Evaluation of the "Bayesian Processor of Forecasts" Algorithm Using GFS Surface-Temperature Reforecasts

Thomas M. Hamill

NOAA Earth System Research Laboratory, Boulder, Colorado

Submitted to *Monthly Weather Review*

12 December 2008

Corresponding author address:

Dr. Thomas M. Hamill
NOAA Earth System Research Laboratory
R/PSD1, 325 Broadway
Boulder, CO USA 80305-3328
tom.hamill@noaa.gov
Phone: (303) 497-3060 ; Fax (303) 497-6449

ABSTRACT

A critical review of the Bayesian Processor of Forecasts (BPF) algorithm for probabilistic weather forecast calibration is provided. The BPF algorithm of Krszysztofowicz and Evans (2008) produces a probabilistic forecast by implementing Bayes rule, updating a prior probabilistic estimate to a new forecast datum, producing a "posterior" probabilistic estimate. How the forecast datum is used is controlled by the relationship between forecasts and observations in a training data set, the "likelihood" function. Several algorithmic complexities are introduced in the BPF algorithm to transform the problem to where data is distributed normally and an analytic solution can be applied. Since it leverages prior climatological information, the BPF algorithm may be appealing for its potentially increased skill when using small training data sets.

The BPF is compared against competing probabilistic forecast methods for surface temperatures using an ensemble reforecast data set. Forecasts are evaluated in the fall of 2005 at a set of observation locations in North America. Two major deficiencies of the BPF are noted. First, forecast and observed are transformed independently to Gaussian distributions under the BPF algorithm. Linear relationships between the forecast and observed are assumedly preserved after the transformation. In fact they are sometimes lost, resulting in a poor likelihood function model and the occasional pathologically poor forecast. BPF forecast skill can be improved substantially by requiring the same transformation for the forecast data as was used for the observations, preserving linear relationships between forecast and observations in the original training data. Second, if the prior estimate is inaccurate and biased, the posterior will be inaccurate and biased, especially at long leads where forecast-observation correlations are small. With this

particular test data set the prior was strongly biased, since the fall of 2005 was much warmer than the climatological estimates from 1980-2004 data.

1. Introduction

Recently, Krzysztofowicz and Evans (2008; hereafter KE08) proposed a novel method for the production and calibration of probabilistic weather forecasts that they called the "Bayesian Processor of Forecasts," or "BPF". This method is based upon the well-known Bayes Rule

$$\phi(w|x) = \frac{f(x|w)g(w)}{\int f(x|w)g(w)\,dw} \quad . \tag{1}$$

Here $\phi(w|x)$ is the "posterior" probability density, *w* refers to a possible value for the observed, and *x* to the forecast. $f(x|w)$ is called the "likelihood function" and describes the conditional probability density of the forecast given an observed value. $g(w)$ is the "prior" probability density, typically estimated from independent information such as a long-term observation climatology. Commonly, pairs of consistent forecasts and observations may be available only for the recent past, perhaps a month to a year or two, whereas the observations alone may be available for a much longer period of time. Hence, rather than attempting to calibrate the forecast solely based on the recent forecast-observation pairs, KE08 propose that a more accurate and stable method may be to update a climatological prior determined from a long time series of observations with a likelihood based on the smaller set of recent forecasts and observations.

Such a method is potentially very attractive to operational forecast centers; recent articles (e.g., Hamill et al. 2006; Hamill and Whitaker 2006; Wilks and Hamill 2007; Hagedorn et al. 2008; Hamill et al. 2008) have discussed the improvements in probabilistic forecasts that can be obtained through calibration using a long, stable training data set of "reforecasts." Incorporating reforecasts into operations has drawbacks for numerical weather prediction (NWP) centers, however. If the reforecast data set is pre-computed, then the operational center must continue to run the model version used to generate the reforecast to ensure consistency between training and real-time forecasts. Rapid changes to the forecast model are thus less desirable, as changes require the expensive computation of a new reforecast data set. Alternatively, reforecasts could be computed in real time, using whatever version of the forecast model is used operationally. Such an approach is currently used at ECMWF (Hagedorn, 2008). However, this approach uses computer time that might otherwise be allocated to increasing model resolution or ensemble size. A calibration method that is able to produce good results with small training data sets is thus very attractive.

While KE08 provided a proof-of-concept for the BPF algorithm, the method should be more rigorously tested and compared to more established methods in order to provide a better understanding of its comparative strengths and weaknesses. Have KE08 provided an algorithm that will obviate the need for lengthy reforecast training data sets? Accordingly, this article will test and critically examine the hypothesis that the BPF is more skillful than existing calibration techniques, especially for small training data sets. This experiment will use surface-temperature data from the same Global Forecast System (GFS) reforecast data set used in the prior experiments cited above. Forecasts will be validated over a set of stations in North America for 1 September to 1 December 2005.

Below, section 2 will provide a brief review of the BPF algorithm, highlighting a few important modifications to the algorithm described in KE08. Section 3 discusses the implementation of the BPF and competing methods, and the supporting data sets. Section 4 provides results, and section 5 provides a discussion and conclusions.

2. The Bayesian Processing of Forecasts algorithm.

a. The simplest case: Gaussian prior and likelihood.

Evaluation of (1) is relatively simple when the prior and likelihood are normally distributed; then the posterior is normally distributed as well and (1) can be evaluated analytically. Suppose the prior is $g(w) \sim N(\mu_w, \sigma_w^2)$, a Gaussian (normal) distribution with mean μ_w and variance σ_w^2. Similarly, assume the likelihood is $f(x|w) \sim N(\mu_x, \sigma_x^2)$. Then the posterior distribution is

$$\phi(w|x) \sim N\left(\mu_x \frac{\sigma_w^2}{\sigma_x^2 + \sigma_w^2} + \mu_w \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2}, \left(\sigma_x^{-2} + \sigma_w^{-2}\right)^{-1}\right) \quad . \tag{2}$$

The mean of the posterior probability density function (pdf) is a weighted sum of the prior and likelihood means, with the weights determined by the respective variances (Daley 1991, section 2.2).

A more realistic model, however, would allow the likelihood to be developed from a regression relationship between past forecasts and observations. Suppose

$$f(x|w) \sim N(a_x w + b_x, \sigma_x^2), \tag{3}$$

where a_x and b_x are regression coefficients and σ_x^2 is the regression model error variance. Then

$$\phi\left(w|x\right) \sim N\left(Ax+B,T^2\right) \tag{4}$$

where

$$\begin{aligned} A &= \frac{a_x\sigma_w^2}{\sigma_x^2+a_x^2\sigma_w^2} \\ B &= \frac{\mu_w\sigma_x^2-a_xb_x\sigma_w^2}{\sigma_x^2+a_x^2\sigma_w^2} \\ T^2 &= \frac{\sigma_x^2\sigma_w^2}{\sigma_x^2+a_x^2\sigma_w^2} \end{aligned} \quad . \tag{5}$$

This is outlined in Krzysztofowicz (1987) and employs the theory of the conjugate family of distributions (Degroot 1970). These equations are very similar in form to those presented in eq. (24) of KE08, though in that article the climatological prior was transformed to a standard-normal distribution, ~ $N(0,1)$, with mean 0 and standard deviation 1.

An example of a Gaussian update is illustrated in Fig. 1. The black curve denotes the prior. The small sample of joint forecasts and observations (black dots) are used to develop a regression relationship (3) to predict the probability density for the forecast given the observations. The parallel red lines show the regression relationship and the regression plus and minus 1 and 2 σ_x. Equation (3) can then be used to evaluate the likelihood function across the range of *w*, here shown when the forecast is +20. Eqs. (4) – (5) are then used to evaluate the posterior distribution, shown in the dashed blue curve.

Results of an array of simple tests are shown in Fig. 2 that illustrate the potential of the BPF method. Here, the climatological event probabilities were estimated from a normal distribution whose parameters were estimated from 10,000 independent samples drawn from a ~ N(0,1) distribution. Next, the training sample size and forecast-observed correlation were specified. Independent paired batches of forecasts and observations of this sample size were drawn from a standard normal distribution with the specified forecast-observed correlation. Each time a batch of paired samples of a given sample size was drawn, a linear regression model was developed that estimated a_x, b_x, and σ_x^2 of (3). A synthetic real-time forecast and verifying observation were generated, drawn again from a standard normal distribution with the same specified forecast-observation correlation. The BPF algorithm of (4) – (5) was then applied, and the subsequent probabilistic forecast was scored with the continuous ranked probability score (*CRPS*; Wilks 2006, section 7.5), defined as

$$CRPS = \int_{-\infty}^{\infty} \left[F(x) - F^o(x) \right]^2 dx \qquad (6)$$

where $F(x)$ denotes the cumulative density function of the forecast at the value x and $F^o(x)$ denotes the cumulative density of the observed, which is 1.0 for values greater than or equal to the observed and 0.0 for values less than the observed.

The *CRPS* of the climatological prior was evaluated as well. This process was repeated 40,000 times, and finally, the continuous ranked probability skill score (*CRPSS*) was evaluated, where

$$CRPSS = 1. - \frac{\overline{CRPS(\text{BPF})}}{\overline{CRPS(\text{climatology})}}, \qquad (7)$$

and the overbar indicates the arithmetic average over the 40,000 cases. This whole process was repeated for training sample sizes of 5, 10, 30, 60, 120, 240, 480, and 960. Forecast-observation correlations were tested in a range between 0.25 and 0.99.

For a standard of comparison, the *CRPSS* was also determined for a linear regression model based only upon the forecast-observation training data, with no knowledge of the long-term observed climatology. The regression equation predicted the expected observation value given the forecast value; a probabilistic forecast could be made implicitly from this by producing a pdf centered on the predicted observation, with its variance determined from the regression analysis.

Figure 2a shows the *CRPSS* of the BPF forecasts as functions of sample size and forecast-observed correlation. For small sample sizes and low correlations, there is little skill. BPF skill increases quickly as the forecast-observed correlation increases. Skill appears to saturate at ~ 50 -100 samples. Linear regression (Fig. 2b) shows generally similar characteristics, though the comparison of the two (Fig. 2c) shows that the skill difference, BPF minus the linear regression, is uniformly positive and especially large when the correlation and/or sample sizes are small. Together, Fig. 2 illustrates the potential of BPF, especially with little training data and marginal forecast quality. However, it should be noted that this experiment has stripped almost all of the complexities of the weather forecast application from the problem. In reality, distributions are likely to be non-stationary and non-Gaussian; samples may have auto-correlated errors and are subject to slowly varying, state-dependent model bias.

Son et al. (2008) have also performed an examination of BPF for bias correction with synthetic data under somewhat more realistic conditions, including effects such as the temporal autocorrelation of data. They similarly have found the competitive superiority of BPF and the ability for the algorithm to achieve the same results with small training data sets as other methods achieve with large training data sets. Like the test above, their experimental design assumed errors were Gaussian.

b. BPF when distributions are non-Gaussianl.

When the prior or likelihood is non-Gaussian, a direct analytical solution is often impossible. Several alternatives may then be considered. One possibility is a "brute-force" approach; the prior and likelihood functions might, for example, be estimated using non-parametric techniques such as mixture models (Hastie et al. 2001, section 6.8) and the posterior evaluated from prior and likelihood estimates at a finely discretized set of grid point values. Another possibility is to use the BPF framework but assume the distributions are Gaussian, which may be an acceptable compromise if departures from Gaussianity are moderate.

KE08 follow a different approach. Generally following their notational conventions and definitions, let W denote a random observation variable that takes on the specific value w. Let X denote a random forecast variable taking on the specific value x. The objective of KE08 is to permit the general use of (2) – (5) across a wide range of situations, where W and/or X may have distribution functions of any form, and the relationships between X and W need not be linear and homoscedastic, as they are assumed in regression analysis. They propose, then, a "*Bayesian meta-Gaussian model*" whereby the prior distribution and the distribution of forecasts in the training data set are modeled with parametric distributions, presumably non-Gaussian, but then transformations are applied so that the transformed data is approximately Gaussian, and relationships between X and W become remain linear and become more homoscedastic.

To make this algorithm more concrete, let P denote the probability. The posterior cumulative distribution function (CDF) is defined as

$$\Phi(w|x) = \int_{-\infty}^{w} \phi(u|x)\,du = P(W \leq w|X = x) \quad . \tag{8}$$

Define the inverse distribution function $\Phi^{-1}(\bullet|x)$, also known as the "posterior quantile function," such that for any number p, $0 < p < 1$ and a deterministic forecast $X = x$, the posterior quantile of the predictand W is the quantity w_p such that $\Phi(w_p|x) = p$. Given this, define

$$w_p = \Phi^{-1}(p|x) \ . \tag{9}$$

Let $K(X)$ denote the CDF for the forecast, and $G(W)$ the CDF for the observation. At the heart of KE08, then, is the "*normal quantile transform*" (NQT):

$$V = Q^{-1}\left[G(W)\right], \tag{10}$$

and

$$Z = Q^{-1}\left[K(X)\right], \tag{11}$$

where Q is the standard normal CDF and Q^{-1} its inverse. V is a random variable whose particular instance is v, and Z is a random variable whose particular instance is z. In the space of the transformed deviates (Z,V), each distribution is then Gaussian (KE08).

The overall strategy of KE08 is thus as follows: (a) collect a long time series of observations. Determine the parameters of the CDF $G(W)$ that best fits the observations for this time of the year. (b) Collect a shorter time series of joint forecast-observation pairs, which will be called the "training data." (c) Allowing for the possibility that forecasts may have a different CDF than observations, determine a parametric best-fit CDF to the forecast training data, $K(X)$. (d) Using (10) and (11), transform the training data (x,w) to standard normal deviates (z, ν). (e) Perform a linear regression analysis using the transformed training data to determine a, b, and σ^2 such that

$$E\left(Z\middle|V=\nu\right)=a\nu+b \quad , \tag{12a}$$

and

$$Var\left(Z\middle|V=\nu\right)=\sigma^2 \quad , \tag{12b}$$

where E(•) denotes the expected value and Var(•) denote the predicted variance. (f) Apply (5) to determine A, B, and T^2, remembering that since the observations have been transformed to a standard normal distribution, $\mu_w \equiv 0$ and $\sigma_w^2 \equiv 1$. (g) Given today's forecast value x, predict the posterior's cumulative distribution, quantile, and density functions. Following KE08, these can be written compactly as

$$\Phi(w|x)=Q\left(\frac{1}{T}\left[Q^{-1}\left(G(w)\right)-AQ^{-1}\left(K(x)\right)-B\right]\right), \tag{13}$$

$$w_p=G^{-1}\left(Q\left(AQ^{-1}\left(K(x)\right)+B+TQ^{-1}(p)\right)\right), \tag{14}$$

and

$$\phi(w|x) = \frac{1}{T}\exp\left(\frac{1}{2}\left\{\left[Q^{-1}\left(G(w)\right)\right]^2 - \left[Q^{-1}\left(\Phi(w|x)\right)\right]^2\right\}\right)g(w) \quad . \qquad (15)$$

This is the basic algorithm that is to be examined in this article. KE08 provide much more detail, describing how they used the Weibull distribution, for example, to fit the parameters of $K(X)$ and $G(W)$.

3. Experimental configuration

The efficacy of BPF will be tested using 2-meter temperature observations predicted using ensemble-mean forecasts and will be compared against other probabilistic forecast methods.

a. Observations

0000 UTC and 1200 UTC 2-meter temperature observations were extracted over much of North America from the National Center for Atmospheric Research (NCAR) data set DS472.0, the same observational data set used in Hagedorn et al. (2008). Additionally, only the stations that had 96 percent or more of the observations present over the 20-year period were considered. A plot of these station locations is provided in Fig. 1 of Hagedorn et al. (2008).

b. The GFS reforecast data set.

The GFS reforecast data set utilized a T62, 28 sigma-level, circa-1998 version of the National Centers for Environmental Prediction (NCEP) GFS (Hamill et al. 2006). Fifteen-member forecasts were available to 15 days lead for every day from 1979 to current. Ensemble forecasts were started each day from 0000 UTC initial conditions, and forecast information was archived on a 2.5-degree global grid. GFS lowest-sigma level forecast data was also bi-linearly interpolated to surface-observation locations. Forecasts are tested for leads of 0.5 to 10.0 days.

In this experiment, forecasts were evaluated whose initial conditions ranged from 0000 UTC 1 September to 0000 UTC 1 December 2005. Two training data sets were considered; the first was a small training data set comprised of the previous 30 days of forecasts, as in Hagedorn et al. (2008). The second was a 26-year (1979-2004) × 31 day sample of reforecasts, +/− 15 days around the date of interest. For the 30-day training data set, an actual forecast was made and evaluated if and only if at least 20 of the days had available forecasts and observations. For the large reforecast data set, forecasts were evaluated only if at least 75 percent of the forecast and observations were available.

c. Fitting a parametric distribution to the data samples.

The Weibull distribution used by KE08 and the related Generalized Extreme Value (GEV) distribution (Wilks 2006; section 4.4.5) exhibit an undesirable characteristic that may constrain their usefulness for general probabilistic forecast calibration of temperatures; the distributions do not provide unbounded support. That is, the random Weibull or GEV variable with parameters fit to the observations will have nonzero probability only above or below some threshold η (see appendix, KE08). Commonly, it was found that the parameters for the Weibull or GEV distribution that were chosen as the best overall fit to the sample of observations would assign zero probability assigned in one tail where there were actual observations (Fig. 3a). The consequences of this are illustrated in Fig. 4. Since Bayes rule can be evaluated from the normalized product of the prior and likelihood functions, if the prior or likelihood have zero probability, the posterior will have zero probability as well. However cold today's forecast may be, if the fitted prior distribution modeled zero probability for the prior at this temperature, the posterior would be assigned zero probability at that temperature. Hence, for the extreme events that are most important, the BPF using a Weibull or GEV distribution would produce a poor forecast.

Another class of parametric distributions was sought that provided unbounded support. After experimentation, it was found that the observation distribution could be adequately fit across a range of climatologies with standard-normal distribution that had been rescaled and/or power transformed. The algorithm used hereafter to determine a parametric distribution for the observations was as follows:

(1) For a batch of climatological data, determine that batch's mean μ_1 and standard deviation σ_1. The batch was then standardized so it had zero mean and unit standard deviation by subtraction of the mean from each sample and division by the standard deviation.

(2) A range of power transformations and further standardizations were tested. The power transformations that were tested were the set

$$\lambda = [\ 0.25, 0.4, 0.6, 0.8, 0.9, 1.0, 1.1, 1.2, 1.4, 1.6, 1.8, 2.0, 3.0\ ].$$

For a sample *y* from the standardized batch from step (1), the transformed sample ψ , following Yeo and Johnson (2000), was

$$\psi(\lambda,y)=\begin{cases}\left\{(y+1)^{\lambda}-1\right\}/\lambda & (y\geq 0,\lambda\neq 0)\\ \log(y+1) & (y\geq 0,\lambda=0)\\ -\left\{(1-y)^{2-\lambda}-1\right\}/(2-\lambda) & (y<0,\lambda\neq 2)\\ -\log(y+1) & (y<0,\lambda=2)\end{cases} \quad . \tag{16}$$

This class of power transformations was useful in that the transformation accepted both positive and negative values of *y* and was smooth as the sample value changed sign (ibid). So, for a given exponent λ to be tested, the following sub-steps were performed: (a) transform the standardized batch from step 1 above according to (16). (b) Determine this batch's mean μ_2 and standard deviation σ_2. (c) Perform a second standardization of the resulting ψ 's, subtracting μ_2 and dividing by standard deviation σ_2. (d) Generate a test statistic, the maximum absolute difference, or MAD (e.g., KE08, eq. (9)) between the empirical CDF for this scaled and transformed batch of data and the CDF of a standard normal distribution. Note the μ_1 , σ_1, ψ , μ_2 , σ_2 associated with this test statistic.

(3) Select the parameters associated with the test statistic with the smallest MAD.

In summary, then an observed value *w* can be converted to a standard normal deviate *w*' under the following transformation:

$$w'=\frac{S\left(\dfrac{w-\mu_1}{\sigma_1}\right)-\mu_2}{\sigma_2}\ , \tag{17}$$

where *S* represents the appropriate power transformation in (16) given the value of λ and $y = (w - \mu_1) / \sigma_1$. The best-fit power transformation shown in Fig. 3(b) used an exponent of 1.6 to fit the data.

One valid criticism of this approach is that it is more complex than it needs to be; say, if *w* is a temperature in Kelvins, could not one achieve the same through a simple power transformation w^{λ}, and then a standardization? The approach outlined above was chosen because this conceptually simpler approach would in practice require testing a much larger number of possible exponents for the power transformation to find an approximately ideal transformation. The more involved algorithm described above was thus much more computationally efficient to implement, despite being more algorithmically complex.

d. Definition of the observation climatology

The climatological probability distributions were determined by fitting the power-transformed normal distributions as described in the previous section to a data set of observed 2-meter temperatures for the 25-year period of 1980-2004. Different distributions were fit for each station and each day of the year. When fitting the observation's distribution for a given day of the year, samples from +/− 20 days were used. When no data was missing, this provided 41 × 25 = 1025 samples.

e. Methods of producing probabilistic forecasts.

Five basic methods for producing probabilistic forecasts will be evaluated. The first method is "raw" probability forecasts; here, the probability forecasts will be estimated from the 15-member ensemble forecast. For example, if 5 of 15 members indicate a temperature greater than 280 K, the probability for this event is estimated to be 1/3. The remaining methods, all involving statistical calibration, will be evaluated with both the 30-day and full reforecast training data sets described in section 3b.

The second method is the "standard BPF" of section 2b, using the power transformations as described in Section 3c. The third method is a variant of the basic BPF method, whereby the power transformation applied to the forecast data is forced to be the same as the transformation used for the observed; the reason for testing this will be made clear in the results section. This method will be named "BPF ($\lambda_x = \lambda_w$)." The fourth method is another BPF variant, here applying the BPF algorithm under the assumption that all distributions are Gaussian, i.e., only applying step 1 of the algorithm of distribution fitting in step 3c. This will be referred to as the "Gaussian BPF."

The remaining method, used as a calibration reference, is the "non-homogeneous Gaussian regression" or "NGR" technique of Gneiting et al. (2005) and used with reforecast data in Hagedorn et al. (2008). A Gaussian distribution is forecast using a regression model that uses the ensemble mean and variance estimate as predictors. A post-processed pdf is produced:

$$f^{CAL}\left(\overline{\mathbf{x}}, \sigma^2\right) \sim N\left(a + b\overline{\mathbf{x}}, c + d\sigma^2\right). \qquad (18)$$

The regression model fits coefficients *a* and *b*, which permit a bias correction of the mean, and coefficients *c* and *d*, which produce a variance correction. In testing (not shown), it was found that this NGR reference was somewhat more skillful than linear regression; interestingly, this difference was *not* due primarily to its ability to incorporate the variance information from the ensemble. Rather, the increased skill resulted from the NGR algorithm's internal optimization to minimize the CRPS score (see Gneiting et al. 2005, eq. 12) as opposed to linear regression's choosing parameters that maximize a quadratic likelihood function; when the linear regression approach was adapted to minimize CRPS, the scores were improved to be nearly as small as NGR.

f. Method of forecast evaluation.

Forecasts will primarily be evaluated with the continuous ranked probability skill score, or "*CRPSS*." A modification of the standard *CRPSS* calculation is used, following Hamill and Juras (2006) and Hamill and Whitaker (2007). The modification ensures that the score does not inappropriately inflate forecast skill as a result of combining samples that have different error statistics of the reference. To achieve this, the set of forecast samples for a particular lead from all stations and dates was divided into subgroups; in each subgroup, the associated standard deviation of the climatological samples varied only through a small range. Here *NC=8* subgroups were used, and equal numbers of samples assigned were assigned to each subgroup. The *CRPSS* was determined for each subgroup, and then the final *CRPSS* was determined as a weighted average of the subgroups' *CRPSS*. More explicitly, let $\overline{CRPS}^{f}\left(s\right)$ denote the average forecast *CRPS* for the *s*th subgroup, and $\overline{CRPS}^{c}\left(s\right)$ denote the average *CRPS* of the climatological reference forecast for this subgroup, determined from the climatological distribution fitted as discussed in section 3d. Then the overall *CRPSS* was calculated as

$$CRPSS = \frac{1}{NC}\sum_{s=1}^{NC}\left(1-\frac{\overline{CRPS}^{f}(s)}{\overline{CRPS}^{c}(s)}\right) \quad . \tag{19}$$

Confidence intervals were calculated via a block bootstrap approach (Hamill 1999, Hagedorn et al. 2008). Because of the large sample size, these intervals were very small, from approximately 0.033 at the half-day lead to 0.02 at the 10-day lead; accordingly, they were not plotted.

4. Results

Figure 5 shows the *CRPSS* of the various forecast algorithms, both for 30-day training data sets (panel a) and the reforecast training data set (panel b).[1] Unlike the simple, Gaussian experiment results shown in Fig. 2 and discussed in section 2b, surprisingly the standard BPF algorithm consistently performed worse than the NGR method. For the 30-day training data set, the skill of the standard BPF decreased rapidly with increasing forecast lead and was less skillful than climatology by 4 days lead.

Why were the 30-day training results of the standard BPF algorithm on this data set especially poor? There are at least two major causes, one that is a consequence of the standard BPF formulation, another due to the unusual climatological conditions in 2005.

Figure 6 provides an example of a pathological standard BPF forecast, initialized on the day associated with the climatological distributions of Fig. 3, thus demonstrating that a misfit of the prior climatological distribution was not the primary source of the

[1] The skill of raw forecasts in this figure are much higher than the skill of the comparable raw forecasts in Fig. 4 of Hagedorn et al. (2008). In that article, we defined the reference climatology using dependent data from 2005. In this article's Fig. 5, climatology was defined using independent 1980-2004 data.

problem. Here, the posterior distribution is neither consistent with the prior, nor consistent with the ensemble-mean forecast that day, denoted by the vertical solid line. By process of elimination, something must have been wrong with the likelihood function. Figure 7(a) shows the 30-day training data as well as that day's forecast and verifying observation. While that day's forecast-observation sample was outside the span of the training data, it was at least linearly consistent with it. However, the regression model was not fit to this training data, but rather to transformations of the data that were presumed to make the data more Gaussian. In the case of the observations, the applied power transformation was 1.6, indicating that the original data had a heavy tail on the cold side of the distribution. However, the 30-sample forecast training data was fit individually from the observations, following KE08. In this case, the four forecast samples with temperatures of 305-310 K in Fig. 7a caused the distribution-fitting algorithm to assume that the forecast data had a heavy tail on its warm side, and the chosen exponent for the power transformation that best achieved approximate normality was 0.25. Figure 7(b) shows the resulting transformed data and the linear regression model. While the regression model appeared to be a reasonable model for the training data, the distinct transformation applied separately to observations and forecasts made the forecast-observation pair from that day grossly inconsistent with the regression model. Figure 8 synthesizes how the standard BPF made such a poor forecast in this instance. The regression model, once transformed into the coordinates of the original data, provided a poor fit to the real-time datum. When considering the likelihood model, i.e., the probability density for the forecast across the range of possible observed values, the horizontal black line through the forecast shows that at high observed values, the regression model predicted low probability density, and as the observed value decreased, the probability density increased. Hence, in this case the posterior, the normalized multiplication of the prior and likelihood (i.e., the application of Bayes Rule), produced an anomalous maximum at approximately 242K.

For this case, the large mis-estimation of the posterior pdf can be traced to the application of separate power transformations to the observation and the forecast data. Though the original forecast-observed pair was outside the range of the training data, it was still consistent with the training data, in that the bias was similar to the bias from the training data. This consistency was lost when the separate power transformations were applied. The usage of markedly different power transformations for the forecasts and observations in fact happened quite frequently (Fig. 9), and the average CRPS was often large in instances when the forecast transformation exponent was small while the observation exponent was larger. Also, note that while this example used power transformations to fit the data rather than the Weibull distributions of KE08, there is no reason to expect a qualitatively different result with the Weibull; the essence of the problem is treating forecast data differently than observed data, not the class of parametric distributions used.

Suppose the same power transformation that was applied to the observations was also applied to the forecasts. Figure 5 shows the impact of enforcing this consistency. The yellow area denotes the change of skill of BPF forecasts under this revision, so that the top of the yellow area denotes the skill of the BPF ($\lambda_x = \lambda_w$). With the short training data set, forecast skill is improved markedly, ~ +1 day lead. With the reforecast data set, the BPF ($\lambda_x = \lambda_w$) skill becomes effectively the same as the NGR algorithm. Figure 5 also shows how well the calibration method performed if the distributions were assumed to be Gaussian. Again, there is an improvement relative to the reference BPF algorithm; the improvement from more carefully fitting non-normal distributions in this instance is outweighed by the consequences of employing different transformations for forecasts and observations, destroying the linear relationships. Also, note that by comparing the Gaussian BPF to the BPF ($\lambda_x = \lambda_w$) in Fig. 5a, the improvement from use of non-normal distributions can be quantified. In general they are quite small, especially at short leads.

Note that with the large reforecast training data set in Fig. 5b, the two variants on the standard BPF scored nearly equivalently to the NGR method. All three of these calibration methods are slightly better than the standard BPF at short leads and comparable at long leads. Contrary to the original hypothesis that BPF would be particularly useful for small training data sets, in fact the flaws in the standard BPF method become more apparent. With a smaller training data set, two unfortunate situations happen more frequently. First, overfitting happens more readily; there may not be enough training data to provide a detailed fit of a non-normal distribution to the forecast training data, and an inappropriate power for the transformation may be driven by a few outliers. Second, there will be a larger fraction of situations where the real-time forecast datum lies outside the span of the training data; the extrapolation of a bad regression fit, as shown in Figs. 6-8, may result in very large forecast errors.

Examining Fig. 5a, note that the standard BPF forecasts decreased in skill relatively quickly as forecast lead increased, while NGR and raw forecasts decreased more slowly. A primary difference is that the BPF forecasts at long leads increasingly reflected the climatological prior, as the likelihood pdf broadened with decreased forecast-observation correlation. So, was the prior a bad estimate in the fall of 2005? Indeed, it was. September to November temperatures were much above normal in much of the US[2]. Figure 10 shows how much warmer the fall 2005 observations were relative to the 1980-2004 average; at many locations the forecasts were more than 2K warmer, especially at 0000 UTC. If the BPF results had been calculated for years that were much closer to the climatological prior, the scores would have been substantially improved.

5. **Discussion and conclusions**

This article provided a critical review of the Bayesian Processor of Forecast (BPF) algorithm of KE08 for producing and calibrating probabilistic forecasts. Unlike most other methods, the BPF algorithm leverages a prior estimate of the probability density, taken in KE08 from a long-term observation climatology. The prior is updated with information provided by a new forecast datum; the character of this modification is controlled by a model of the relationship between forecasts and observations in a training data set, the "likelihood" function. The BPF algorithm has been proposed to be attractive for calibrating weather forecast model output, for perhaps the leveraging of a long-term climatology will reduce the need for large training data sets, which typically require freezing the forecast model in a stable configuration.

The results presented here show the potential benefits and drawbacks of the BPF algorithm for producing calibrated probabilistic forecasts. Tests in a simple model where the data were sampled from normal distributions showed that the BPF algorithm was

[2] See, for example, http://www.ncdc.noaa.gov/oa/climate/research/2005/ann/us-summary.html

clearly superior for small training sample sizes and low forecast-observed correlations. However, tests of surface-temperature forecast accuracy in the fall of 2005 using the ensemble-mean temperature from a reforecast data set as a predictor were inferior to the existing calibration methods.

Section 5 discussed two potential deficiencies of the BPF algorithm as illuminated with this data set. First, the standard BPF algorithm requires that the forecasts and observations be converted to distributions that are approximately Gaussian. However, there is an inherent tension between the requirement that data be Gaussian and the requirement of ensuring linear relationships between forecasts and observations (as the likelihood model is a simple linear regression). This relationship may be lost after they are separately transformed to Gaussian distributions. Probabilistic forecasts were improved significantly by requiring the forecast data to be transformed in the same way as the observed data. This indicates that preserving linear relationships between forecasts and observations in the raw data is more important to algorithmic accuracy than deviations from normality.

The second drawback illustrated here, very much in evidence for the anomalously warm fall 2005 season, is that if the prior distribution is consistently biased, the BPF algorithm will suffer in accuracy. For more normal years, the BPF can be expected to improve from use of the climatology.

In this study, following KE08, the prior was estimated strictly from a long-term climatology. However, the prior could also have been estimated from persistence, or a blend of climatology and persistence, or perhaps some more sophisticated model, such as one based on a canonical correlation analysis (Wilks 2006, section 12.1.2, and references therein). This may improve the accuracy and reduce the bias in the prior. Such techniques were not explored here, however.

There are several other reasons why one may choose methods other than BPF for calibrating and producing probabilistic forecasts. The BPF as currently designed works with a single estimate of a forecast; how it could leverage possible spread-skill relationships from an ensemble, as are done in the NGR method, is not clear. More generally, the BPF has not yet been formulated to permit multiple forecast predictors, as in Model Output Statistics (MOS; Glahn and Lowry 1972; Carter et al. 1989). The additional predictors in MOS build in some additional weather-specific dependency; for example, perhaps forecast cloud optical depth is a useful additional predictor, so that the relationship between forecast and observed temperature changes as cloud optical depth increases or decreases. How BPF can be applied to bounded, strictly positive distributions such as precipitation is unclear, too.

In summary, the BPF algorithm is amazingly elegant in concept. It is a novel calibration algorithm that is truly Bayesian in character, modifying a prior independent estimate of the pdf. However, significant further research is needed to overcome algorithmic difficulties with the initially proposed version. These difficulties will need to be overcome for BPF to be accepted as a standard calibration method in weather services. Regrettably, the development of BPF has apparently not lessened the need for large training data sets such as reforecasts.

ACKNOWLEDGMENTS

D. J. Seo, John Schaake, Bob Glahn, and Zoltan Toth are thanked for their constructive criticism on a prior presentation of this material.

REFERENCES

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401-412.

Daley, R. 1991: *Atmospheric Data Analysis*. Cambridge Press, 457 pp.

DeGroot, M. H., 1970: *Optimal Statistical Decisions*. McGraw Hill, 480 pp.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098-1118.

Hagedorn, R, T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: 2-meter temperature. *Mon. Wea. Rev.*, **136**, 2608-2619.

Hagedorn, R, 2008: Using the ECMWF reforecast data set to calibrate EPS forecasts. *ECMWF Newsletter*, **117**, 8-13. Available at www.ecmwf.int/publications/newsletters.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

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.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134**, 3209-3229.

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

Hamill, T. M., and J. S. Whitaker, 2006: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. *Mon. Wea. Rev.*, **135**, 3273-3280.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part II: precipitation. *Mon. Wea. Rev.*, **136**, 2620-2632.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Press, 533 pp.

Krszysztofowicz, R., 1987: Markovian forecast processes. *J. Amer. Stat. Assoc.*, **82**, 31-37.

Krszysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting*, **23**, 270-289.

Son, J., D. Hou, and Z. Toth, 2008: An assessment of Bayesian bias estimator for numerical weather prediction. *Nonlinear Processes in Geophysics*, accepted. Available from Dingchen.Hou@noaa.gov.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd Ed., Academic Press, 627 pp.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379-2390.

Yeo, I.-K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954-959.

FIGURE CAPTIONS

Figure 1: Illustration of application of Bayes Rule when prior is normally distributed and likelihood function is determined from a standard linear regression relationships. The solid black line denotes the prior pdf. Black dots provide observation and forecast data points from a short training data set. Parallel red lines indicate the regression relationship predicting the forecast given the observation (heavy line) and + / - 1 and 2 standard deviations (lighter lines). Today's forecast is assumed to be +20 degrees. The likelihood function is plotted in the red curve.

Figure 2. (a) *CRPSS* of BPF forecast algorithm for experiment described in section 2.a. (b) *CRPSS* of linear regression, and (c) *CRPSS* difference (BPF minus linear regression).

Figure 3: (a) Illustration of the bounded support when GEV distribution is used to model climatological distribution. Observed climatological distribution of 0000 UTC 2-meter temperatures for Omaha, NE on 24 October (using data for + / - 20 days around 24 October, 1979-2004) shown in dots. The fitted GEV distribution is shown with the light-grey curve. Notice zero probability beyond ~ 302K. (b) As in (a), but using a power-transformed Gaussian distribution, with unbounded support.

Figure 4: Illustration of the posterior distribution in an application of Bayes rule where the prior has bounded support.

Figure 5: CRPSS for (a) 30-day training data set, and (b) reforecast training data set. Dashed black line denotes skill of NGR forecast, dotted black line the skill of BPF when the algorithm is forced to assume distributions are Gaussian. Solid black line denotes

skill of raw forecast, dotted red line denotes skill of standard BPF forecast. Yellow area indicates how much forecast has improved when power transformation of forecast training data is forced to be the same as the transformation for the observed data.

Figure 6: Illustration of a pathologically bad BPF forecast, here a 1-day forecast from 0000 UTC 24 October 2005. Prior pdf is solid line; ensemble-mean forecast is vertical solid line; posterior pdf is dotted curve; observation the dashed line.

Figure 7: (a) 30-day training data set (small dots) and real-time forecast and observation (large dot) for the 1-day forecast from 0000 UTC 24 October 2005. (b) Training data set after power transformations and rescalings are applied to the training data set to make the training data more normally distributed. Regression line (thick dashed line) and the +/- 1 and 2 standard deviations (thinner dashed lines) are also plotted.

Figure 8: Illustration of how pathological BPF forecast in Fig. 6 was produced. Training data are small black dots in the upper-right corner; real-time forecast and observation is the large black dot. Prior distribution plotted in heavy black curve, with probability density scale on the right. Posterior plotted in blue curve, likelihood in red curve. The regression model remapped to the original data coordinates (thick red dashed line) and +/- 2, 4, 6, and 8 standard deviations (thinner red dashed line) are also plotted. Horizontal black line through the forecast value shows that the distribution of forecast given the observation will be larger as the observed temperature decreases.

Figure 9: Average CRPS of 1-day forecasts as a function of the power of the observation transformation and forecast transformation. Area of square indicates relative proportion

of the sample that used this particular combination of observation and forecast transformations.

Figure 10: The 1 September – 1 December temperature difference (2005 average value minus 1980-2004 average value). (a) 0000 UTC observations, and (b) 1200 UTC observations.

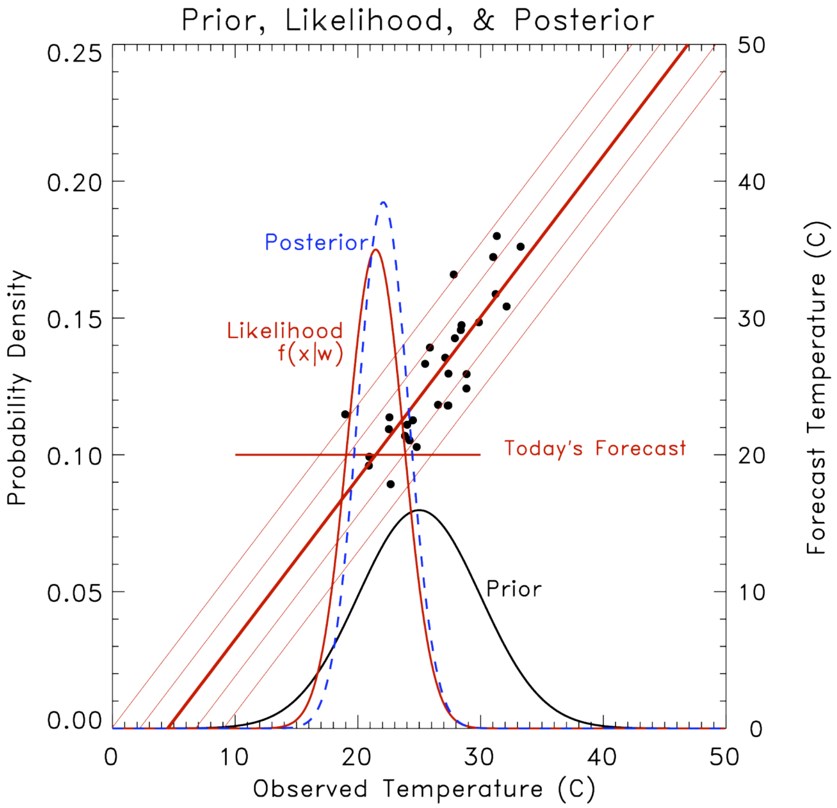


Figure 1: Illustration of application of Bayes Rule when prior is normally distributed and likelihood function is determined from a standard linear regression relationships. The solid black line denotes the prior pdf. Black dots provide observation and forecast data points from a short training data set. Parallel red lines indicate the regression relationship predicting the forecast given the observation (heavy line) and + / - 1 and 2 standard deviations (lighter lines). Today's forecast is assumed to be +20 degrees. The likelihood function is plotted in the red curve.

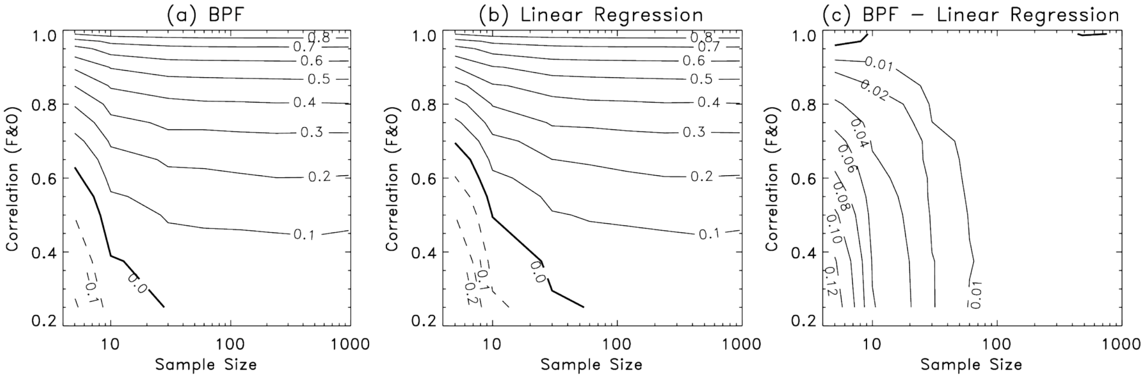


Figure 2. (a) *CRPSS* of BPF forecast algorithm for experiment described in section 2.a. (b) *CRPSS* of linear regression, and (c) *CRPSS* difference (BPF minus linear regression).

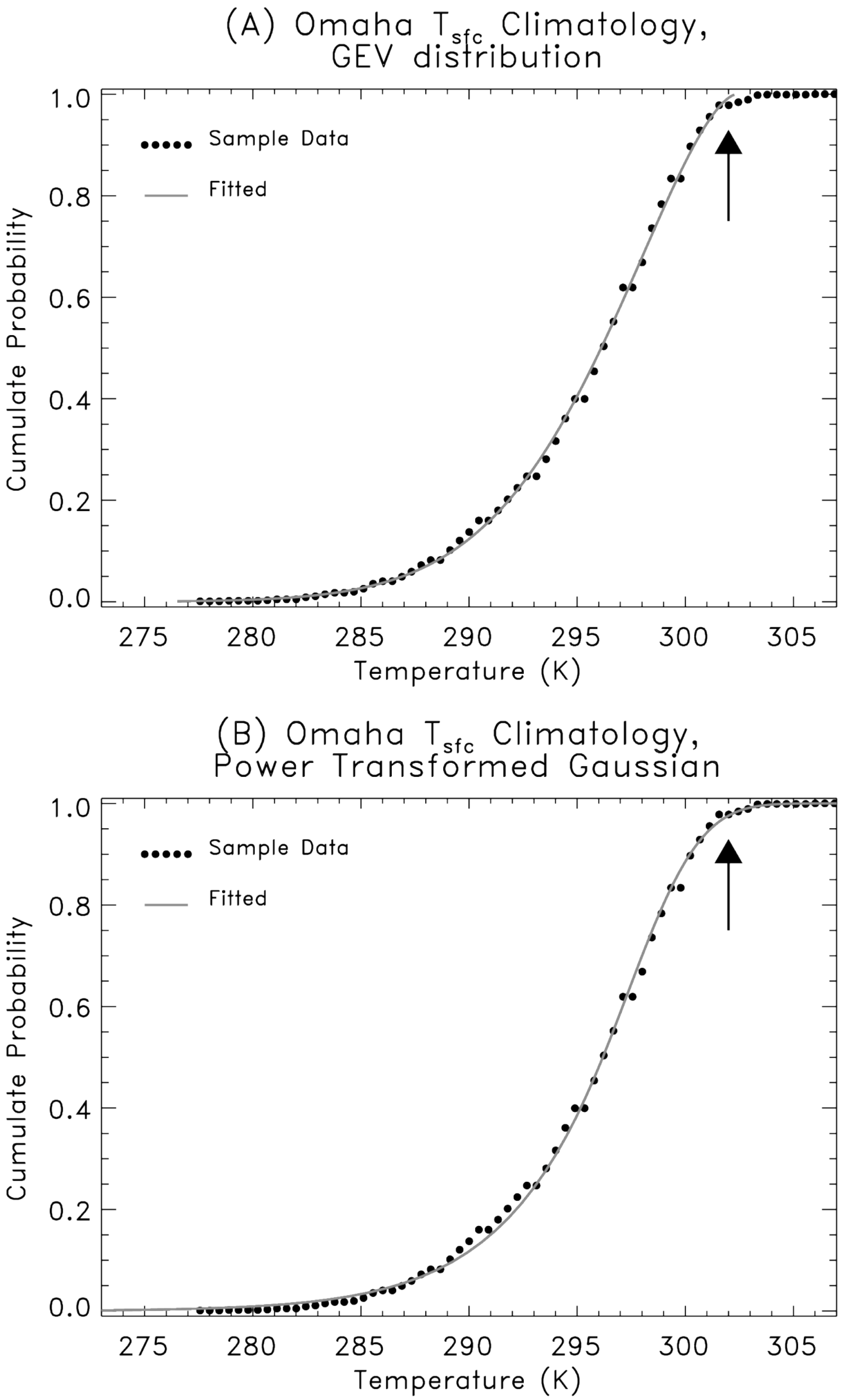


Figure 3: (a) Illustration of the bounded support when GEV distribution is used to model climatological distribution. Observed climatological distribution of 0000 UTC 2-meter temperatures for Omaha, NE on 24 October (using data for + / - 20 days around 24 October, 1979-2004) shown in dots. The fitted GEV distribution is shown with the light-grey curve. Notice zero probability beyond ~ 302K. (b) As in (a), but using a power-transformed Gaussian distribution, with unbounded support.

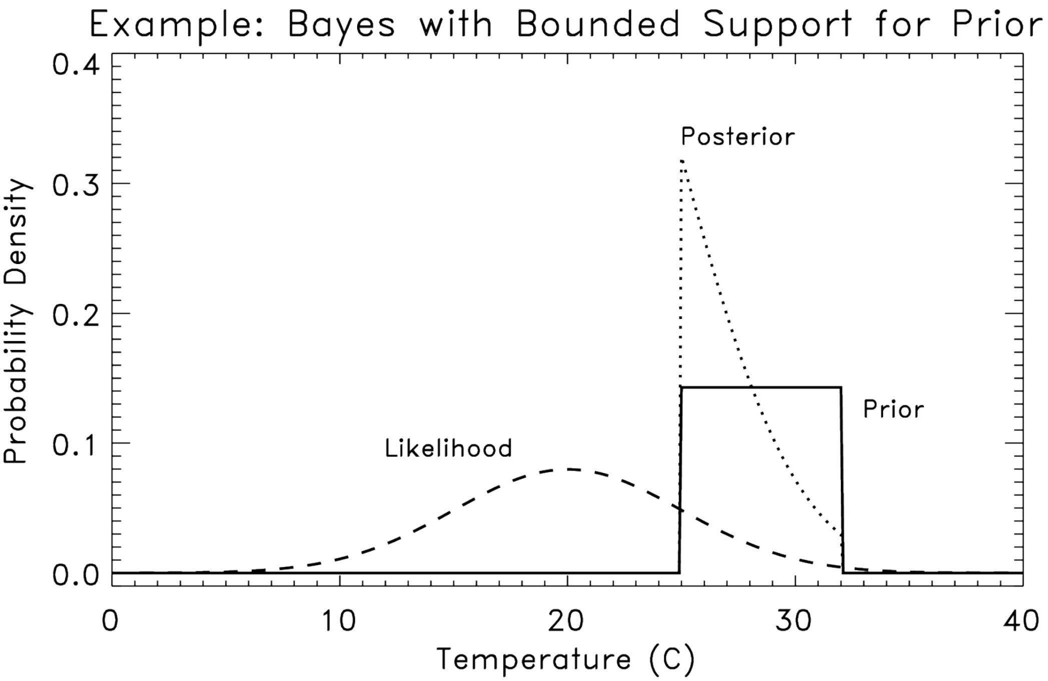
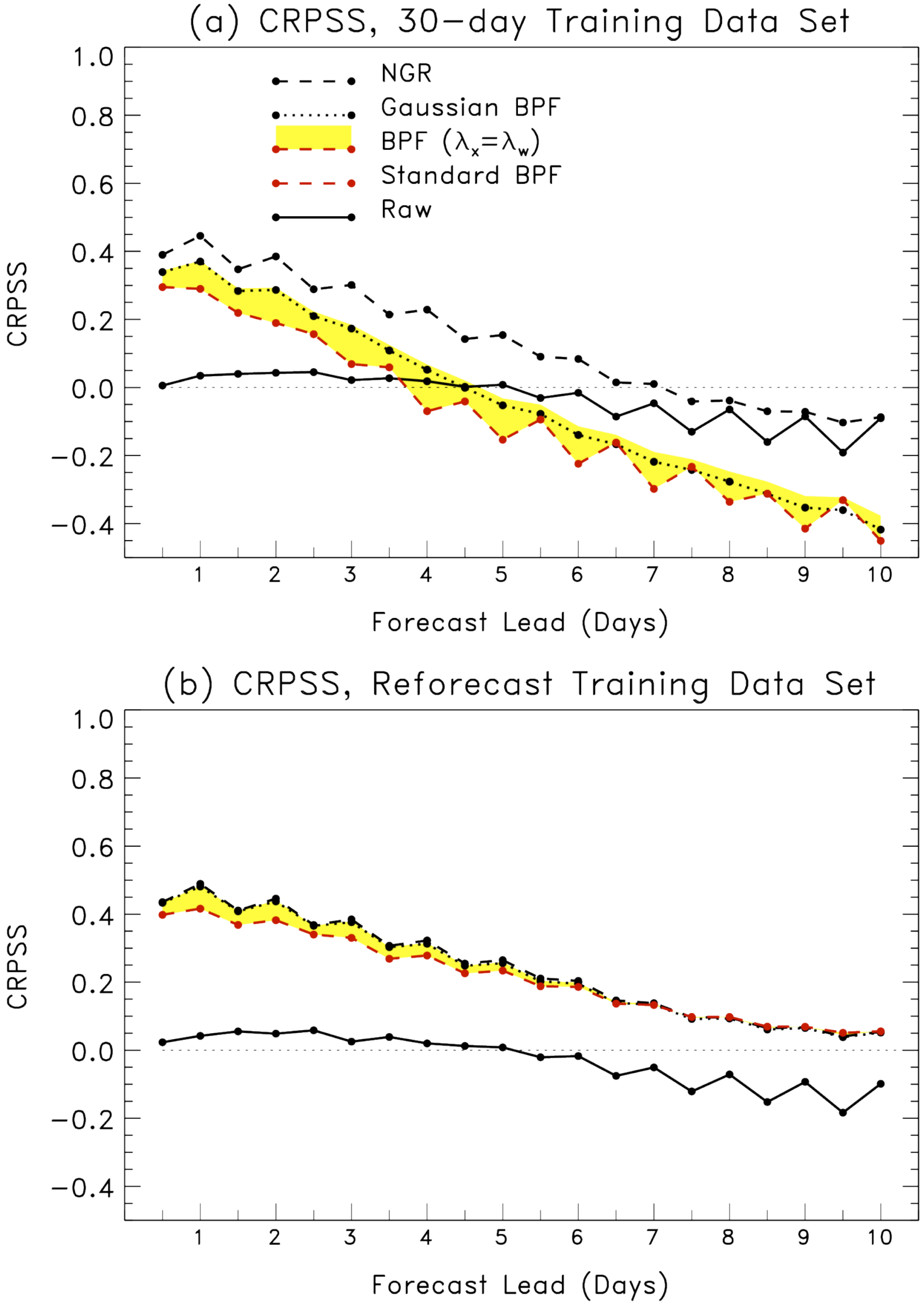


Figure 4: Illustration of the posterior distribution in an application of Bayes rule where the prior has bounded support.



(a) CRPSS, 30−day Training Data Set

(b) CRPSS, Reforecast Training Data Set

Figure 5: CRPSS for (a) 30-day training data set, and (b) reforecast training data set. Dashed black line denotes skill of NGR forecast, dotted black line the skill of BPF when the algorithm is forced to assume distributions are Gaussian. Solid black line denotes skill of raw forecast, dotted red line denotes skill of standard BPF forecast. Yellow area indicates how much forecast has improved when power transformation of forecast training data is forced to be the same as the transformation for the observed data.

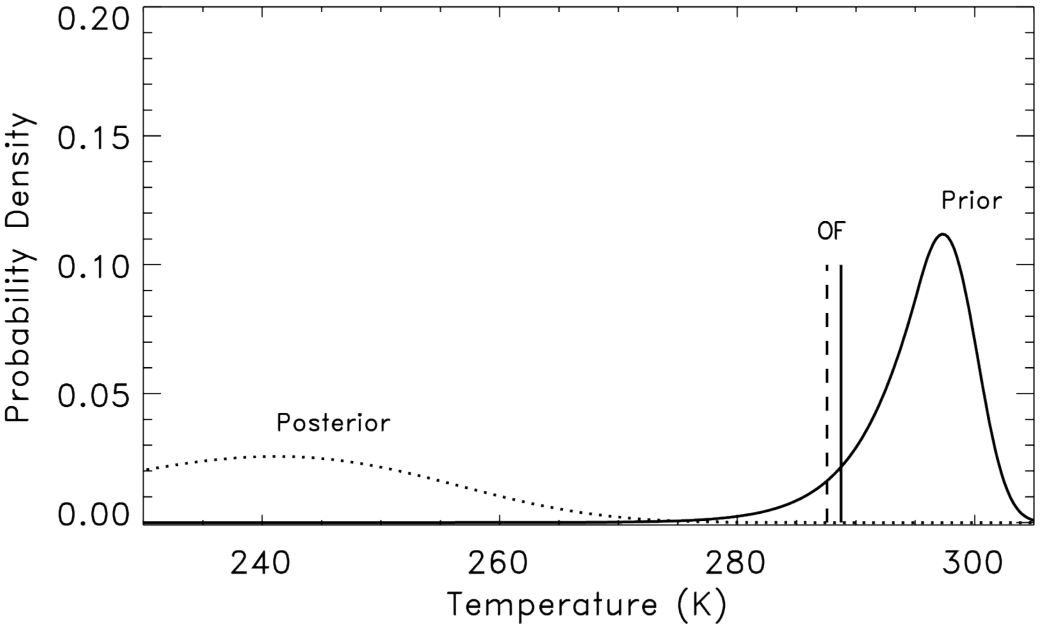


Figure 6: Illustration of a pathologically bad BPF forecast, here a 1-day forecast from 0000 UTC 24 October 2005. Prior pdf is solid line; ensemble-mean forecast is vertical solid line; posterior pdf is dotted curve; observation the dashed line.

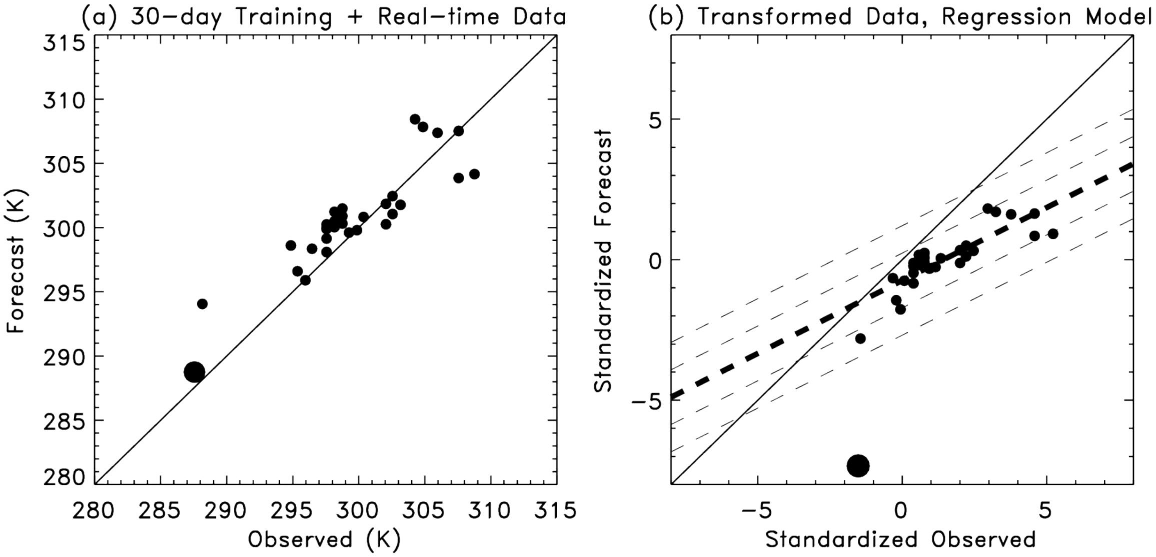


Figure 7: (a) 30-day training data set (small dots) and real-time forecast and observation (large dot) for the 1-day forecast from 0000 UTC 24 October 2005. (b) Training data set after power transformations and rescalings are applied to the training data set to make the training data more normally distributed. Regression line (thick dashed line) and the +/- 1 and 2 standard deviations (thinner dashed lines) are also plotted.

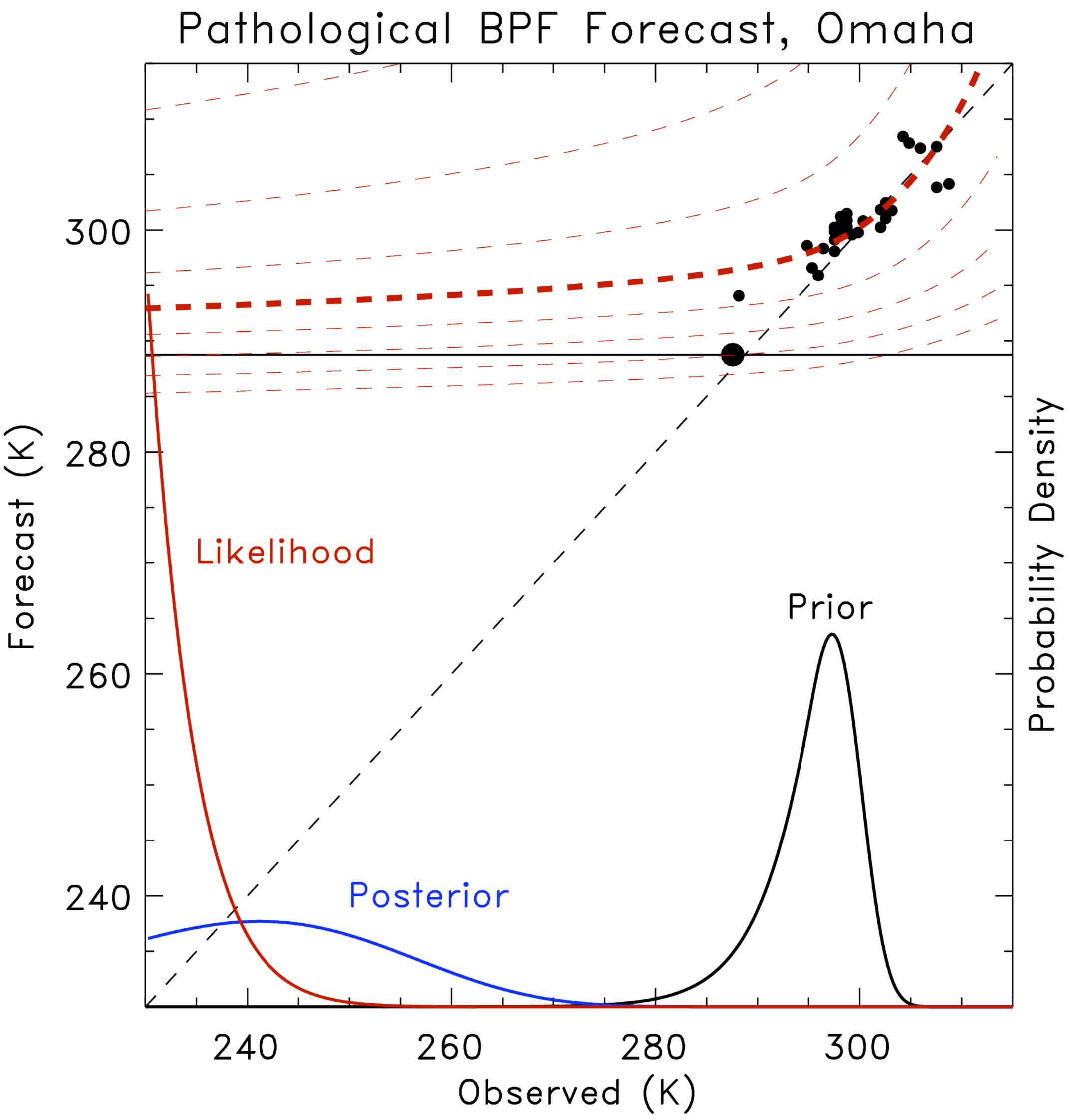


Figure 8: Illustration of how pathological BPF forecast in Fig. 6 was produced. Training data are small black dots in the upper-right corner; real-time forecast and observation is the large black dot. Prior distribution plotted in heavy black curve, with probability density scale on the right. Posterior plotted in blue curve, likelihood in red curve. The regression model remapped to the original data coordinates (thick red dashed line) and +/- 2, 4, 6, and 8 standard deviations (thinner red dashed line) are also plotted. Horizontal black line through the forecast value shows that the distribution of forecast given the observed will be larger as the observed temperature decreases.

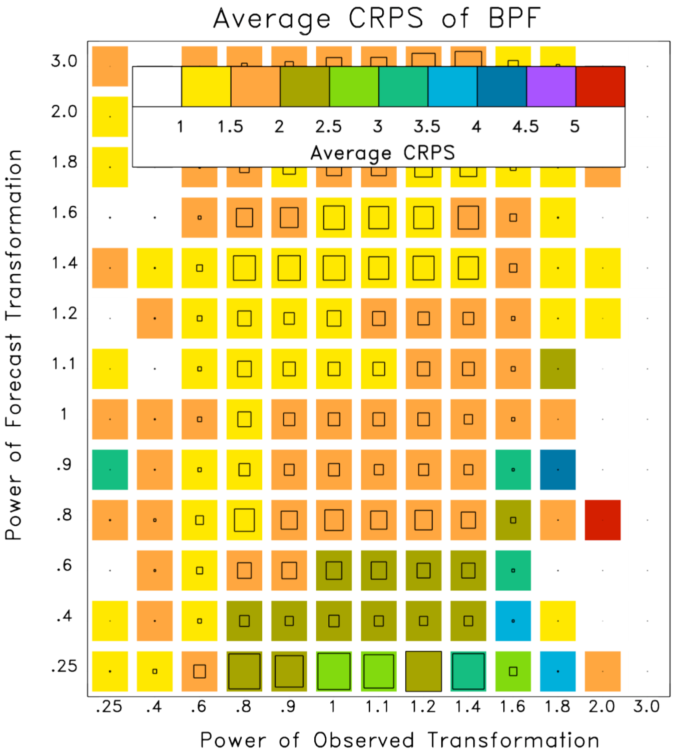


Figure 9: Average CRPS of 1-day forecasts as a function of the power of the observed transformation and forecast transformation. Area of square indicates relative proportion of the sample that used this particular combination of observed and forecast transformations.

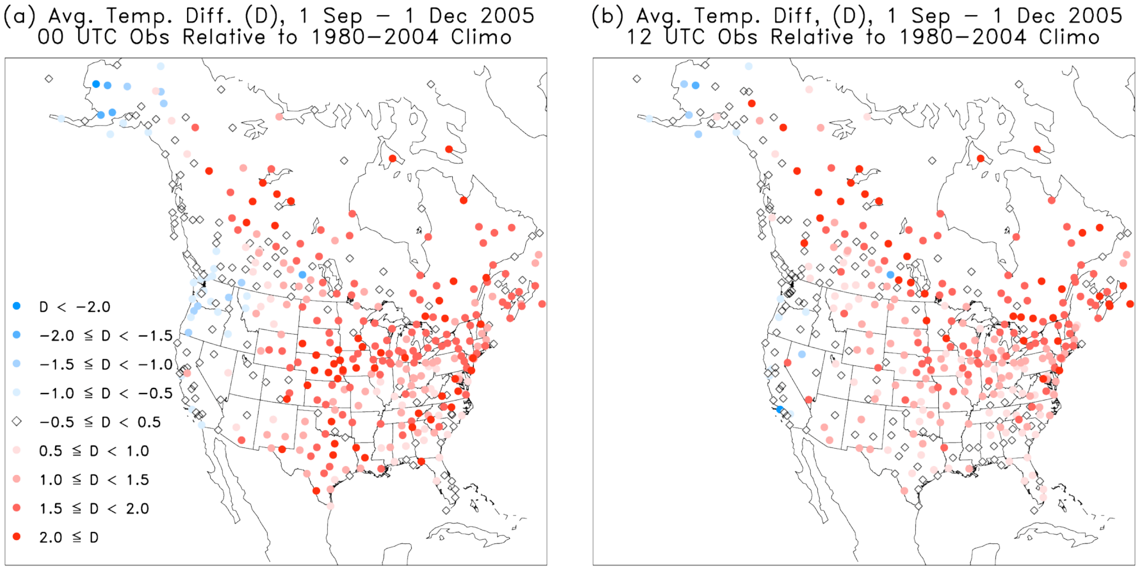

(a) Avg. Temp. Diff. (D), 1 Sep – 1 Dec 2005
00 UTC Obs Relative to 1980–2004 Climo
(b) Avg. Temp. Diff, (D), 1 Sep – 1 Dec 2005
12 UTC Obs Relative to 1980–2004 Climo
D < −2.0
−2.0 ≦ D < −1.5
−1.5 ≦ D < −1.0
−1.0 ≦ D < −0.5
−0.5 ≦ D < 0.5
0.5 ≦ D < 1.0
1.0 ≦ D < 1.5
1.5 ≦ D < 2.0
2.0 ≦ D

Figure 10: The 1 September – 1 December temperature difference (2005 average value minus 1980-2004 average value). (a) 0000 UTC observations, and (b) 1200 UTC observations.