Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2011), 63A, 858–875
Printed in Singapore. All rights reserved

TELLUS

Calibrating probabilistic forecasts from an NWP ensemble

By THOMAS NIPEN* and ROLAND STULL, *University of British Columbia, Vancouver, Canada*

(Manuscript received 12 December 2010; in final form 21 June 2011)

ABSTRACT

A post-processing method for calibrating probabilistic forecasts of continuous weather variables is presented. The method takes an existing probability distribution and adjusts it such that it becomes calibrated in the long run. The original probability distributions can be ones such as are generated from a numerical weather prediction (NWP) ensemble combined with a description of how uncertainty is represented by this ensemble. The method uses a calibration function to relabel raw cumulative probabilities into calibrated cumulative probabilities based on where past observations verified on past raw probability forecasts. Applying the calibration method to existing probabilistic forecasts can be beneficial in cases where the underlying assumptions used to construct the probabilistic forecast are not in line with nature's generating process of the ensemble and corresponding observation. The method was tested on a forecast data set with five different forecast variables and was verified against the corresponding analyses. The calibration method reduced the calibration deficiency of the forecasts down to the level expected for perfectly calibrated forecasts. When the raw forecasts exhibited calibration deficiencies, the calibration method improved the ignorance score significantly. It was also found that the ensemble-uncertainty model used to create the original probability distribution affected the ignorance score.

1. Introduction

If forecasts were perfect, then we would not need probabilistic forecasts. For uncertain forecasts, information on the probability of different forecast outcomes can allow end users to make decisions that optimize their budget and safety (AMS, 2008). However, such optimization is possible only if the probability information provided is useful. Developing methods for producing useful probabilistic forecasts from an ensemble of weather forecasts is an area of active research.

Throughout this paper we take the view that creating probabilistic forecasts follows a two-step process, as shown in Fig. 1. The first step takes an ensemble of deterministic forecasts as input and models how this ensemble conveys forecast uncertainty. The second step is a simple post-processing step that ensures that the probabilistic forecast generated by the uncertainty model exhibits the desirable statistical property of being calibrated.

The uncertainty model is an algorithm that prescribes probability density to each of the possible values that the forecast variable can take. Ensemble uncertainty can be modelled in much the same way that radiation or precipitation is modelled in a weather model. For example, we could decide to place more confidence where ensemble members are clustered, or we could decide to place most of the confidence near the ensemble mean. The number of other algorithms for placing confidence given a certain arrangement of the input forecasts is endless.

The uncertainty model will inevitably contain assumptions about how nature generates ensemble members and the corresponding observation. For example, a Gaussian probability distribution could be centred on the ensemble mean, where the spread of the distribution is a tuning parameter. When the Gaussian assumption of uncertainty is valid we get calibrated (or reliable) forecasts. That is, a weather event that is forecast to occur with probability p will indeed be observed a fraction p of the time over many forecast periods.

However, in many cases the uncertainty model used can make assumptions that are not in line with how ensembles and observations are generated. In these cases, the uncertainty model may produce uncalibrated probabilistic forecasts. In the previous example, if the observations are in fact drawn from a non-Gaussian distribution, no value for the tuning parameter of the Gaussian distribution will generate calibrated probabilistic forecasts. The calibration step can then be used to remove this calibration deficiency, thereby improving the probabilistic forecast.

Separating the tasks of determining an uncertainty model and ensuring probabilistic calibration allows one to focus efforts to improve probabilistic forecasts. Perfecting the uncertainty model helps concentrate probability mass in the correct area and perfecting the calibration step increases the reliability of the

*Corresponding author.
e-mail: tnipen@eos.ubc.ca
DOI: 10.1111/j.1600-0870.2011.00535.x

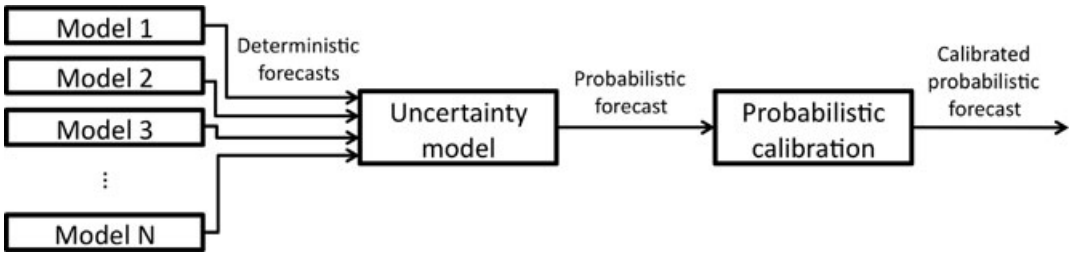

Fig. 1. A two-step process of generating probabilistic forecasts from an ensemble of forecasts. A set of deterministic forecasts from weather models feed into a system that models how uncertainty is conveyed by the ensemble. The resulting probabilistic forecast is fed to a calibration scheme that generates calibrated probabilistic forecasts.

forecast. Requiring a probabilistic method to model the uncertainty and ensure probabilistic calibration simultaneously can therefore be avoided.

The goal of this paper is to present a calibration scheme that takes an existing probability forecast and ensures that it becomes calibrated regardless of the uncertainty model used and regardless of whether or not this distribution accurately models the ensemble uncertainty.

The calibration method proposed relabels the cumulative probabilities of some initial probability distribution into calibrated cumulative probabilities that are based on how often and where observations in the past verified on the initial probability distributions. As will be shown, the initial probability distribution may very well be calibrated to begin with, in which case the calibration step is redundant. However, for cases where the uncertainty model used fails to generate calibrated forecasts, the method can improve the probabilistic forecasts.

In this paper, we consider both continuous meteorological variables (such as temperature) and bounded mixed discrete-continuous variables (such as relative humidity) that can have finite probability mass at one or both boundaries.

The remainder of the paper is organized as follows: First, we summarize some of the ways to represent uncertainty. Next, in Section 3, we discuss the metrics used to evaluate the quality of probability forecasts—important for measuring if and by how much the calibration method can cause improvement. In Section 4 we present the proposed calibration method. Section 5 describes case-study data from a 4-yr period with five forecast variables, a 14-member ensemble and 1225 grid locations. Those case-study data will be used in Section 6 to evaluate the calibration method for the uncertainty models from Section 2. Implications of this approach are summarized in Section 7.

2. Methods for representing uncertainty

A number of methods have previously been devised with the goal of producing calibrated probabilistic forecasts. Each of these methods, however, use widely different ways to describe how uncertainty is expressed by an ensemble. Different uncertainty descriptions arise because the methods make different assumptions about how forecasts and observations are realized.

To set up a framework for probabilistic forecasts, let $f_t(x)$ be the forecast probability density function (PDF) of a meteorological variable x for time t. The corresponding forecast cumulative distribution function (CDF) $F_t(x)$ is

$$F_t(x) = \int_{-\infty}^{x} f_t(s)\,\mathrm{d}s. \tag{1}$$

Thus, $F_t(x)$ gives the probability that the meteorological variable is forecasted to have any value less than x.

Let the actual observed value of the variable at time t be x_t. The observed value can be represented by an observed CDF $G_t(x)$ that we model as a step function.

$$G_t(x) = H(x - x_t), \tag{2}$$

where $H(s)$ is the Heaviside function defined by

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases}. \tag{3}$$

That is, the observed distribution is an infinitesimally wide region of finite probability mass at the observed value.

We denote an ensemble of K forecasts of some meteorological variable as $\xi_{t,k}$, where t represents a time point and k is an index between 1 and K. At time t, the ensemble mean is denoted by $\bar{\xi}_t$ and the ensemble spread is denoted by s_t.

2.1. Binned probability ensemble (BPE)

A very common way to model uncertainty is to assume that each ensemble member and the corresponding observation are realizations of the same unknown probability distribution. For this situation, the rank of the verifying observation when pooled with the ensemble should be a random integer between 1 and $K + 1$. Here rank is defined as the integer position of an element in a sorted array of values. Thus, each bin has the same probability of capturing the observation, where a bin is the region between two consecutive ensemble members. This is often referred to as the BPE technique (Anderson, 1996).

To convert this description to a probabilistic forecast, one assigns a constant probability mass $(K + 1)^{-1}$ between each consecutive ensemble member. Ensemble members spread further apart will have a lower density between them compared to

members that are closer together. The effect is that an ensemble that has all of its members close together represents a more certain forecast than one where all members are spread out.

The CDF values at each ensemble member location are set to $k(K+1)^{-1}$, where k is the rank of the ensemble member and is linearly interpolated between members.

The CDF below and above the ensemble must also be specified. For precipitation forecasts, Hamill and Colucci (1998) used a linear function below the lowest ensemble member, and a Gumbel distribution above in order to estimate extreme precipitation events. With this modification, the BPE probabilistic forecast $F_t(x)$ becomes

$$F_t(x) = \begin{cases} \frac{1}{K+1} A(\xi_{t,1} - x) & x \le \xi_{t,1} \\ \frac{k}{K+1} + \frac{1}{K+1} \frac{x - \xi_{t,k}}{\xi_{t,k+1} - \xi_{t,k}} & \xi_{t,k} < x \le \xi_{t,k+1} \\ 1 - \frac{K}{K+1} B(x - \xi_{t,K}) & \xi_{t,K} < x \end{cases}, \tag{4}$$

where $\xi_{t,k}$ represents the kth sorted ensemble member, and $A(s)$ and $B(s)$ are monotonic functions equal to 1 when $s = 0$, and drop off towards 0 for high values of s.

2.2. Method of moments

A Gaussian distribution $\mathcal{N}$ can be used to represent a probability distribution as follows:

$$F_t \sim \mathcal{N}\left(\bar{\xi}_t - \mu_{\mathcal{T}}, a_{\mathcal{T}} s_t^2 + b_{\mathcal{T}}\right). \tag{5}$$

The first parameter of $\mathcal{N}$ represents the mean of the distribution, and corresponds to the bias-corrected ensemble mean. The second parameter represents the spread of the distribution, given by a linear regression fit to the variance of the ensemble (s_t^2).

$\mu_{\mathcal{T}}$ can be computed from the first moment of past forecast errors.

$$\mu_{\mathcal{T}} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\bar{\xi}_t - x_t). \tag{6}$$

Here, $\mathcal{T}$ represents a set of time points over which the mean is computed, and $\|\mathcal{T}\|$ represents the size of this set. Past values of the square of the error of the bias-corrected ensemble mean $(\bar{\xi}_t - \mu_{\mathcal{T}} - x_t)^2$ (for all t in training period $\mathcal{T}$) is used to estimate $a_{\mathcal{T}}$ and $b_{\mathcal{T}}$ using least-squares linear regression. That is, the spread of the forecast distribution is dependent on the spread of the ensemble (provided that $a_{\mathcal{T}} \neq 0$).

As historical moments of the forecast errors are used to generated the probabilistic forecasts, this method is often called the method of moments (MM, Jewson et al., 2005).

2.3. Bayesian model averaging (BMA)

Another way to model the uncertainty is to assume that the true state is distributed according to one of several candidate distributions, although it is not known which candidate is the true one. The candidate distributions are formed by fixing an a priori specified probability distribution to each ensemble member. The total distribution is the sum of each individual distribution, weighted by the likelihood that each candidate distribution is the true one.

This technique is referred to as BMA (Hoeting et al., 1999). The use of BMA was suggested by Raftery et al. (2005) as a method for producing calibrated probabilistic weather forecasts. This method and variants thereof have been applied successfully for a number of cases (Raftery et al., 2005; Sloughter et al., 2007; Wilson et al., 2007; Johnson and Swinbank, 2009). By training on data, BMA can weight the various candidate distributions based on their performance in the past. If the underlying assumption is valid, then the predictive (weighted) BMA distribution will converge to the true distribution, given a sufficiently large data set. For temperature and sea level pressure, a Gaussian distribution centred on the bias-corrected value of the ensemble member has been used (Raftery et al., 2005).

Given a set of forecasts $\xi_{t,k}$ (where k, unlike for BPE, no longer represents a sorted index), the BMA predictive distribution is

$$F_t(x) = \sum_{k=1}^{K} w_k F_{t,k}(x), \tag{7}$$

where w_k are non-negative weights and $F_{t,k}(x)$ are the predictive distributions for each ensemble member given by

$$F_{t,k}(x) \sim \mathcal{N}(\xi_{t,k} - \mu_{\mathcal{T},k}, \sigma_{\mathcal{T}}^2) \tag{8}$$

$$\mu_{\mathcal{T},k} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\xi_{t,k} - x_t). \tag{9}$$

As before, $\mathcal{T}$ represents the training period. Raftery et al. (2005) used a common $\sigma_{\mathcal{T}}$ for all ensemble members to reduce the number of parameters, and still found good results. $\mu_{\mathcal{T},k}$ is a bias correction term specific to each ensemble member.

To compute the weights and standard deviation, Raftery et al. (2005) use the expectation maximization (EM) algorithm, an iterative process given by

$$z_{t,k}^{(j)} = \frac{w_k^{(j-1)} f_{t,k}^{(j-1)}(x_t)}{\sum_{i=1}^{K} w_i^{(j-1)} f_{t,i}^{(j-1)}(x_t)}, \tag{10}$$

$$w_k^{(j)} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} z_{t,k}^{(j)}, \tag{11}$$

$$\sigma_{\mathcal{T}}^{2(j)} = \frac{1}{\|\mathcal{T}\|} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}} z_{t,k}^{(j)} (x_t - \xi_{t,k} - \mu_{\mathcal{T},k})^2, \tag{12}$$

$$f_{t,k}^{(j)}(x) \sim \mathcal{N}\left(\xi_{t,k} - \mu_{\mathcal{T},k}, \sigma_{\mathcal{T}}^{2(j)}\right), \tag{13}$$

where (j) as a superscript represents the value after iteration j. This iteration is continued until the parameters change by less than some small tolerance. $z_{t,k}^{(j)}$ are intermediate values on the interval [0, 1] that represent the extent to which member k is the best member of the ensemble for time t.

2.4. Climatology

Finally, one can completely ignore the guidance of the ensemble and describe the uncertainty based only on the distribution of past observations. This is referred to as a climatology forecast and can be computed by

$$F_{\text{clim}}(x) = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} H(x - x_t). \tag{14}$$

Thus the climatology forecast for a given threshold is the frequency of past observations that have fallen below that threshold.

Climatology forecasts are independent of any NWP model output, and require only past observations. Therefore, we will use these probabilistic forecasts as a baseline against which the other probabilistic forecasting methods will be compared.

The climatology forecast is heavily dependent on the definition of $\mathcal{T}$. A very coarse climatology would define $\mathcal{T}$ to be all days of the year. A more refined climatology would only include observations from days that are from roughly the same time of the year as the desired forecast time point. We will use this more refined climatology as our baseline.

2.5. Comparison of these uncertainty models

We have discussed four ways of representing uncertainty. These can be summarized as follows:

(i) BPE: fixing a constant probability mass between each pair of consecutive ranked ensemble members.

(ii) MM: fixing a shape function to the bias-corrected ensemble mean.

(iii) BMA: fixing a shape function to each bias-corrected ensemble member.

(iv) Climatology: fixing a constant-in-time shape function directly onto forecast-variable x.

Figure 2 illustrates these different methods schematically. Each method behaves differently depending on whether the ensemble spread is small (top row) or large (bottom row). The probability density produced by BPE scales linearly with the spread of each pair of consecutive ensemble members. Forecasts produced by MM also generally scale with the spread of the ensemble, however they are independent of the particular way that ensemble members are organized. BMA, unlike MM, is able to represent multimodal distributions due to the individual Gaussian distributions, however, compared to BPE, its peaks are less sensitive to the exact positions of the ensemble members.

3. Metrics of probabilistic-forecast quality

There are two performance characteristics of probabilistic forecasts that we will investigate. The first, calibration, concerns the statistical consistency between the probabilistic forecasts and observations. The second, ignorance score, measures the extent to which probability has not been concentrated in the correct areas.

3.1. Calibration deviation

Probabilistic calibration, or reliability (Murphy, 1973), is a measure of correspondence between forecast probabilities and the frequency of occurrence of observed values. Events forecasted with probability p should occur a fraction p of the time, when

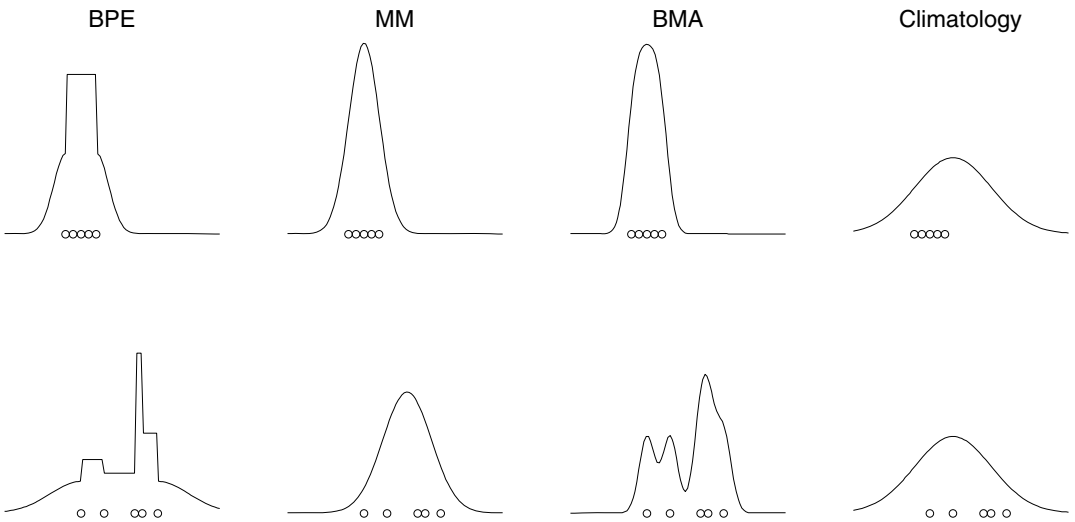

Fig. 2. Schematic PDF diagram of four methods for representing ensemble uncertainty. Here probability density curves for binned probability ensemble (BPE), method of moments (MM), Bayesian model averaging (BMA) and climatology are shown for an ensemble of size five, with the variable of interest in the abscissa and the probability density in the ordinate. Circles represent the five ensemble member forecasts. The top and bottom rows represent two different forecast times having different ensemble distributions.

evaluated over a set of times $\mathcal{T}$. Here, an event is defined as an observation being less than some threshold value x_a. The probability of this event occurring is forecasted by $F(x_a)$.

Calibration can be assessed by checking the distribution of probability integral transform (PIT) values (Gneiting et al., 2007). PIT values p_t are the values of the cumulative forecast distribution F_t corresponding to the observation; i.e., $p_t = F_t(x_t)$. Gneiting et al. (2007) define the set of forecasts $F_t(x)$ to be probabilistically calibrated relative to $G_t(x)$ for all t within $\mathcal{T}$ if

$$\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} G_t\left(F_t^{-1}(p)\right) = p, \tag{15}$$

where probability p is a real number between 0 and 1 and F_t^{-1} is the inverse of F_t. Using the definition of G_t in eq. (2), eq. (15) can be rewritten to show that probabilistic forecasts are calibrated if

$$\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H(p - p_t) = p. \tag{16}$$

Thus, probabilistic calibration requires that, for a given p on the interval [0, 1], a fraction p of the PIT values lie below p. Asymptotically over an infinite sample size, eq. (16) can be shown to be a necessary and sufficient condition for probabilistic calibration (Gneiting et al., 2007).

A forecast that is calibrated at all instances in time (i.e., $F_t(x) = G_t(x)$ for all t) is said to exhibit complete probabilistic calibration (Gneiting et al., 2007). As pointed out by Hamill (2001), uniformly distributed PIT values do not necessarily imply that the forecast exhibits complete probabilistic calibration, because the forecast can have distributional bias during various subintervals of $\mathcal{T}$. For example, uncalibrated forecast distributions during the first half of $\mathcal{T}$ and different uncalibrated forecast distributions during the second half can cancel out when evaluated over the whole time period $\mathcal{T}$. Furthermore, by defining the observational distribution to be a step function as in eq. (2), F can never exhibit complete probabilistic calibration unless $F_t(x) = H(x - x_t)$ for all t, which is the case of a perfect deterministic forecast. Therefore, when referring to calibration, we will always specify a time period over which the calibration is computed, and we will not require the forecast to exhibit calibration at smaller timescales.

To better visualize the degree of calibration using PIT, one can create a histogram of PIT values. For a perfectly calibrated forecast, each equally sized bin will contain the same number of PIT values thereby giving a flat histogram. Deviations from a flat histogram can be used to diagnose problems with calibration. For example, a U-shaped histogram indicates that the observation verifies low or high on the CDF curve too often, an indication that the probability distribution is too narrow.

A PIT histogram is the generalization of a rank histogram, the latter of which is used for determining reliability when BPE is used to model uncertainty. The rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997) shows the frequency of the observations taking on various ranks when pooled with the ensemble, and the number of bins used is $K + 1$. For a PIT histogram the number of bins used can be arbitrary, since we are looking at numbers on the real line as opposed to integers between 1 and $K + 1$. For our PIT histogram, we separate the interval [0, 1] into 20 equally sized bins.

Denote by b_i the bin count for bin i, where i is an integer between 1 and the number of bins B. Bin frequencies are then given by $b_i\|\mathcal{T}\|^{-1}$. We use the standard deviation of the bin frequencies as a summary metric for the reliability of a forecast. Low variability in the bin frequency is indicative of a PIT histogram that is flat. The calibration deviation metric is computed as follows:

$$D = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{b_i}{\|\mathcal{T}\|} - \frac{1}{B}\right)^2}. \tag{17}$$

Low values of D are preferred.

Sampling error will cause even perfectly calibrated forecasts to exhibit calibration error (Bröcker and Smith, 2007; Pinson et al., 2010). That is, PIT values from a perfectly calibrated system will likely not generate a perfectly flat PIT histogram. The bin counts b_i of a perfectly calibrated forecasting system will be multinomially distributed with variance $\|\mathcal{T}\|B^{-1}(1 - B^{-1})$. The expected value of the calibration deviation D_{perfect} of perfectly calibrated forecasts is therefore

$$E[D_{\text{perfect}}] = \sqrt{\frac{1 - B^{-1}}{\|\mathcal{T}\|B}}. \tag{18}$$

3.2. *Ignorance score*

A forecast must be more than just calibrated in order to be useful. For example, a vague climatology forecast can be perfectly calibrated, but might lack the desired concentration of probability needed to make informed decisions.

The ignorance score (Roulston and Smith, 2002), originally defined as the logarithmic score by Good (1952), is a metric that measures the extent to which a probabilistic forecast is not concentrated in the correct areas. The ignorance score is defined as follows:

$$IGN = -\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} \log_2(f_t(x_t)). \tag{19}$$

Lower values of the ignorance score are desired. The ignorance score rewards forecasts that places high confidence in regions where the verifying observation falls and disregards the probability density placed elsewhere.

Due to the use of the logarithm in the definition of the ignorance score, arithmetic differences between two ignorance scores is more relevant than the ratio of the scores. A change of units in the forecast variable for example, will cause scores to be changed by an additive constant.

The ignorance score has a very natural interpretation in estimating expected gambling returns. When placing bets on the future outcome x_t, the optimal strategy for distributing one's current wealth is to distribute wealth to each possible outcome weighted by the probability density. Users with forecasts that have lower ignorance scores than their betting competitor can expect to increase their wealth in the long run.

Given a probabilistic forecast A and a reference forecast with ignorance scores IGN_{A} and IGN_{ref} respectively, users of forecast A can expect to double their wealth against a user of the reference forecasts in N_{bets} bets, where N_{bets} is computed by

$$N_{\mathrm{bets}} = \frac{1}{IGN_{\mathrm{ref}} - IGN_{\mathrm{A}}}, \tag{20}$$

provided that $IGN_{\mathrm{A}} < IGN_{\mathrm{ref}}$. N_{bets} gives a more intuitive feel for the quality of the probability forecast than numeric values of the ignorance score. Smaller N_{bets} values are better.

4. Calibration method

Section 2 identified four ways to create probabilistic forecasts. In many cases, the forecasts produced by these methods are already calibrated. Calibration deficiencies can arise, however, when the underlying assumption of how uncertainty is represented by the ensemble is not in line with how nature generates ensemble members and observations. For these situations, a calibration method may be used to adjust the forecasted distributions such that they are calibrated. Such a calibration method is presented next.

4.1. Basic principles

We propose a calibration method that takes an existing probability distribution $F_t(x)$ and relabels the CDF values to form a new distribution $\hat{F}_t(x)$. The relabelling is done by a calibration function Φ as follows:

$$\hat{F}_t(x) = \Phi_{\mathcal{T}}(F_t(x)). \tag{21}$$

$\Phi_{\mathcal{T}}$ is based on the distribution of past PIT values from the set of time points $\mathcal{T}$. For example, if 30% of past PIT values have values less than 25%, then it seems natural that we should relabel future 25% CDF values to be 30% instead. For the purposes of this paper, we term $F_t(x)$ the *raw* distribution, and $\hat{F}_t(x)$ the *calibrated* distribution.

For the set of probabilistic forecasts $\hat{F}_t(x)$ (for all $t \in \mathcal{T}$) to be calibrated, eq. (16) requires that $\Phi_{\mathcal{T}}(p)$ be generated as follows:

$$\Phi_{\mathcal{T}}(p) = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} H(p - F_t(x_t)). \tag{22}$$

This equation states that $\Phi_{\mathcal{T}}(p)$ is the empirical cumulative frequency distribution of the PIT values $F_t(x_t)$. This calibration function would generate perfectly reliable forecasts since we have invoked the definition of calibration directly in its formulation. However, since x_t is unknown to us when forecasting $\hat{F}_t(x)$, we must approximate $\Phi_{\mathcal{T}}(p)$ based on data accumulated during some previous time period $\mathcal{T}'$, known as the training period.

The approximation $\Phi_{\mathcal{T}} \approx \Phi_{\mathcal{T}'}$ is valid as long as the statistical properties of F do not change much between $\mathcal{T}$ and $\mathcal{T}'$ (i.e. between the actual forecast period and the training period); namely, the statistics are stationary.

We will denote *raw* PIT values originating from a raw forecast as p_t and *calibrated* PIT values from a calibrated forecast as $\hat{p}_t = \hat{F}_t(x_t)$. If the calibrated forecasts have been properly calibrated, the sorted $\hat{p}_t$ values will be distributed evenly on the interval [0, 1].

Combining eqs. (1) and (21) and using the chain rule, gives the following property for the calibrated PDF.

$$\hat{f}_t(x) = \Psi_{\mathcal{T}}(F_t(x)) f_t(x), \tag{23}$$

where we have defined $\Psi_{\mathcal{T}}(p)$ to be the derivative of the calibration function $\Phi_{\mathcal{T}}(p)$.

$$\Psi_{\mathcal{T}}(p) = \frac{\mathrm{d}\Phi_{\mathcal{T}}(p)}{\mathrm{d}p}. \tag{24}$$

$\Psi_{\mathcal{T}}(p)$ acts as an amplification function to the raw PDF. The calibrated PDF will have higher density in regions where $\Psi_{\mathcal{T}}(F_t(x)) > 1$ and lower density where $\Psi_{\mathcal{T}}(F_t(x)) < 1$.

Note that $\Psi_{\mathcal{T}}(p)$ is also the probability density function for observing a raw PIT value of p if the distribution of raw PIT values is stationary over time. This has the consequence that future PIT values are more likely to occur where the probability density of the calibrated forecast has been increased compared to the raw.

The basic calibration principles described earlier can be applied directly to unbounded continuous variables such as temperature. These same principles can be used for bounded variables, as described next.

4.2. Bounded mixed discrete-continuous distributions

Some variables, such as relative humidity, are bounded; that is, there are minimum and/or maximum values that the variables can take. Relative humidity for example has a minimum of 0% and maximum of 100%.[1]

Often these bounds represent values that have discrete probability. That is, they are values that can have non-zero probability within an infinitesimally narrow region. The finite probabilities at these points are called probability mass instead of probability density. Thus, variables such as relative humidity are best modelled by mixed discrete-continuous probability distributions

[1] Temperature is technically speaking also a bounded variable with a minimum of 0 K, but the commonly occurring temperature values are so far away from the boundary that it can be treated as an unbounded variable.

where finite probability masses are used at the bounds, and probability densities are used elsewhere.

Mixed discrete-continuous distributions can be devised that model this behaviour. Sloughter et al. (2007), for example, showed how mixed discrete-continuous distributions can be forecasted within the BMA framework (by separately modelling the discrete part and the continuous part of the distribution).

An alternative to modelling these boundaries is to use the aforementioned uncertainty models to generate CDFs that naturally spill over the boundaries. These distributions can be truncated at the boundaries so that the CDF is 0 below the bottom boundary and the probability mass at the lower boundary is set to the original CDF at the lower boundary. A similar treatment is performed on the upper boundary. The lower and upper boundaries are denoted by $x_{\min}$ and $x_{\max}$, respectively. The truncated CDFs $F(x)$ can be created from the original non-truncated distribution $F^*(x)$ as follows:

$$F(x) = \begin{cases} 0 & x < x_{\min} \\ F^*(x) & x_{\min} \leq x < x_{\max} \\ 1 & x_{\max} \leq x \end{cases} . \tag{25}$$

The PDF becomes

$$f(x) = \begin{cases} 0 & x < x_{\min} \\ F^*(x_{\min}) & x = x_{\min} \\ f^*(x) & x_{\min} < x < x_{\max} \\ 1 - F^*(x_{\max}) & x = x_{\max} \\ 1 & x_{\max} < x \end{cases} , \tag{26}$$

where the values at the boundaries are probability masses and the rest are densities.

This treatment of the boundaries may result in raw forecasts that are uncalibrated. However, by using the calibration method proposed in Section 4.1, the CDF can be adjusted so that even the CDFs that frequently lie on the boundaries become calibrated. In this sense, the calibration method can be used to create calibrated forecasts without having to determine a suitable model for the boundaries.

When generating the calibration function Φ for mixed discrete-continuous variables, care must be taken when the verifying value equals $x_{\min}$ or $x_{\max}$, since the PIT value is not uniquely defined. We follow the approach of Sloughter et al. (2007) by picking a random value on the intervals $[0, F(x_{\min})]$ and $[F(x_{\max}), 1]$ for each of these cases, respectively.

4.3. Implementation approach

There are a few issues that must be addressed when implementing the proposed calibration scheme. First, the distribution of past PIT values is subject to sampling errors. These sampling errors cause problems when evaluating Ψ, which is required when computing the PDF. The sampling errors are especially troublesome because a derivative is computed. For example, when two PIT values coincidentally are very close to one another, an unrealistic spike appears in Ψ. We have therefore used a smoothing technique on the calibration curve Φ. Greater smoothing reduces the chance of spikes in Ψ due to sampling, however increases the risk of removing real features in the calibration curve.

Cubic splines with nine points were used as this represents a good balance between representing features and smoothing out noise. An example of an initial cumulative distribution of 365 past PIT values from the MM method are shown in Fig. 3a. The points used for the spline were the lowest and highest PIT values as well as seven intermediate values distributed as evenly as possible through the sample.

Calibration curves for the BPE method often have sharp changes where $p = (K + 1)^{-1}$ and $p = K(K + 1)^{-1}$, as these correspond to the lower and upper boundary of the ensemble, respectively. To preserve this feature, a concatenation of three splines were used for calibrating BPE, where the slope of the splines are no longer forced to be continuous at the two boundary points between the three splines (Fig. 3b).

Other options for smoothing the calibration curve exist (such as simply resampling the curve combined with linear

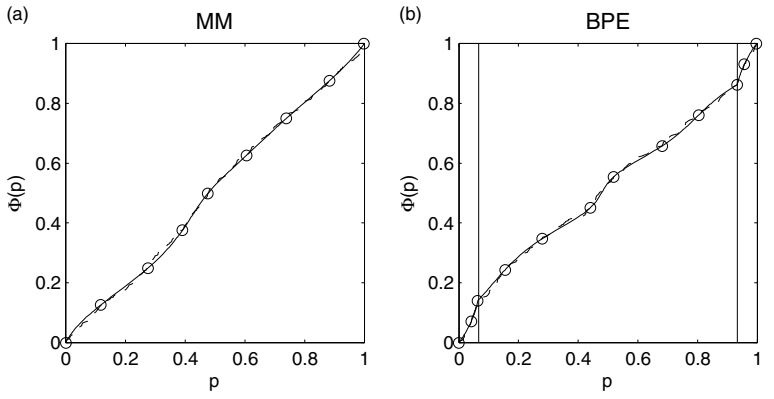

Fig. 3. Sample calibration curve for MM and BPE. Dashed lines show the actual cumulative distribution of PIT values, whereas the solid represent the smoothed curve using cubic splines. The circles represent the interpolation points for the splines. Three separate splines were used for BPE, where the separation is shown by the two vertical lines.

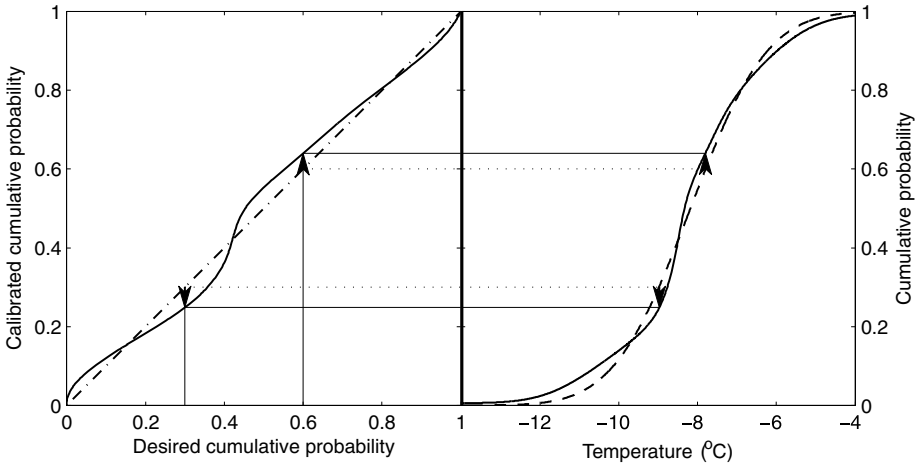

Fig. 4. Illustrative example of the calibration of a probabilistic forecast by the method presented in the text. The figure shows a probabilistic temperature forecast created using MM. The left half shows the calibration curve (solid line) and a one-to-one line (dash-dotted line). The right half shows the raw forecast (dashed line) and the calibrated forecast (solid line). The cumulative probabilities 0.3 and 0.6 are adjusted as shown by the thin solid lines and arrows. The horizontal dotted lines show the forecast without calibration adjustment.

interpolation) and will in general produce similar results. We chose the approach based on splines as we found this to be a stable way to generate a curve with continuous derivatives for a wide range of forecast variables.

A sliding window on the past data was used to empirically estimate the calibration curve Φ. For any given forecast day at a given location, all dates with available forecast and observation pairs for that location from the previous 365 d comprised the training period $\mathcal{T}'$.

Picking the training period for calibration should be a trade-off between capturing calibration deviations that vary in the short term and having enough data to robustly create the calibration. However, we have opted for a longer training period of 365 d as we found calibration curves based on much shorter training periods tended to overfit the calibration deviation. The optimal training period will likely depend on the application it is used for, but we have found that in general the performance is not very sensitive to its length provided that the training period consists of at least on the order of 100 past PIT values.

Figure 4 illustrates how a probabilistic temperature forecast is calibrated. The raw forecast (dashed line on the right) is adjusted to a calibrated forecast (thick solid line on the right) according to the calibration curve shown on the left.

4.4. Impact of calibration on verification metrics

Here we discuss the expected impact of the calibration scheme as evaluated using the metrics discussed in Section 3. First, the calibration scheme relabels CDF values such that future calibrated PIT values will be evenly distributed. We therefore expect the scheme to lower the calibration deviation D down to that expected for perfectly reliable forecasts $E[D_{\text{perfect}}]$.

Second, calibrating a forecast can also have benefits in terms of the ignorance score. Using eq. (23), the ignorance score of a set of raw forecasts $f = \{f_t \text{ for all } t \in \mathcal{T}\}$ can be decomposed into two terms as follows:

$$IGN(f) = IGN(\hat{f}) + \frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} \log\big(\Psi(p_t)\big). \tag{27}$$

The first term on the right-hand side is the ignorance score of perfectly calibrated forecasts $\hat{f}$, and the second term on the right-hand side is the extra ignorance caused by the lack of calibration. When a raw forecast is uncalibrated, and if the distribution of PIT values is stationary over time, then the right-most term will be positive. That is, the raw forecast will have a higher (worse) ignorance score than the calibrated forecast. This is because, as mentioned earlier, PIT values are more likely to fall where $\Psi(p)$ is greater than 1, since $\Psi(p)$ is also the probability density function for raw PIT values.

Reducing the ignorance score of f can be done by (1) improving the quality of the ensemble forecasts or using a more suitable uncertainty model, thereby reducing the first term on the right-hand side; (2) calibrating the forecast in a post-processing manner such as the calibration scheme presented, thereby reducing the last term.

For variables that are mixed discrete-continuous, one must compute the ignorance score differently for the discrete parts than for the continuous part. The probability mass is used in the calculation for the discrete parts, whereas the probability density is used for the continuous part. An overall ignorance score can still be computed as the sum of the discrete and the continuous ignorance scores, even though these represent the ignorance score for different probability entities. This may seem unintuitive at first, but since the score is logarithmic, any arbitrary weighting

between the probability entities will factor out as an additive constant. This additive constant cancels out when differences between ignorance scores are used.

4.5. Comparison with other calibration schemes

The BPE method by itself often produces unreliable probabilistic forecasts when the ensemble members and the observation are not drawn from the same distribution. Hamill and Colucci (1998) suggested a calibration scheme where the probability mass between each pair of consecutive ensemble members is adjusted by the frequency of historical observations falling in each bin. Eckel and Walters (1998) referred to this as the weighted ranks (WR) method. The CDF at each ensemble member is shifted to the frequency of historical observations that fall below that ensemble member rank. This WR calibration scheme is relevant only for the BPE uncertainty model as it makes adjustments based on ensemble counts and not on probabilities. The calibration scheme presented in this paper is a generalization of the WR scheme for any system that generates forecast probabilities, regardless if these were determined by ensemble ranks or otherwise.

Quantile-to-quantile mapping (Hopson and Webster, 2010) and similarly the bias-corrected relative frequency technique (Hamill and Whitaker, 2006) have been used to calibrate ensemble forecasts. Here, the value of each ensemble member $\xi_{t,k}$ is adjusted to new values $\hat{\xi}_{t,k}$, based on past statistics as follows:

$$\hat{\xi}_{t,k} = \mathcal{G}^{-1}\big(\mathcal{F}_k(\xi_{t,k})\big). \tag{28}$$

$\mathcal{G}$ and $\mathcal{F}_k$ are historical CDFs of the observations and kth ensemble forecasts respectively given by

$$\mathcal{G}(x) = \frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H(x - x_t), \tag{29}$$

$$\mathcal{F}_k(x) = \frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H(x - \xi_{t,k}), \tag{30}$$

where again $\mathcal{T}$ represents the training period, and where appropriate smoothing must be performed on $\mathcal{G}$ in order to make it invertible. The calibrated ensemble members will then have the same climatology as the observation and can then be used as input to a probabilistic method. The calibration method proposed in this paper differs from the quantile-to-quantile correction method in that it adjusts probabilities (output of an uncertainty model) instead of adjusting forecast values (inputs to an uncertainty model).

Finally, the concept of relabelling probabilities based on eq. (21) has been used in other forecasting studies (see e.g. Nielsen et al., 2006; Bremnes, 2007). The relabelling approach taken here differs in that sorted historical PIT values are used to create a non-parametric calibration curve Φ instead of using separate regression equations to calibrate each quantile of the forecast distribution.

5. Case-study data

To test the four different uncertainty models (BPE, MM, BMA and climatology) and the effect of the calibration method, we use data from the reforecast data set described in Hamill et al. (2006). This includes the forecasts from a 15-member ensemble using the NCEP's Medium Range Forecast (MRF) model as well as verifying analyses. The control forecast was removed and the remaining 14 bred members (which are assumed to be equally skillful) were used. We used an excerpt from the global grid centred on North America with 25 north–south points and 49 east–west points for a total of 1225 grid points, as shown in Fig. 5. The model was initialized at 00 UTC and forecasts for the 48-h offset were used. These were verified against the analysis valid at that time.

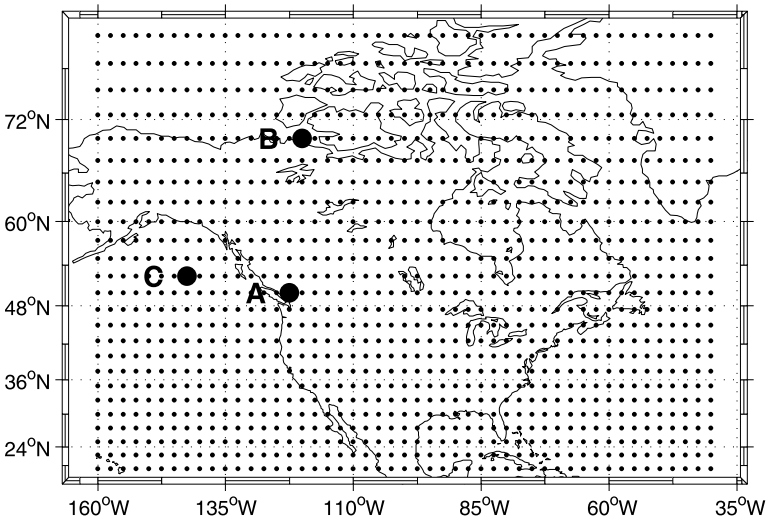

Fig. 5. Geographical locations of the 1225 grid locations used in the study. Point A represents the grid point nearest Vancouver, Canada, point B represents a grid point in the Northwest Territories, Canada, and point C represents a grid point at the Gulf of Alaska in the Pacific Ocean.

Five meteorological variables (with their abbreviations and units) were used: 2-m temperature (T2M, °C), mean sea level pressure (PRMSL, Pa), 10-m u-component of wind (U10M, m s^{-1}), precipitable water (PWAT, kg m^{-2}), and 70-kPa relative humidity (RHUM, %). We tested the raw versions of BPE, MM, BMA and climatology, as well as BPE, MM and BMA after the calibration scheme was applied. Daily data from runs initialized on 1 January 2001 to 31 December 2004 were used.

We used a 40-d sliding window to train the parameters for MM and BMA distributions, with each window ending prior to each forecast date. The parameters were computed separately for each grid location. Training periods of similar lengths have been used in other studies of BMA probabilistic forecasts (Raftery et al., 2005; Sloughter et al., 2007). For the calibration curve, raw PIT values from the 365 d prior to the forecast date were used. The 40-d sliding window and the 365 d of calibration required a warm-up period of 405 d, before the first forecasts for evaluation could be computed. A total of 1039 d of probabilistic forecasts for evaluation were produced.

Both MM and BMA bias correct the ensemble based on the training period. To get a fairer comparison, we also bias corrected each ensemble member for BPE using the same bias-correction method and sliding window approach as for BMA (see eq. 9).

For RHUM, to ensure that the bias correction in MM, BMA and BPE did not create impossible values, we truncated the values to be within 0% and 100%. Too low values were assigned the value 0%, and too high values were assigned 100%. Also, values above 99.9% were rounded to 100% and values below 0.01% were rounded to 0%.

As suggested by Hamill (2007), BMA weights for ensemble members that are assumed to be equally skillful can be constrained to be equal. This changes the weights w_k in eqs. (7) and (11) to be K^{-1}, but leaves the rest of the EM steps as is. The EM iteration was stopped when the largest change in the standard deviation $\sigma_{\mathcal{T}}$ was less than a tolerance of 10^{-4}, which resulted in around 20 iterations on average.

The 'refined' climatological forecasts for a given day were based on analyses that were within 15 d of the same day-of-year as the forecast day. For example the climatology for 15 April 2003 includes analyses from all of April 2001, 2002, 2003 and 2004. This means the climatology was produced in-sample, but since climatology is only used as a reference forecast to gauge the other methods, we hypothesize that this is acceptable. Separate climatologies were produced for each forecast grid point. The climatology was implemented by spreading a fixed Gaussian distribution across the range of the variable and then using the calibration method to adjust the probabilities. This was done to

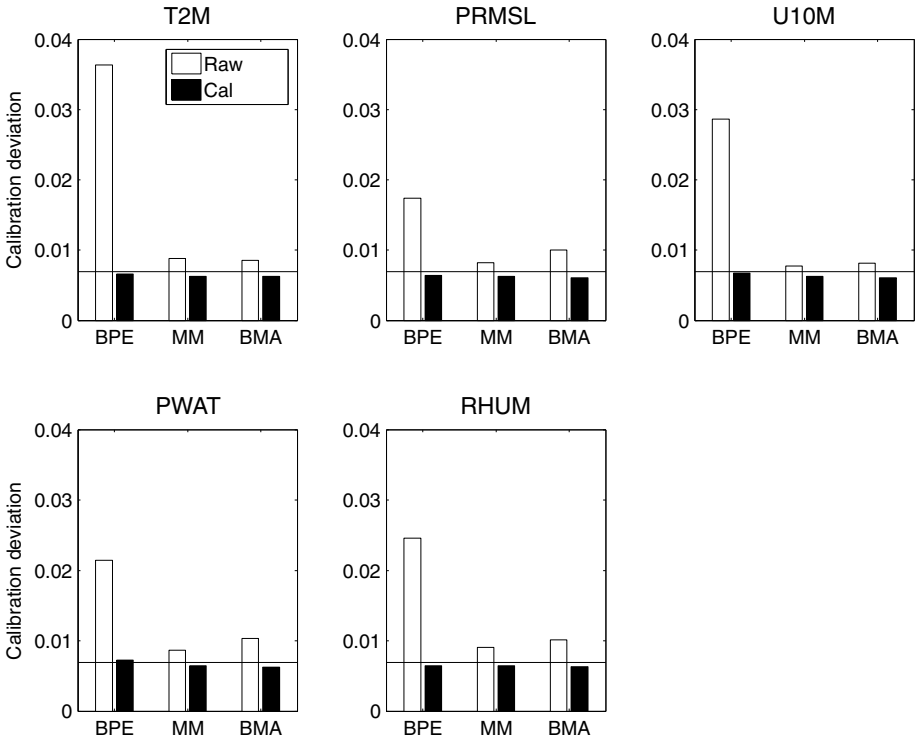

Fig. 6. Calibration deviation is shown for five forecast variables. Deviation of raw forecasts is shown by white bars, and deviation of calibrated forecasts is shown by black bars. The solid horizontal line shows the expected deviation of a perfectly calibrated forecast. T2M is 2-m temperature, PRMSL is mean sea level pressure, U10M is the 10-m u-component of the wind, PWAT is precipitable water and RHUM is 70-kPa relative humidity. Taller bars are indicative of forecasts exhibiting calibration deficiencies.

smooth the climatology, as the climatology is based on a finite sample of past values. Different smoothing approaches would likely give similar results.

For the BPE method, we used Gaussian distributions for $A(s)$ and $B(s)$ in eq. (4), with mean 0 and variance computed by

$$\sigma_{\mathcal{T}}^2 = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\bar{\xi}_t - \mu_{\mathcal{T}} - x_t)^2, \tag{31}$$

where $\mu_{\mathcal{T}}$ is computed by eq. (6). That is, we have used the second (central) moment of the forecast error of the bias-corrected ensemble mean to determine the drop off in probability outside the ensemble. The Guassian distributions must be multiplied by a factor of 2, so that $A(0)$ and $B(0)$ are 1. By using a function that stretches as the ensemble stretches for $A(s)$ and $B(s)$ we maintain the perfect spread-skill assumption that BPE already has for the interior of the ensemble.

With these data, we next evaluate the quality of the raw and calibrated probabilistic forecasts using the metrics from Section 3.

6. Results and analysis

6.1. General effects of the calibration

Figure 6 shows the calibration deviation for each variable (shown by different panels) and each uncertainty model (shown by each

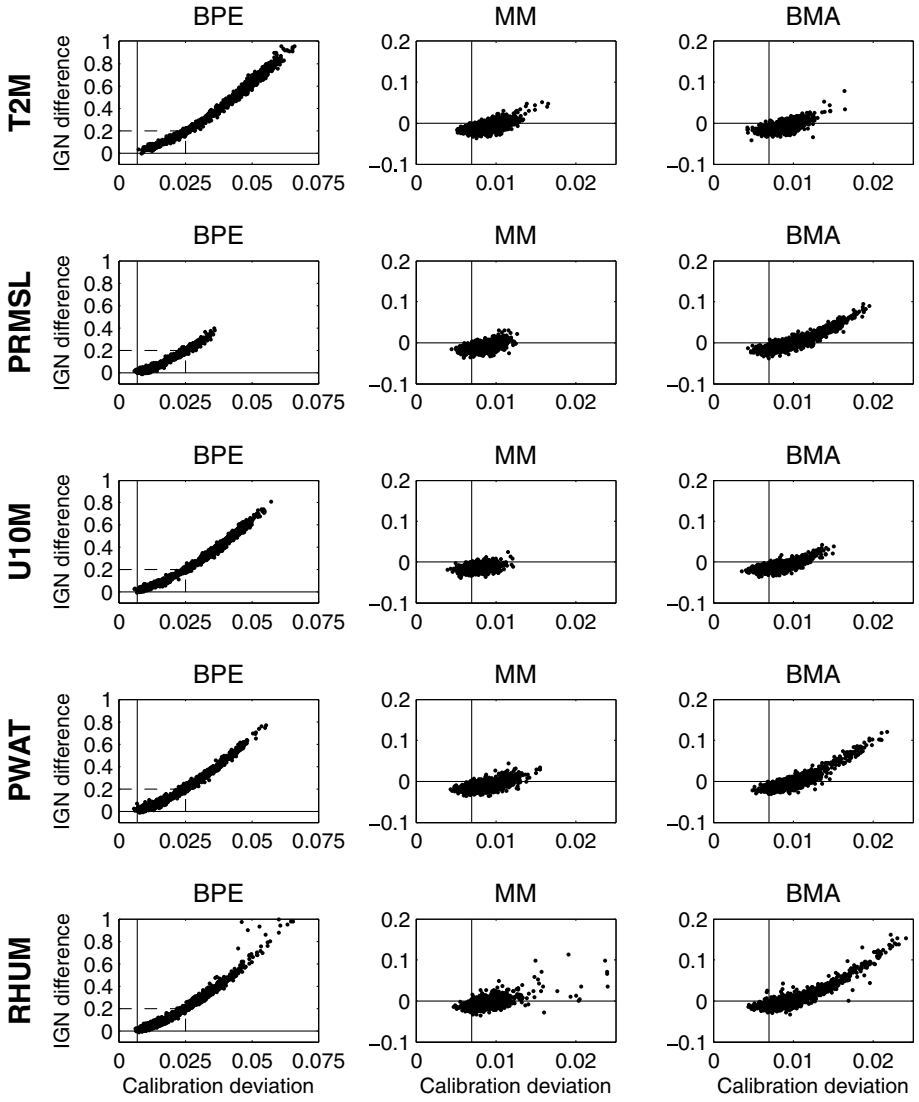

Fig. 7. The difference of ignorance scores between raw and calibrated forecasts is shown as a function of the calibration deviation of the raw forecast. Positive ignorance score differences indicate that the calibration method improves the ignorance score. Each dot represents a separate grid point from Fig. 5. Each row represents a variable and each column represents an uncertainty model. The vertical solid line represents the expected calibration deviation of perfectly calibrated forecasts. The dashed box in the left-most column shows the scale of the axes for the other two columns.

set of bars). Calibration deviation for the raw forecasts are shown by white bars, whereas those for which the calibration step has been applied are shown by black bars. The calibration deviation was computed for the 1039 forecast days separately for each grid location, and then averaged. The solid horizontal line indicates the expected deviation for perfectly calibrated forecasts as given by eq. (18). The figure shows that the raw forecasts have calibration deviations that are above that expected of perfectly calibrated forecasts. The calibration method reduces this deviation in all cases down to the level expected for perfectly calibrated forecasts. Also, the calibration deviation is much greater for the raw BPE forecasts than for MM and BMA. These results are evident for all five variables.

Figure 7 shows how the calibration method improves the ignorance score when the raw forecast exhibits calibration deviation. For cases where the calibration deviation of the raw forecast is high, the calibration method reduces the ignorance score significantly, as predicted by eq. (27). However, the calibration method actually increases the ignorance score slightly when the calibration deviation of the raw forecast is near that of perfectly calibrated forecasts (as seen by the dots below the horizontal line that are also close to the vertical line). This is because in these cases there is no calibration deficiency in the raw forecast for the calibration method to correct. The correction is then based on a calibration curve that has been fitted to a noisy signal of past PIT values. Ignorance is not reduced in this case, despite eq. (27), because the assumption of stationary PIT statistics no longer holds. For BPE, all variables show large potential for reducing the ignorance score through calibration. For MM, RHUM shows the greatest potential, and for BMA PRMSL, PWAT, and RHUM all have great potential for reducing the ignorance score via the proposed calibration method.

Figure 8 shows the average difference of the ignorance score of the uncertainty models compared to climatology. Positive values indicate that the probabilistic forecast has a lower (better) ignorance score than climatology. White bars show the difference of the raw forecasts to climatology whereas the black bars show the difference for calibrated forecasts. The figure shows that BPE yields smaller improvements over climatology when compared to MM and BMA. This is true for both the raw and calibrated forecasts. Black bars that are taller than their corresponding white bars indicate that the calibration method improved the ignorance score overall. For BPE, this is the case for all variables. Although improvements in the ignorance score were noted in Fig. 7 for MM and BMA for cases with high calibration deviation, the increase in the ignorance score for near-calibrated raw forecasts caused the average improvement of the ignorance score to be roughly negligible.

The use of the calibration method then relies on a priori identifying locations where calibration deficiencies are known to be present. For locations where the forecasts are already close to calibrated, the raw forecasts are best left unadjusted. We

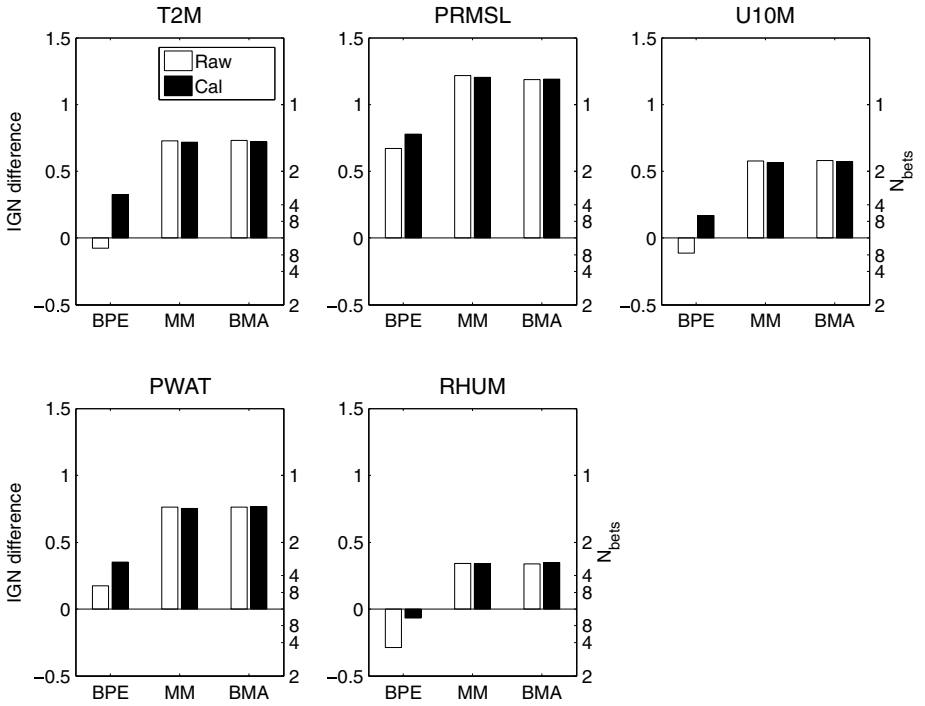

Fig. 8. Overall difference in ignorance scores between climatology and the uncertainty models is shown by the left axis. The expected number of bets required to double wealth when a forecast model is used against climatology is shown on the right axis. Differences to climatology are shown for raw forecasts (white bars) and calibrated forecasts (black bars).

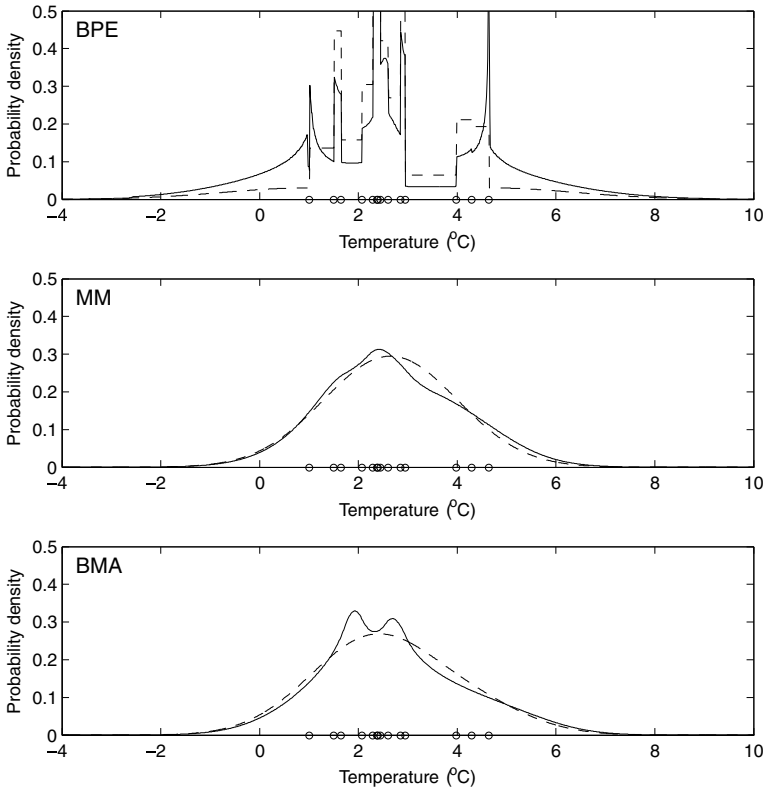

Fig. 9. Temperature PDF forecast for the Vancouver location for 3 January 2003 for BPE, MM and BMA uncertainty models. The raw forecast is shown by the dashed lines and the calibrated by the solid lines. Circles represent the bias-corrected ensemble members.

speculate that an alternative to using the calibration method for MM and BMA for cases with calibration deficiencies would likely be to find a non-Gaussian distribution that fits better. This distribution would also be tuned based on past statistics. However, the appropriate distribution would have to be determined for each location separately since different locations may have different types of calibration deficiencies. The calibration method on the other hand automatically determines a suitable fit to each location through the calibration curve.

6.2. *Performance of BPE*

A striking feature of Fig. 8 is that BPE gave forecasts with markedly larger ignorance scores than MM and BMA. Investigating the forecast PDFs reveals that BPE produces spikes of probability where two ensemble members are close in value. For example, Fig. 9 shows PDF forecasts for temperature on 3 January 2003 for location 'A' in Fig. 5. Large spikes in the BPE forecast are located where ensemble members are close for both raw (dashed lines) and calibrated (solid line) forecasts. These spikes are not present in MM and BMA.

The problem is that two close ensemble members are close only by coincidence and not because there is higher probability of observing a value in that region. That is, the spikes are unlikely to have any physical meaning and are purely a product of having a finite number of ensemble members that inadequately sample the true distribution.

This flaw can be traced to an underlying assumption behind BPE—that the observation rank is a random number between

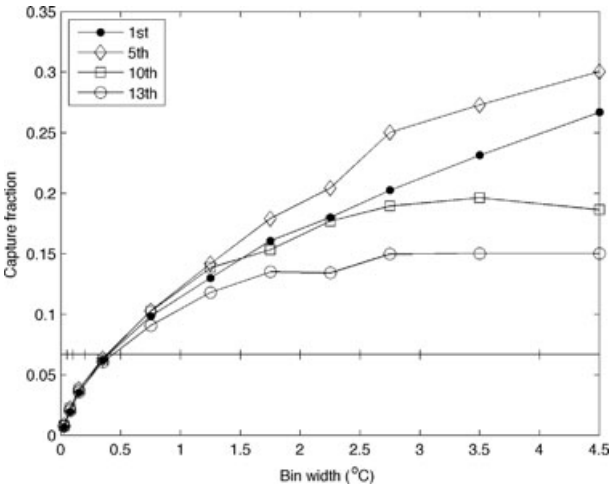

Fig. 10. Fraction of verifying temperature analyses captured by an ensemble bin as a function of bin width is shown for four different bins. The predicted capture fraction by BPE is shown by the horizontal solid line. The ticks along the horizontal line show the separations used when binning the bin widths.

1 and $K + 1$. This assumption is valid only prior to the instant when values of the ensemble are revealed. As soon as these values are known, however, the rank is no longer a random number. In general, members that are spaced *farther* apart will more likely capture the verifying value.

To test this assertion, the capture fractions of different pairs of ordered ensemble members for T2M as a function of their separation distance are shown in Fig. 10. Data from all grid locations and all available days were pooled together. BPE predicts a constant capture fraction of $(K + 1)^{-1}$ for every bin, shown by the horizontal line. However, the plot clearly shows that capture fraction increases with bin width. That is, when two ensemble members are spaced further apart, the likelihood of the analysis falling between them is higher. This causes the BPE technique to produce greater ignorance scores since narrow bins are given too high probability density despite their low probability of capture. Similarly, wider bins are given too low a probability density. Since the ignorance score is a proper skill score (Gneiting and Raftery, 2007), issuing a probability that we know a priori is biased will result in greater ignorance scores.

The calibration method lowers the calibration error compared to the raw BPE forecasts and thereby significantly improves the ignorance score. BPE exhibits calibration deficiencies in general because the analysis does not fall evenly between the ensemble members. However, further reduction in the ignorance score, closer to that of MM and BMA, is not possible since the calibration method cannot remove the spikes that BPE produces. For spikes to be removed, they would have to appear frequently enough in the same ensemble bin, such that the calibration function could identify that the CDFs associated with that bin happened too frequently.

6.3. *Examples of large calibration deviations*

Figure 11 shows the spatial pattern of average calibration deviation of raw and calibrated forecasts from MM for precipitable water. Figure 12 shows the same information for BMA. For a significant portion of the area, the calibration deviations for the raw forecasts are small, however there are large regions of large calibration deviation as shown by the darker colours. The

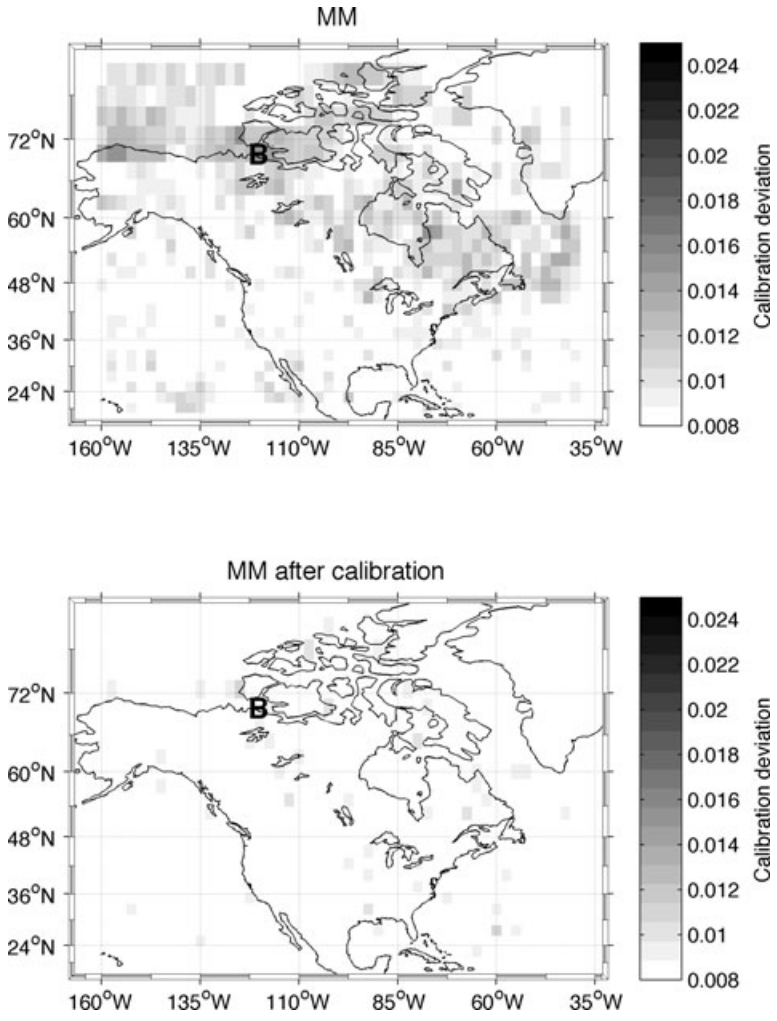

Fig. 11. Spatial pattern of calibration deviation for raw and calibrated forecasts from MM for precipitable water. Smaller calibration deviation is better. The letter 'B' is centred on the Northwest Territories location identified in Fig. 5.

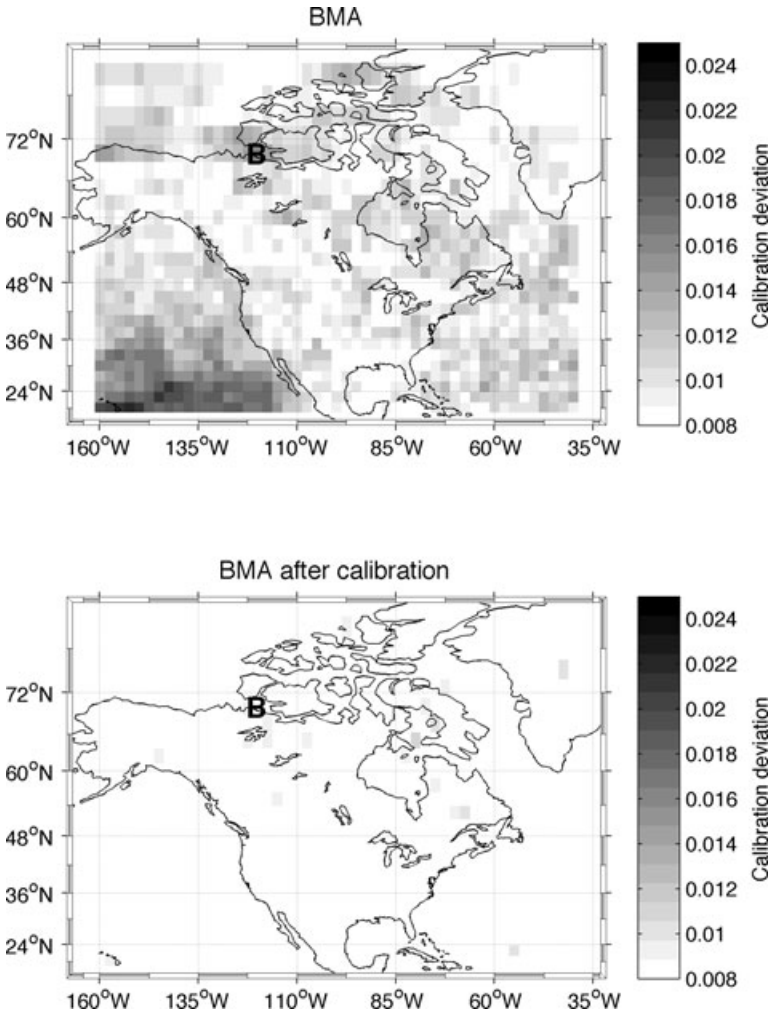

Fig. 12. Same as Fig. 11 except for BMA.

calibration deviation for the calibrated forecasts are all low. Both MM and BMA have large calibration deficiencies at the location marked by 'B', in the Northwest Territories, Canada.

The reasons for this can be diagnosed in Fig. 13, which shows PIT histograms for location 'B'. The raw BPE forecasts give distributions that are underdispersed as indicated by the high bin counts at the extremes. The raw MM forecasts have too many counts at the extremes and the middle, suggesting the Gaussian distribution with its one spread parameter cannot model a distribution with thicker tails, a taller middle and reduced probabilities elsewhere. The raw BMA forecasts have the same issue. The calibrated forecasts have smaller calibration errors, close to $E[D_{\text{perfect}}] = 0.0068$.

Figure 14 shows precipitable water PDF forecasts for 2 July 2002 for the same location as the PIT histogram in Fig. 13. The calibration method alters the shape of the raw PDF for both MM and BMA to be taller in the middle, have thicker tails and have lower probabilities elsewhere to correct the calibration deficiency. For BPE, the calibration increases the width of the tails and lowers the density in the middle.

Figure 15 shows relative humidity PDF forecasts for 16 May 2004 for the Pacific Ocean location marked by 'C' in Fig. 5. The probability mass at the boundaries are shown by the white bar for the raw forecast and by the black bar for the calibrated forecast, and uses the scale on the right-hand side. We again see that the calibration function changes the shape of the raw forecast distribution, including the probability mass at the upper boundary.

6.4. *Comparison between BMA and MM*

MM uses a simpler method to represent uncertainty than BMA. Unlike BMA, MM does not allow for multimodal probability distributions. Despite this, we found no large differences in the overall performance of these two methods. We speculate that the ability of the ensemble to correctly identify cases where multimodal uncertainty is appropriate was weak enough that

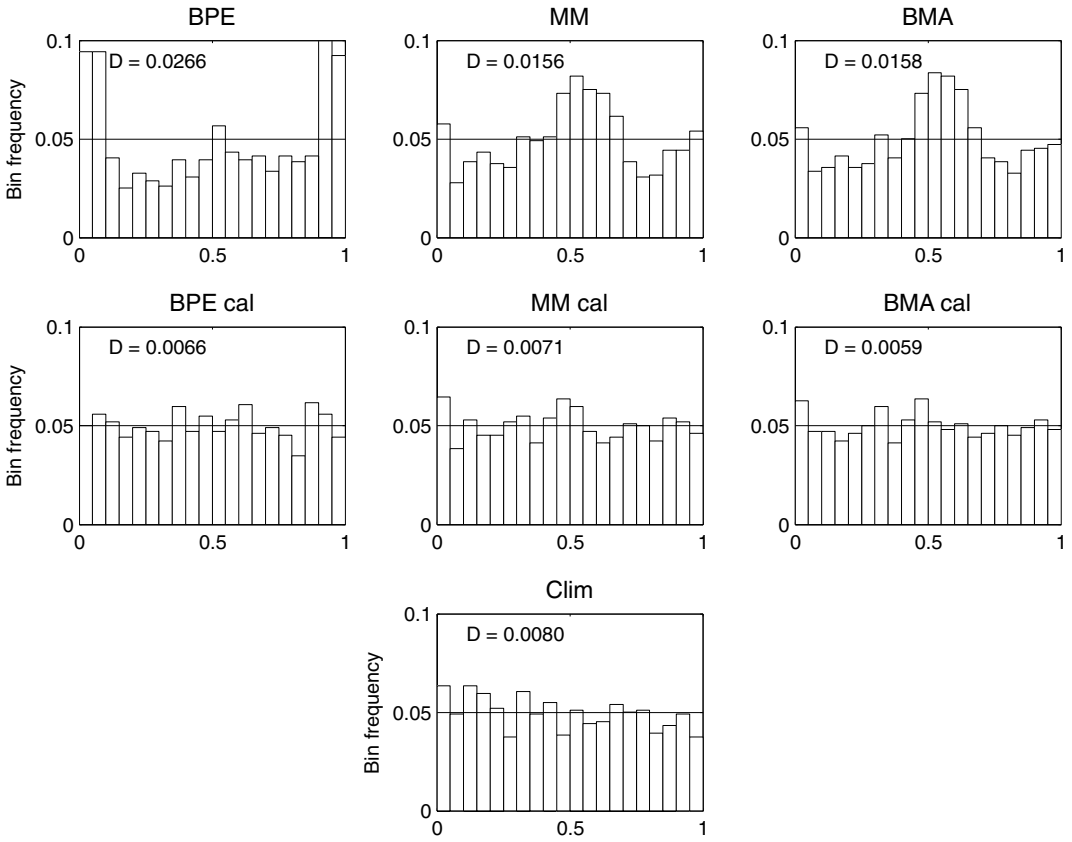

Fig. 13. PIT histogram of precipitable water forecasts for the Northwest Territories location evaluated for all 1039 forecast d. The calibration deviation score is indicated in the top of each histogram. D values of 0.0068 represent the expected level of calibration deviation for perfectly calibrated forecasts.

BMA could not take advantage of it. This may not necessarily be the case for other ensemble systems or forecast variables.

7. Conclusions and further work

We have presented a general approach for calibrating probabilistic forecasts of continuous variables and tested it on a data set with five variables, 1225 grid locations and an ensemble of 14 members that are assumed to be equally skillful. When trained with appropriate data, this method produces calibrated forecasts regardless of the underlying assumption of the uncertainty of the ensemble.

The method relabels the CDF values of an existing probability distribution according to eq. (21). The relabelling is done by the calibration curve given by eq. (22), which is based on which CDF values the past observations verified on. The calibration curve must be appropriately smoothed, such as by spline interpolation.

The method reduces calibration deviation down to the level expected by perfectly calibrated forecasts. When the deviation of the raw forecasts are large, the method significantly reduces the forecasts' ignorance score. The method can therefore yield benefits in both calibration and ignorance when the forecast location is known to have calibration deficiencies. Benefits in terms of calibration are due to adjustments made by the calibration curve Φ and benefits in terms of the ignorance score are due to adjustments made by the amplification factor Ψ. When the uncertainty model already produces calibrated forecasts, the redundant calibration step actually increases the ignorance score slightly due to the added overhead. In these cases, the original forecasts are best left unadjusted.

The quality of probabilistic forecasts is not only a function of the quality of the ensemble forecast used, but also a function of what uncertainty model is used. We found that, in general, BMA and MM produced forecasts with comparable ignorance scores, but both significantly outperformed forecasts produced by BPE, which is due to what we believe is a flaw in the uncertainty assumption in BPE.

Future work includes finding and evaluating new uncertainty models—not discussed here. Also, better smoothing mechanisms for the calibration curve may help reduce overfitting of the calibration method when the raw forecasts are already nearly calibrated. Finally, investigating the performance of the

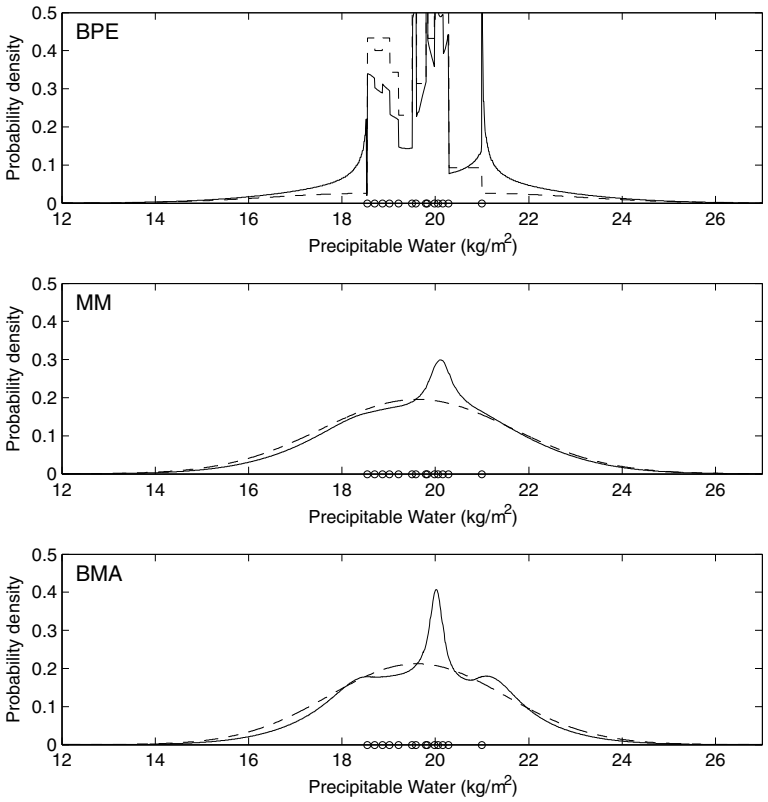

Fig. 14. Same as Fig. 9 except for precipitable water and for the Northwest Territories location for 2 July 2002.

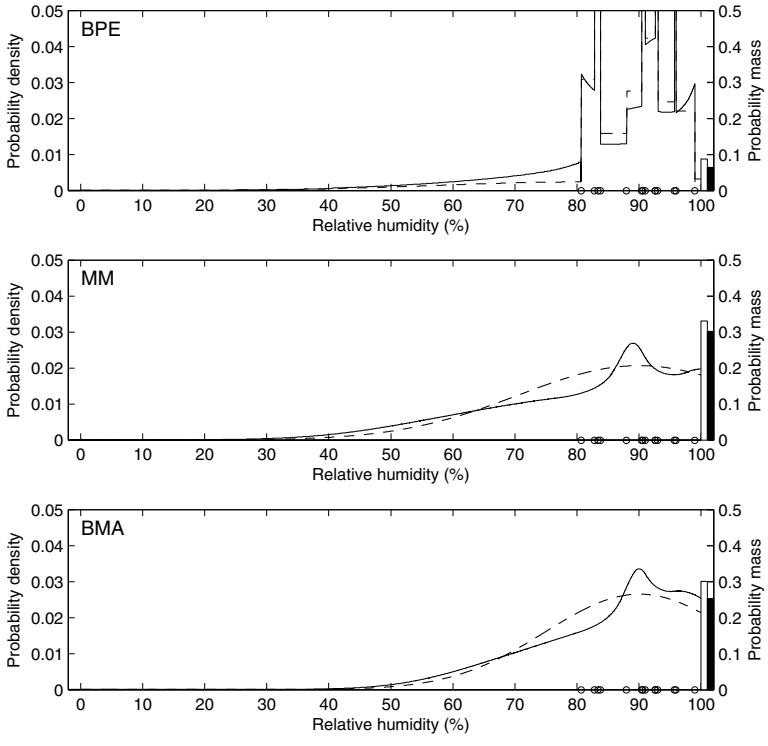

Fig. 15. Same as Fig. 9 except for relative humidity and for the Pacific Ocean location for 16 May 2004. White bars represent the probability mass assigned by the raw forecasts at 100% relative humidity, and black bars represent the same quantity for calibrated forecasts.

different uncertainty models for ensembles with members of unequal skill would be interesting.

8. Acknowledgments

This research was made possible by funding from the Canadian Natural Science and Engineering Research Council, the Canadian Foundation for Climate and Atmospheric Science, and the BC Hydro and Power Authority. We also thank May Wong, Doug McCollor, Greg West and two anonymous reviewers for their helpful comments and suggestions.

References

AMS 2008. Enhancing weather information with probability forecasts. *Bull. Am. Meteor. Soc.* **89**, 1049–1053.

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

Bremnes, J. B. 2007. Improved calibration of precipitation forecasts using ensemble techniques. Part 2: statistical calibration methods, Technical report, Norwegian Meteorological Institute.

Bröcker, J. and Smith, L. A. 2007. Increasing the reliability of reliability diagrams. *Wea. Forecasting* **22**, 651–661.

Eckel, F. A. and Walters, M. K. 1998. Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting* **13**, 1132–1147.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc., B* **69**, 243–268.

Gneiting, T. and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378.

Good, I. J. 1952. Rational decisions. *J. R. Stat. Soc., B* **14**, 107–114.

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

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.

Hamill, T. M. and Colucci, S. J. 1997. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.* **125**, 1312–1327.

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

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

Hamill, T. M., Whitaker, J. S. and Mullen, S. L. 2006. Reforecasts: an important dataset for improving weather predictions. *Bull. Am. Meteor. Soc.* **87**, 33–46.

Hoeting, J. A., Madigan, M., Raftery, A. E. and Volinsky, C. T. 1999. Bayesian model averaging: a tutorial. *Stat. Sci.* **14**, 382–401.

Hopson, T. M. and Webster, P. J. 2010. A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: forecasting severe floods of 2003–07. *J. Hydrometeor.* **11**, 618–641.

Jewson, S., Brix, A. and Ziehmann, C. 2005. *Weather Derivative Valuation* (eds. Jewson, S., Brix, A. and Ziehmann, C.). Cambridge, Cambridge University Press.

Johnson, C. and Swinbank, R. 2009. Medium-range multimodel ensemble combination and calibration. *Q. J. R. Meteor. Soc.* **135**, 777–794.

Murphy, A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteor.* **12**, 595–600.

Nielsen, H. A., Nielsen, T. S., Madsen, H., Giebel, G., Badger, J. and co-authors. 2006. From wind ensembles to probabilistic information about future wind power production: results from an actual application, In: *9th International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden.

Pinson, P., McSharry, P. and Madsen, H. 2010. Reliability diagrams for non-parametric density forecasts of continuous variables: accounting for serial correlation. *Q. J. R. Meteor. Soc.* **136**, 77–90.

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

Roulston, M. S. and Smith, L. A. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.* **130**, 1653–1660.

Sloughter, J. M., Raftery, A. E. and Gneiting, T. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.* **135**, 3209–3220.

Talagrand, O., Vautard, R. and Strauss, B. 1997. Evaluation of probabilistic prediction systems, In: Proc. ECMWF Workshop on Predictability, Reading, United Kingdom, ECMWF, pp. 1–25.

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