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

Thomas M. Hamill

NOAA Earth System Research Laboratory, Physical Sciences Division

Boulder, Colorado

Submitted to *Monthly Weather Review*

17 October 2006

(revised)

Corresponding author address

Dr. Thomas M. Hamill
NOAA Earth System Research Laboratory
Physical Sciences Division
R/PSD 1, 325 Broadway
Boulder, CO 80305

1. **Introduction**.

Wilson et al. (2006, hereafter W06) 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 W06 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 Estimation-Maximization (EM) algorithm (Dempster et al. 1997). The final probability density function (pdf) was a sum of the weighted kernels.

W06 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 counter-intuitive; 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 W06 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 sample size is small[1]; specifically, the radically unequal weights of W06 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

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

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 W06, 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 W06's methodology, a reforecast data set was used. This was comprised of more than two decades of daily ensemble forecasts with perturbed initial conditions, all from a single forecast model. This large data set permitted a comparison of BMA properties based on small and large training samples. This reforecast data set used a T62, circa-1998 version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS). A 15-member forecast, consisting of a control and 7 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 data set, please see Hamill et al. (2006). The verification data was the NCEP/NCAR 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 parameter 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 $\delta << 1.0$.

To illustrate the tendency for the BMA EM to over-fit when trained with small sample sizes, consider 4-day 850 hPa temperature ensemble forecasts for a grid point near Montreal, Canada. Forecasts were produced and validated for the 23 years × 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 2004) 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 W06 to be an acceptably long training period. We shall refer to this as the "40-day training" data set. In addition, the BMA algorithm was also trained with a very long training data set in a cross-validated manner using 22 years × 91 days of data, with the 91 days centered on the Julian day of the forecast. This will be referred to as the "22-year training data set."

The BMA algorithm was coded generally following the algorithm used in the R05 and W06 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 (CRPS; Hersbach 2000), as suggested in R05. Performing the refinement increased the computational expense but had minimal impact on forecast skill. Second, W06'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 $\overline{x}^f$, and a regression-corrected ensemble-mean forecast $\left(a + b\overline{x}^f\right)$, the member forecast was replaced with a forecast that was the sum of the initial perturbation from the ensemble mean and the corrected forecast:

$$x_i^f \leftarrow \left(x_i^f - \overline{x}^f\right) + \left(a + b\overline{x}^f\right) \quad , \tag{1}$$

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 co-linearity 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 non-uniform weights produced by the BMA were a generic property of a short training data set or whether they were artificially induced from the increased co-linearity induced by the regression analyses.

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 $\delta << 0.01$ (Fig. 3a). With the 22-years 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. However, we can also track the fit to the validation data. Figures 4 a-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 data set, 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 wasn't surprising, given that the E-M 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 W06 proposition that the radical differences in weights may also be in part a consequence of the co-linearity of members' errors in the training data. What is clear here is that this co-linearity was not properly estimated from small samples, which led to the inappropriate de-weighting 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 over-estimating skill; 0.0 = the skill of

[2] Fig. 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.

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 0.00003, the resulting CRPSS was smaller, 0.35. When the 22-year training data was 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 de-weighting some members with small samples? If co-linearity 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 establishes 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, p. 36, eq. 2.2.3) under assumptions of normality of errors was

$$w_i = \frac{1}{s_i^2} \Bigg/ \sum_{j=1}^{n} \frac{1}{s_j^2} \quad . \qquad (2)$$

The advantage of this method for setting weights, also, was that if there truly was a strong co-linearity 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 co-linear 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, $1.0\times x_i^f + 0.0\times x_j^f \cong 0.0\times x_i^f + 1.0\times x_j^f \cong 0.5\times x_i^f + 0.5\times x_j^f$.

4. **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 data set. 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 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., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.

----------, and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, in press. Available at http://www.cdc.noaa.gov/people/tom.hamill/skill_overforecast_QJ_v2.pdf

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**, 437-472.

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, 2004: 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* (2^{nd} Edition). Academic Press, 627 pp.

--------------- , 2006b: Comparison of ensemble-MOS methods in the Lorenz ’96 setting. *Meteor. Apps*., in press. Available from dsw5@cornell.edu.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2006: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging. *Mon. Wea. Rev*., in press. Available from lawrence.wilson@ec.gc.ca

FIGURE CAPTIONS

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

Figure 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.

Figure 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*365/2)th rank-ordered ratio among the 23*365 sample days. (a) 40-day training period, (b) 22-year cross-validated training period.

Figure 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-year cross-validated training data.

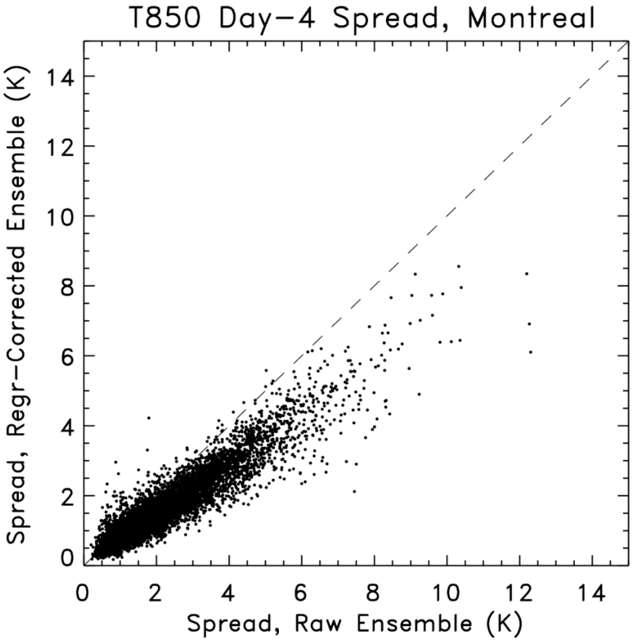

Figure 1: Spread of a regression-corrected ensemble of Day-4 forecasts of 850 hPa temperature at Montreal, CA (using a 40-day training data) vs. the spread of the raw ensemble forecasts.

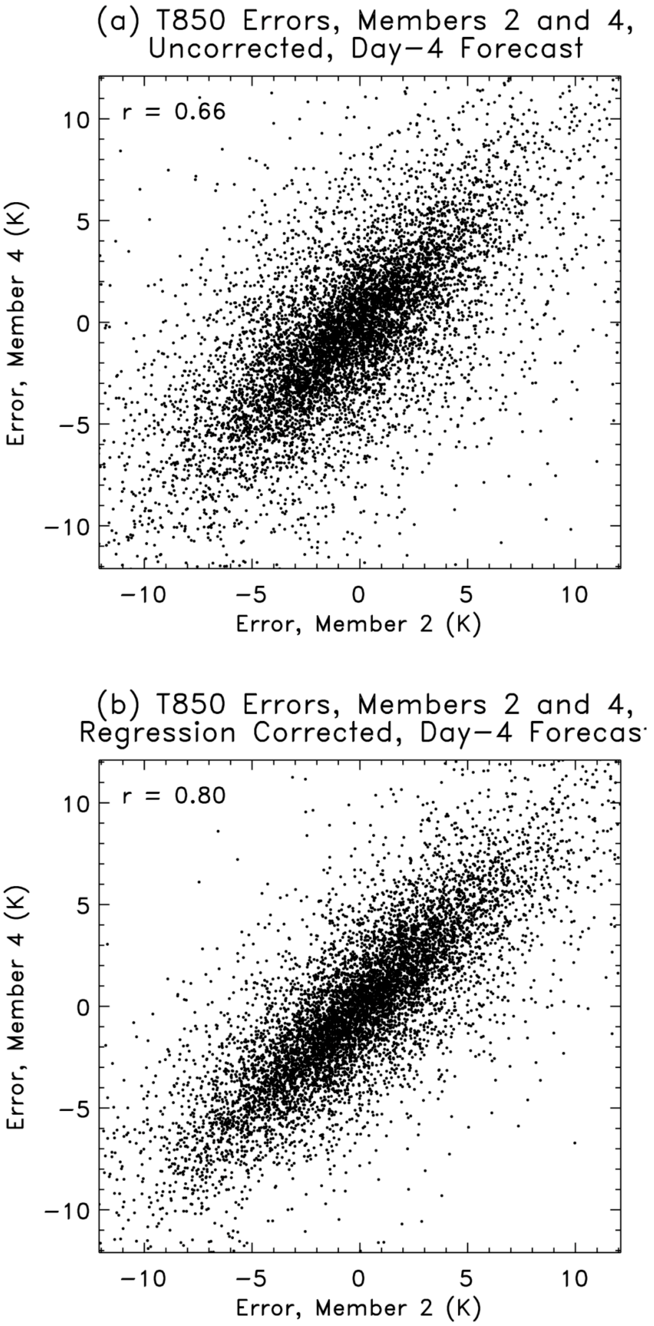

Figure 2: Errors of Day-4 850 hPa temperature forecasts for members 2 and 4 of the ensemble (a) before a regression correction of the member errors using the prior 40 days for training, and (b) after the regression correction. Correlation coefficient (r) noted in the upper-left corner.

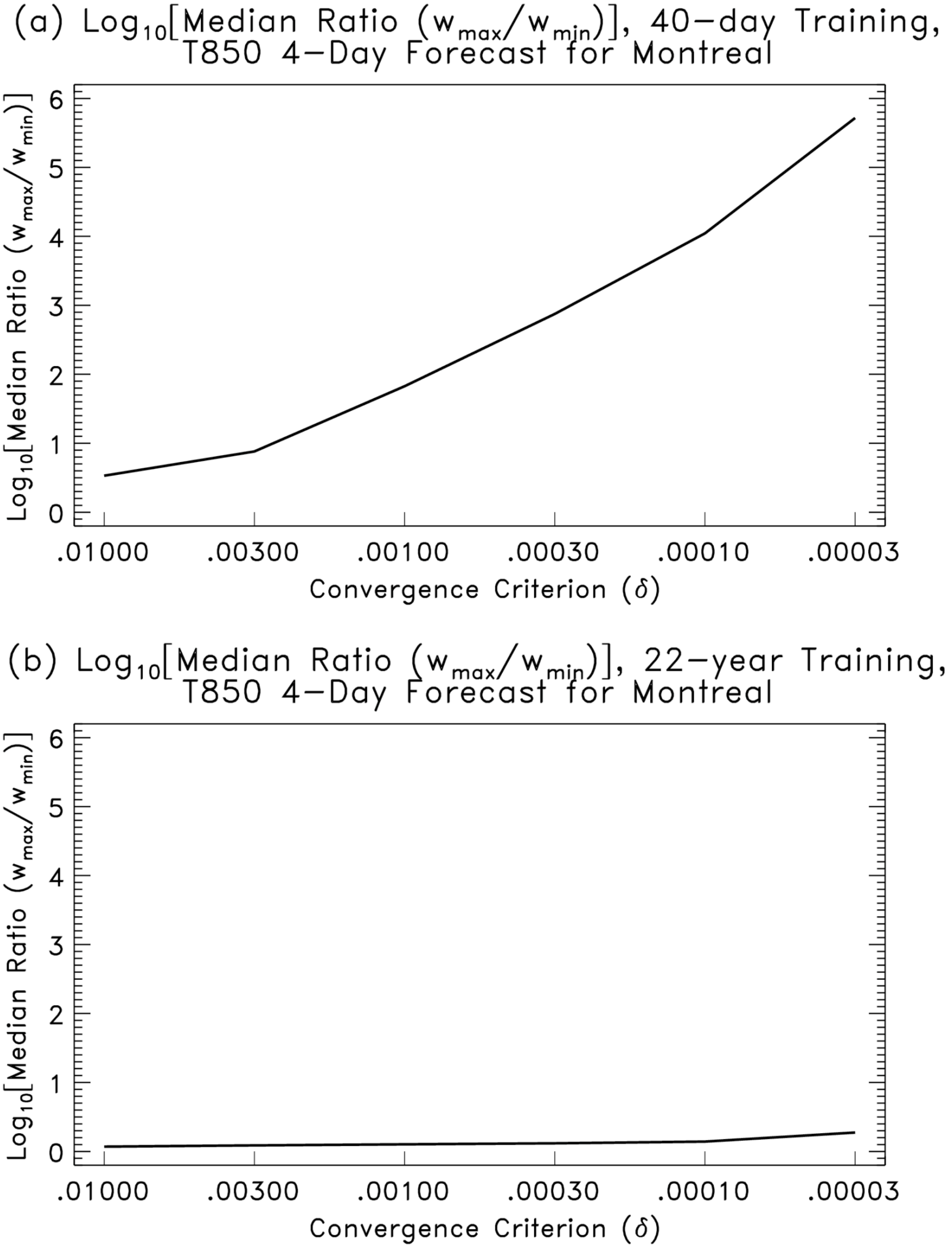

Figure 3: Log$_{10}$ of the median sample's maximum member weight divided by minimum member weight, as a function of the EM convergence criterion. The median represents the (23*365/2)th rank-ordered ratio among the 23*365 sample days. (a) 40-day training period, (b) 22-year cross-validated training period.

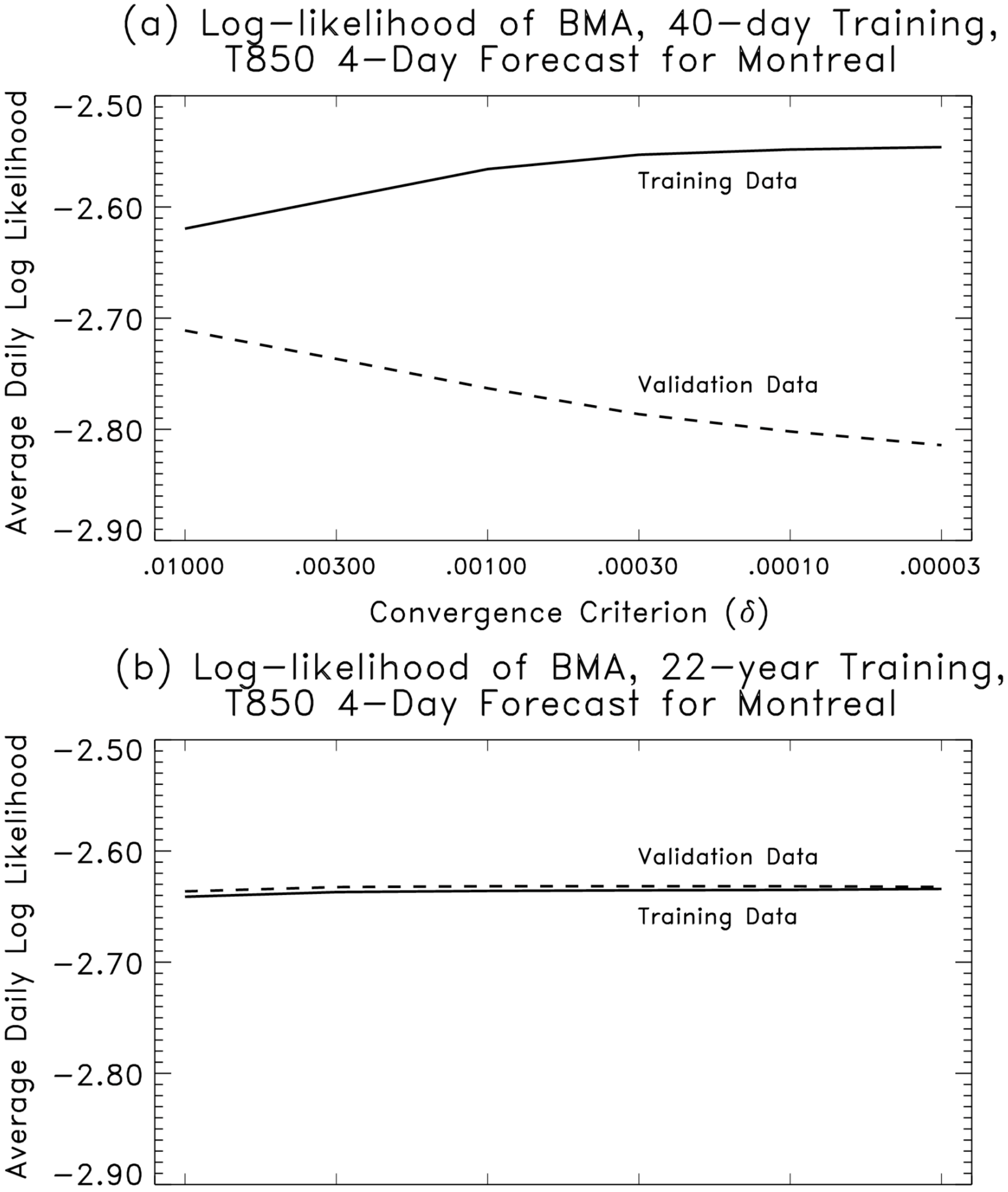

Figure 4: Log likelihood (per unit day) of training and validation data as a function of the convergence criterion. (a) 40-day training data, and (b) 22-year cross-validated training data.