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

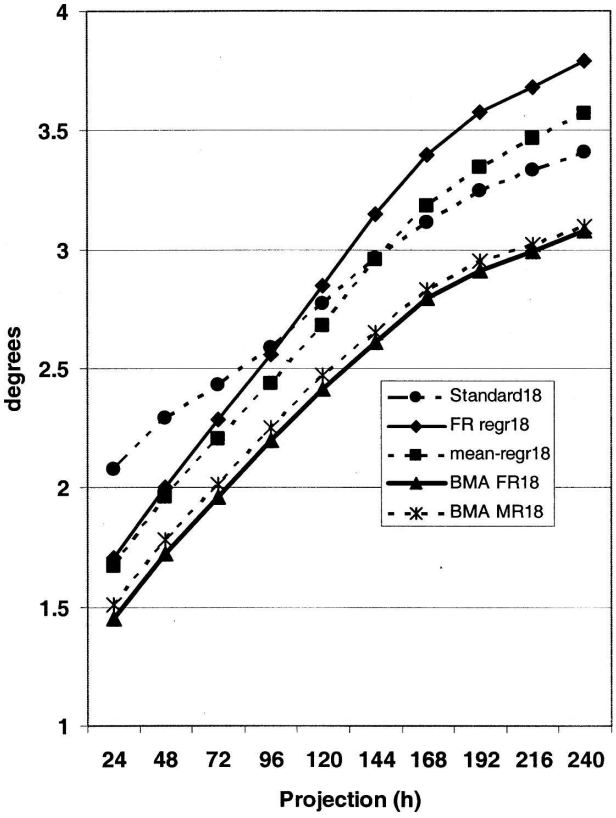

FIG. 1. CRPS as a function of projection time for two types of bias correction and for corresponding BMA-calibrated forecasts for 18-member ensembles. The curves are original ensemble (Standard18), FR bias-corrected ensemble (FR regr18), bias correction of mean only (mean-regr18), BMA for full regression 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

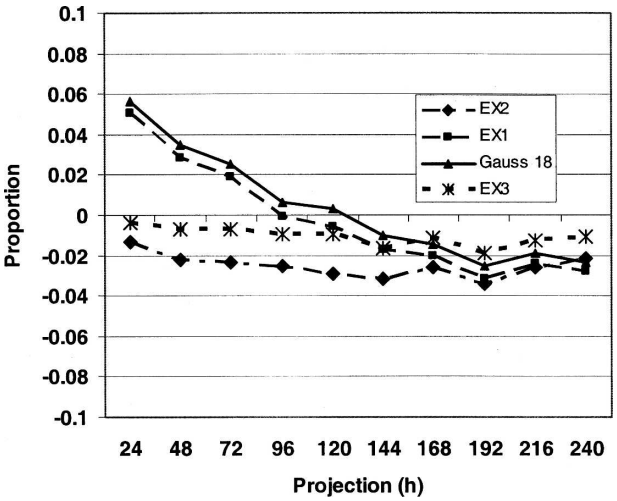

FIG. 2. CRPS results for the three BMA experiments EX1, EX2, and EX3 compared with results from the original analysis (W07). Positive indicates superior results for the original system.

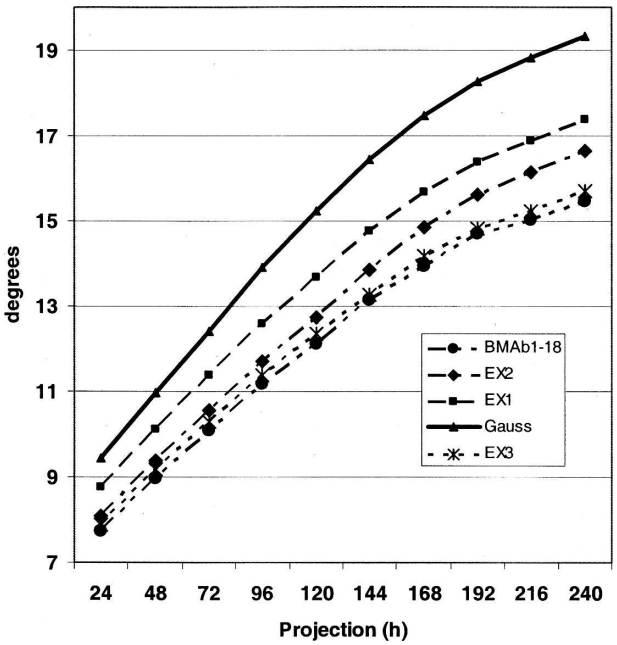

FIG. 3. The 90% prediction interval in degrees for further tests of BMA compared with the original 18-member ensemble results (W07) and the simple Gaussian (W07) as a function of forecast projection.

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

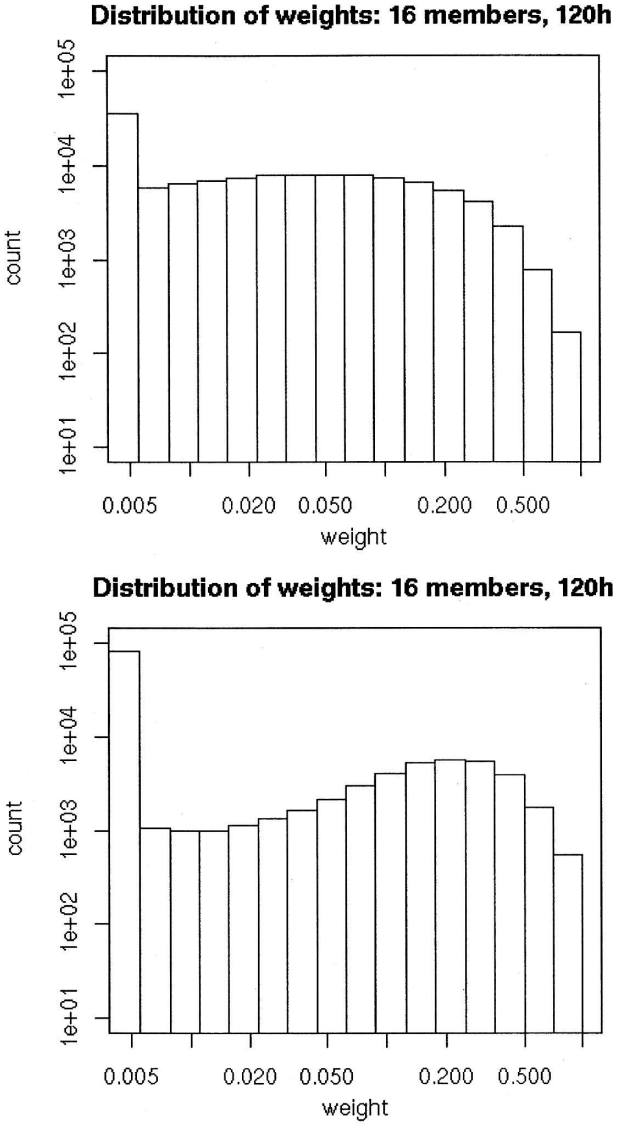

FIG. 4. Histograms of coefficients for the 16-member ensemble BMA, 120-h forecasts, 21 stations, 366 days, 40-day training period. EM cutoff criterion of (top) 0.01 and (bottom) 0.0001.

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.