficients are averaged over all stations for the full year. The top row of Fig. 3 shows this for a training period of 40 days, for BMA carried out after FR bias removal, for 24-, 120-, and 216-h projections. At the short and perhaps medium forecast range, there is a slight tendency for the coefficients to favor the GEM model components of the ensemble on average, models 9–16. At the longest range, 216 h, average coefficient values are distributed nearly equally across all ensemble members.

Although there is not much variation in the yearly average coefficient values among the ensemble members, stratification by season reveals greater differences. Figure 3 illustrates this also, showing winter (December–February) and summer (June–August) average coefficients in the second and third rows. The eight GEM model members are strongly favored in summer at the shortest ranges, but not so much in winter. At the medium range, there is not a strong preference shown for either model, but individual members are preferred in different seasons. For example, SEF member 8 carries the highest average weight in winter, but GEM member 16 dominates in summer. At the longer ranges, there is not a clear seasonal signal. This is not surprising, since none of the individual forecast models are believed to have significant predictive skill at the longer-range projections, although there may still be useful forecast information in the distribution obtained from the ensemble. As the predictive accuracy decreases, one would expect the coefficients and model selection in each case to approach a random selection from a uniform distribution, and the histogram should therefore tend toward a flat distribution. Noticeable departures from uniformity, for example members 2 and 4 at 216 h in the summer, would warrant further investigation. Although based on averages over relatively large samples (21 stations for 90-day periods, sample sizes in the range of 1800 cases), these results may not be significant; further testing on additional seasons would be of value to see whether there is a tendency for the preferred models to change according to seasonal average flow regimes for example.

These results are potentially useful diagnostics for model performance. It would be interesting to investigate whether there is any relationship between the characteristics of the models favored with relatively high coefficients and specific types of synoptic flow regimes. It might also be informative to relate variations in the seasonal average coefficients of individual models to the strengths and weaknesses of their formulations vis-à-vis seasonal differences in average atmospheric structure. It should be noted, however, that the weights are not directly related to the performance of

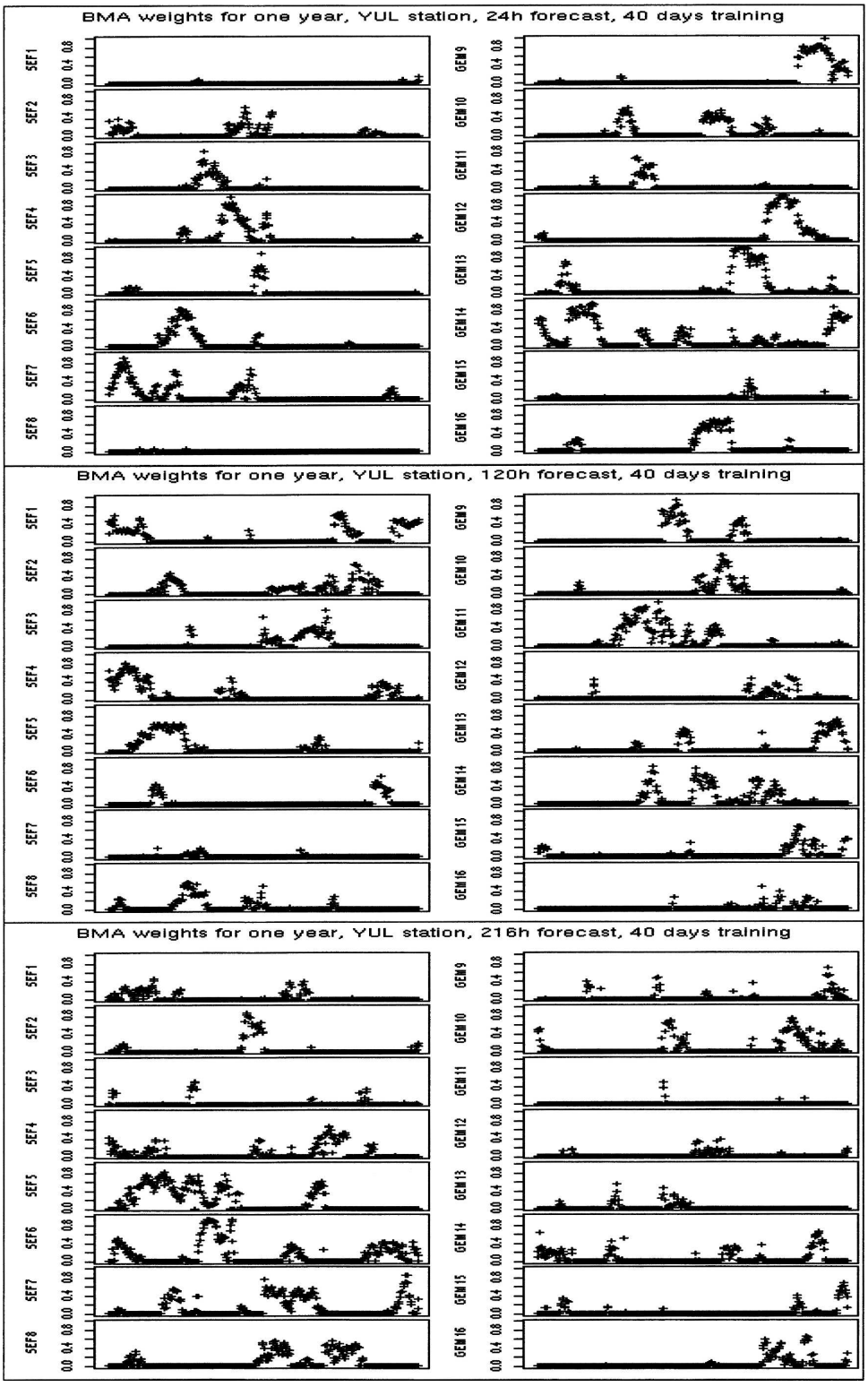

FIG. 2. Time series plots of BMA weights for the 1-yr development period for Montreal, based on 40-day training period: (top) 24, (middle) 120, and (bottom) 216 h with (left) SEF members and (right) GEM members. The range for each series is 0–1, and each plot consists of 366 points covering the year from 28 Sep 2003 to 27 Sep 2004.

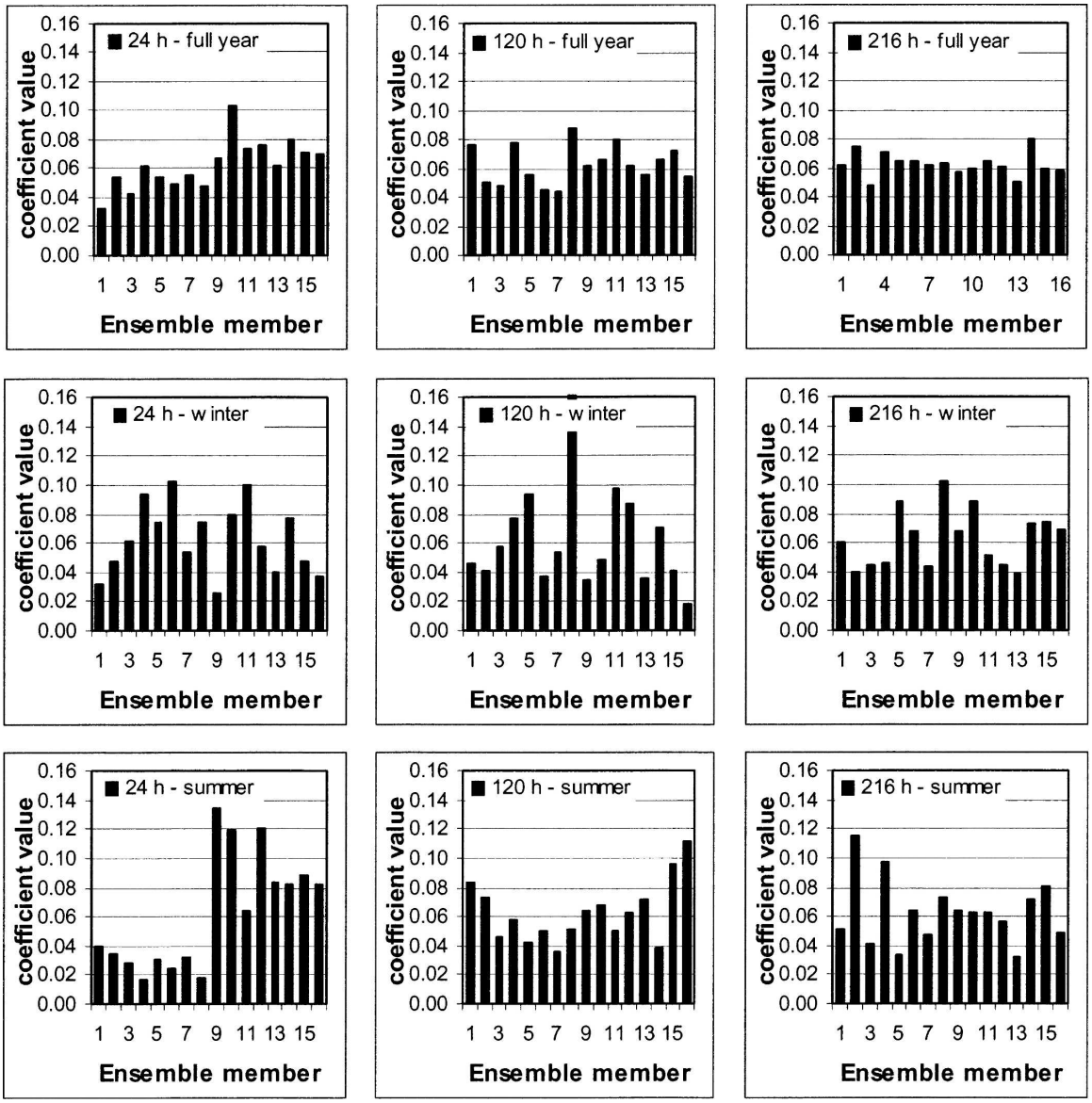

FIG. 3. BMA weights averaged over the 21 stations for (left) 24-, (middle) 120-, and (right) 216-h forecasts averaged over (top) the full year, (middle) December–February, and (bottom) June–August.

the individual models; each model is evaluated in light of the predictive information from all models. Thus, if two models perform similarly during the training period, but neither adds much predictive skill with respect to the other, BMA will tend to select the better one and practically ignore the other. This occurs, for example, when two forecasts are very highly correlated during the training period. The magnitude of the coefficients relates both to the accuracy of the model and its ability to bring unique predictive information to the ensemble.

Figure 4 shows an example of how the BMA works on a single forecast. In this case, a full regression has been used prior to the BMA, and the error variance was pooled over all models and applied to each in the averaged pdf. The BMA pdf is shown by the heavy line. This is an example where the forecast predicted cooler temperatures than long-term climatology (dash–dot line), and the regression has increased the temperatures. The component model pdfs, shown by the lighter Gaussian curves, tend to spread the distribution somewhat: one of the models near the cold tail of the distribution carries the highest weight, while two others with lower weights, predict near the climatological mean. Of the 16 models offered to BMA, only 3 of them carry any significant weight in this case. This is consistent with the

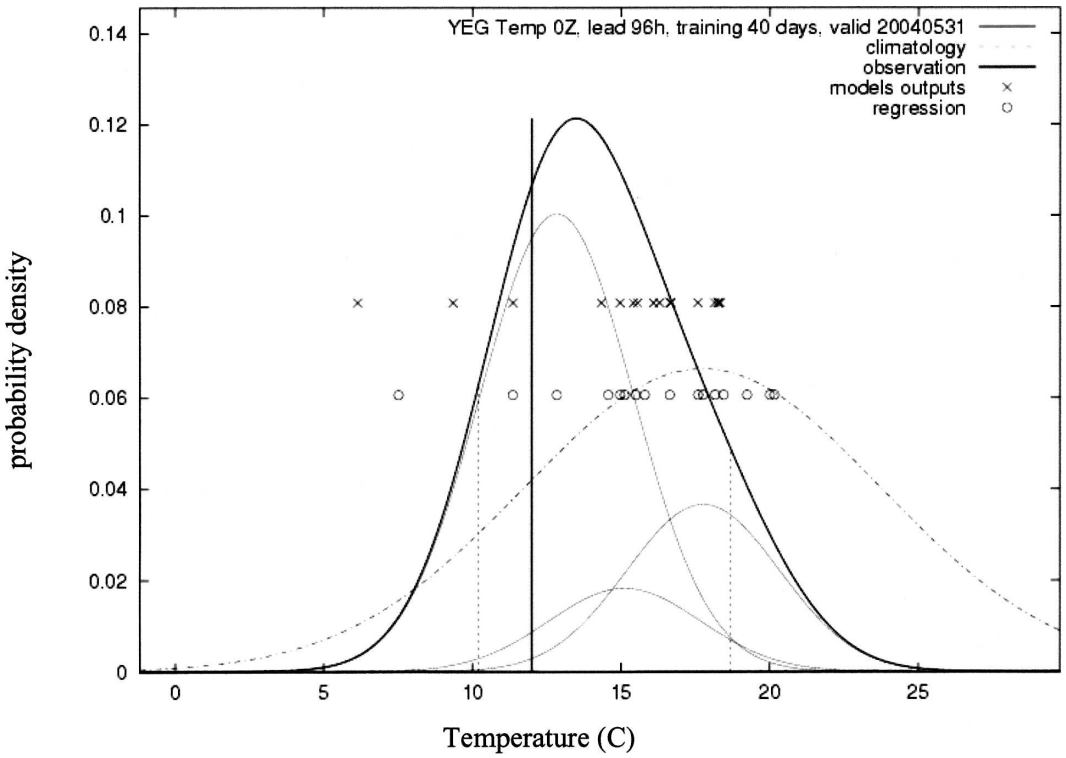

FIG. 4. Example of a single forecast, with a lead time of 96 h, for Edmonton. The original ensemble is shown as crosses; the bias-corrected ensemble is shown as circles, dashed curves are the pdfs for the individual ensemble members, the heavy curve is the BMA pdf, and the dash–dot curve is the sample climatological distribution for the valid time. The heavy vertical line represents the observation and the vertical dotted lines indicate the 80% probability interval from the BMA pdf.

coefficient trends shown above, and was typical of all the results for all stations.

The tendency of the BMA to assign significant weights to only a few of the models might seem surprising. In this way it objectively mimics the "model of the day" approach used by forecasters, where the model judged to give the best solution on a particular day is chosen, perhaps modified, and all other solutions are rejected. It is possible that nonzero coefficients would be spread out over more members with larger sample sizes, especially if the samples were kept seasonally homogeneous. A comparison of the coefficient statistics for 40- and 80-day training samples gave no evidence of this however, and the much larger sample sizes used in Raftery et al. (2005) also produced negligibly small coefficients for some models. A more likely explanation of this outcome is that there is considerable colinearity in the training samples, which means that relatively few of the models are needed to explain most of the explainable variance in the observations.

To further investigate this issue, histograms of the coefficient values were examined. Of the approximately 120 000 weights for the 16-member, 21-station 1-yr dataset, as many as 75% took on small values, and up to 500 or so took on relatively large values near one. There was a tendency for the number of very small coefficients and the number of very large coefficient values to decrease with longer projections, indicating spreading of the weights over more of the models at these projections. This is consistent with expectation since, at longer projections the accuracy of all the models is low, and the best model for a particular training period moves closer to a random selection from all the models. Even at 10 days though, a significant majority of the weights took on small values, reflecting colinearity in the forecasts as discussed above. The tendency toward significant numbers of very large weights is of more concern especially at longer projections, for it indicates that there could be overfitting in the results. Accordingly, we tested this too, by rerunning the BMA analysis on the full dataset with a considerably relaxed convergence criterion for the EM algorithm, differences between successive iterations of 10^{-2} instead of 10^{-4}. This did indeed significantly reduce the number of high-valued weights and spread the weights more evenly over more models. Also, these results slightly

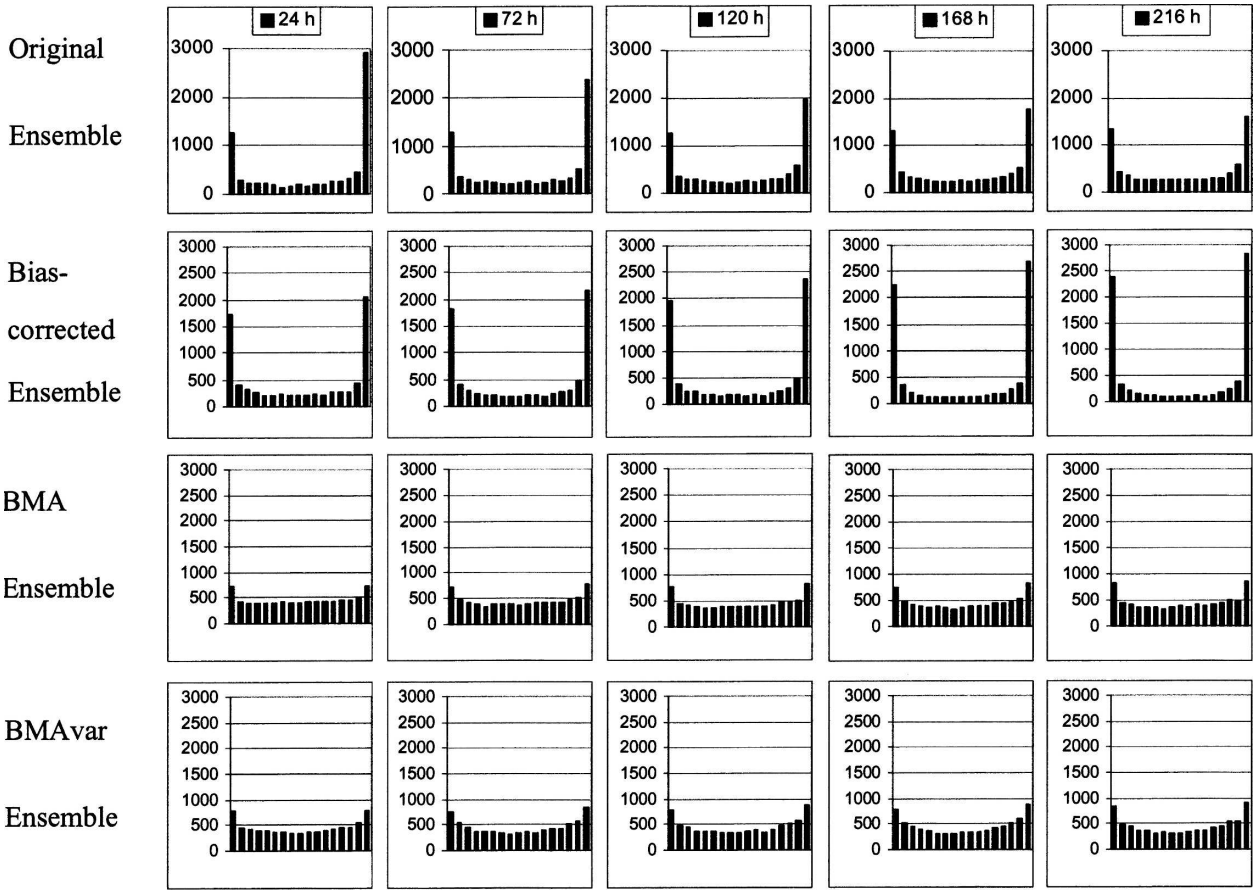

FIG. 5. Rank histograms for temperature forecasts for 1 yr of independent data, 21 Canadian stations. (top to bottom) Original ensemble forecasts, regression-corrected forecasts, BMA forecasts, and BMA forecasts with variable variance. Columns are for 1-, 3-, 5-, 7-, and 9-day forecasts. The height of the bars in each histogram indicates the number of cases for which the observation fell in each of the 17 bins.

improved performance on the independent dataset, which is also consistent with overfitting in the original results. However, the effects were not significant, improving the CRPS only by about 0.05° on average, and a significant majority of the weights remained near zero. Furthermore, the relaxed criterion produced some undesirable effects, such as decreasing the spreading of the weights at longer projections, which is counterintuitive. We therefore report the original results in the following sections, and note that further exploration is needed to optimize the convergence criterion in the BMA for smaller samples.

b. Rank histogram and CRPS results

The rank histogram is a way of assessing the ability of the BMA to correct the tendency of the ensemble distribution toward underdispersion. Rank histograms were computed over all the independent forecasts, for the original ensemble, the bias-corrected ensemble (both FR and b1), BMA, and BMAvar. BMA and BMAvar equations were run on both versions of the bias-corrected forecasts.

Figure 5 shows a grid of rank histograms, with the rows representing the stages of the calibration, from the original ensembles at the top to BMAvar on the bottom. The columns are for the different projections, each 2 days from 1 to 9 days left to right. These results were obtained with a training period of 40 days and bias removal using FR. The original ensembles show the underdispersion that is characteristic of ensemble forecasts, especially for surface weather variables. This underdispersion decreases with increasing projection time, which reflects the tendency for the ensemble spread to approach the spread of the climatological distribution at longer projections. The original ensembles also show a cold bias on average; the observation more often occurs in the highest (warmest) bin than in the lowest (coldest) bin of the ensemble distribution. The second row shows that the FR reduces the bias, especially at the early projections, but also increases the

underdispersion, especially at the longer projections. This is characteristic of regression; as the quality of the fit decreases for longer projections, the regression predicts more toward the mean of the predictand, reducing the predicted variance in the forecasts.

The third row in Fig. 5 shows the results for the BMA with constant variance over all models. The forecasts are essentially unbiased and the dispersion has been increased noticeably by BMA, as indicated by the near flatness of the rank histograms. The correction effected by BMA is nearly uniform across all projections; the BMA has corrected the underdispersion even when fed the reduced-dispersion longer-range FR-based forecasts. There remains some underdispersion, however, which might be attributable to differences between dependent and independent sample characteristics. The fourth row of rank histograms, for BMAvar forecasts, indicates performance rather similar to the BMA results, although perhaps not quite as good. To examine this further, we quantified the departure from flatness of the rank histogram following the method of Candille and Talagrand (2005):

$$\mathrm{RMSD} = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(s_k - \frac{M}{N+1}\right)^2},$$

where RMSD is the root-mean-square difference from flatness, expressed as number of cases, M is the total sample size on which the rank histogram is computed, N is the number of ensemble members, and s_k is the number of occurrences in the kth interval of the histogram. Candille and Talagrand (2005) point out that, due to the finite sample size, the expected value of the sum is $[MN/(N + 1)]$. This means the expected value of the RMSD is $\sqrt{[MN/(N + 1)^2]}$, which for the sample sizes considered here is about 20. The RMSD for the raw ensemble ranged from almost 700 at 24 h to just under 400 at 240 h. The lowest value we could obtain in our analyses was a little less than 100 cases confirming that there remains some systematic underdispersion in all the results, although we have removed most of the underdispersion in the raw ensemble. In other words, if the rank histogram were flat, sampling variation would allow a RMSD of about 20 cases over the full histogram. The best we managed in our analysis was an RMSD between 80 and 100, which indicates more variation in the heights of the bars than expected over the rank histogram. The RMSD was computed for the original ensembles, bias correction by FR and bl, and the subsequent BMA and BMAvar analyses.

Figure 6 shows these results, for original, regression, BMA, and BMAvar, using both full-regression and regression constant-based bias correction only. The figure

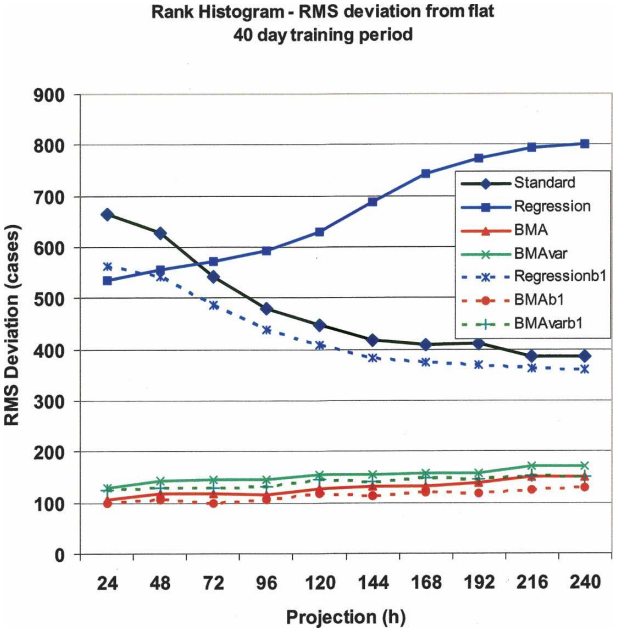

FIG. 6. RMS deviation from a flat rank histogram as a function of forecast projection for original ensemble (black), bias-corrected (blue); BMA (red) and BMAvar (green) ensemble forecasts (1-yr sample) for all 21 stations. BMA following FR is shown as a solid line and the BMA following the b1 bias correction is shown as the dashed lines.

confirms that BMAvar is not quite as well calibrated as the BMA. This may be due to the additional parameters that must be estimated for the BMAvar, 16 variances instead of one. Training samples of 40 days might not be long enough to provide stable estimates of these additional parameters. BMAvar performance is slightly worse than that of BMA also for the b1 equations. Using only an additive bias removal does improve the BMA results overall in comparison to the FR. This is likely due to seasonality problems in the fitting of the full regression, an issue explored more fully below.

The figure also shows that the departure from flatness of the rank histogram is greater after the FR than it is for the original ensemble after day 2 of the forecast. This is due to the tendency of the full regression to decrease the ensemble spread for the longer projections. The b1 results show a fairly small, but consistent improvement in the rank histogram over all projections compared to the original ensemble. This improvement can be attributed mainly to the tendency of the bias correction to "equalize" the frequency of occurrence of the observation in the two extreme bins of the histogram.

Figure 7 shows a comparison of the RMS departure from flatness of rank histograms for b1 equations, for 40- and 60-day training periods. The curves for the 40-

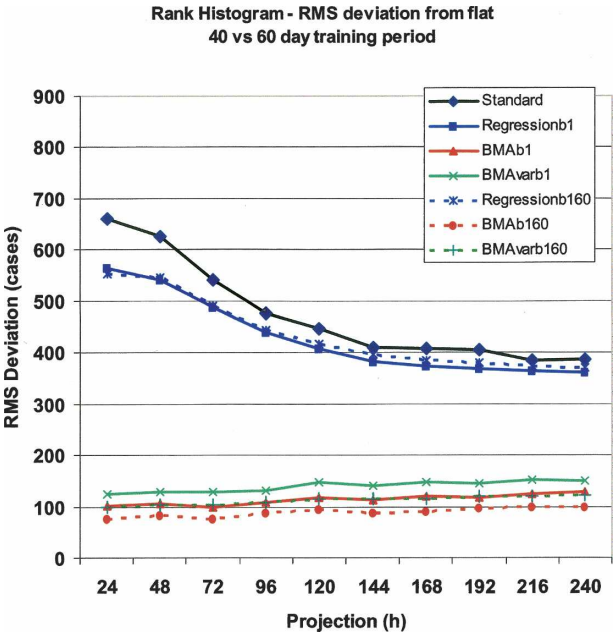

FIG. 7. Same as in Fig. 6, but for b1 bias-correction results only, 40- (solid) and 60-day training periods (dashed).

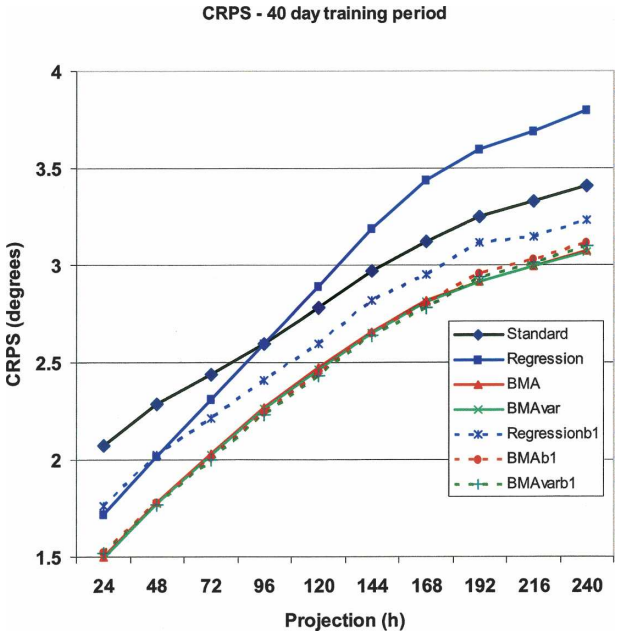

FIG. 8. Same as in Fig. 6, but for the CRPS.

day training period are the same as in Fig. 6, to facilitate comparison. As would be expected, the bias correction itself is not quite as effective for 60-day training periods as for 40 days, especially after day 3. The extra 20 cases included in the 60-day training periods occur earlier in time compared to the independent case on which the forecasts are tested. Thus, the training sample as a whole is likely to be less representative of the independent data than for shorter training periods. This effect is most pronounced for the longer forecast projections. While this problem might be avoided by using cross validation for the development and testing, this is not possible in operations, where the latest updated equations must be used to prepare the next forecast.

Despite a slightly poorer bias correction, the longer training period does result in a small improvement in calibration of the BMA-corrected forecasts. The longer training period presumably provides a more stable estimate of the weights using larger samples, which translates to slightly less noise in the rank histogram distribution. For BMAvar, where more coefficients must be estimated, the difference with the larger sample is slightly more pronounced, but there is still more variation from flatness in the BMAvar than in the BMA results. In numerical terms, BMAvar with a 60-day training sample is about as well calibrated as BMA with a 40-day training sample. In summary, these last two figures suggest that longer training periods with simple bias removal (bl) and BMA with constant variance provide the most reliable correction of the underdispersion in the ensemble temperature forecasts.

We now turn to the CRPS results for further exploration of the different options for BMA analysis. The CRPS summarizes the accuracy of the ensemble distribution, by determining the integrated distance of the cdf from the observation represented as a cdf. The CRPS has the same dimension as the variable in question, temperature, and is understandable as the equivalent of mean absolute error for a single deterministic forecast.

Figure 8 corresponds to Fig. 6, except it shows the CRPS results for a 40-day training period rather than the RMS departure of the rank histogram from flatness. First of all, the FR bias removal improves on the original ensemble forecast only for the first 3 days of the forecast. By reducing the dispersion in the ensemble as forecast projection time increases, the FR, which is essentially model output statistics with one predictor, produces forecast distributions that are too sharp (too overconfident) in the face of decreasing accuracy of the individual members. This leads to an increased frequency of situations where the bulk of the probability density is far from the observation location.

Second, the b1 bias correction leads to a reduction (improvement) in the CRPS of about 0.2°, with slightly larger values at the short forecast ranges. The BMA improves the CRPS by another 0.2°. It is interesting that the BMA result for the FR is nearly indistinguishable from the result for the b1 bias correction. Essen-

tially, the BMA has "made up" for the effects of the decrease in dispersion at the longer forecast ranges, which suggests that, at least for a 40-day training period, either FR or b1 could be used to correct bias, as long as a BMA analysis follows. And finally, the BMAvar results are similar to the BMA results.

Figure 9 shows the yearly summary results for a longer training period of 80 days. The levels of error indicated differ very little from the corresponding results for 40 days, though there is a slight tendency for these results to be worse overall. The only noticeable difference is now that, for medium and long ranges, there is a tendency for the b1 bias removal to perform better than the FR version. This tendency was traced to the failure of the FR to fully remove the bias with respect to the independent data, as shown below.

To explore the question of the best training period using the available data, Fig. 10 shows the CRPS for BMA for independent data as a function of the length of the training period, for both FR and b1. In general, a minimum in the CRPS (best performance) occurs around 35–50 days, depending on the projection. However, all the curves are rather flat, suggesting that the dependence of the performance on the length of the training period is rather weak within the range tested here. There is ample evidence, however that 25 days is too short. For the shortest forecast ranges, there is little difference between FR and b1. At medium and longer ranges, the minimum CRPS occurs for shorter training periods in the FR than for the b1 version. This minimum in the FR results may be artificial in some sense: the relatively rapid rise in the CRPS for longer training periods and longer forecast projections for the FR is caused by its failure to remove the bias in the spring and autumn. Rank histograms for 216 h (Fig. 11) show a cold bias in the spring and warm bias in the fall, while the corresponding results for b1 are relatively bias free. In the figure, the cold bias is indicated by a large number of occurrences of the observation in the 17th bin, which means all the ensemble members are forecasting lower temperatures than observed. Conversely, a warm bias is indicated when the observed temperature most frequently occurs in the lowest bin of the histogram. When the extreme bins of the histogram are approximately equally populated, this indicates an unbiased ensemble. When presented with biased forecasts following the regression step, the BMA improves the spread in the forecasts, as before, but they remain biased (Fig. 12). The FR may fail to remove the bias in spring and fall because the slope coefficient attempts to fit the seasonal trend in the data. The independent case, which for a 8-, 9-, or 10-day forecast will verify at least 8 days after the end of the training period, may be seen

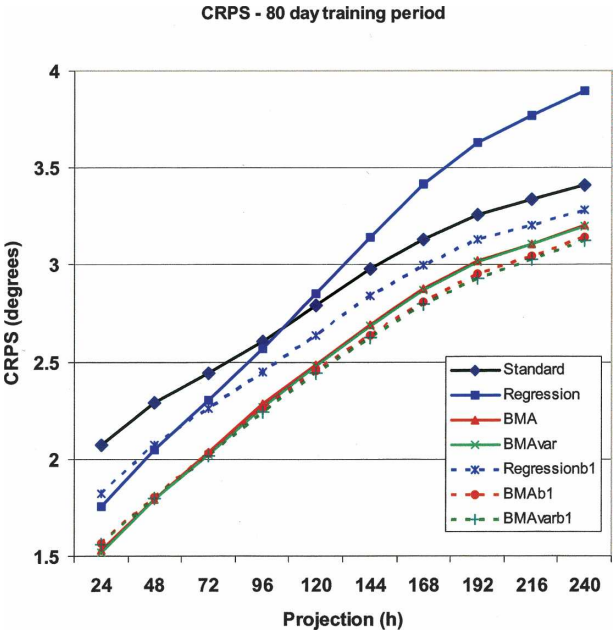

FIG. 9. Same as in Fig. 8, but for an 80-day training period.

as an outlier with respect to the training sample. This effect would be greatest for longer training periods and longer forecast projections, which is consistent with the results. Another factor may be the tendency for overlap of spring and autumn training periods with the previous seasons, which would lead to underfitting of the seasonal trend. Again, this would be most pronounced for longer training periods and longer forecast projections. Since the biases in the spring and fall results are in the opposite sense, cold in the spring and warm in the fall, they tend to partially cancel each other when the rank histogram is accumulated over the whole year. This is an example of a limitation of the rank histogram, which was pointed out by Hamill (2001).

In summary, these results indicate that a training period of about 40 days is a good choice for general use. BMA does a good job of correcting the underdispersion for either a full-regression bias removal or an additive bias correction with training periods of this length. Shorter training periods are too short for reliable estimates of the weights, while for longer training periods, best results are obtainable only if the simpler b1 bias correction is used, because of difficulties fitting seasonal variations in the training sample. Allowing the variance to vary among the component models does not improve the results.

c. 16- versus 18-member ensembles

One advantage of BMA is that it is a simple matter to include any available models in the analysis, and the

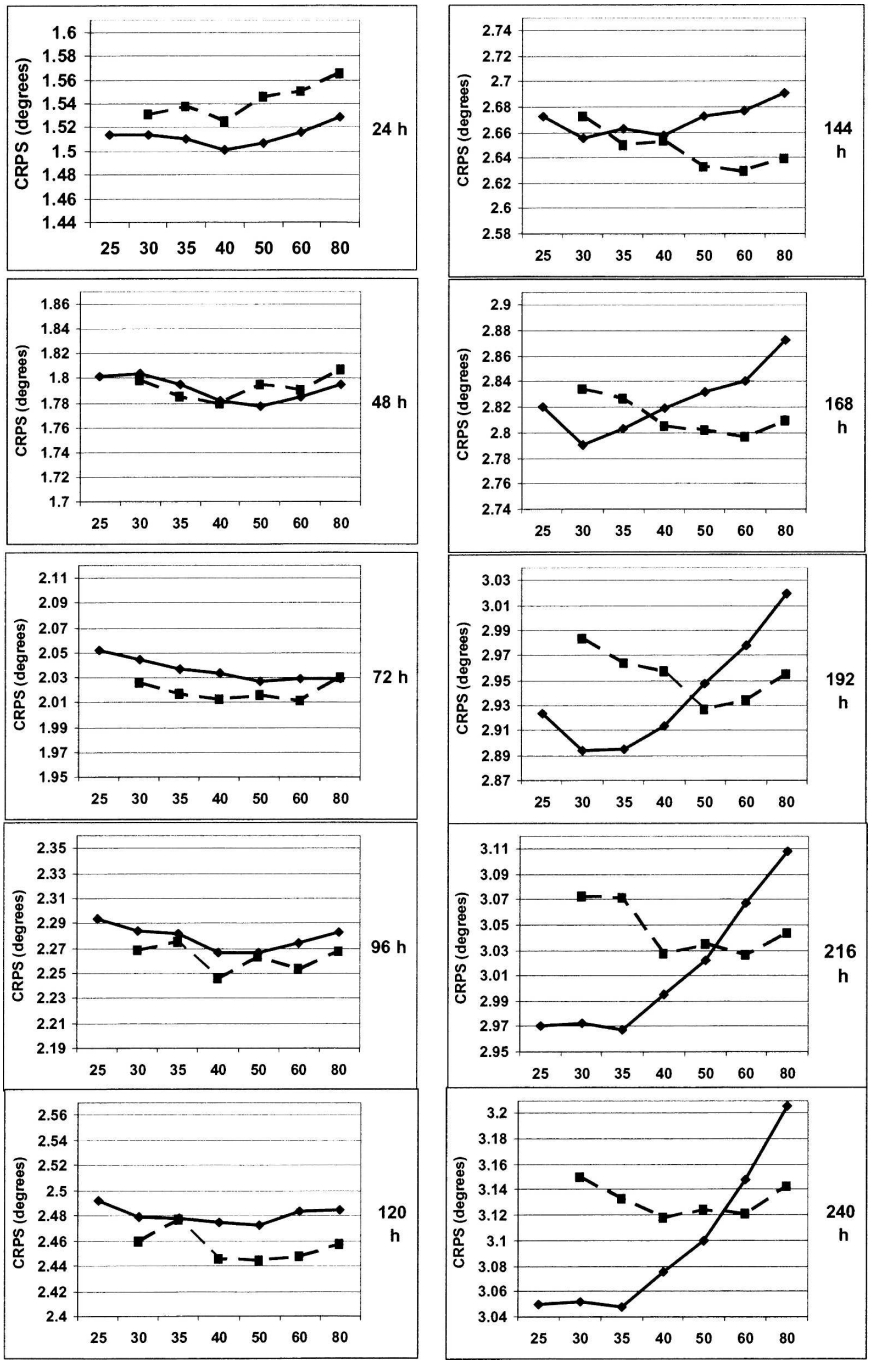

FIG. 10. CRPS for 1 yr of independent BMA forecasts (all stations) as a function of training sample size. The FR bias removal is shown by solid lines and the b1 bias removal is shown by dashed lines.

BMA procedure will combine and calibrate forecast distributions based on all available models. In adding the unperturbed control and the full-resolution global model to the analysis, we expected that the additional predictive information would improve the forecast output. In general this was true, but the improvements were relatively modest.

Figure 13 shows the summary CRPS results for the full year independent dataset for the 18-member ensemble, for both the FR and b1 bias corrections, plotted

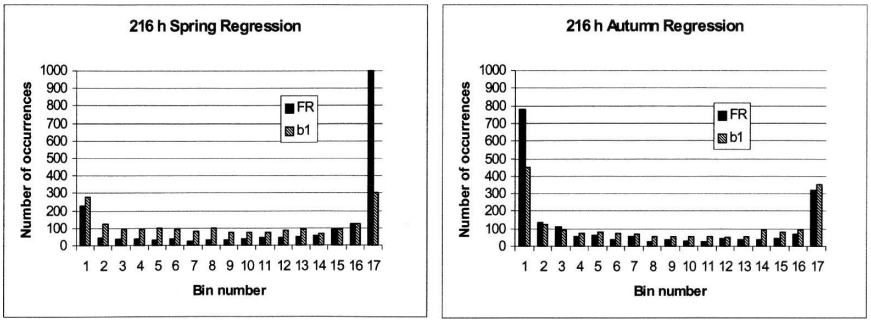

FIG. 11. Rank histograms for (left) spring and (right) autumn ensemble forecasts following FR (solid) and b1 (hatched) bias removal for the 216-h forecast projection. The samples are for 3-month periods at all 21 stations.

as a function of both the training sample length and the projection. The differences shown with respect to the 16-member ensemble are positive everywhere except for the shortest training period and longest forecast ranges. The impact of the additional members reaches a maximum of nearly 0.1° for 96-h forecasts, then drops off for the longer-range projections. To put that in perspective, the BMA typically improves the CRPS by about 0.2° with respect to the bias removal only, (see Fig. 8) which means the maximum achieved improvement is about 50% of the basic BMA improvement. For forecasts of 8 days or more it is safe to say that there is no meaningful advantage to the additional ensemble members for any training period; neither the full-resolution model nor the control forecast adds any predictive information. The patterns for the two forms of bias removal are very similar.

Additional diagnostic information on the relative predictive ability of the different models is provided by looking at the average weights for the 18-member ensemble BMA. Figure 14 shows another grid of histograms with the columns representing the short-, medium-, and longer-range projections. The top row is for the full year, and the other four rows are for the four 3-month seasons. In all cases, the control forecast is the first (leftmost) and the full-resolution global model is the 18th (rightmost) member. This grid of charts makes several points. First, at the shortest forecast range, the full-resolution model is by far the most important member of the ensemble. This is not surprising, indicating the advantage of the higher resolution at short range, even though the model resolution is only slightly finer (100 km) than the 130 km of the GEM members of the ensemble. The relative importance of the global model can be seen in all seasons except summer, where all the GEM members are relatively important. Second, the relative importance of the global model decreases with increasing projection; by 216 h, it is "just one of the members." This is consistent with the CRPS results in Fig. 13. Third, the control model carries, along with the GEM model, relative importance at the medium range. This is interesting, suggesting that the perturbations may have a tendency to degrade the forecast relative to an unperturbed model of the same resolution. The relative importance of the control model is usually less than the full-resolution model, and it too becomes "one of the pack" by 216 h.

Overall, these results do give the impression that

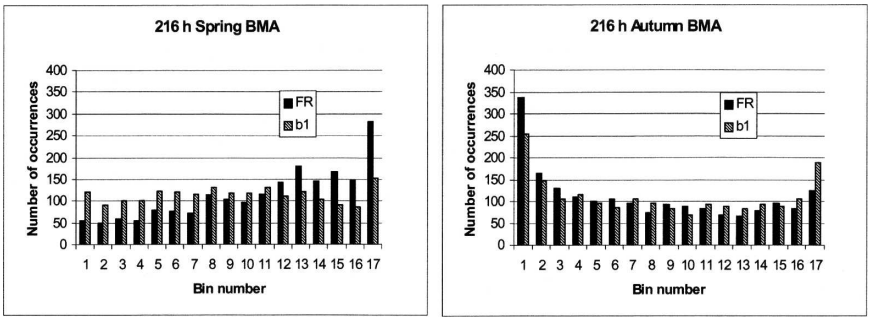

FIG. 12. Same as in Fig. 11, but for 216-h forecasts. Samples are for 3 months for all 21 stations.

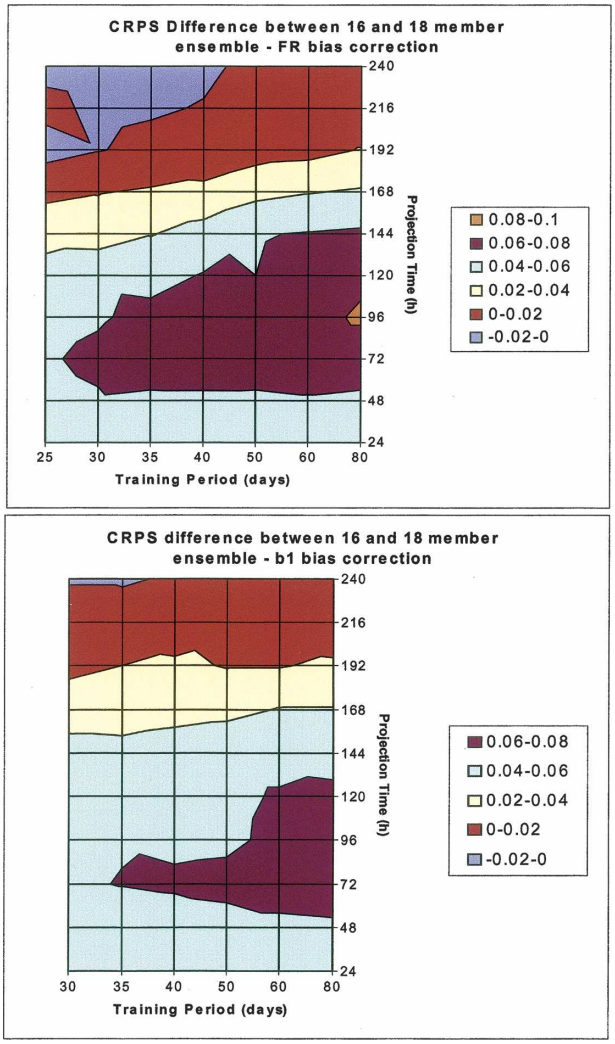

FIG. 13. CRPS difference between 16- and 18-member ensemble forecasts after BMA calibration, as a function of training period and forecast projection. Positive values indicate better (lower) CRPS values for 18 members.

both the control model and the full-resolution model bring predictive information that is unique compared to the ensemble members. This is potentially significant, for it suggests that there may be an advantage to running BMA on ensembles composed of models from different sources.

d. CRPSS results

All of the results discussed so far relate to the comparison of the accuracy of the independent sample forecasts for different configurations of the calibration. The CRPSS shows the skill of the forecasts against an unskilled standard forecast, that is, the station-specific daily climatological temperature distribution. The climatology is derived from 30 or more years of observations for each station and each day of the year. The skill score expresses the difference between the CRPS achieved by climatology and the CRPS achieved by the forecast, normalized by the seasonal average climatological score. Details of the score computation are given in the appendix.

Figure 15 summarizes the skill characteristics of the original, regression and BMA-calibrated forecasts, over the 1-yr independent sample. Original and BMA-calibrated results are also shown for the 18-member ensembles.

The skill curves for the original, bias-corrected, and BMA forecasts for 16-member ensembles are essentially mirror images of the corresponding curves in Fig. 6. This is expected, for the underlying climatology is the same for all these curves; it is only the CRPS values that change. The original forecasts show positive skill to about day 8 while the FR bias correction moves the 0 skill point back to about 156 h. The simple additive bias removal is sufficient to extend the positive skill to the end of the forecast period and the BMA improves on the skill of the bias-corrected forecasts.

Regarding the results for the 16- versus 18-member ensembles, there is essentially no difference in the skill of the original ensemble forecasts. However, the BMA-calibrated forecasts show slightly higher skill for the 18-member ensembles than for the 16-member ensembles, out to day 7. This is consistent with Fig. 14, where the full-resolution model is favored with higher weights for the short- and medium-range forecasts.

e. Comparison of BMA with a simple Gaussian pdf

Finally, we try to answer the question of whether it is worth going to the trouble of assigning different weights to the different members of the ensemble. Perhaps a simple pdf estimated using the errors in the training sample would perform as well as the BMA. To test this, a Gaussian pdf was created for each forecast, using the bias-corrected ensemble mean as the mean and the root-mean-squared error of the ensemble mean based on the training period as the standard deviation. The Gaussian pdfs were then evaluated on the independent data using the CRPS in exactly the same way as the BMA forecasts were evaluated.

The results of this evaluation are presented in Figs. 16 and 17. Figure 16 shows the overall CRPS results for the 16- and 18-member ensembles, and the corresponding Gaussian pdfs, all based on the independent data. There is little difference in performance between the BMA and the Gaussian overall; the Gaussian proves to be a good competitor. Without the ability to assign higher weights to the better performing ensemble mem-

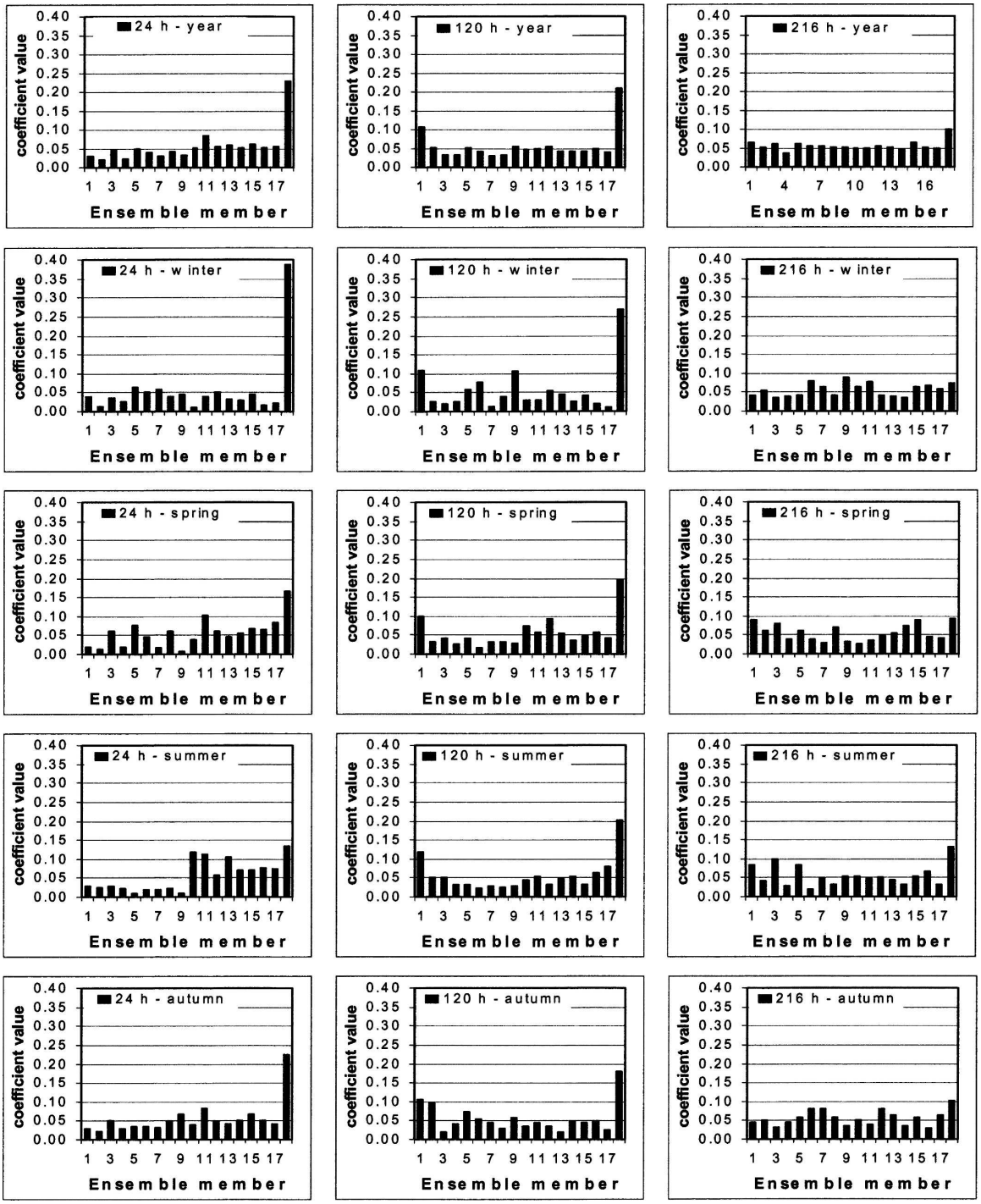

FIG. 14. Average BMA weights for 18-member ensembles for all 21 stations, for (top) the full year and (second to bottom) 3-month seasons. (left to right) Columns are for 24, 120, and 216 h.

bers, the Gaussian results are nearly identical for 16 and 18 members. On the other hand, the BMA is able to take advantage of the better performance of the full-resolution model in the first 5 days to give it a slight performance edge over the Gaussian and the 16-member ensemble BMA. This suggests that, in multimodel situations where there are large differences between the models, the ability to weight the individual members becomes more important.

Figure 16 also shows the mean absolute error (MAE)

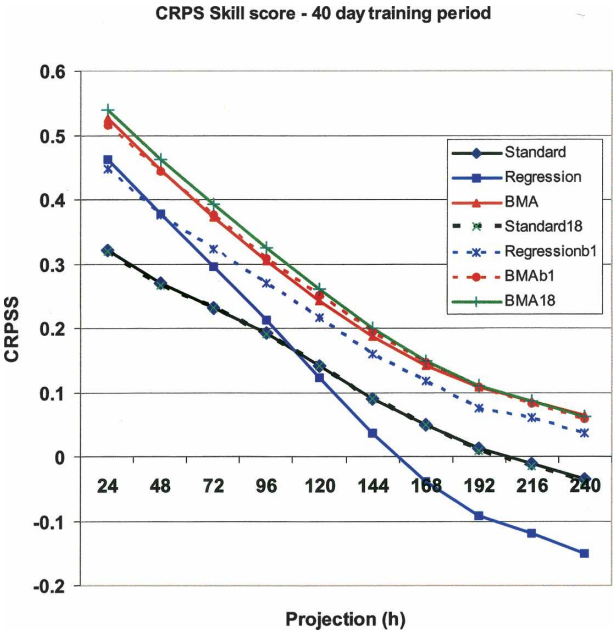

FIG. 15. CRPSS with respect to climatology for 16- and 18-member ensembles based on 1 yr of independent data. Original ensemble results are in black, bias-corrected in blue, BMA for 16 members in red, and BMA for 18 members in green.

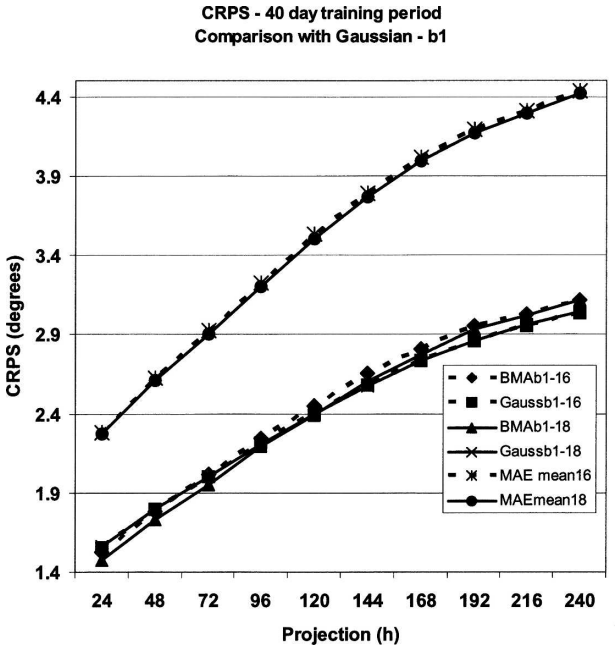

FIG. 16. CRPS for 1 yr of independent BMA and Gaussian pdf forecasts (40-day training period) for 16- (dashed) and 18-member ensembles (solid). Also shown is the MAE for the corresponding ensemble mean forecasts.

of the ensemble mean, which is much higher than the CRPS for either the BMA or the Gaussian. Since the MAE is consistent with the CRPS in the sense that the CRPS reduces to the MAE for a deterministic forecast, this effectively shows the error levels that would be obtained if the spread of the ensemble were to be reduced to zero.

The true advantage of the BMA over the Gaussian in the present results is revealed in Fig. 17, which shows the average 90% prediction interval for Gaussian and BMA pdfs, for both 16- and 18-member ensembles. There is little difference between the 16- and 18-member ensemble results, but the BMA has significantly reduced the 90% prediction interval compared to the Gaussian, by 20%–25% over the whole 10-day prediction range. This is significant if the forecast pdfs are to be used for credible interval temperature forecasting, for example. A 25% more precise interval is important when it can be achieved without loss of accuracy (Fig. 16) and with ensembles that are not markedly underdispersed (see Fig. 5).

6. Discussion

This paper describes results of experiments with Bayesian model averaging as a tool for the calibration of ensemble forecasts from the operational Canadian ensemble forecast system. 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, then applied to the next forecast. The BMA was applied to bias-corrected forecasts only, and two methods were tested for bias removal prior to the BMA analysis. Several different training sample sizes were tested, ranging from 25 to 80 days. The benefits of allowing the variance of the component models to vary, and the effect of adding the control forecast and the full-resolution model forecast to the ensemble were also evaluated. And finally, the fundamental question of whether the BMA improves upon a very simple Gaussian pdf with equal weights was examined.

The main results of the study can be summarized by the following statements:

1) The BMA faithfully removed most, but not all of the underdispersion exhibited by the original ensemble. It proved to be capable of this even for the longest-range forecasts, where the underdispersion had been increased by the regression-based bias removal procedure. The remaining underdispersion seems to be systematic but small and may be related to differences between the training sample and the independent data on which the BMA was evaluated.
2) On the basis of the 1 yr of data available, the results suggest that training periods of 40 days give the best

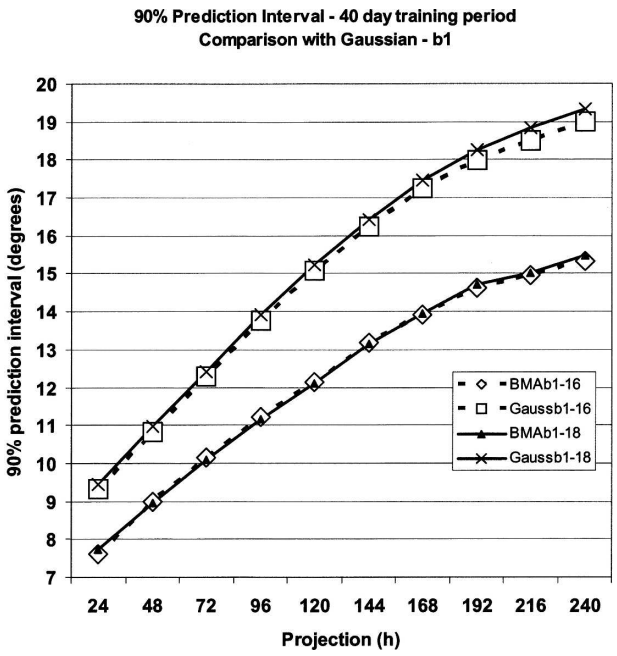

FIG. 17. 90% prediction interval for 1 yr of independent BMA and Gaussian ensemble forecasts (40-day training period) for 16 (dashed) and 18-member ensembles (solid).

results, though we found that forecast accuracy on the independent data did not vary strongly with the training period. However, 25 days seemed to be clearly too short a period on which to fit the BMA weights, and there was some evidence that longer training periods with only an additive bias removal gave a more stable BMA analysis.

3) Bias removal using simple linear regression seemed to work about as well as bias removal by correcting only for the mean error up until about 7 days, and for training periods up to about 50 days. For longer training periods and longer projections, performance for regression-based bias removal was poorer than for additive bias removal apparently because the bias was not successfully removed by the regression for the spring and fall seasons.
4) Allowing the variance to vary among the component models in the BMA did not improve the performance; in fact, there was a tendency to slightly degrade the performance on the independent data. This probably means that the training samples were not large enough to obtain reliable estimates of the additional coefficients.
5) Addition of the control forecast and the full-resolution model to the ensemble improved the results modestly, up until about 7 days. Beyond that, the additional models do not have enough skill to contribute to the accuracy of the calibrated forecasts.
6) Examination of the BMA weights is extremely useful in a diagnostic sense, to identify models that contribute less to the ensemble than other models, and to identify synoptic situations that are handled particularly well or poorly by the individual model formulations. The fact that the full-resolution model carried higher weight on average for shorter-range forecasts indicates that the extra resolution is beneficial, while the relatively high importance of the control model suggests that, at least for some seasons and projections, the perturbations increase the error levels of the individual models.
7) The BMA achieved approximately equal accuracy in terms of the CRPS overall, but produced significantly sharper pdfs when compared to a simple Gaussian pdf with standard deviation equal to the RMSE of the ensemble mean based on the same training period. The 90% prediction interval was reduced by more than 20% over the 10-day forecast range in the independent sample. These results also suggest that the ability to weight the component models of an ensemble is more important when there are significant systematic differences in the structure of the ensemble members.

These results have dealt with surface temperature only, for which the assumption of a normal error distribution for each model is reasonable. For other variables such as precipitation and wind, the gamma distribution might be more appropriate, but there are problems to work out such as how to handle calm winds and 0 precipitation events in the combined distribution. This is the next major step of our work in applying BMA to Canadian ensemble prediction.

One might also speculate whether a system that assigns different weights to the events of the training sample would enhance the performance. On the assumption that more recent realizations of the forecast are better indicators of the current performance than earlier realizations, perhaps a weighting scheme could be added to the BMA weight calculation. Such a scheme might also improve the fitting of the seasonal trends.

Another enhancement one might consider is the use of a full screening multipredictor MOS fit to remove the bias. In that case, seasonal variations in the training sample could be accounted for by the addition of specific predictors, and/or a weighting scheme could be used.

Both the bias removal and the BMA steps might also benefit from the use of longer training periods of data from the ensemble models, such as would be available from reforecasting projects (e.g., Hamill et al. 2004). If

sufficient data were available, one could stratify the training sample by season, using the corresponding period from more than 1 yr. Considering the evidence presented here that seasonal variations within the training dataset led to poorer performance on independent data, the use of multiyear, seasonally stratified training datasets might produce higher-quality predictions than were possible in this study. The disadvantage of such an approach is that any significant change to the ensemble model would mean that the reforecasting would have to be redone to produce a new statistically representative training sample. Or, perhaps an updating system could be devised that would ensure a smooth transition from calibration with respect to an old system to calibration with respect to a new system.

When BMA is applied to small samples, as in this study, care must be taken to set the parameters of the EM algorithm to avoid overfitting, while at the same time integrating the algorithm far enough to achieve meaningful results. Further work is needed on this issue.

In this study, we applied BMA in a way that is consistent with previous applications, that is, to combine models that are distinct in order to generate a consensus pdf that is calibrated. The question arises how BMA might be applied to other operational ensemble forecasts that do not consist of separate distinct models. BMA could also be used for a single-model ensemble system where only the initial conditions are perturbed. In that case, error distributions over a training sample would be expected to be as random draws from the same distribution, since it is the same model that is run each day. In that case, the weights ω_k in (5) and (7) would be constrained to be equal, so that $\omega = 1/k$ for each k. The EM algorithm can easily be modified for this case.

BMA should be a valuable method for the calibration of any multimodel ensemble system, whether it consists of individual discrete models, as in a "poor-man's" system, or of combinations of ensembles from different centers. The latter is becoming more important with the initiatives of the North American Ensemble Forecast System (NAEFS) and the global ensemble initiative of The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; Richardson et al. 2005). In such a system, which will be made up of single-model ensembles along with multimodel ensembles, BMA could be applied to simultaneously combine and calibrate them. If the members of a component ensemble are not distinct, the weights ω_k for the members of that ensemble would be constrained to be equal. Some very preliminary work has been done to apply BMA to combined Meteorological Service of Canada and the National Centers for Environmental Prediction ensembles, using a total of 26 members all initialized at the same time. Results of this work are promising so far.

Finally, it is the flexibility of BMA that makes it most attractive for use in multimodel ensemble systems: one can add any model for which data is available during the training period and produce a calibrated combination of models. This means it can be used to diagnose the added information that the ensemble brings to the deterministic forecast, which is of importance to all centers that run both an ensemble system and one or more deterministic models.

Acknowledgments. The authors wish to thank Dr. Tilmann Gneiting for helpful discussions on this paper, and Dr. Eric Grimit, Michael Polakowski, and J. McLean Sloughter for contributing software. Adrian Raftery's work was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. We also wish to thank Dr. Tom Hamill and an anonymous reviewer for their comments, which have greatly improved the paper.

APPENDIX

Verification Measures

a. Rank histogram

The rank histogram (Anderson 1996) is formed by first ranking the values from each ensemble forecast from lowest to highest. These ordered values then are used as thresholds to define $N + 1$ ranges or bins of the predictand, where N is the ensemble size. The two extreme bins are open ended. The histogram is formed by tallying the number of occurrences of the observation in each bin over the verification sample.

Under the assumption that the observation is equally likely to fall in each bin, a "perfect" rank histogram is one which is flat, indicating that, on average, the ensemble spread covers the variability in the predictand. The U-shaped histograms indicate that ensemble spread is too small on average (the observation too often lies outside the ensemble), and asymmetric histograms indicate biases in the forecasts.

Rank histograms are meaningful only for relatively large verification sample sizes.

b. Continuous rank probability score

Following the notation of Hersbach (2000), consider an ensemble forecast pdf $\rho(x)$ and a verifying observed value x_a. The CRPS is defined by

$$\mathrm{CRPS}(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 \, dx, \quad \text{(A1)}$$

where P and P_a are cumulative distributions (cdf),

$$P(x) = \int_{-\infty}^{x} \rho(y) \, dy, \quad \text{and} \quad \text{(A2)}$$

$$P_a(x) = H(x - x_a). \quad \text{(A3)}$$

Here $H(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$ is the Heaviside function. The CRPS is thus the difference between the predicted cdf and the observation expressed as a cdf. It is negatively oriented (smaller is better) and the perfect score of 0 is achieved only for a perfect deterministic forecast. The CRPS has dimensions of the variable x. The CRPS reduces to the mean absolute error for a deterministic forecast, and therefore can be considered as a mean absolute error for a probability distribution.

c. Continuous rank probability skill score

The CRPSS is a skill score in the usual format:

$$\mathrm{CRPSS} = \frac{\mathrm{CRPS}_c - \mathrm{CRPS}_f}{\mathrm{CRPS}_c}, \quad \text{(A4)}$$

where CRPS_c is the standard score, in this case for a climatological forecast and CRPS_f is the score for the forecast. To avoid the inclusion of artificial skill, the climatology reference score was computed using the long-term climatological distribution applicable to the day of the year and the station. Scores for each station i and each of four seasons s were then computed as

$$\mathrm{CRPSS}_{is} = \frac{\sum_{j=1}^{n_{is}} (\mathrm{CRPS}_{cisj} - \mathrm{CRPS}_{fisj})}{\sum_{j=1}^{n_{is}} \mathrm{CRPS}_{cisj}}, \quad \text{(A5)}$$

and the total score is the average of the individual station/season scores, weighted by the number of cases for the station and season:

$$\mathrm{CRPSS}_{\mathrm{TOT}} = \frac{\sum_{i,s} n_{is} \mathrm{CRPSS}_{is}}{N}. \quad \text{(A6)}$$

The score has a range of $-\infty$ to 1, with positive values indicating skill with respect to the standard score. Division by the sum of climatological scores over some number of cases (1 season in about 90 cases) is necessary to stabilize the score and prevent near-zero divisors when the observed temperature is near the climatological mean.

REFERENCES

Anderson, J., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **125,** 99–119.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.,* **131,** 231–250.

Coté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational CMC/MRB Global Environmental Multiscale (GEM) Model. Part I:–Design considerations and formulation. *Mon. Wea. Rev.,* **126,** 1373–1395.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.,* **39B,** 1–39.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **222A,** 309–368.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1202–1211.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

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

——, and S. J. Colucci, 1997: Verification of ETA-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensembles. *Science,* **285,** 1548–1550.

Lefaivre, L., P. L. Houtekamer, A. Bergeron, and R. Verret, 1997: The CMC ensemble prediction system. *Proc. Sixth Workshop on Meteorological Operational Systems,* Reading, United Kingdom, ECMWF, 31–44.

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions*. Wiley, 274 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Pellerin, G., L. Lefaivre, P. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes Geophys.,* **10,** 463–468.

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

Richardson, D. S., R. Buizza, and R. Hagedorn, 2005: Report of the 1st Workshop on the THORPEX Interactive Grand

Global Ensemble (TIGGE). ECMWF, 34 pp. [Available online at http://www.ecmwf.int/newsevents/meetings/workshops/2005/TIGGE/TIGGE_report.pdf.]

Ritchie, H., 1991: Application of the semi-Lagrangian method to a multi-level spectral primitive-equations model. *Quart. J. Roy. Meteor. Soc.,* **117,** 91–106.

Simonsen, C., 1991: Self-adaptive model output statistics based on Kalman filtering. Lectures and papers presented at the WMO training workshop on the interpretation of NWP products in terms of local weather phenomena and their verification, Wageningen, Netherlands, WMO PSMP Research Rep. Series 34, XX-33–XX-37.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting,* **16,** 463–477.

——, and Coauthors, 2005: The North American Ensemble Forecast System. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., CD-ROM, 11A.1.

Vallée, M., L. Wilson, and P. Bourgouin, 1996: New statistical methods for the interpretation of NWP output at the Canadian Meteorological Center. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 37–44.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

Wilson, L. J., and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206–222.

——, and ——, 2003: The Canadian updateable model output statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting,* **18,** 288–302.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus,* **52A,** 280–299.

Comments on "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging"

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 26 July 2006, in final form 18 October 2006)

1. Introduction

Wilson et al. (2007, hereafter W07) recently described the application of the Bayesian model averaging (BMA; Raftery et al. 2005, hereafter R05) calibration technique to surface temperature forecasts using the Canadian ensemble prediction system. The BMA technique as applied in W07 produced an adjusted probabilistic forecast from an ensemble through a two-step procedure. The first step was the correction of biases of individual members through regression analyses. The second step was the fitting of a Gaussian kernel around each bias-corrected member of the ensemble. The amount of weight applied to each member's kernel and the width of the kernel(s) were set through an expectation maximization (EM) algorithm (Dempster et al. 1977). The final probability density function (pdf) was a sum of the weighted kernels.

W07 reported (their Fig. 2) that at any given instant, a majority of the ensemble members were typically assigned zero weight, while a few select members received the majority of the weight. Which members received large weights varied from one day to the next. These results were counterintuitive. Why effectively discard the information from so many ensemble members? Why should one member have positive weight one day and none the next?

This comment to W07 will show that BMA where the EM is permitted to adjust the weights individually for each member is not an appropriate application of the technique when the sample size is small;[1] specifically, the radically unequal weights of W07 exemplify an "overfitting" (Wilks 2006a, p. 207) to the training data. A symptom of overfitting is an improved fitted relationship to the training data but a worsened relationship with independent data. This may happen when the statistician attempts to fit a large number of parameters using a relatively small training sample. In W07, the EM algorithm was required to set the weights of 16 individual ensemble members and a kernel standard deviation with between 25 and 80 days of data.

To illustrate the problem of overfitting in W07's methodology, a reforecast dataset was used. This was composed of more than two decades of daily ensemble forecasts with perturbed initial conditions, all from a single forecast model. This large dataset permitted a comparison of BMA properties based on small and large training samples. This reforecast dataset used a T62, circa 1998 version of the National Centers for Environmental Prediction (NCEP) Global Forecast System. A 15-member forecast, consisting of a control and seven bred pairs (Toth and Kalnay 1997) was integrated to 15 days lead for every day from 1979 to current. For more details on this reforecast dataset, please see Hamill et al. (2006). The verification data were from the NCEP–National Center for Atmospheric Research reanalysis (Kalnay et al. 1996).

2. Overfitting with the BMA–EM algorithm

EM is an iterative algorithm that adjusts the BMA model parameters through a two-step procedure of pa-

Corresponding author address: Dr. Thomas M. Hamill, NOAA/Earth System Research Laboratory, Physical Sciences Division, R/PSD 1, 325 Broadway, Boulder, CO 80305.
E-mail: tom.hamill@noaa.gov

[1] This is not meant to imply that BMA and the EM method are inappropriate, merely that the methods can be inappropriately applied.

DOI: 10.1175/2007MWR1963.1

rameter estimation and maximization. R05 [Eqs. (5)–(6) and accompanying text] provides more detail. The algorithm iterates to convergence, stopping when the change in log-likelihood function from one iteration to the next is less than a cutoff δ. The magnitude of δ may be chosen by the user, but it can be assumed $\delta \ll 1.0$.

To illustrate the tendency for the BMA EM to overfit when trained with small sample sizes, consider 4-day 850-hPa temperature ensemble forecasts for a grid point near Montreal, Quebec, Canada. Forecasts were produced and validated for 23 yr $\times$ 365 days $-$ 40 days $=$ 8355 cases. Because we would like to assume in this example a priori that the member weights should be equal, the 15-member ensemble was thinned, eliminating the slightly more accurate control member. The remaining 14 bred members can be assumed to have identically distributed (but not independent; see Wang and Bishop 2003) errors and hence should have been assigned equal weights. The BMA algorithm was then trained using the remaining 14 identically distributed bred members and only the prior 40-days forecasts and analyses, posited in W07 to be an acceptably long training period. We shall refer to this as the "40-day training" dataset. In addition, the BMA algorithm was also trained with a very long training dataset in a cross-validated manner using 22 yr $\times$ 91 days of data, with the 91 days centered on the Julian day of the forecast. This will be referred to as the "22-yr training dataset."

The BMA algorithm was coded generally following the algorithm used in the R05 and W07 articles. Two adjustments were used, however. First, no refinement of the fitted standard deviation was performed in order to maximize the continuous ranked probability score (Hersbach 2000), as suggested in R05. Performing the refinement increased the computational expense but had minimal impact on forecast skill. Second, W07's proposed regression correction was here applied only to the ensemble mean, while the original deviation of each member about the mean was preserved. More concretely, given an ensemble member x_i^f, an ensemble mean $\bar{x}^f$, and a regression-corrected ensemble mean forecast $(a + b\bar{x}^f)$, the member forecast was replaced with a forecast that was the sum of the initial perturbation from the ensemble mean and the corrected forecast:

$$x_i^f \leftarrow (x_i^f - \bar{x}^f) + (a + b\bar{x}^f), \tag{1}$$

where $\leftarrow$ denotes the replacement operation. This modified regression correction was used because when every member was regressed separately, as forecast lead increased and skill decreased, all the members were increasingly regressed toward the training sample mean of the observed. Consequently, the ensemble spread of adjusted members shrank (Fig. 1; see also Wilks 2006b) and colinearity of errors among members was accentuated (Fig. 2). These were clearly undesirable properties; the spread should asymptotically approach the climatological spread of the ensemble forecast, and ideally, member forecasts should have independent errors. Had the regression correction of each member been applied, there may have been some confusion as to whether the subsequent highly nonuniform weights produced by the BMA were a generic property of a short training dataset or whether they were artificially induced from the increased colinearity induced by the regression analyses.

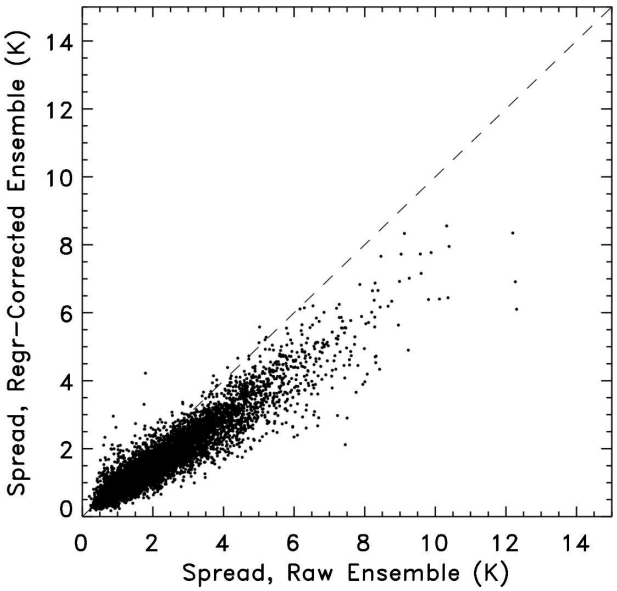

FIG. 1. Spread of a regression-corrected ensemble of day 4 forecasts of 850-hPa temperature at Montreal (using 40-day training data) vs the spread of the raw ensemble forecasts.

We now consider the properties of the EM algorithm for this application. The initial guess for all member weights was 1/14. Keeping track of the ratio of maximum to minimum BMA member weights after EM convergence for each of the 8355 cases, these ratios were sorted, and the median ratio was plotted as the EM convergence criterion δ was varied. For the 40-day training period, when $\delta = 0.01$, the largest and smallest weights were much more similar compared to when $\delta \ll 0.01$ (Fig. 3a). With the 22-yr training data, the weights stayed much more equal as δ was decreased (Fig. 3b).

Could the unequal weightings with the 40-day training set and tight δ actually be appropriate? As mentioned in R05, as the EM iterates, the log likelihood *of the fit to the training data* is guaranteed to increase.

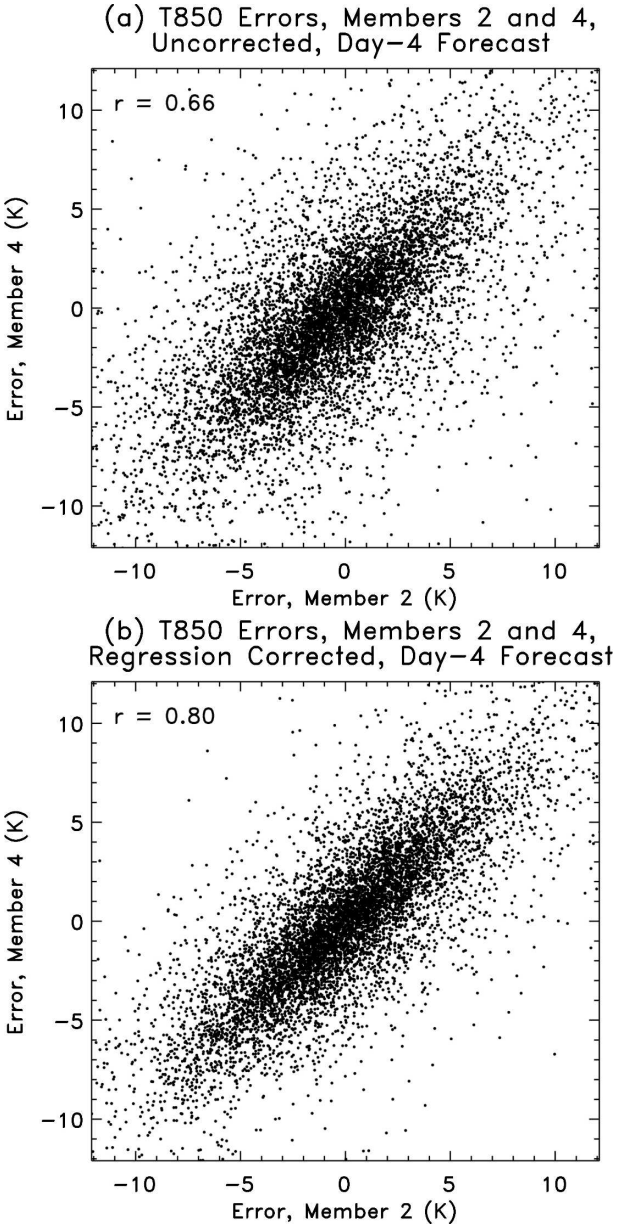

FIG. 2. Errors of day 4 850-hPa temperature forecasts for members 2 and 4 of the ensemble (a) before a regression correction of the member errors using the prior 40 days for training and (b) after the regression correction. Correlation coefficient (r) noted in the upper-left corner.

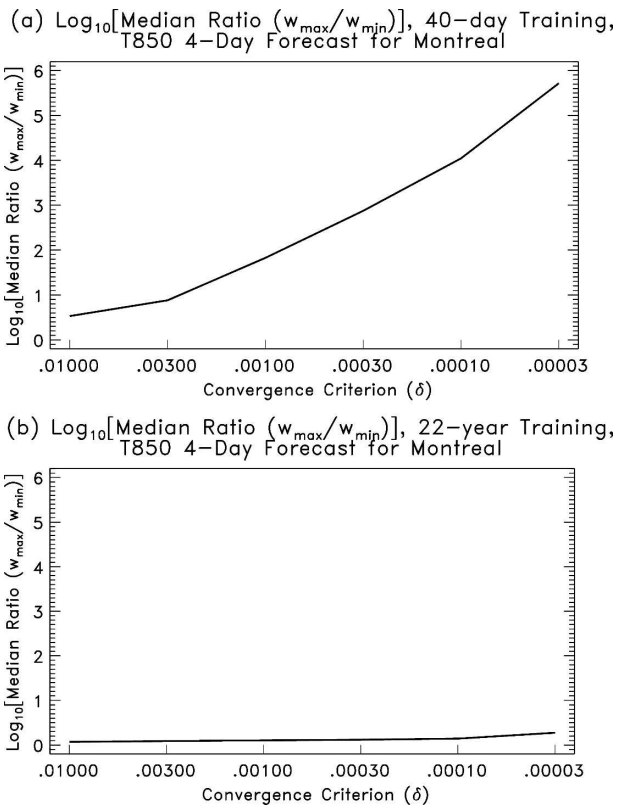

FIG. 3. Log_{10} of the median sample's maximum member weight divided by minimum member weight, as a function of the EM convergence criterion. The median represents the $(23 \times 365/2)$th rank-ordered ratio among the 23×365 sample days. (a) 40-day training period and (b) 22-yr cross-validated training period.

However, we can also track the fit to the validation data. Figures 4a,b show the average training and validation log likelihoods (per forecast day) for the small and large training data sizes. Notice that for the small sample size, the validation data log likelihood decreased as the convergence criterion was tightened, a sign that the unequal weights were not realistic. The same effect was hardly noticed with the large training dataset, where the weights remained nearly equal as the convergence criterion was tightened.[2] This demonstrates that the highly variable weights with the 40-day training were most likely an artifact of overfitting. Perhaps this was not surprising, given that the EM algorithm was expected to fit 15 parameters here (14 weights plus a standard deviation) with the 40 samples. Further, the effective sample size (Wilks 2006a, p. 144) may actually have been smaller than 40; perhaps the assumption of independence of forecast errors in space and time (R05, p. 1159) was badly violated with these ensemble forecasts. Also, we agree with the W07 proposition that the radical differences in weights may also be in part a consequence of the colinearity of members' errors in the training data. What is clear here is

[2] Figure 4b does display one oddity, namely, that the fit to the validation data is slightly closer than the fit to the training data. We expect that this small difference can be attributed to sampling variability.

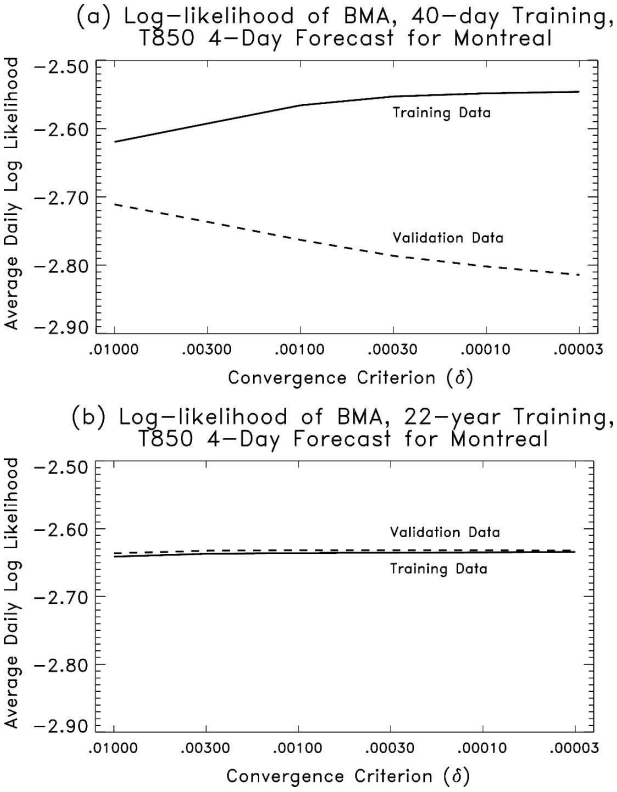

FIG. 4. Log likelihood (per unit day) of training and validation data as a function of the convergence criterion. (a) 40-day training data and (b) 22-yr cross-validated training data.

that this colinearity was not properly estimated from small samples, which led to the inappropriate deweighting and exclusion of information from some members.

When the BMA weights were enforced to be equal and 40-day training was used, the resulting continuous ranked probability skill score [CRPSS; calculated in the manner suggested in Hamill and Juras (2006) to avoid overestimating skill; 0.0 = the skill of climatology, 1.0 = perfect forecast] was 0.38. When the individual weights were allowed to be estimated by the EM and the convergence criterion was 0.000 03, the resulting CRPSS was smaller, 0.35. When the 22-yr training data were used, the CRPSS was 0.410, regardless of whether the weights were enforced to be equal or allowed to vary.

Is there a way of setting the BMA weights to avoid radically deweighting some members with small samples? If colinearity of member errors in the training data were essentially zero, then the weights would resemble those set in a weighted least squares process. Suppose the training data establish that the estimated root-mean-square errors for the bias-corrected members were $s_1, \ldots, s_n$. The weights that would have produced the minimum variance estimate of the mean state [e.g., Daley 1986, p. 36, Eq. (2.2.3)] under assumptions of normality of errors were

$$w_i = \frac{1}{s_i^2} \Big/ \sum_{j=1}^{n} \frac{1}{s_j^2}. \tag{2}$$

The advantage of this method for setting weights, also, was that if there truly was a strong colinearity of member errors, the BMA pdf should not have been worse as a consequence of using the more equal weights of Eq. (2) rather than the unequal weights from a highly iterated EM. This can be demonstrated simply by considering two member highly colinear forecasts with similar errors and biases, so $x_i^f \cong x_j^f$. Then the weighted sums are similar, regardless of the partitioning of the weights. For example,

$$\begin{aligned} 1.0 \times x_i^f + 0.0 \times x_j^f &\cong 0.0 \times x_i^f + 1.0 \times x_j^f \\ &\cong 0.5 \times x_i^f + 0.5 \times x_j^f. \end{aligned}$$

3. Conclusions

While the BMA technique is theoretically appealing for ensemble forecast calibration, the BMA and the EM technique cannot be expected to set realistic weights for each member when using a short training dataset. Enforcing more similar weights among BMA members [Eq. (2)] may work as well or better than allowing the EM method to estimate variable weights for each member.

REFERENCES

Daley, R., 1986: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.,* **39B,** 1–39.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

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

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

Wilks, D. S., 2006a: *Statistical Methods in the Atmospheric Sciences*. 2d ed. Academic Press, 627 pp.

——, 2006b: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.,* **13,** 243–256.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

Reply

LAURENCE J. WILSON

Meteorological Research Division, Environment Canada, Dorval, Québec, Canada

STÉPHANE BEAUREGARD

Canadian Meteorological Center, Meteorological Service of Canada, Dorval, Québec, Canada

ADRIAN E. RAFTERY

Department of Statistics, University of Washington, Seattle, Washington

RICHARD VERRET

Canadian Meteorological Center, Meteorological Service of Canada, Dorval, Québec, Canada

(Manuscript received 11 January 2007, in final form 28 February 2007)

1. Introduction

In his comments to our paper, Wilson et al. (2007, hereafter W07), Hamill (2007, hereafter H07) argues that our application of Bayesian model averaging (BMA) to short training samples leads to overfitting and represents an inappropriate use of the technique. He further contends that assignment of near-zero weight to a significant proportion of the ensemble members amounts to throwing out potentially useful information from the ensemble. To demonstrate his points, he has tested BMA in a fashion similar to the tests reported in W07, using an ensemble of 14 interchangeable members, which should a priori be expected to be weighted equally.

The additional tests we carried out, shown in this reply, do not indicate that assigning low weights to some poorer performing members in an ensemble of noninterchangeable members is an undesirable effect of integrating the expectation maximization (EM) algorithm to near convergence. Instead, it has the effect of finely tuning the probability density function (pdf) prediction intervals compared with the alternatives we tested at little or no cost to overall accuracy.

Corresponding author address: Laurence J. Wilson, Environment Canada, 2121 Transcanada Highway, 5th Floor, Dorval, QC H9P 1J3, Canada.
E-mail: lawrence.wilson@ec.gc.ca

DOI: 10.1175/2007MWR2138.1

Overfitting manifests itself by a good fit of the BMA pdf to training data and poor predictive performance on test data; forecast performance is the main criterion for determining whether there is overfitting. The conclusions in our paper were based on the predictive performance of BMA, not its fit to training data. The results were clear: BMA yielded probabilistic forecasts that were much better calibrated than the raw ensemble and performed better by a number of measures, including verification histograms and the continuous rank probability score (CRPS).

H07 does not really suggest an alternative to BMA but rather variant implementations of the method to alleviate what he sees as an overfitting problem. These are

1) using a different, typically more equal, set of weights, given by his Eq. (2);
2) using a reforecast dataset and much longer training set; and
3) stopping the EM algorithm early without iterating it to full convergence.

Before discussing the results of the additional experiments, we offer the following comments on H07. First, there are significant differences in the 40-day training samples used in H07 compared with the samples used in W07. The former are extracted from a low-resolution upper-air analysis, while the latter are station-specific surface observations. There is likely to be a higher serial correlation in the upper-air data than in the surface

values, perhaps much higher. Statistically, this means more degrees of freedom exist in a 40-day sample of W07 data than in 40 days of H07 data. H07 refers to this point; it is our view that the overfitting effect on 40-day samples would be stronger, perhaps significantly stronger, in the H07 results than in the W07 results, because of higher spatial and temporal correlation of errors.

Second, it must be remembered that the Canadian ensemble used in the W07 experiments is made up of noninterchangeable members, and therefore they cannot be assumed to be extracted from the same (unknown) distribution on each occasion. If an ensemble is constructed of interchangeable members as, for example, that used in H07, then there is no reason to apply BMA as if the members are separate. In that case, one should use the a priori knowledge and constrain the weights to be equal, using the BMA to estimate the standard deviation of the kernels. We tried this on the Canadian data, and the results are shown below.

Third, a clarification is needed regarding Fig. 4 in H07 and its interpretation. The difference in performance between independent samples and training samples depends on the differences in statistical characteristics between the samples. Figure 4 was constructed using two different analysis methods. For Fig. 4b, cross validation was used, a method that would tend to ensure close agreement between dependent and independent samples because each case of the development sample is used in turn as an independent case. Figure 4a was constructed using the same method as in W07, which was chosen partly on the basis of operational feasibility. In that case, the independent case immediately follows the training sample in temporal sequence. This would be expected to lead to a systematic difference in dependent and independent sample characteristics. Therefore, some portion of the difference shown in Fig. 4a of H07 is surely related to such systematic differences in samples (bias difference, e.g., as illustrated in W07 for the spring and autumn seasons) rather than overfitting. We agree with H07 that *changes* in the accuracy on independent data as a function of changes in the cutoff criterion may indicate overfitting, but the differences shown on the left-hand side of Fig. 4a for a suitably relaxed criterion are more likely due to systematic differences in the dependent and independent samples. We were able to compare some of the points of Fig. 4a of H07 using our data and found results that are consistent with the analysis discussed below: changes in log likelihoods as a function of cutoff criterion were smaller for both dependent and independent samples.

Fourth, the H07 suggestion to remove the bias by correcting the ensemble mean only is an interesting alternative to the two bias correction methods we assessed. In W07, we showed that correcting each member with a bivariate regression (denoted "FR" for "full regression" in W07) led to a decrease in the ensemble spread with forecast projection and is an undesirable property of that method, as mentioned by H07. For that reason, we preferred our other method, which was to correct only the mean error on the training period (obtained by setting the slope coefficient to 1, called "b1" in W07), again for each member. This method removes the bias but also corrects the variation between the means of the individual members and the mean observation for the training sample. Thus this method could also result in reduced ensemble spread for longer projections, although the reduction was much smaller than it was using the FR method. H07's suggestion is a third alternative, where only the ensemble mean bias is corrected, thus preserving the spread of the ensemble. This could be a preferred method, especially for ensembles of interchangeable members, but also might work for an ensemble of noninterchangeable members. This is not a feature of the BMA itself but might have an impact on the performance of the BMA. In the results presented below, we refer to this method as "MR" bias removal.

Last, the argument of H07 in favor of setting weights using Eq. (2), and the reference (Daley 1986), assumes independence of members, which is inconsistent with the description of the 14-member ensemble of H07, where it is pointed out that the members are not independent.

2. Further experiments using BMA

Inspired by Hamill's comments, we conducted some further experiments with BMA using the same dataset that was used in W07. In all tests, we used a 40-day training period and evaluated the results on independent data over the full year sample of 21 stations for 366 days. Thus, each result represents an average over approximately 7500 forecasts. For the first test, we compared the FR bias correction method with the MR method suggested in H07, using the 18-member ensemble consisting of the 16 members plus the unperturbed control forecast and the full-resolution global model forecast.

Figure 1 shows the results of this test in terms of the CRPS. There are five curves in the figure: the original uncorrected ensemble CRPS values, the CRPS for the FR-corrected ensemble, the CRPS for BMA-calibrated forecasts based on the FR-debiased ensembles, the CRPS for MR bias-corrected ensembles, and finally the CRPS for BMA-calibrated forecasts based on the MR

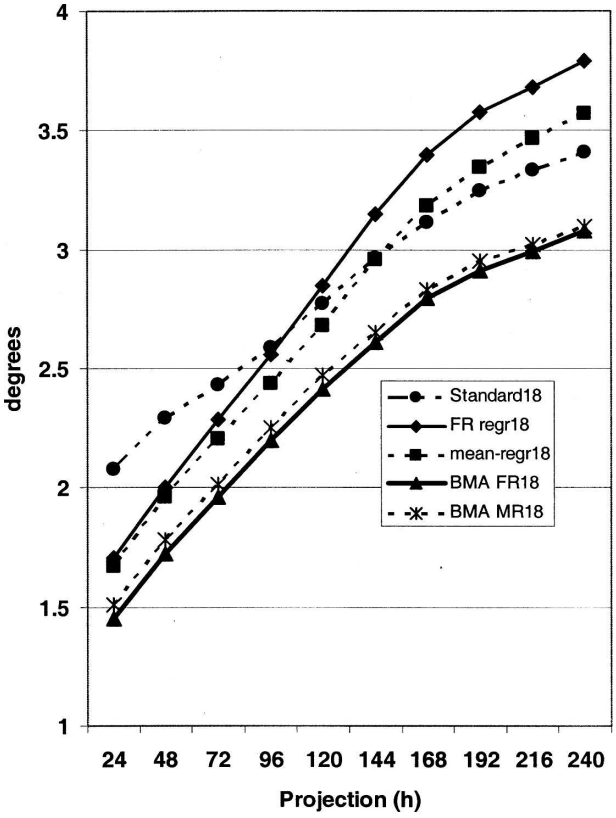

FIG. 1. CRPS as a function of projection time for two types of bias correction and for corresponding BMA-calibrated forecasts for 18-member ensembles. The curves are original ensemble (Standard18), FR bias-corrected ensemble (FR regr18), bias correction of mean only (mean-regr18), BMA for full regression bias corrected ensembles (BMA FR18), and BMA for ensembles with correction of only the ensemble mean (BMA MR18). See text for more details. Independent sample of approximately 7500 cases.

bias removal. The result for the FR bias-corrected ensembles is similar to the corresponding result for the 16-member ensembles in W07 (Fig. 8): the CRPS increases rapidly with increasing projection, reflecting the tendency toward a decrease in the ensemble spread at longer ranges. For the FR forecasts, the CRPS is higher (worse) than the original ensemble CRPS beyond day 4. The CRPS for the MR-corrected forecasts is about equal to that for FR bias removal at the shortest ranges but increases more slowly with projection time and improves on the original ensembles until day 5. An examination of some of the rank histograms (not shown) confirmed that the MR method exhibits a smaller tendency toward reducing the ensemble spread for longer projections, which is consistent with the better performance on the longer range forecasts. There was, however, still some tendency to enhance the underdispersion at the longest forecast ranges compared with the original ensembles; this behavior would warrant further investigation. While the MR bias correction performed better than the FR method, its performance deteriorates more rapidly with projection time than the b1 method described in W07, which would seem to be preferred for bias correction on ensembles of noninterchangeable members.

Figure 1 also shows that a BMA calibration following the original FR bias removal performs slightly better than a BMA calibration following the MR bias removal. This is opposite to what might be expected given the performance of the bias-corrected forecasts. Although the difference may not be significant, it seems clear from these results that the BMA can perform well even if fed a seriously underdispersed ensemble, consistent with the results shown in W07.

For the other three experiments, the bias was corrected using the b1 method described in W07. These three variants of BMA were as follows:

1) EX1: Running the BMA analysis with all coefficients constrained to be equal. We used the 18-member ensemble; thus the weights were all set to 1/18 in this test. The BMA analysis was limited to determining the standard deviation of the kernels from the errors in the training sample.
2) EX2: Running the BMA analysis with coefficients constrained to be equal for four subensembles, the eight members from the SEF model; the eight members from the GEM model; the control forecast; and the full-resolution model. The latter two are one-member subensembles. This is an illustration of the use of BMA for mixed ensembles that contain subensembles of interchangeable members.
3) EX3: Stopping the EM algorithm early. In the original tests in W07, we used a stopping criterion of 0.0001, which is a fairly restrictive value. In this test, we used 0.01, which corresponds to the left-hand side of Fig. 4 in H07. If overfitting is a significant problem, this radical change in stopping criterion would result in significantly different results.

The EX1 variant of BMA was implemented by modifying the EM algorithm as described in Raftery et al. (2005, p. 1159, hereafter R05) as follows. The ensemble weights are equal, and so $w_k = 1/K$. The expectation (E) step is still given by Eq. (6) of R05 but with the $w_k^{(j)}$ set equal to $1/K$. Of the two equations defining the maximization (M) step, the first is no longer necessary and the second is unchanged. The output of the EM algorithm is then just the maximum likelihood estimate of σ^2.

The EX2 variant of BMA was implemented as follows. Suppose that K ensemble members are partitioned into M subensembles such that the ensemble

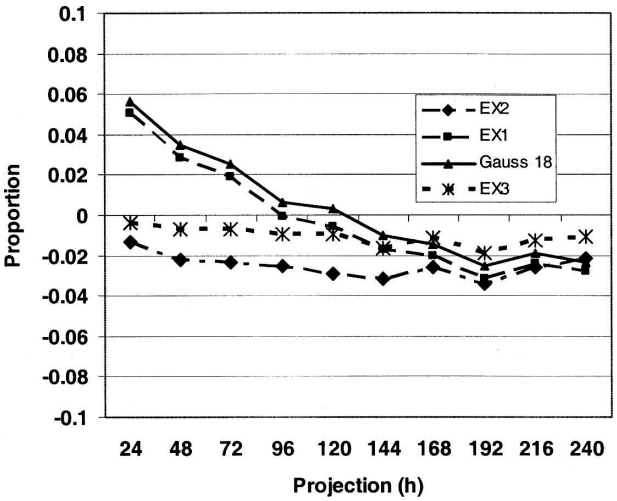

FIG. 2. CRPS results for the three BMA experiments EX1, EX2, and EX3 compared with results from the original analysis (W07). Positive indicates superior results for the original system.

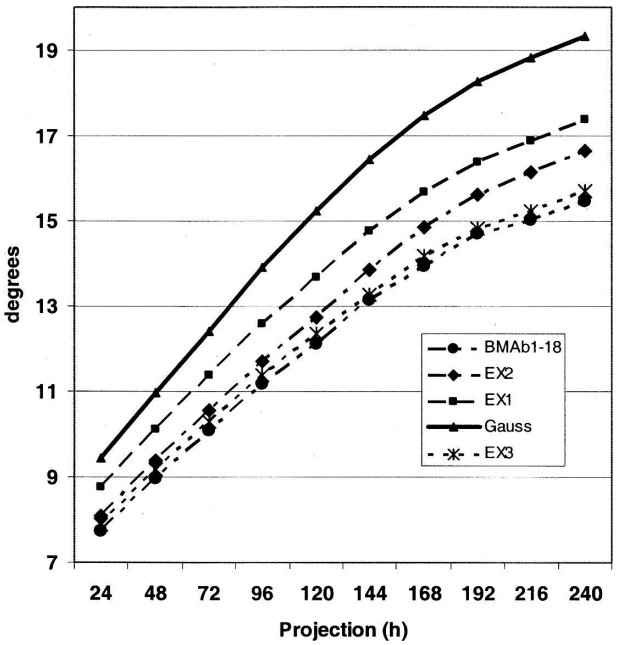

FIG. 3. The 90% prediction interval in degrees for further tests of BMA compared with the original 18-member ensemble results (W07) and the simple Gaussian (W07) as a function of forecast projection.

members in the same block are exchangeable. In the EX2 variant there are $M = 4$ subensembles as described above. Let $B(k)$ denote the subensemble to which the kth ensemble member belongs, so that $B(k) = m$ if the kth ensemble member is in the mth subensemble. Let N_m be the number of members in the mth subensemble. Thus $N_{B(k)}$ is the number of members in the subensemble to which the kth member belongs. Then the E step is unchanged and is still given by R05, their Eq. (6). The part of the M step that updates σ^2 is also unchanged. Only the part of the M step that updates the weights changes, as follows:

$$w_k^{(j)} = \frac{1}{N_{B(k)}} \sum_{\ell:B(\ell)=B(k)} \frac{1}{n} \sum_{s,t} \hat{z}_{\ell st}^{(j)}.$$

Note that in this equation the weights for ensemble members in the same subensemble will be the same throughout the EM algorithm.

We show results of these experiments in comparison to the corresponding original results reported in W07.

Figure 2 shows the CRPS difference between the three experiments and the original results for the 18-member ensembles. For reference, the differences with respect to the simple Gaussian described in section 5e of W07 are also shown. The CRPS differences are expressed as fractions of the original value of the CRPS and are plotted as (experiment value − original value) so that positive values indicate the original results score better. The first result to note is that all differences in CRPS are rather small, amounting to at most 5% of the original value, averaged over the approximately 7500 BMA analyses. Constraining all coefficients to be equal results in CRPS values marginally better than the simple Gaussian was able to produce but poorer than the original BMA results for the first 4 days. Treating all the members as equal takes away the ability of the BMA to reward higher-quality members with higher weights, which is especially important in the shorter ranges of the forecast. The improvement over the Gaussian results must be because the BMA distribution is not constrained to have a Gaussian shape, since the two estimates are otherwise similar.

The best performer according to Fig. 2 is EX2, where we have used the a priori knowledge of the makeup of the 18-member ensemble to constrain the weights to be equal for subensembles containing members that are expected to be most alike. For this experiment, the BMA needed to estimate only five coefficients using the 40-day training period rather than the 19 required for the original experiment. These results are slightly superior to the original BMA at all forecast projections, by amounts ranging from 2% to 4%. Stopping the EM algorithm early, before full convergence (EX3), also improved the CRPS on independent data slightly compared with the original results, but by a lesser amount than shown by EX2.

Figure 3 shows the 90% prediction interval in degrees, averaged over all the forecasts of the independent sample. As pointed out in W07, the BMA has reduced the width of this interval by as much as 25% compared with the simple Gaussian. Results for the

three additional experiments are intermediate; the original BMA still produces the sharpest prediction intervals, followed closely by EX3, then by EX2 and EX1.

Taken together, Figs. 2 and 3 suggest there is some overfitting because CRPS results on independent data are degraded a little compared with stopping the EM algorithm before full convergence. However, the impact is small and amounts to a choice of a slightly narrower prediction interval at small cost in terms of overall accuracy of the pdf. This is akin to the common trade-off between a smooth forecast, which scores well using quadratic scoring rules, and a sharper forecast, which might be more useful. The two figures also suggest that the option of stopping the EM algorithm before full convergence may be attractive: one can improve the CRPS modestly, while retaining almost all of the sharpness of the full integration. Also, stopping early saves computation time. The idea of stopping the EM algorithm early has been proposed previously by Vardi et al. (1985, p. 17), who suggested that "a limited number of iterations (our experience suggests about 50 iterations) gives very good [results]."

But are the radically different weights assigned to the different members of the ensemble due to running the EM algorithm to convergence rather than stopping early? To examine this, we compared weights from the original experiment in W07 (16-member ensembles this time) with weights obtained by stopping the EM algorithm at a 0.01 tolerance level. We have chosen to display the statistics of the weights in a different way from H07 (his Fig. 3). In H07, the median ratio of highest to lowest over all the BMA analyses is plotted as a function of the stopping criterion. Such a ratio gives undue importance to differences in the smallest weights: the ratio changes by an order of magnitude if the smallest weight changes from 0.01 to 0.001, but both are effectively zero when the weights are constrained to add to one for each analysis. We chose instead to construct histograms of all the weights, with bin widths equal to half-powers of 2, centered on 2^{-4}, the expected weight if all are equal for the 16-member ensemble. There are approximately 120 000 weights produced for our 7500 BMA analyses; we use an exponential ordinate to clarify the shape of the distribution. If the weights are equal, we would expect the histogram to show a single mode at the central bin.

Figure 4 shows the weight distribution using the two stopping criteria. Certainly, the relaxed criterion results in more coefficient values near the central value of 1/16. There is also a noticeable decrease in the number of coefficients in the highest and lowest bins. Nevertheless, the distribution for the relaxed cutoff indicates that

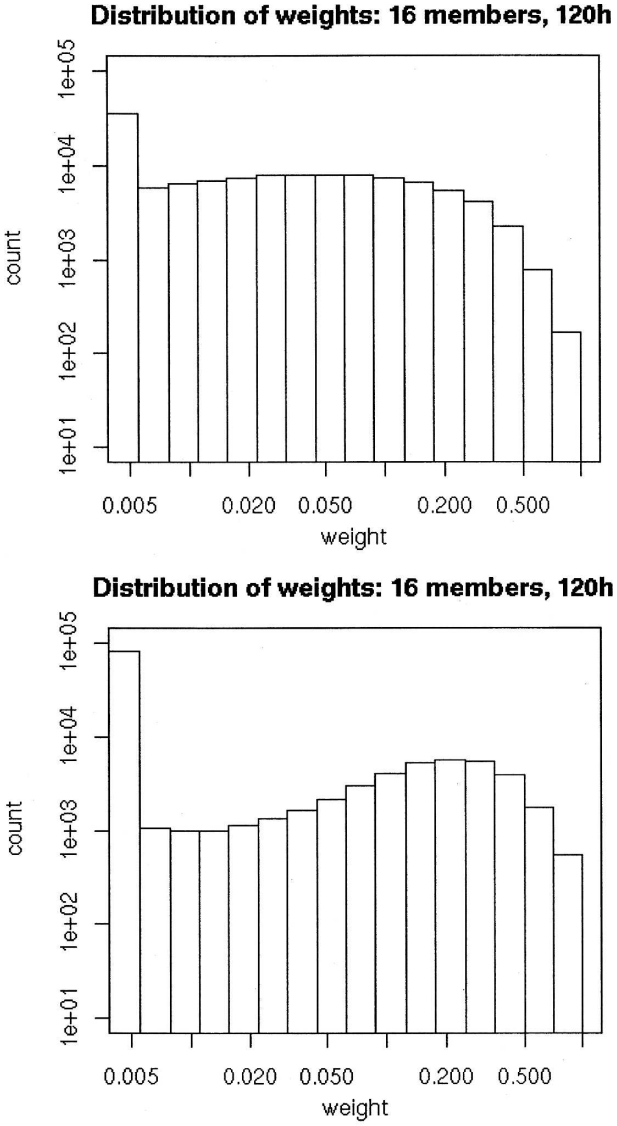

FIG. 4. Histograms of coefficients for the 16-member ensemble BMA, 120-h forecasts, 21 stations, 366 days, 40-day training period. EM cutoff criterion of (top) 0.01 and (bottom) 0.0001.

more than 25% of the weights remain in the lowest bin, suggesting that on average at least four members are given essentially zero weight in each analysis. These results do not support the contention in H07 that the unequal weights are the result of overfitting. However, Fig. 4 supports the evidence in Fig. 2 that there may be a small degree of overfitting in the original results. Of most concern is the right-hand bin, which identifies cases where one member was given most of the weight. Intuitively, this would suggest overconfidence or overtuning to a specific member. With the relaxed cutoff, the number of such cases is reduced from about 500 to 200, out of the total of 7500 analyses. This is perhaps a

small but desirable change and supports the use of a more relaxed cutoff criterion than we used in W07.

3. Discussion

Our use of BMA to calibrate Canadian ensemble forecasts based on recent performance statistics is a valid and useful application of the technique, as shown by these results. H07's primary contention is that BMA involves overfitting of the pdf to the training data. Overfitting manifests itself by a good fit to training data and poor predictions on independent cases. Our evaluations in the original paper, W07, and in this reply use independent samples, and these are good, indicating that overfitting was not an issue for BMA in terms of the probabilistic forecasts issued: it provided significantly improved, nearly perfectly calibrated, and sharp predictive distributions. We were not able to substantially improve the performance on independent data either by constraining some or all of the weights to be equal or by stopping the EM algorithm early. Based on these results, we do recommend careful attention to the cutoff criterion used with the EM algorithm, especially if training samples are small.

Our results also suggest that modest improvements can be obtained by constraining coefficients of the ensemble or of the subensembles to be equal if the corresponding members are interchangeable. For the Canadian ensemble, the most competitive results were obtained when the 18-member ensemble was treated as if it consisted of four separate subensembles. This also had the effect of reducing any overfitting, since only five parameters needed to be fit.

Of course, in statistical development, it is always desirable to have a large, homogeneous sample for training. Long reforecast datasets, such as those used in H07, represent an ideal that is unfortunately unachievable in practice, because of frequent changes to operational ensemble systems. Nevertheless, shorter reforecast datasets are planned in some centers, including our own. The addition of even 1 or 2 yr of data for calibration should improve the performance of BMA and permit its full potential to be realized. With a larger representative sample, the BMA can be extended to allow different values of the variance parameter σ^2 for different members, for example. The samples in W07 were too small to make effective use of this feature.

The results shown here also indicate that BMA is a flexible method, which can effectively calibrate and extract predictive information not only from ensembles of noninterchangeable members, such as the W07 application, but also from mixed ensembles and from ensembles of interchangeable members. The results obtained by considering the Canadian ensemble to be made up of four distinct members or subensembles are particularly interesting, because we obtained the best performance in terms of the CRPS for this configuration. This shows the potential for the use of BMA to calibrate mixed ensembles, such as those from the North American Ensemble Forecast System and The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble.

Finally, regarding the issue of "effectively discard-[ing] information from so many ensemble members" (H07) by assigning low weights to one or more members, the concern seems to be that a member that has been rejected on the basis of a relatively short training sample may well be the member that uniquely forecasts an extreme event the next time around. The BMA analysis tends to reject members that perform poorly during the training period and/or are highly correlated with better-performing members (see R05, 1161–1162). In this context, a member rejected by the BMA would be highly unlikely to suddenly correctly forecast an extreme event and be the only member to do so. The cost of retaining all the members in the forecast pdf with approximately equal weights is a probability distribution with larger prediction intervals, arguably a less useful pdf for forecast application.

Acknowledgments. The authors thank Tilmann Gneiting and Ken Mylne for stimulating and useful discussions during the preparation of this response.

REFERENCES

Daley, R., 1986: *Atmospheric Data Anlaysis.* Cambridge University Press, 457 pp.

Hamill, T. M., 2007: Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging." *Mon. Wea. Rev.,* **135,** 4226–4230.

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

Vardi, Y., L. A. Shepp, and L. Kaufman, 1985: A statistical model for positron emission tomography. *J. Amer. Stat. Assoc.,* **80,** 8–20.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.