A component-based probabilistic weather forecasting system for operational usage

by

Thomas N. Nipen

B.Sc. Hons. Computer and Atmospheric Science, The University of British Columbia, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES
(Atmospheric Science)

The University of British Columbia
(Vancouver)

April 2012

Abstract

This dissertation presents a probabilistic weather prediction system for operational (real-time) usage. The proposed system provides complete probability distributions for both continuous weather variables, such as temperature, and mixed discrete-continuous variables like precipitation accumulations.

The proposed system decomposes the process of generating probabilistic forecasts into a series of sequential steps, each of which is important in the overall goal of providing probabilistic forecasts of high quality. Starting with an ensemble of input predictors generated by numerical weather prediction models, the system uses the following four components: 1) correction; 2) uncertainty modeling; 3) calibration; and 4) updating. The correction component bias-corrects the input predictors. The uncertainty model converts these predictors into a suitable probability distribution. The calibration component improves this distribution by removing any distributional bias. The update component further improves the forecast by incorporating recently made observations of the true state.

The system is designed to be modular. Namely, different implementations of each component can be used interchangeably with any combination of implementations for the other components. This allows future research into probabilistic forecasting to be focused on any one component and also allows new methods to be easily incorporated into the system.

The system uses a number of existing correction and uncertainty models, but the dissertation also presents two new methods: Firstly, a new method for calibrating probabilistic forecasts is created. This method is shown to improve probabilistic forecasts that exhibit distributional bias. Secondly, a new method for incorporating recently made observations to existing probabilistic forecasts is developed.

The system and its components are tested using meteorological data from daily operational runs of ensemble numerical weather prediction models and their verifying observations from surface weather stations in North America. Each component's contribution to overall forecast quality is analysed.

Preface

With the exception of the introductory and concluding chapters, this dissertation is composed entirely of work from the following two published and one submitted journal manuscripts. The material in these articles have been reformatted to conform to the dissertation formating requirements. A small number of editing changes have also been made, but the content is otherwise unaltered.

Chapter 2

T. N. Nipen and R. Stull. Calibrating probabilistic forecasts from an NWP ensemble. *Tellus*, **63**:858–875, 2011. Copyright 2011 Swedish Geophysical Society, published by Blackwell Publishing.

The need for calibrating probabilistic forecasts was identified by Dr. Stull, and the project was collaboratively designed by T. N. Nipen and Dr. Stull. T. N. Nipen conducted the research, analysed the results, and wrote the original journal manuscript, with editing provided by Dr. Stull. Research funding for T. N. Nipen was provided by Dr. Stull.

The work in this paper started as an extension to T. N. Nipen's undergraduate honours thesis titled "A percentile calibration method for probabilistic weather forecasts". However, the method used and its evaluation differ substantially from the original work.

Chapter 3

T. N. Nipen, G. West, and R. Stull. Updating short-term probabilistic weather forecasts of continuous variables using recent observations. *Wea. Forecasting*, **24**:564–571, 2011. Copyright 2011 American Meteorological Society.

T. N. Nipen identified the need for updating probabilistic forecasts, and designed the research project with input from Dr. West. T. N. Nipen conducted the research, analysed the research data, and wrote the manuscript for publication, with editing provided by Dr. Stull and Dr. West. Research funding for T. N. Nipen and Dr. West was provided by Dr. Stull.

Chapter 4

T. N. Nipen and R. Stull. A modular operational probabilistic weather forecasting system. Submitted for publication on 20. Feb 2012.

The co-author contributions were as follows: T. N. Nipen identified the need for this research topic, with input from Dr. Stull. T. N. Nipen designed and conducted the research, analysed the research data, and wrote the manuscript for publication, with editing provided by Dr. Stull. Research funding for T. N. Nipen was provided by Dr. Stull.

Table of contents

List of tables

List of figures

Mathematical symbols and abbreviations

Symbols

Below is a summary of all mathematical variables used throughout the dissertation:

	Lowercase Greek
α	Gamma shape parameter (Chapter 4)
β	Gamma scale parameter (Chapter 4)
δ	Dirac delta function (Chapter 3)
δ	Fraction of ensemble members predicting no precipitation (Chapter 4)
ϵ	Threshold for occurrence of precipitation (Chapter 4)
μ	Mean
μ_i	Member-specific mean (Chapter 4)
$\mu_{\mathcal{T},k}$	Bias-correction term for ensemble member k (Chapter 2)
ϕ	PDF of a Gaussian distribution
σ	Standard deviation
σ_0^2	Variance of PIT step sizes (Chapter 3)
$\sigma_{\hat{\xi}}^2$	Variance of corrected ensemble
$\sigma_{t,\hat{t}}^2$	Variance of PIT step sizes for run index t and offset index $\hat{t}$ (Chapter 4)
τ	E-folding length (Chapter 4)
θ	Parameter (Chapter 4)
θ_t	Parameter for time t (Chapter 4)
θ_t^*	Parameter for time t (Chapter 4)
$\boldsymbol{\theta}$	Vector of parameter (Chapter 4)
ξ	Ensemble mean (Chapter 4)
ξ_i	Ensemble member i (Chapter 4)
ξ_t	Ensemble mean at time t (Chapter 4)
$\xi_{t,k}$	Ensemble member k at time t (Chapter 2)
$\hat{\xi}$, $\hat{\xi}_i$, $\hat{\xi}_t$	Corrected ensemble member/mean (Chapter 4)
$\bar{\xi}_t$	Ensemble mean at time t (Chapter 2)

Uppercase Greek

Γ	CDF of a Gamma function (Chapter 4)
Φ	Calibration function
Φ	CDF of a Gaussian distribution (Chapter 3)
Φ_0	Calibration function for discrete part of distribution (Chapter 4)
Φ_n	Calibration function n hours after most recent PIT value (Chapter 3)
Φ_p	Calibration point (Chapter 4)
Ψ	Amplification function (Chapter 2)
Ψ_n	Amplification function n hours after most recent PIT value (Chapter 2)

Calligraphic

$\mathcal{F}_k$	Historical CDF of kth ensemble member
$\mathcal{G}$	Historical CDF of observations
$\mathcal{N}$	Gaussian distributed
$\mathcal{T}$	Set of time points
$\mathcal{T}_0$	Set of time points for cases where no precipitation occurred (Chapter 4)
$\lvert\mathcal{T}\rvert$	Size of set $\mathcal{T}$ (Chapter 4)
$\lVert\mathcal{T}\rVert$	Size of set $\mathcal{T}$ (Chapter 2)

Lowercase Roman

a, b, c_0, c_1, c_2, c_3	Constants (Chapter 4)
$a_\mathcal{T}, b_\mathcal{T}$	Constants (Chapter 2)
b_i	Number of PIT values in bin i (Chapter 2)
$f_t(x)$	Probability density function at time t
$f_{t,k}(x)$	BMA ensemble member PDF (Chapter 2)
f^*	Non-truncated PDF (Chapter 2)
$\hat{f}_t(x)$	Calibrated PDF (Chapter 2)
$^{(j)}$	Value at iteration j (Chapter 2)
n	Number of hours since a recent observation was made
p_t	PIT value at time t
$p_{t,\hat{t}}$	PIT value for run index t and offset $\hat{t}$ (Chapter 4)
s_t^2	Ensemble variance (Chapter 2)
t	Time point
t_{obs}	Time point when most recent observation was made (Chapter 3)
$\hat{t}$	Time point within a certain forecast run (Chapter 4)

x	Value of weather variable
x_{max}	Upper boundary of variable (Chapter 2)
x_{min}	Lower boundary of variable (Chapter 2)
x_t	Observation at time t
w_k	BMA weight (Chapter 2)
z_k	BMA intermediate value (Chapter 2)

Uppercase Roman

$A(s)$	CDF for below the ensemble (Chapter 2)
B	Number of PIT-histogram bins (Chapter 2)
$B(s)$	CDF for above the ensemble (Chapter 2)
$BS(0\,\mathrm{mm})$	Brier score with threshold $0\,\mathrm{mm}$ (Chapter 4)
CRPS	Continuous ranked probability score
D	Calibration deviation (Chapter 2)
$D_{perfect}$	Calibration deviation of perfectly calibrated forecasts (Chapter 2)
$E[]$	Expectation operator (Chapter 2)
$F_t(x)$	Probabilistic forecast at time t in CDF form
$\hat{F}_t(x)$	Calibrated probabilistic forecast in CDF form (Chapter 2 and Chapter 4)
$\hat{F}_t(x)$	Updated probabilistic forecast in CDF form (Chapter 3)
$F_c(x)$	CDF of the continuous portion of a mixed discrete-continuous distribution (Chapter 4)
$F_{t\|t-n}(x)$	Updated probabilistic forecast in CDF form (Chapter 4)
F_{clim}	Climatological CDF (Chapter 2)
F_{raw}	Forecast CDF before updating (Chapter 3)
$F_{updated}$	Forecast CDF after updating (Chapter 3)
F^*	Non-truncated CDF (Chapter 2)
$G_t(x)$	CDF of observation for time t
$H(s)$	Heaviside function
IGN	Ignorance score
IGN$(0\,\mathrm{mm})$	Ignorance score based on binary variable with threshold $0\,\mathrm{mm}$ (Chapter 4)
IGN$_{ref}$	Ignorance score of reference forecast (Chapter 2)
IGN$_{pot}$	Potential ignorance score (Chapter 4)
IGN$_{uncal}$	Added ignorance score due to calibration deviation (Chapter 4)
K	Number of ensemble members (Chapter 2)
L	Likelihood function (Chapter 4)

MAE	Mean absolute error
N	Number of ensemble members (Chapter 4)
N_{bets}	Number of bets (Chapter 2)
P	Probability mass (Chapter 4)
$\boldsymbol{R}$	Covariance matrix (Chapter 4)
$S(p, q)$	Transition function (Chapter 3)
U	Update function (Chapter 4)
X	Weather variable

Abbreviations

AMS	American meteorological society
BMA	Bayesian model averaging
BPE	Binned probability ensemble
BS	Brier score
CDF	Cumulative distribution function
CMC	Canadian meteorological centre
CRPS	Continuous ranked probability score
EM	Expectation maximization
EnKF	Ensemble Kalman filter
EPS	Ensemble prediction system
GFS	Global forecast system
MAE	Mean absolute error
MAXT	24-h maximum temperature
MC2	Mesoscale compressible community model
MINT	24-h minimum temperature
MM	Method of moments
MM5	Penn state/NCAR mesoscale model
MOS	Model output statistics
MRF	NCEP medium range forecast
NAEFS	North American ensemble forecasting system
NAM	North American mesoscale model
NCEP	National centers for environmental prediction
NWP	Numerical weather prediction

OO	Object oriented
PCP	24h accumulated precipitation
PDF	Probability density function
PIT	Probability integral transform
PRMSL	Mean sea-level pressure
PWAT	Precipitable water
RHUM	Relative humidity
RUC	Rapid update cycle
SREF	Short-range ensemble forecasts
THOUR	Hourly temperature
UBC	University of British Columbia
WFRT	Weather forecast research team
WR	Weighted ranks
WRF	Weather research and forecasting model
WRFG	WRF with GFS initialization
WRFN	WRF with NAM initialization

Acknowledgments

Firstly, I would like to acknowledge the endless support and guidance of my supervisor, Dr. Roland Stull. His enthusiasm for weather research has been incredibly inspiring, resulting in a truly enjoyable research experience. I would also like to express my thanks to the other members of my supervisory committee, Dr. Susan Allen and Dr. Steve Wilton, for their input and expertise.

I am grateful to all my wonderful colleagues at the Weather Forecast Research Team (WFRT). In particular Atoossa Bakhshaii, George Hicks, Dr. Henryk Modzelewski, and May Wong were responsible for the daily operational runs of the UBC forecasting system, which provided input data for this dissertation. I also appreciate the support, friendliness, and helpful comments and suggestions on my work by Dominique Bourdin, Daniel Casanova, Rosie Howard, Bruce Thomson, and Katelyn Wells.

I thank Dr. Doug McCollor and Dr. Greg West at BC Hydro for providing very valuable input on the requirements from an end user's perspective, ensuring that the research in this dissertation can easily be applied in an operational settings.

I thank John Michalakes (NREL) and Dr. Luca Delle Monache (NCAR/RAL) for the opportunity to visit NCAR on research exchanges. I thank Dr. Luca Delle Monache and Dr. Thomas Hopson (NCAR/RAL) for interesting discussions on calibrating probabilistic forecasts, and Kristian Soltesz for his expertise on system control theory inspiring work on updating probabilistic forecasts. I am also grateful for the many helpful comments and suggestions by several anonymous reviewers and editors, which resulted in improvements to the journal manuscripts.

Finally, I thank my partner Louise as well as my parents and siblings for support and inspiration throughout my university studies.

Funding for this research was provided by the Canadian Natural Science and Engineering Research Council, the Canadian Foundation for Climate and Atmospheric Science, and the BC Hydro and Power Authority. Computational resources were provided by the Geophysical Disaster Computational Fluid Dynamics Centre at the University of British Columbia, which was funded in part by the Canadian Foundation for Innovation.

Chapter 1

Introduction

1.1 The need for probabilistic forecasts

Weather forecasts are typically stated in deterministic form. That is, forecasts are given by a single value representing the forecaster or model's best estimate of the weather in the future. An example is "the overnight low temperature for tonight will be $4\,^{\circ}\mathrm{C}$" (Figure 1.1).

However, no estimate is complete without an estimate of its error (Hirschberg et al., 2011). Even after removing known systematic biases, deterministic forecasts are rarely perfectly accurate. The accuracy of a deterministic weather forecast depends greatly on factors such as forecast lead-time, location, season, and the availability of a dense observing network nearby. For example, the forecast error of modern numerical weather prediction (NWP) models generally increases with increasing lead-time to the point where the forecast no longer provides better guidance than climatological values (Figure 1.2).

Indicating the amount of forecast uncertainty is therefore important since it greatly affects the end user's confidence in, for example, the occurrence or non-occurrence of freezing temperatures in Figure 1.1. Forecast uncertainty can be expressed by a probability distribution, which (unlike a deterministic forecast) indicates the likelihood of occurrence of each temperature value.

For example, suppose an orchard owner learns the low is forecasted to be $4\,^{\circ}\mathrm{C}$ overnight. The owner might take no precautions, but could lose $200,000 in ruined fruit if the forecast is wrong and the low is actually $-1\,^{\circ}\mathrm{C}$. But suppose the owner could spend $2,000 running orchard fans, smudge pots, or water sprays to prevent damage to the fruit crop. Should the owner spend the $2,000 in preventative costs, knowing that there is a large chance that the low temperature will remain above freezing, but a small non-zero chance that the temperature could be below freezing? This is the motivation for probability forecasts, which would allow the orchard owner to make cost-loss decisions (Murphy, 1977; Richardson, 2000) that minimize her expenditures over the course of many possible freeze events.

The move towards providing weather forecasts in probabilistic form is endorsed by the American Meteorological Society (AMS, 2008), and the potential value that these forecasts can provide

has been well documented (Richardson, 2000; Palmer, 2000; Zhu et al., 2002). In fact, probabilistic weather forecasts have been applied in a wide variety of applications such as hydroelectric power management (McCollor and Stull, 2008b), road maintenance applications (Berrocal et al., 2010), and visibility at airports (Chmielecki and Raftery, 2010).

This dissertation focuses on improving probabilistic forecasts that are based on the output of NWP model runs.

1.2 Current probabilistic forecasting practices

In this section the current methods and approaches to probabilistic forecasting are reviewed, with a focus on the use of statistical methods.

1.2.1 Statistical post-processing

It is important to note the distinction between physical and statistical research in meteorology. Improvements in NWP forecasts typically result from research at two fronts: physical and statistical. Physical improvements are due to the development of physics-based models that better describe how the atmosphere behaves. Statistical methods, which is the focus of this dissertation, improve forecasts by recognizing statistical relationships between forecasts and observations.

NWP models frequently exhibit biases due to the limited resolution of the discretized grid, systematic errors in initialization and boundary conditions, or problems with the physical parameterizations used (Eckel and Mass, 2005). Models are often found to exhibit systematic biases for certain locations or under certain weather conditions. For example, in a case study of forecasting for the 2002 Winter Olympics, Hart et al. (2004) found that surface temperatures were consistently overpredicted during cold-pool events, due to the model's difficulty in simulating the strength of the cold pools. Also, in mountainous terrain, the elevation of the observing station may be significantly different than the modeled (smoothed) terrain height, resulting in surface-temperature biases.

These and other model biases can be corrected by employing statistical post-processing methods such as model output statistics (MOS; Glahn and Lowry, 1972), Kalman filtering (Homleid, 1995), neural networks (Yuval and Hsieh, 2002; Marzban, 2003), analog methods (Delle Monache et al., 2011), and gene-expression programming (Bakhshaii and Stull, 2009). These methods improve forecasts by removing the systematic error based on historical comparisons between forecasts and observations.

1.2.2 Ensemble forecasting

NWP models stray from reality due to the limited resolving ability of the model (discretization error), errors in the initialization and boundary conditions, and error in the physics parameterizations used. To specify forecast uncertainty one must account for these errors. For simple dynamical systems, a specified error distribution can explicitly be evolved forward in time by the continuity equation for probabilities (Liouville equation; Ehrendorfer, 1994), for example by using the stochastic dynamic prediction approach of Epstein (1969).

However, evolving forward such a distribution is computationally prohibitive for an NWP model with millions of variables, and therefore ensemble methods (Leith, 1974) are used instead. Ensemble forecasting samples the error distribution by using a finite number of ensemble members and then evolves each of them forward in time. If the ensemble members are sampled from the true probability density function (PDF) of the error distribution, then each member represents an equally likely evolution of the atmosphere. Provided also that enough ensemble members are used, the spread (or disagreement) among the members is indicative of the uncertainty of the forecast.

Ensembles of NWP-model runs are typically created by perturbing initial conditions (Molteni and Palmer, 1993; Toth and Kalnay, 1993), using several model runs with different model physics (Krishnamurti et al., 1999), or some combination of both. Due to the chaotic nature of the atmosphere (Lorenz, 1963), these initially similar ensemble members eventually diverge over time.

To get probability information from the ensemble, the binned probability ensemble (BPE) technique (Anderson, 1996) is often used (see for example Hamill and Colucci, 1998). When ensemble members are assumed to be a random sample from the same distribution as the verifying observation, the cumulative probability for a given threshold can be determined by the fraction of ensemble members that are below this threshold.

Ensemble forecasts often suffer from two major problems. Firstly, ensembles are often found to be underdispersive (Hamill and Colucci, 1998; Buizza et al., 2005; Raftery et al., 2005). That is, the observation verifies outside the ensemble range more often than would be expected of an ensemble that samples the error distribution perfectly. Secondly, correctly sampling the error distribution implies the existence of a spread-skill relationship. That is, the spread of the ensemble should be related to the accuracy (or skill) of the mean of the ensemble. However, the value of the ensemble spread as a predictor of forecast skill has been mixed, with some studies showing little or no value (Hamill and Colucci, 1998; Stensrud et al., 1999) and others showing some value (Grimit and Mass, 2002; Stensrud and Yussouf, 2003; Scherrer et al., 2004).

1.2.3 Probabilistic methods

These deficiencies of ensembles have led to the development of statistical methods that do not require the ensemble members to sample the true PDF of errors. Instead, probability distributions (such as a Gaussian distribution) are used, where the parameters of these distributions are adjusted based on empirical relationships found between various attributes of the ensemble and the verifying observations.

Ensemble MOS methods (EMOS; Gneiting et al., 2005) or moment-based methods (Jewson et al., 2005) fit Gaussian distributions by performing linear regression on empirical moments of the error of the ensemble mean. These methods can account for underdispersion or overdispersion of the ensemble by adjusting the variance of the Gaussian distribution. They can also account for the strength of the spread-skill relationship, by using the coefficients found from regression between ensemble variance and ensemble mean error.

Another popular method is Bayesian model averaging (BMA; Hoeting et al., 1999), which has been introduced in the weather prediction field by Raftery et al. (2005). BMA fits weighted distributions to each ensemble member and combine these via Bayes theorem to form a total distribution (see Figure 1.3). BMA has been used successfully to produce probabilistic forecasts for a variety of meteorological variables such as sea-level pressure (Raftery et al., 2005), precipitation (Sloughter et al., 2007), surface temperature (Wilson et al., 2007), and recently visibility (Chmielecki and Raftery, 2010).

Just like the deterministic post-processing methods of Section 1.2.1, these probabilistic methods improve forecasts through statistical means, as opposed to through improved physical modeling.

1.2.4 Evaluating probabilistic forecasts

To evaluate probabilistic forecasts, the correspondence between forecasts and observations are investigated. A large variety of metrics are available, but for probabilistic forecasts, two commonly used metrics include the ignorance score (Good, 1952; Roulston and Smith, 2002) and the continuous ranked probability score (CRPS; Hersbach, 2000). The ignorance score uses the negative logarithm of the PDF corresponding to the observation (Figure 1.4a) and therefore rewards forecasts that place high probability density at the value of the observation. The CRPS is sensitive to the area under the curve in Figure 1.4b, and rewards forecasts that are sharp (narrow) and are centred near the observation.

The PIT-histogram (Gneiting et al., 2005) is another commonly used tool to assess the quality of probabilistic forecasts, which looks at the statistical consistency between forecasts and observations. The probability integral transform (PIT) value is the cumulative probability corresponding to the

observation (Figure 1.4c). A flat PIT-histogram is indicative of evenly distributed PIT values and is a desired attribute referred to as calibration (or reliability). These three metrics will be heavily used to evaluate the probabilistic forecasts in this dissertation.

1.3 Dissertation contributions

The overall goal of this dissertation is to improve probabilistic weather forecasts through the use of statistical methods. To achieve this, new methods are developed and evaluated. These contributions are discussed in more detail next.

1.3.1 Probabilistic calibration

The first contribution of this dissertation is a new calibration scheme presented in Chapter 2.

Calibration refers to the statistical consistency between forecast probabilities and observations. For example, if a set of events are predicted to have a 20% probability of occurrence and 20% of observations confirm the occurrence of the event, the forecasts are said to be *calibrated*. Calibrated probabilities are essential for making informed, risk-based decisions.

For ensemble forecasts, calibration refers to the case when an equal number of observations fall between each pair of consecutive ensemble members. As ensembles typically are underdispersed, probabilities produced by the BPE technique (as was described in Section 1.2.2) can be calibrated by the weighted ranks (WR) method (Hamill and Colucci, 1998; Eckel and Walters, 1998), where probabilities are adjusted based on the rank histogram (Anderson, 1996; Talagrand et al., 1997). Calibration can also be achieved by altering the ensemble members, instead of altering the resulting probabilities (Hamill and Whitaker, 2006; Hopson and Webster, 2010).

Raftery et al. (2005) suggested using BMA as a calibration method for underdispersed forecasts. The variance of the BMA forecast is greater than the variance of the ensemble because, in addition to the between-forecast variance provided by the spread of ensemble members, BMA includes a within-forecast variance term in its formulation for each individual ensemble member (Raftery et al., 2005).

The aforementioned calibration methods operate on a set of ensemble members. A new calibration method is devised in Chapter 2 that instead operates on existing probability distributions. The method ensures calibrated results by removing any distributional bias that the existing probabilistic forecast may have. It is effective in cases such as when a Gaussian (i.e. not-skewed) distribution model is used for cases where the actual error distribution is skewed. Calibrating probability distributions instead of ensemble members has the advantage that it allows for the separation of modeling uncertainty from the aspect of calibrating.

The calibration method corrects probabilistic forecasts by ensuring the uniformity of verifying PIT values (Gneiting et al., 2007), and is analogous to correcting non-uniform rank histogram of ensemble forecasts by the WR method. As the calibration method operates only on the probability distribution, it is therefore independent of the construction of the ensemble.

The calibration method is shown to generate uniform PIT-histograms for a variety of forecast variables. In addition, as a byproduct of improving calibration, the calibration method is shown to reduce the ignorance score for forecast distributions that exhibit distributional bias.

1.3.2 Statistical data assimilation for probabilistic forecasts

The second contribution of this dissertation is a new statistical data assimilation scheme for probabilistic forecasts.

To avoid NWP models straying from reality over time, newly made observations must be used to correct the model's state. This is referred to as the data assimilation cycle, and involves the ingestion of large amounts of observed data from remote sensors, such as satellite and radar, as well as in-situ measurements from aircrafts, ships, buoys, ground-based stations, radiosondes, and dropsondes.

Observations are made continuously, but the data assimilation cycle is typically only performed several times a day, such as every 6h for the global forecast system (GFS), but can be as frequent as every hour as with the rapid update cycle (RUC; Benjamin et al., 2004).

There are several techniques used to assimilate observations into weather models, such as the ensemble Kalman filter (EnKF; Evensen, 1994), variational data assimilation (Lewis and Derber, 1985), and Newtonian relaxation (Anthes, 1974). These methods alter the modeled state of the atmosphere throughout the whole model grid and must ensure that the model maintains dynamic balance, such that unrealistic instabilities are not created.

These assimilation methods provide updated initialization and boundary conditions for the NWP model. The model must then be evolved forward in time again with this new data. To acquire updated probabilistic forecasts the following three-step process would be required: assimilating new observation data into the model initialization, rerunning the ensemble, and regenerating the probabilistic forecasts. This is computationally very expensive and generally not worthwhile if only small amount of new recent data is assimilated.

An alternative is to use only recent observations recorded at the forecast location of interest and directly alter the forecast distribution, without simulating forward the ensemble. To my knowledge there are no such methods currently developed for weather prediction purposes. In Chapter 3, I present such a method, which relies on the verifying PIT values being correlated in time. For example, if the observed state recently verified in the 20th percentile of the distribution, observations in the near future are likely to continue to verify near this percentile. Thus, in the short-term the

forecast distribution can be sharpened significantly. This technique has the advantage that it is computationally much less expensive than the conventional three-step process, and is also much simpler to implement than complex data assimilation schemes.

The method improves the CRPS and the ignorance score of the probabilistic forecasts. Updating probabilistic forecasts can therefore be considered to be another class of methods for improving probabilistic forecasts.

1.3.3 Decomposition of the probabilistic forecasting process

The factors affecting weather are complex. To deal with this complexity, NWP models typically separate various aspects of atmospheric modeling into independent components. Each component can have several alternative implementations called schemes. For example, version 3 of the Weather Research and Forecasting (WRF; Skamarock et al., 2005) model has four alternative shortwave radiation schemes, three longwave radiation schemes, nine microphysics schemes, three surface layer schemes, four land-surface schemes, four boundary layer schemes, and four cumulus parameterization schemes.

With some exceptions, any combination of schemes for each physics category can be used together to form a model configuration. The construction of the configuration is important as a scheme is often optimized for a geographical region, for capturing specific weather phenomena, or for computational speed.

This decomposition is useful for two important reasons: 1) It reduces complexity, as a scheme needs only to model a small subset of atmospheric physics; 2) It allows combinations of schemes to be used, so that the best combinations of schemes can be used for the user's particular forecasting purpose. This decomposition is enforced by the software framework (Michalakes et al., 1999; Skamarock et al., 2005), which specifies the input and output requirements of each component.

Such a decomposition currently does not exist for generating probabilistic forecasts. The third contribution of this dissertation is devising such a decomposition, which is presented in Chapter 4. It conceptually represents a statistical analog to the physically-based decomposition used in NWP modeling. The decomposition includes the two improvement methods from Chapter 2 and Chapter 3.

Chapter 4 will show how the process of producing probabilistic forecasts can be decomposed into four independent components: 1) correction; 2) uncertainty model; 3) calibration; and 4) updating. The correction component bias-corrects the ensemble forecasts. The uncertainty model transforms this set of corrected ensemble members into a probability distribution. The calibration component removes any distributional bias from this probability distribution. Finally, the updating component improves the calibrated distribution by incorporating information from any recently

made observations.

The advantage of viewing probabilistic forecasting in light of this decomposition is that it allows improvement efforts to be focused into independent areas. Also, as with the WRF model, various combinations of schemes can then be tested together to find the optimal combination for a particular use case.

To test the usefulness of the decomposition, a probabilistic forecasting system is implemented, which includes three correction schemes, nine uncertainty models, one calibration scheme, and one update scheme. The implementation is based on an object-oriented programming approach, enabling sufficient abstraction between the components and also allowing for the interchangeability of schemes.

The contributions of the various components to probabilistic-forecast quality are evaluated using the CRPS, the ignorance score, and the PIT-histogram. Forecast data from both short-range and medium-range ensemble prediction systems (EPS) are used for evaluation.

1.4 Dissertation layout

This dissertation uses a manuscript-based format, where the three core chapters are published or submitted journal manuscripts. The material in these articles have been reformatted to conform to the dissertation formating requirements. With the exception of a few minor editing changes, the content is otherwise unaltered.

Chapter 4 presents the proposed system for producing high-quality probabilistic forecasts for operational use. This work has been submitted for peer-review. This system relies on two new probabilistic methods: Chapter 2 presents a new calibration method for reducing distributional bias of probabilistic forecasts, which has been published in Nipen and Stull (2011); Chapter 3 presents a new method for updating probabilistic forecast given recently made observations, which has been published in Nipen et al. (2011). Chapter 5 summarizes the contributions of this dissertation and provides recommendations for future work.

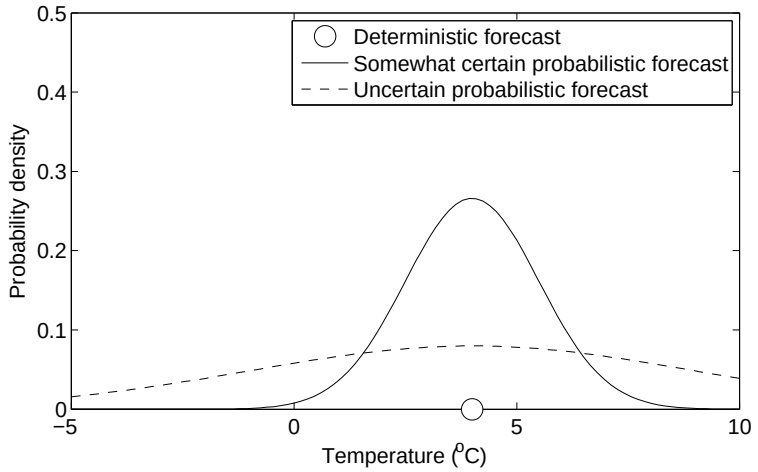

Figure 1.1: A sample weather forecasting scenario showing a deterministic forecast of $4\,^{\circ}\mathrm{C}$ and probability distributions for one somewhat certain and one uncertain probabilistic forecast.

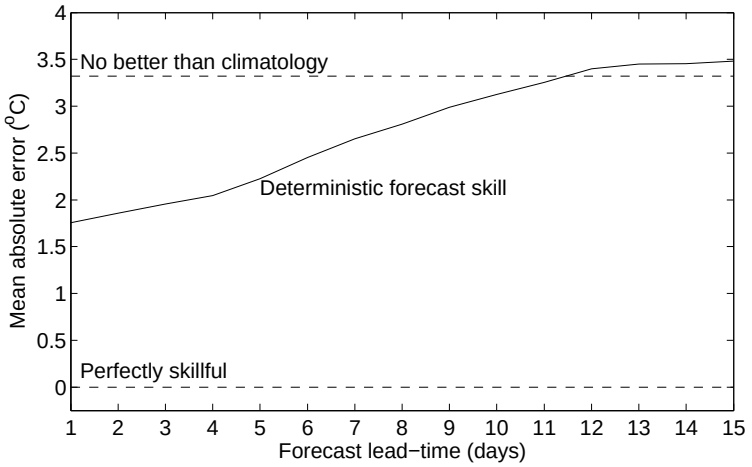

Figure 1.2: Mean absolute error for the bias-corrected NAEFS ensemble mean, averaged over a collection of 15 stations in mountainous Western Canada for the time period 1 Nov 2010 to 31 Aug 2011.

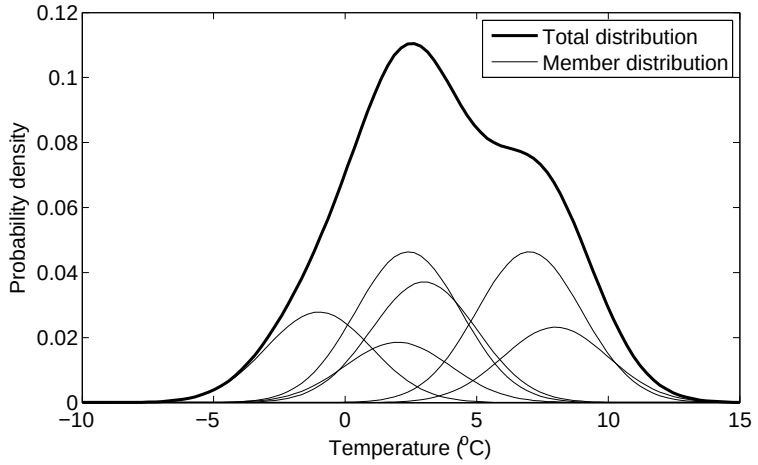

Figure 1.3: Example probabilistic forecast produced by Bayesian model averaging, showing individual member distributions (thin lines) and the total distribution (thick line).

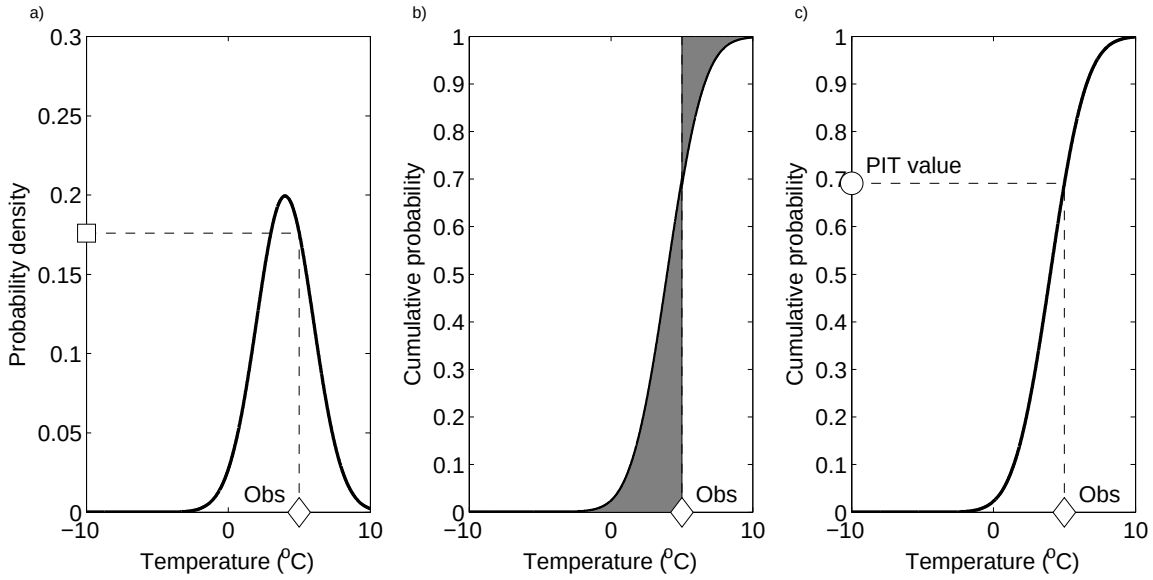

Figure 1.4: Example probabilistic forecast with an observation (diamond) of $4\,^{\circ}\mathrm{C}$. a) Probability density plot showing the verification value used for the ignorance score (square). b) Cumulative probability plot showing the area used in the calculation of the continuous ranked probability score (gray shading). c) Cumulative probability plot showing the corresponding probability integral transform (PIT) value for the observation (circle).

Chapter 2

Calibrating probabilistic forecasts from an NWP ensemble

2.1 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). But 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 chapter we take the view that creating probabilistic forecasts follows a two-step process, as shown in Figure 2.1. 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 modeled in much the same way that radiation or precipitation is modeled 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 centered 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 forecast. Requiring a probabilistic method to model the uncertainty and ensure probabilistic calibration simultaneously can therefore be avoided.

The goal of this chapter 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 chapter, 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 chapter is organized as follows: First, we summarize some of the ways to represent uncertainty. Next, in Section 2.3, 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 2.4 we present the proposed calibration method. Section 2.5 describes case-study data from a four-year period with five forecast variables, a 14-member ensemble, and 1225 grid locations. Those case-study data will be used in Section 2.6 to evaluate the calibration method for the uncertainty models from Section 2.2. Implications of this approach are summarized in Section 2.7.

2.2 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 un-

certainty 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

$$F_t(x) = \int_{-\infty}^{x} f_t(s)ds. \tag{2.1}$$

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:

$$G_t(x) = H(x - x_t), \tag{2.2}$$

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

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases}. \tag{2.3}$$

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 variance is denoted by s_t^2.

2.2.1 Binned probability ensemble

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 binned probability ensemble (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:

$$F_t(x) = \begin{cases} \frac{1}{K+1} A(\xi_{t,1} - x) & x \leq \xi_{t,1} \\ \frac{k}{K+1} + \frac{1}{K+1} \frac{x - \xi_{t,k}}{\xi_{t,k+1} - \xi_{t,k}} & \xi_{t,k} < x \leq \xi_{t,k+1} \\ 1 - \frac{K}{K+1} B(x - \xi_{t,K}) & \xi_{t,K} < x \end{cases}, \tag{2.4}$$

where $\xi_{t,k}$ represents the k^{th} 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.

2.2.2 Method of moments

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

$$F_t \sim \mathcal{N}(\bar{\xi}_t - \mu_{\mathcal{T}}, a_{\mathcal{T}} s_t^2 + b_{\mathcal{T}}). \tag{2.5}$$

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:

$$\mu_{\mathcal{T}} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\bar{\xi}_t - x_t). \tag{2.6}$$

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).

2.2.3 Bayesian model averaging

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 Bayesian model averaging (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 centered 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:

$$F_t(x) = \sum_{k=1}^{K} w_k F_{t,k}(x), \tag{2.7}$$

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

$$F_{t,k}(x) \sim \mathcal{N}(\xi_{t,k} - \mu_{\mathcal{T},k}, \sigma_{\mathcal{T}}^2) \tag{2.8}$$

$$\mu_{\mathcal{T},k} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\xi_{t,k} - x_t). \tag{2.9}$$

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 maxi-

mization (EM) algorithm, an iterative process given by:

$$z_{t,k}^{(j)} = \frac{w_k^{(j-1)} f_{t,k}^{(j-1)}(x_t)}{\sum_{i=1}^{K} w_i^{(j-1)} f_{t,i}^{(j-1)}(x_t)} \tag{2.10}$$

$$w_k^{(j)} = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} z_{t,k}^{(j)} \tag{2.11}$$

$$\sigma_{\mathcal{T}}^{2(j)} = \frac{1}{\|\mathcal{T}\|} \sum_{k=1}^{K} \sum_{t \in \mathcal{T}} z_{t,k}^{(j)} (x_t - \xi_{t,k} - \mu_{\mathcal{T},k})^2 \tag{2.12}$$

$$f_{t,k}^{(j)}(x) \sim \mathcal{N}(\xi_{t,k} + \mu_{\mathcal{T},k}, \sigma_{\mathcal{T}}^{2(j)}), \tag{2.13}$$

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.

2.2.4 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:

$$F_{\text{clim}}(x) = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} H(x - x_t). \tag{2.14}$$

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 year as the desired forecast time point. We will use this more refined climatology as our baseline.

2.2.5 Comparison of these uncertainty models

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

- BPE: Fixing a constant probability mass between each pair of consecutive ranked ensemble members.

- MM: fixing a shape function to the bias-corrected ensemble mean.
- BMA: fixing a shape function to each bias-corrected ensemble member.
- Climatology: fixing a constant-in-time shape function directly onto forecast-variable x.

Figure 2.2 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 multi-modal distributions due to the individual Gaussian distributions, however, compared to BPE, its peaks are less sensitive to the exact positions of the ensemble members.

2.3 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.

2.3.1 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 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

$$\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} G_t\big(F_t^{-1}(p)\big) = p, \tag{2.15}$$

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. (2.2), Eq. (2.15) can be rewritten to show that probabilistic forecasts are

calibrated if

$$\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H\big(p - p_t\big) = p. \tag{2.16}$$

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. (2.16) 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. (2.2), 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:

$$D = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{b_i}{\|\mathcal{T}\|} - \frac{1}{B}\right)^2}. \tag{2.17}$$

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:

$$E[D_{\text{perfect}}] = \sqrt{\frac{1 - B^{-1}}{\|\mathcal{T}\|B}}. \tag{2.18}$$

2.3.2 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:

$$\text{IGN} = -\frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} log_2(f_t(x_t)), \tag{2.19}$$

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:

$$N_{\text{bets}} = \frac{1}{\text{IGN}_{\text{ref}} - \text{IGN}_A}, \tag{2.20}$$

provided that $\text{IGN}_A < \text{IGN}_{\text{ref}}$. 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.

2.4 Calibration method

Section 2.2 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.

2.4.1 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:

$$\hat{F}_t(x) = \Phi_{\mathcal{T}}\big(F_t(x)\big). \tag{2.21}$$

$\Phi_{\mathcal{T}}$ 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 chapter, 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. (2.16) requires that $\Phi_{\mathcal{T}}(p)$ be generated as follows:

$$\Phi_{\mathcal{T}}\big(p\big) = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} H\big(p - F_t(x_t)\big). \tag{2.22}$$

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 Eq. (2.1) and Eq. (2.21) and using the chain rule, gives the following property for the calibrated PDF:

$$\hat{f}_t(x) = \Psi_{\mathcal{T}}\big(F_t(x)\big) f_t(x), \tag{2.23}$$

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

$$\Psi_{\mathcal{T}}(p) = \frac{d\Phi_{\mathcal{T}}(p)}{dp}. \tag{2.24}$$

$\Psi_{\mathcal{T}}(p)$ 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 above can be applied directly to unbounded continuous variables such as temperature. These same principles can be used for bounded variables, as described next.

2.4.2 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 modeled by mixed discrete-continuous probability distributions 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 modeling the discrete part and the continuous part of the distribution).

An alternative to modeling 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:

$$F(x) = \begin{cases} 0 & x < x_{\min} \\ F^*(x) & x_{\min} \leq x \leq x_{\max} \\ 1 & x_{\max} < x \end{cases} . \tag{2.25}$$

The PDF becomes:

$$f(x) = \begin{cases} 0 & x < x_{\min} \\ F^*(x_{\min}) & x = x_{\min} \\ f^*(x) & x_{\min} < x < x_{\max} \\ 1 - F^*(x_{\max}) & x = x_{\max} \\ 1 & x_{\max} < x \end{cases} , \tag{2.26}$$

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 2.4.1, 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

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

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.

2.4.3 Implementation approach

There are a few issues that must be addressed when implementing the proposed calibration scheme. Firstly, 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 Figure 2.3a. 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 (Figure 2.3b).

Other options for smoothing the calibration curve exist (such as simply resampling the curve combined with linear interpolation) and will in general produce similar results. We chose the approach based on splines as we found this to be 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 days 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 days 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 2.4 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.

2.4.4 Impact of calibration on verification metrics

Here we discuss the expected impact of the calibration scheme as evaluated using the metrics discussed in Section 2.3. 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. (2.23), 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:

$$\text{IGN}(f) = \text{IGN}(\hat{f}) + \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} log\big(\Psi(p_t)\big). \tag{2.27}$$

The first term on the right is the ignorance score of perfectly calibrated forecasts $\hat{f}$, and the second term on the right 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.

2.4.5 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 chapter 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:

$$\hat{\xi}_{t,k} = \mathcal{G}^{-1}\big(\mathcal{F}_k(\xi_{t,k})\big). \tag{2.28}$$

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

$$\mathcal{G}(x) = \frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H(x - x_t) \tag{2.29}$$

$$\mathcal{F}_k(x) = \frac{1}{\|\mathcal{T}\|}\sum_{t\in\mathcal{T}} H(x - \xi_{t,k}), \tag{2.30}$$

where again $\mathcal{T}$ represents the training period, and where appropriate smoothing must be performed on 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 chapter 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. (2.21) has been used in other forecasting studies (see for example Bremnes, 2007; Nielsen et al., 2006). 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.

2.5 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 dataset described in Hamill et al. (2006). This includes the forecasts from a 15-member ensemble using the NCEP 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 centered on North America with 25 north-south points and 49 east-west points for a total of 1225 grid points, as shown in Figure 2.5. The model was initialized at 00 UTC and forecasts for the 48-hour offset were used. These were verified against the analysis valid at that time.

Five meteorological variables (with their abbreviations and units) were used: 2-m temperature (T2M, $^{\mathrm{o}}\mathrm{C}$), mean sea-level pressure (PRMSL, Pa), 10-m u-component of wind (U10M, $\mathrm{m\,s^{-1}}$), precipitable water (PWAT, $\mathrm{kg\,m^{-2}}$), 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 through 31 December 2004 were used.

We used a 40-day 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 days prior to the forecast date were used. The 40-day sliding window and the 365 days of calibration required a warm-up period of 405 days, before the first forecasts for evaluation could be computed. A total of 1039 days 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. (2.9)).

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 Eq. (2.7) and Eq. (2.11) to be K^{-1}, but leaves the rest of the expectation maximization (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 days 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 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. (2.4), with mean 0 and variance computed by:

$$\sigma_{\mathcal{T}}^2 = \frac{1}{\|\mathcal{T}\|} \sum_{t \in \mathcal{T}} (\bar{\xi}_t - \mu_{\mathcal{T}} - x_t)^2, \tag{2.31}$$

where $\mu_{\mathcal{T}}$ is computed by Eq. (2.6). 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 Gaussian 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 2.3.

2.6 Results and analysis

2.6.1 General effects of the calibration

Figure 2.6 shows the calibration deviation for each variable (shown by different panels) and each uncertainty model (shown by each 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. (2.18). 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 2.7 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. (2.27). 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. (2.27), 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 2.8 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 Figure 2.7 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 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.

2.6.2 Performance of BPE

A striking feature of Figure 2.8 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, Figure 2.9 shows PDF forecasts for temperature on 3 January 2003 for location "A" in Figure 2.5. 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 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 Figure 2.10. 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.

2.6.3 Examples of large calibration deviations

Figure 2.11 shows the spatial pattern of average calibration deviation of raw and calibrated forecasts from MM for precipitable water. Figure 2.12 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 colors. The 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 Figure 2.13, 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 2.14 shows precipitable water PDF forecasts for 2 July 2002 for the same location as the PIT histogram in Figure 2.13. 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 2.15 shows relative humidity PDF forecasts for 16 May 2004 for the Pacific Ocean location marked by "C" in Figure 2.5. 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.

2.6.4 Comparison between BMA and MM

MM uses a simpler method to represent uncertainty than BMA. Unlike BMA, MM does not allow for multi-modal 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 multi-modal uncertainty is appropriate was weak enough that BMA could not take advantage of it. This may not necessarily be the case for other ensemble systems or forecast variables.

2.7 Conclusions and further work

We have presented a general approach for calibrating probabilistic forecasts of continuous variables and tested it on a dataset with 5 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. (2.21). The relabeling is done by the calibration curve given by Eq. (2.22), 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 increase 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 different uncertainty models for ensembles with members of unequal skill would be interesting.

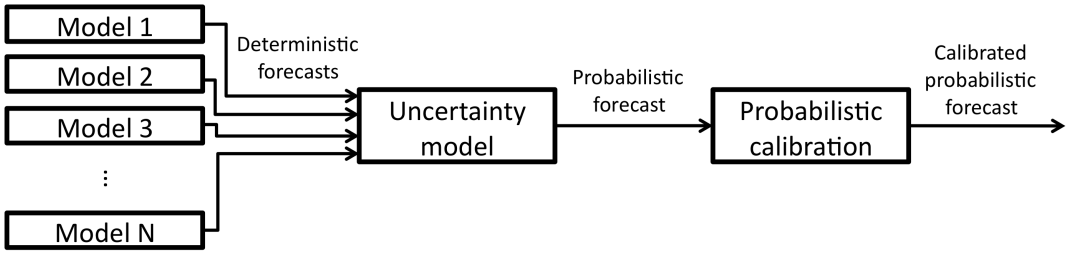

Figure 2.1: A two-step process of generating probabilistic forecasts from an ensemble of forecasts. A set of deterministic forecasts from weather models feed into a system that models how uncertainty is conveyed by the ensemble. The resulting probabilistic forecast is fed to a calibration scheme that generates calibrated probabilistic forecasts.

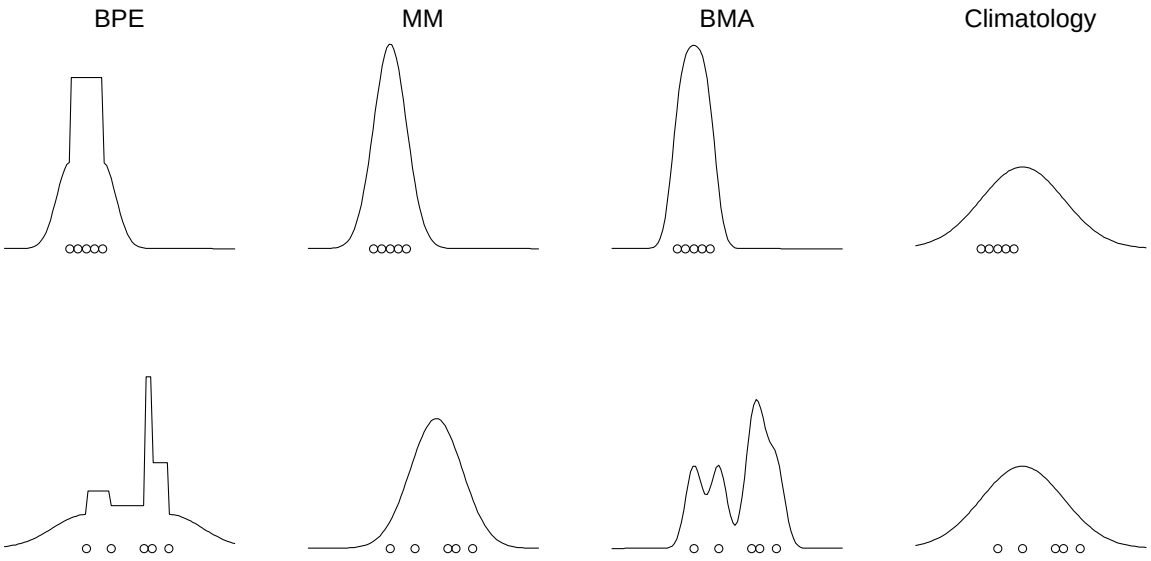

Figure 2.2: Schematic PDF diagram of four methods for representing ensemble uncertainty. Here probability density curves for binned probability ensemble (BPE), method of moments (MM), Bayesian model averaging (BMA), and climatology are shown for an ensemble of size five, with the variable of interest 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.

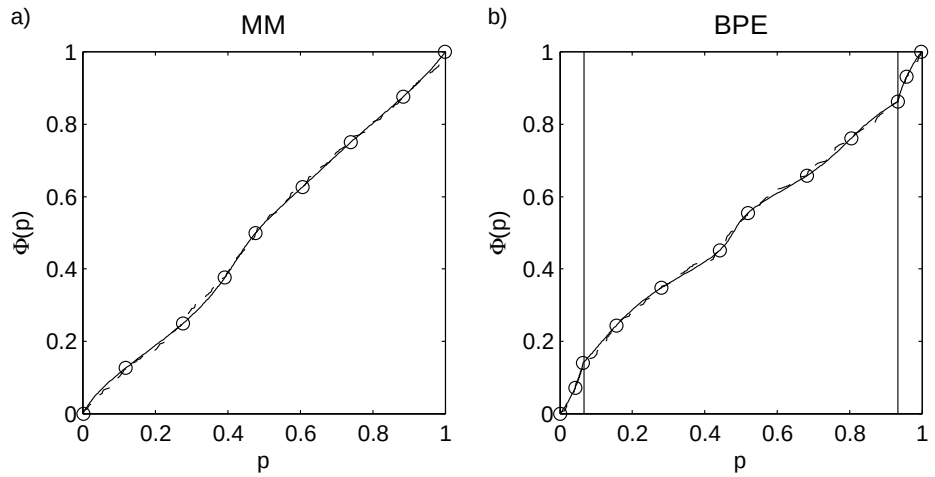

Figure 2.3: Sample calibration curve for MM and BPE. Dashed lines show the actual cumulative distribution of PIT values, whereas the solid represent the smoothed curve using cubic splines. The circles represent the interpolation points for the splines. Three separate splines were used for BPE, where the separation is shown by the two vertical lines.

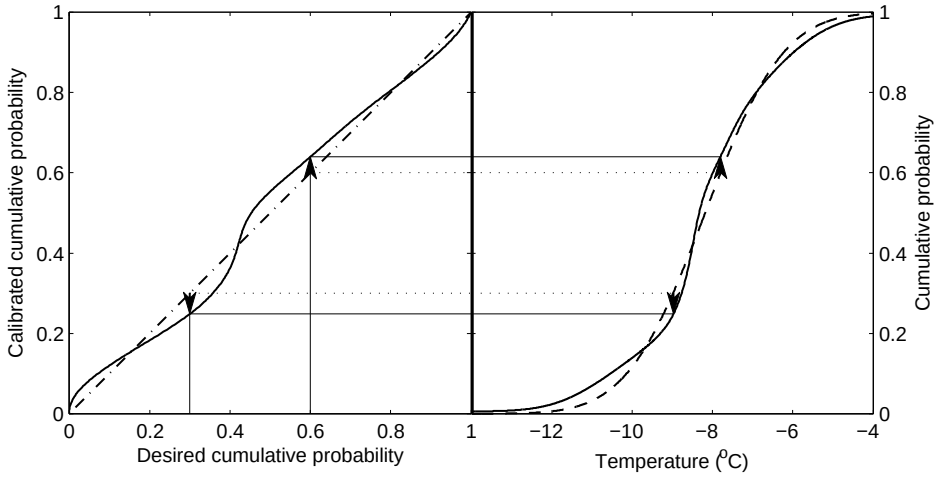

Figure 2.4: Illustrative example of the calibration of a probabilistic forecast by the method presented in the text. The figure shows a probabilistic temperature forecast created using MM. The left half shows the calibration curve (solid line) and a one-to-one line (dash-dotted line). The right half 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.

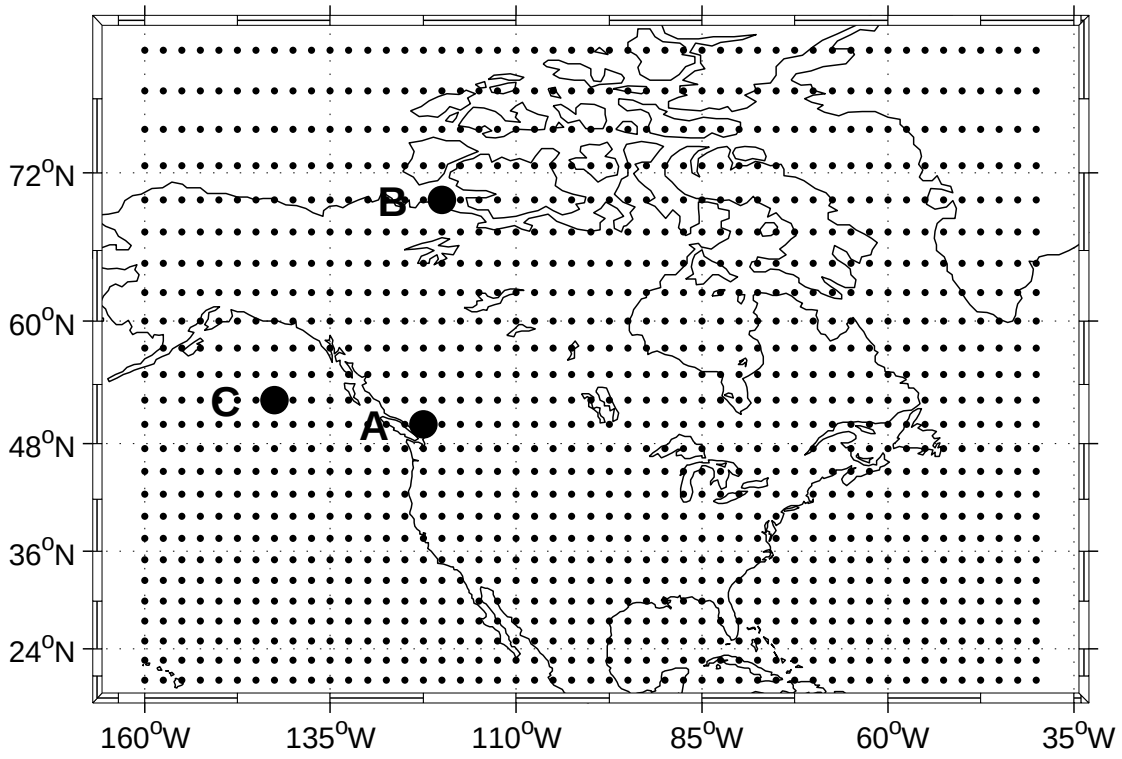

Figure 2.5: Geographical locations of the 1225 grid locations used in the study. Point A represents the grid point nearest Vancouver, Canada, point B represents a grid point in the Northwest Territories, Canada, and point C represents a grid point at the Gulf of Alaska in the Pacific Ocean.

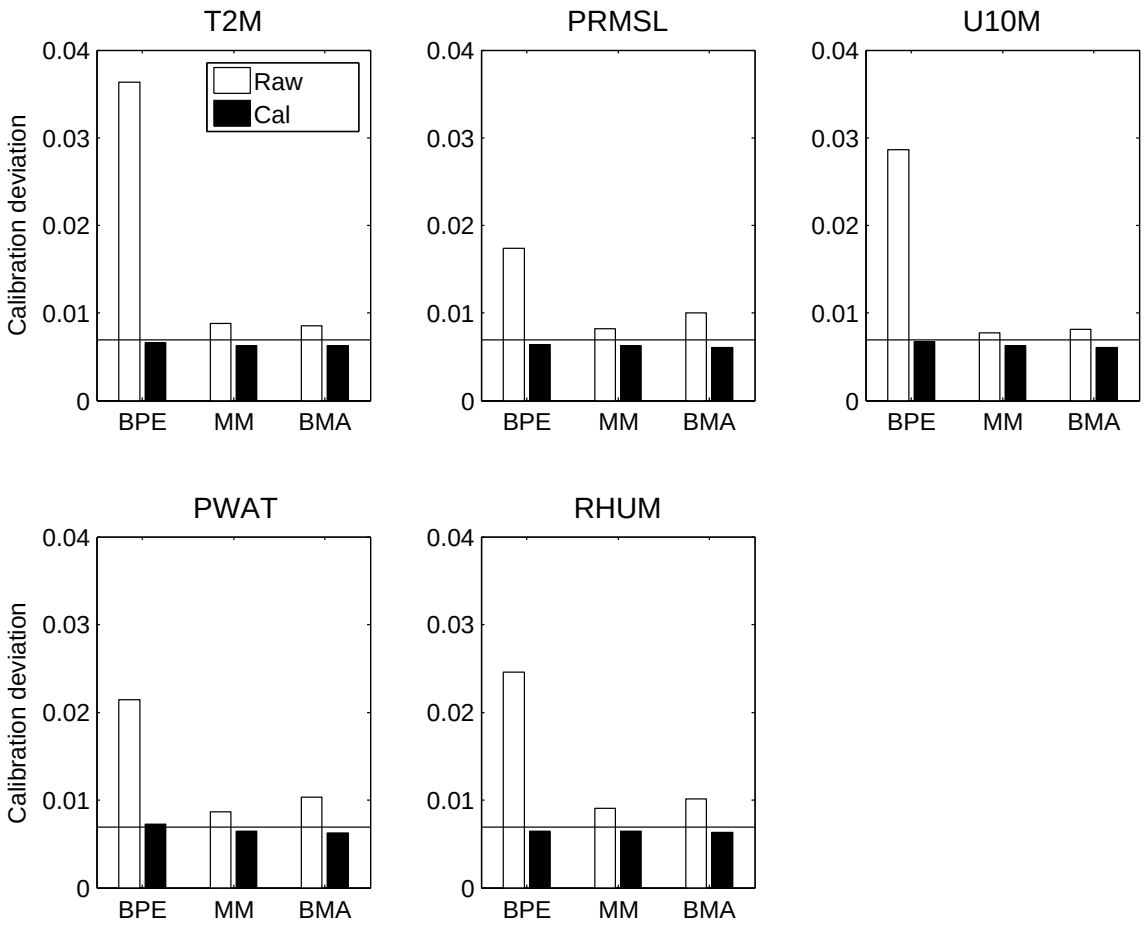

Figure 2.6: Calibration deviation is shown for 5 forecast variables. Deviation of raw forecasts is shown by white bars, and deviation of calibrated forecasts is shown by black bars. The solid horizontal line shows the expected deviation of a perfectly calibrated forecast. T2M is 2-m temperature, PRMSL 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.

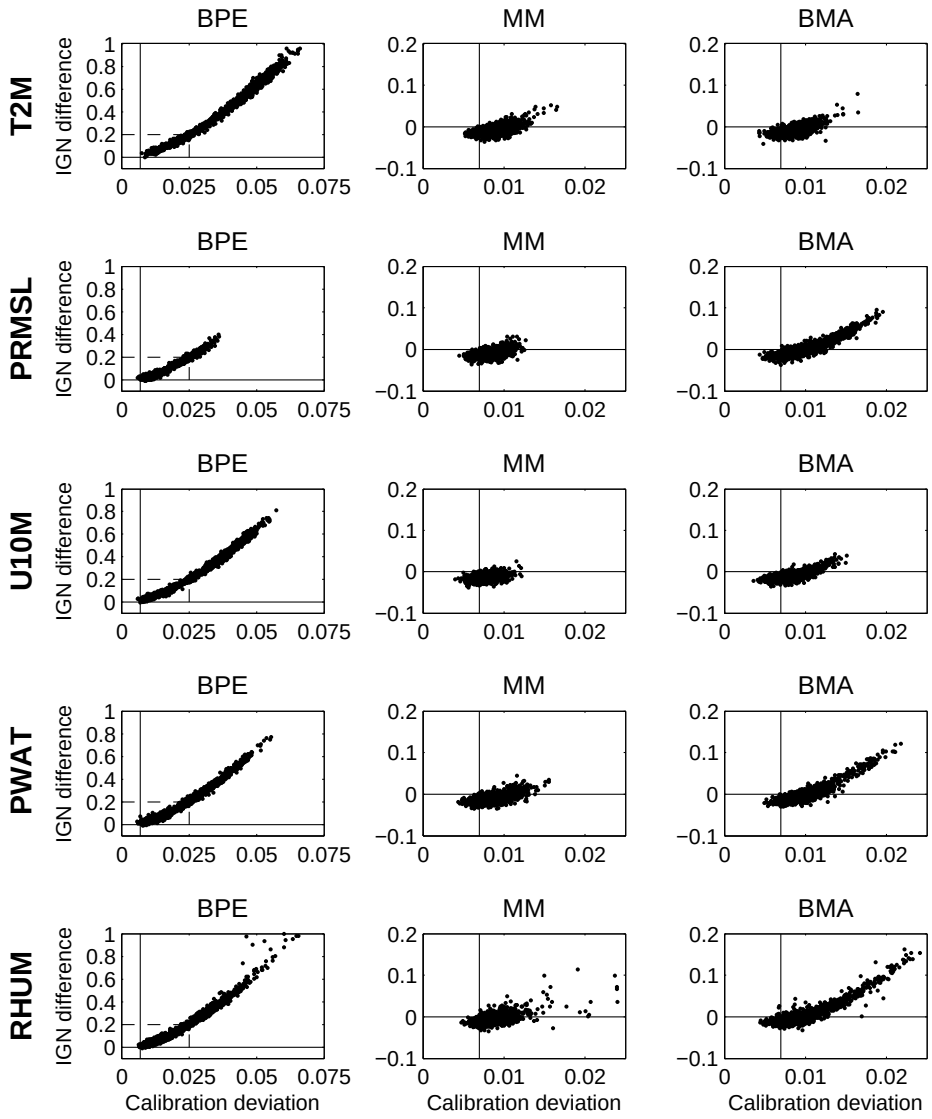

Figure 2.7: The difference of ignorance scores between raw and calibrated forecasts is shown as a function of the calibration deviation of the raw forecast. Positive ignorance score differences indicate that the calibration method improves the ignorance score. Each dot represents a separate grid point from Figure 2.5. 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.

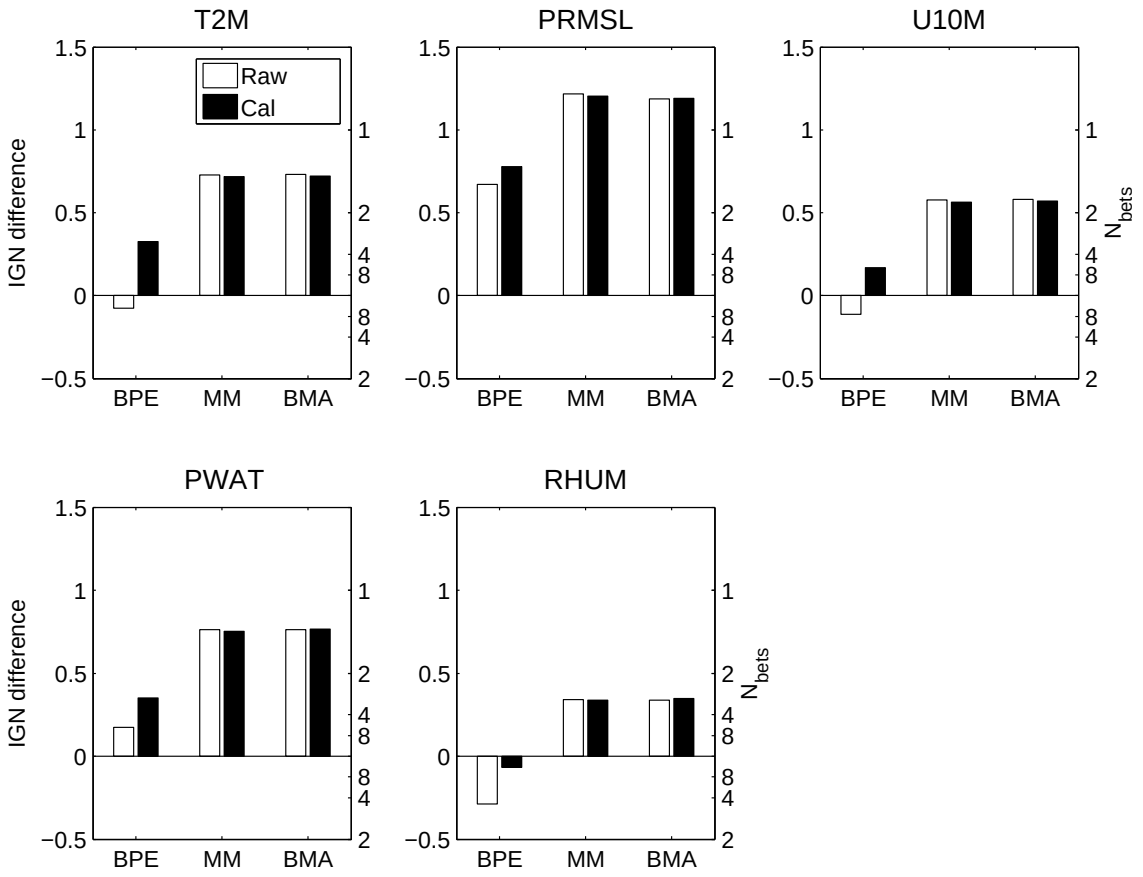

Figure 2.8: Overall difference in ignorance scores between climatology and the uncertainty models is shown by the left axis. The expected number of bets required to double wealth when a forecast model is used against climatology is shown on the right axis. Differences to climatology are shown for raw forecasts (white bars) and calibrated forecasts (black bars).

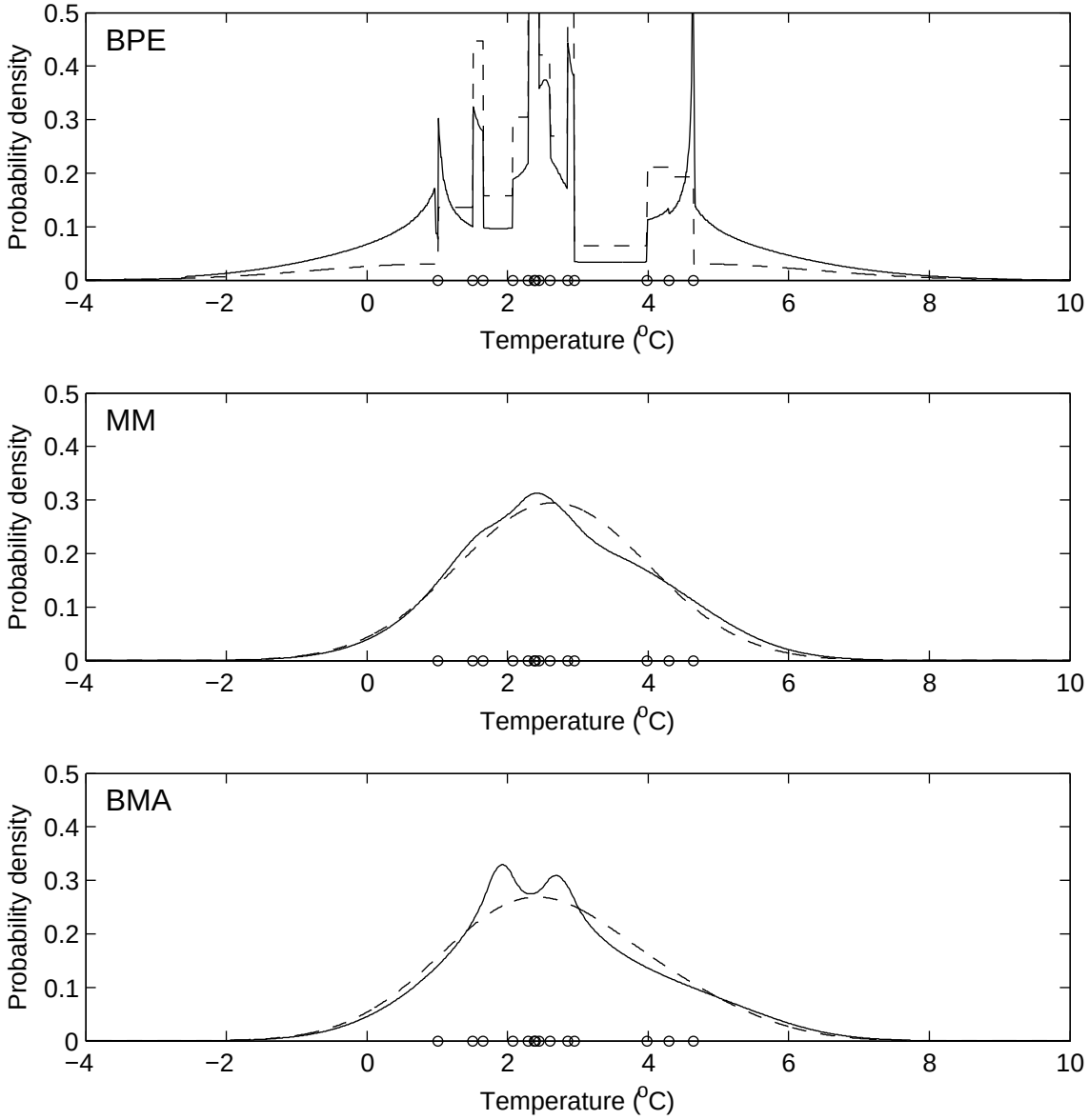

Figure 2.9: Temperature PDF forecast for the Vancouver location for 3 Jan. 2003 for BPE, MM, and BMA uncertainty models. The raw forecast is shown by the dashed lines and the calibrated by the solid lines. Circles represent the bias-corrected ensemble members.

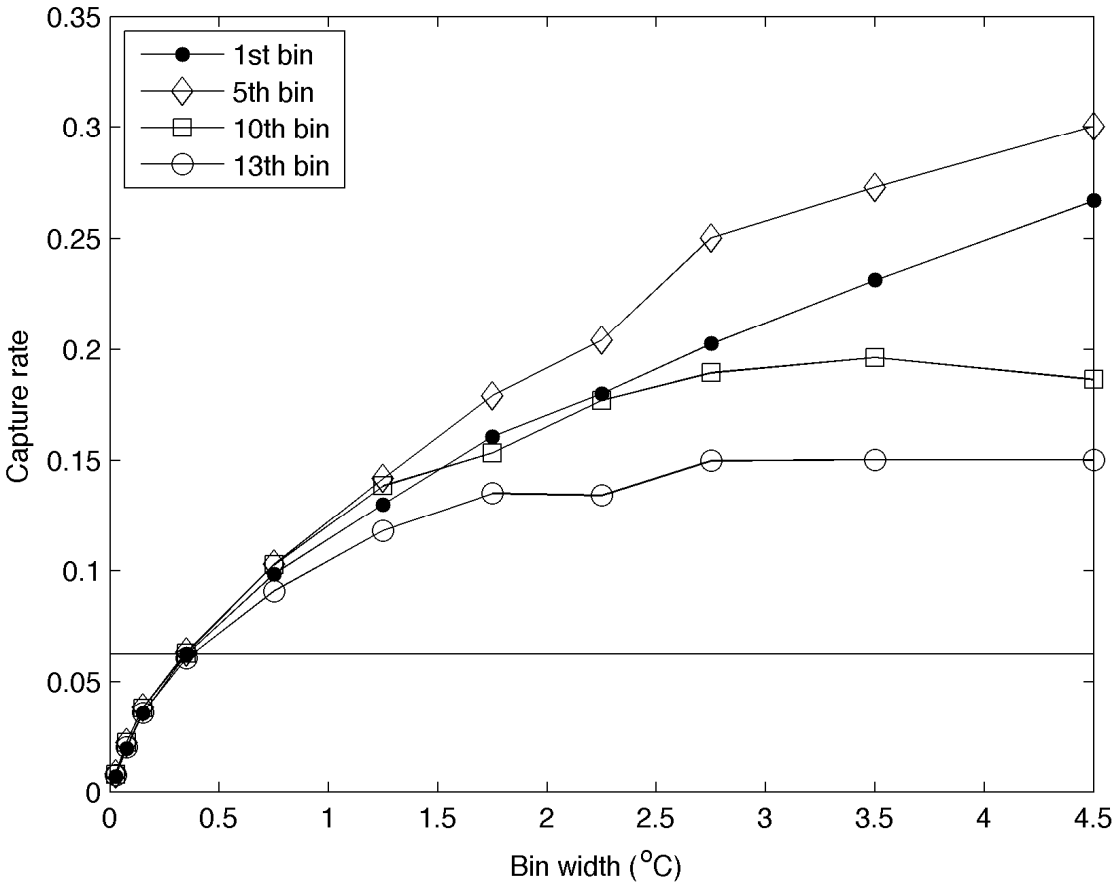

Figure 2.10: Fraction of verifying temperature analyses captured by an ensemble bin as a function of bin width is shown for four different bins. The predicted capture fraction by BPE is shown by the horizontal solid line. The ticks along the horizontal line show the separations used when binning the bin widths.

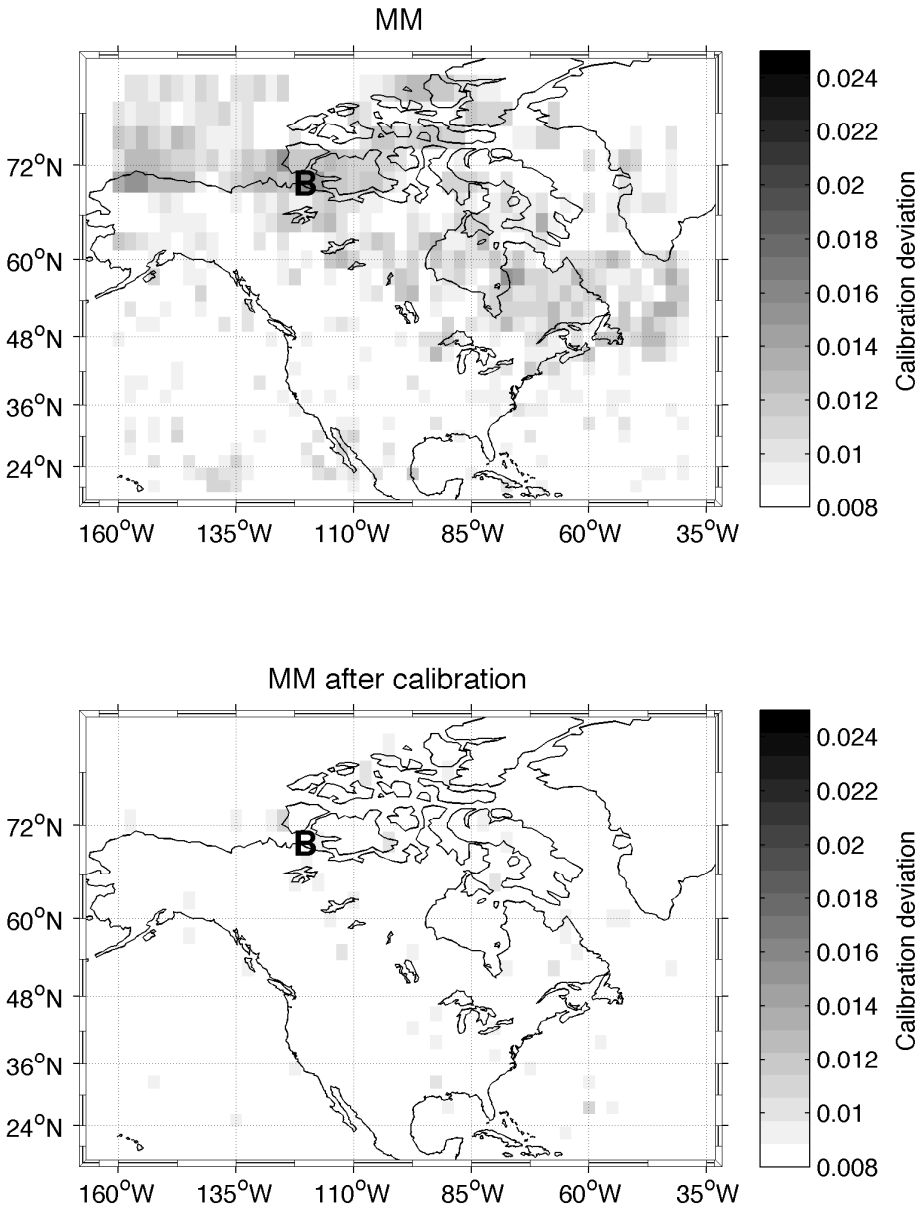

Figure 2.11: Spatial pattern of calibration deviation for raw and calibrated forecasts from MM for precipitable water. Smaller calibration deviation is better. The letter "B" is centered on the Northwest Territories location identified in Figure 2.5.

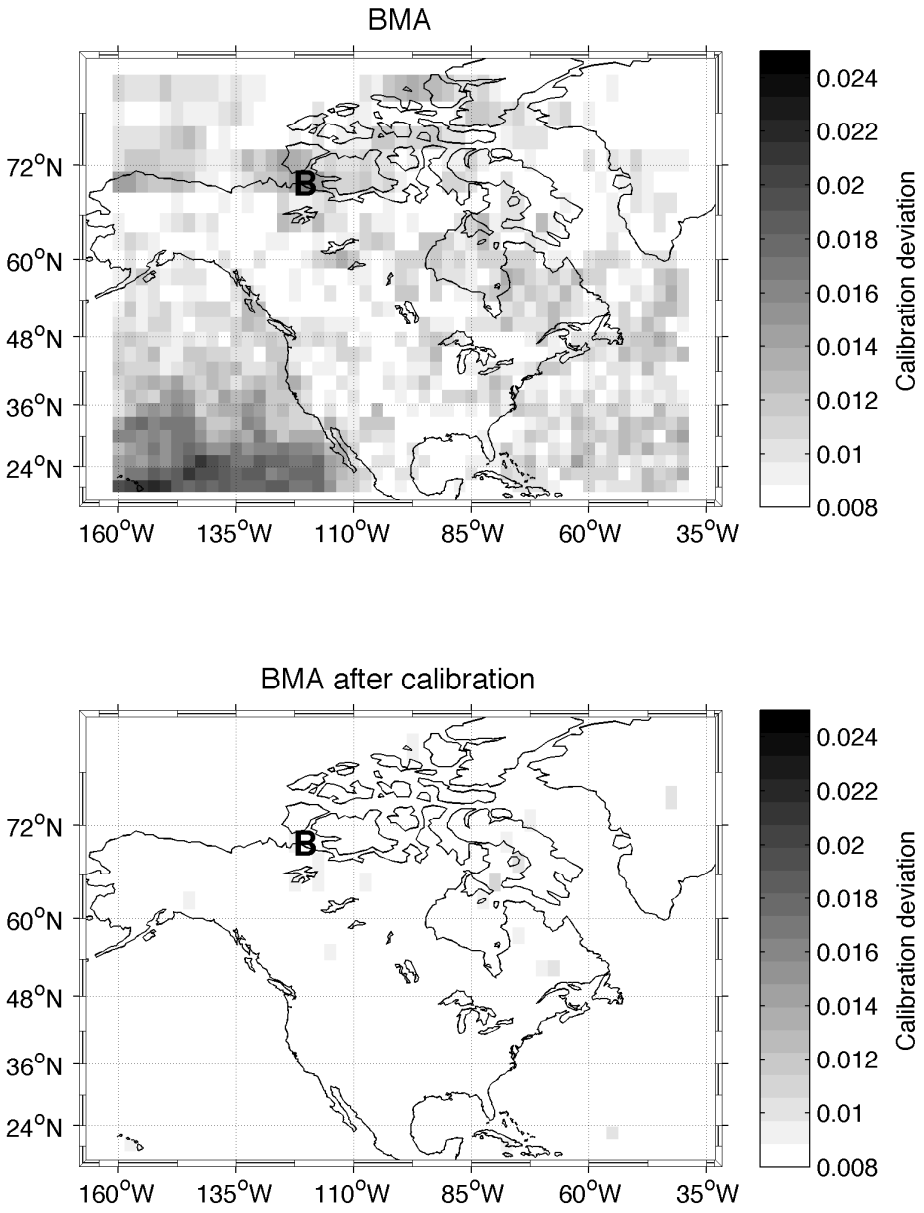

Figure 2.12: Same as Figure 2.11 except for BMA.

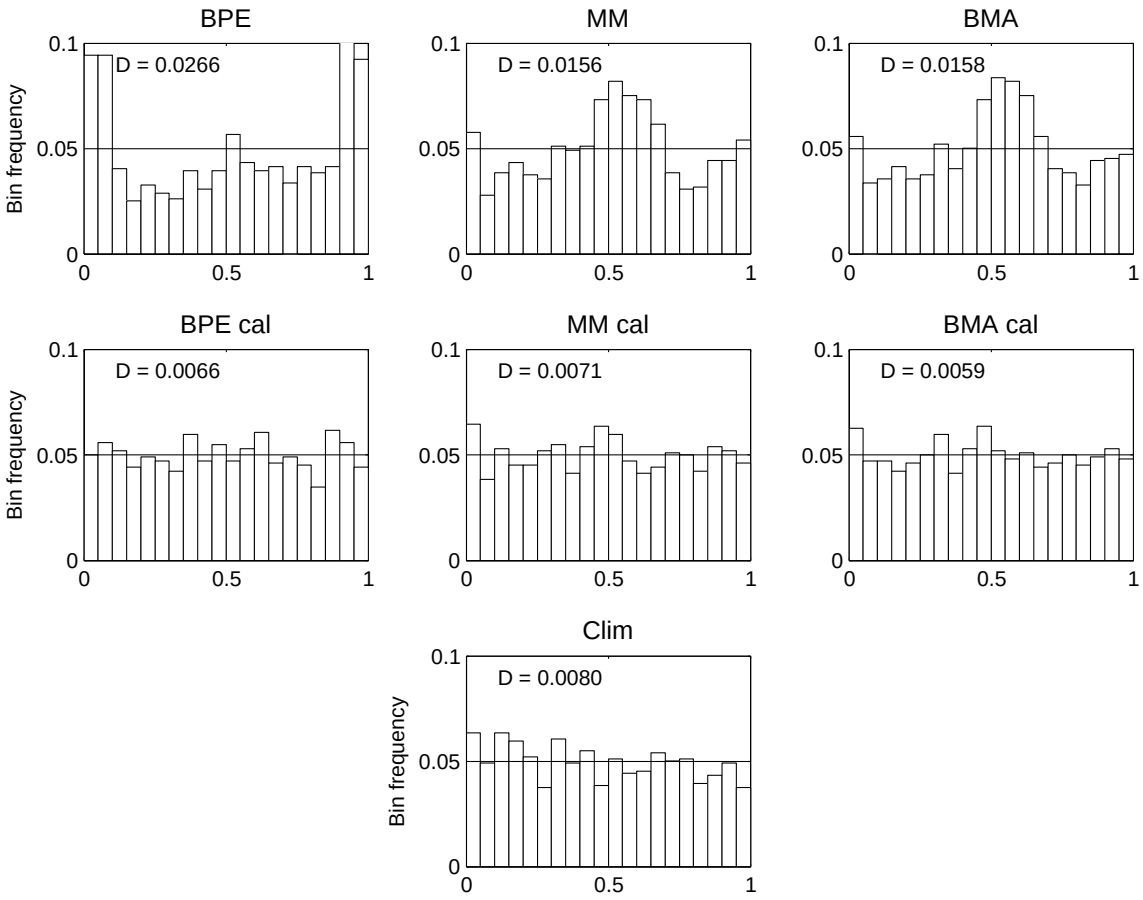

Figure 2.13: PIT histogram of precipitable water forecasts for the Northwest Territories location evaluated for all 1039 forecast days. The calibration deviation score is indicated in the top of each histogram. D values of 0.0068 represent the expected level of calibration deviation for perfectly calibrated forecasts.

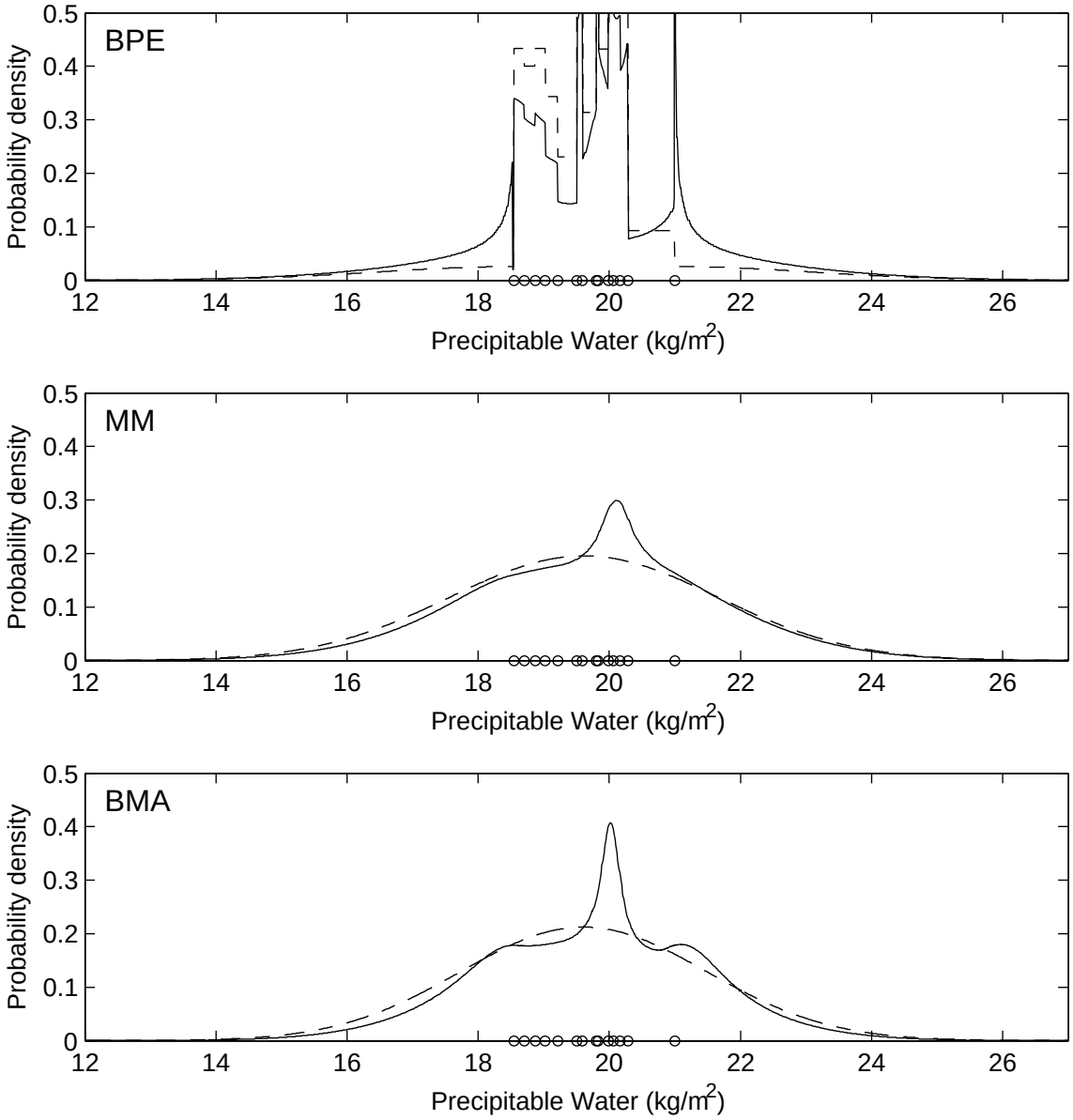

Figure 2.14: Same as Figure 2.9 except for precipitable water and for the Northwest Territories location for 2 Jul. 2002.

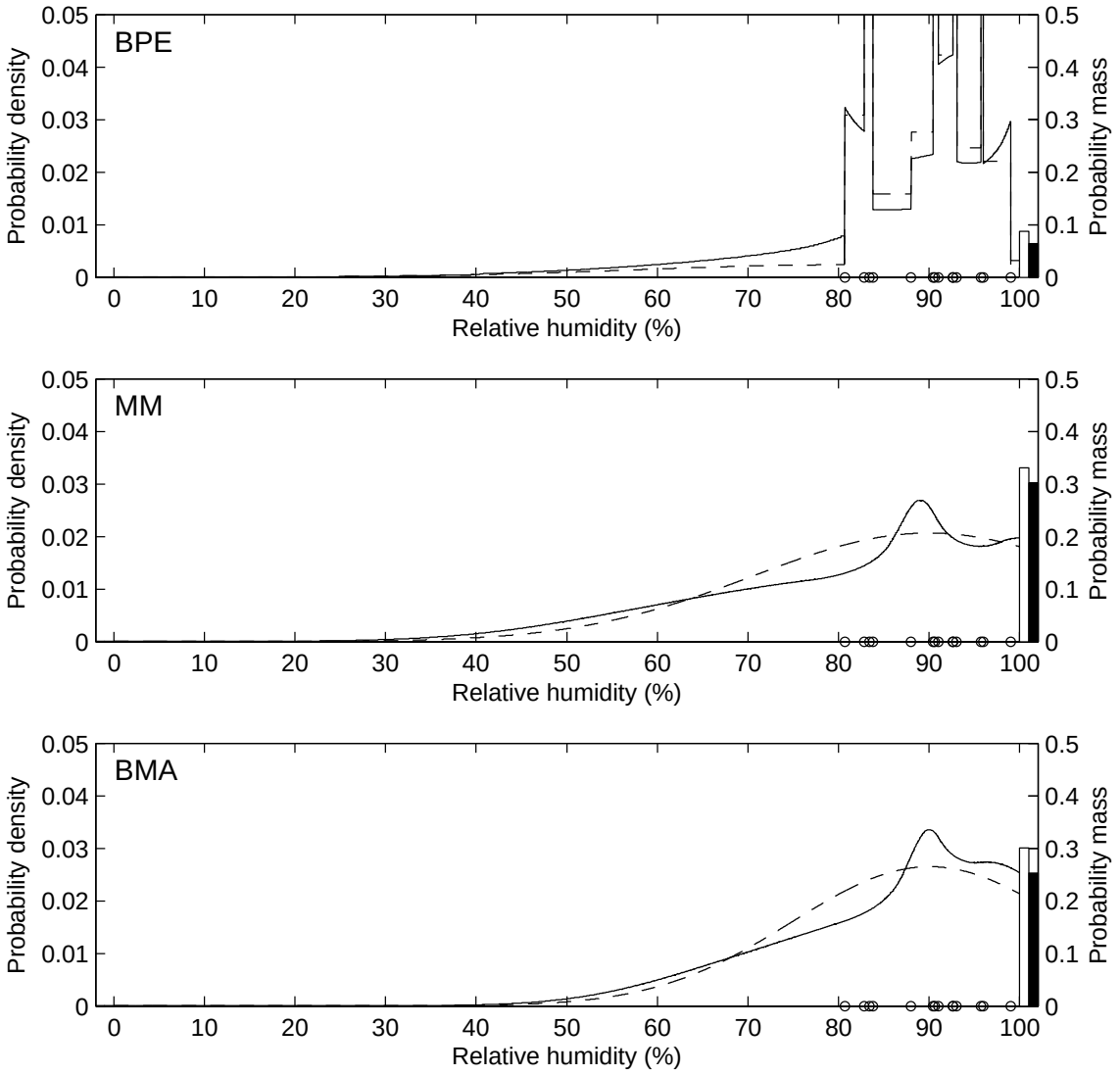

Figure 2.15: Same as Figure 2.9 except for relative humidity and for the Pacific Ocean location for 16 May 2004. White bars represent the probability mass assigned by the raw forecasts at 100% relative humidity, and black bars represent the same quantity for calibrated forecasts.

Chapter 3

Updating short-term probabilistic weather forecasts of continuous variables using recent observations

3.1 Introduction

Correctly predicting forecast uncertainty can bring significant economic benefits to many decision makers (AMS, 2008). Unlike a deterministic forecast, which supplies only the expected weather outcome, a probabilistic forecast gives the likelihood of occurrence of all outcomes. Decisions are based on combining the relative risks of various weather outcomes with the costs and losses corresponding to those outcomes. Thus, probabilistic forecasts are naturally preferred for economic decision making.

Let $f_t(x)$ be the forecasted probability density function (PDF) of a continuous meteorological variable X (such as temperature) valid for time t. One can generate $f_t(x)$ from an ensemble of numerical weather prediction (NWP) models by using methods such as Bayesian model averaging (BMA; Raftery et al., 2005), the binned probability ensemble technique (Anderson, 1996), the method of moments (Jewson et al., 2005), or local quantile regression (Bremnes, 2004).

Let $F_t(x)$ denote the forecasted cumulative distribution function (CDF) given by:

$$F_t(x) = \int_{-\infty}^{x} f_t(s)ds. \tag{3.1}$$

Let x_t denote the observed state of X at time t. Let p_t denote the CDF value corresponding to the observed state:

$$p_t = F_t(x_t). \tag{3.2}$$

p_t is often called the probability integral transform (PIT) value corresponding to the observation.

We will assume an operational ensemble forecasting system initialized at time $t = 0$ that gives

hourly forecasts out to time $t = T$. At times t, where $0 \leq t \leq T$, hourly observations from observing stations are made available, but the models do not incorporate these observations until the next forecast cycle starts.

Figure 3.1a shows a sample temperature CDF forecast for a single location produced from an ensemble. At the time the figure was produced, observations up to 1000 UTC were available. What is clear from the figure is that the CDF value that the observation verifies on (PIT value) is highly correlated in time (Figure 3.1c). Given that the most recent PIT value (at 1000 UTC) is 0.75, the next PIT value (at 1100 UTC) will likely be near 0.75.

The probability distribution can therefore be refined to take into account this new information that was not available at the time the model was initialized. The effect of the most recent observation will diminish for longer lead times. The updated probability distribution will therefore be narrow near the time of the observation and widen back to the original distribution for times in the future (Figure 3.1b).

The goal of this chapter is to present a method for producing an updated probabilistic forecast $\hat{F}_t(x)$ by mapping the original CDF $F_t(x)$ by a function Φ as follows:

$$\hat{F}_t(x) = \Phi\big(F_t(x)\big). \tag{3.3}$$

The mapping will concentrate $\hat{F}$ in a narrower range with the hope of improving short-term verification scores. End-users in need of rapidly-updating probabilistic short-term forecasts at very low computational costs can benefit from this update method.

Post-processing weather forecasts is commonly done to increase the correspondence between forecasts and observations. For deterministic forecasts, methods such as model output statistics (Glahn and Lowry, 1972), Kalman filtering (Homleid, 1995), and analog methods (Delle Monache et al., 2011) are commonly used to reduce forecast error. On the other hand, methods such as ensemble calibration (Hamill and Colucci, 1998) and BMA (Raftery et al., 2005) can be used to improve probabilistic forecasts from an ensemble of deterministic forecasts. The method presented here also aims to improve probabilistic forecasts, but differs in that it is only invoked once observations are available after the raw forecasts are created. It is therefore of most use for operational short-term forecasts.

This chapter is organized as follows: the method for updating probabilistic forecasts is presented in Section 3.2, the data set and verification metric used for testing the method is described in Section 3.3, the performance of the method is evaluated in Section 3.4, and conclusions are drawn in Section 3.5.

3.2 Method

Assume that for a given forecast day, $T+1$ hourly probabilistic forecasts $F_t(x)$ (where $0 \leq t \leq T$) are produced. Let t_{obs} denote the time at which the most recent observation was made. This observation is then used to update all hourly forecasts that are still in the future (that is where $t_{obs} < t \leq T$).

The probabilistic forecast n hours after t_{obs}, that is for time $t = t_{obs} + n$, can be updated according to:

$$\hat{F}_{t_{obs}+n}(x) = \Phi_n\big(F_{t_{obs}+n}(x)\big), \tag{3.4}$$

where $\Phi_n(p)$ will in general be different for each value of n and can be constructed based on forecast and observation data prior to the time t_{obs}. $\Phi_n(p)$ is the probability function that the verifying PIT value of the original forecast will be less than p.

Combining Eq. (3.1) and Eq. (3.4) and using the chain rule gives the following for the updated PDF:

$$\hat{f}_{t_{obs}+n}(x) = \Psi_n\big(F_{t_{obs}+n}(x)\big) f_{t_{obs}+n}(x), \tag{3.5}$$

where $\Psi_n(p)$ is the derivative of $\Phi_n(p)$, and acts as an amplification factor for the original PDF. $\Psi_n(p)$ increases probability density in regions where the PIT value is more likely to occur given the recent observation. That is, $\Psi_n(p)$ is also the probability density of p being the verifying PIT value of the original forecast.

3.2.1 PIT values as a random walk in time

We model the time-sequence of verifying PIT values within one forecast cycle as a random walk in time. Mirror barriers at 0 and 1 are used to handle the fact that PIT values are bounded on the interval $[0, 1]$. That is, any random steps across the boundaries are reflected back into the domain (Figure 3.2). Mirror barriers are commonly used to describe stochastic processes in other areas of modeling (Karlin and Taylor 1981; See also Rose 1995 for applications in economics).

Let $p_{t_{obs}}$ be the PIT value of the most recent observation, and let $\Psi_n(p)$ be the probability density function of the verifying PIT value being p at n hours after t_{obs}. When $n = 0$, the PIT value is fully known and can therefore be described by:

$$\Psi_0(p) = \delta(p - p_{t_{obs}}), \tag{3.6}$$

where δ is the Dirac delta function defined by:

$$\delta(s) = \begin{cases} +\infty, & s = 0 \\ 0, & s \neq 0 \end{cases} \tag{3.7}$$

$$\int_{-\infty}^{\infty} \delta(s)ds = 1. \tag{3.8}$$

Let $S(p, q)$ represents the probability density of arriving at a PIT value of p, given that the previous PIT value was q. Since our stochastic model for PIT values is a first-order Markov model, the probability of a certain PIT at time n can be found from all transitions to that PIT from time $n - 1$. The probability density after a transition can therefore be determined by the following recursive equation:

$$\Psi_n(p) = \int_0^1 S(p, q)\Psi_{n-1}(q)dq \tag{3.9}$$

3.2.2 Determining the transition function

We assume that the step-length from one PIT to the next is Gaussian distributed with mean 0 and variance σ^2. That is, the transition function S can be constructed as follows:

$$S(p, q) = \phi(p; q; \sigma^2) + \phi(-p; q; \sigma^2) + \phi(2 - p; q; \sigma^2) + ... \tag{3.10}$$

$$= \sum_{i=-\infty}^{+\infty} [\phi(p + 2i; q; \sigma^2) + \phi(-p + 2i; q; \sigma^2)], \tag{3.11}$$

where $\phi(x; \mu; \sigma^2)$ is a Gaussian PDF with mean μ and variance σ^2. The first term in Eq. (3.10) comes from steps within the domain, the second comes from steps reflected across 0, and the third term comes from steps reflected across 1. Eq. (3.11) includes all possible steps, including steps that cross both boundaries one or more times.

A transition function that combines n number of steps can also be constructed, and is denoted by S_n. The variance of multiple steps (under the assumed model) increases linearly with time and S_n can therefore be computed by:

$$S_n(p, q) = \sum_{i=-\infty}^{+\infty} [\phi(p + 2i; q; n\sigma^2) + \phi(-p + 2i; q; n\sigma^2)], \tag{3.12}$$

Since σ is small in our study (around 0.15), and we use values of n no larger than 24 we restrict the summation to $i \in [-10, 10]$. A wider range for i may be required for large σ and n values.

Constructing S_n allows us to simplify Eq. (3.9) to the following:

$$\Psi_n(p) = \int_0^1 S_n(p,q)\Psi_0(q)dq \tag{3.13}$$

$$= S_n(p, p_{t_{obs}}), \tag{3.14}$$

where again $p_{t_{obs}}$ is the verifying PIT value at time t_{obs}. This simplification avoids the need to recursively compute Ψ_n (as in Eq. (3.9)). Note that for forecast variables that require a non-Gaussian transition function, it is possible that Eq. (3.12) cannot be constructed analytically in which case the above simplification may not be possible.

Figure 3.3 shows an example sequence of $\Psi_n(p)$ for various values of n. The PIT value distribution clearly widens as time goes on, indicative of the disappearing effect of the last observed PIT value.

3.2.3 Parameter estimation

In order to create the updated forecasts, an estimate of σ^2 is needed by Eq. (3.12). The variance of the step sizes of past PIT values (σ_0^2) can be used:

$$\sigma_0^2 = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\left(p_{t+1} - p_t\right)^2, \tag{3.15}$$

where $\mathcal{T}$ represents a set of time points from past forecast cycles comprising the training period, and where $|\mathcal{T}|$ is the size of this training set. σ_0^2 will in general underestimate σ^2 since some steps will appear to be short steps when in fact they are longer steps that have reflected across a boundary.

For a given σ, the expected value of σ_0 can be computed by the integral over all possible PIT transitions from p to q:

$$\sigma_0^2 = \sum_{i=-\infty}^{\infty}\int_0^1\int_0^1 [\phi(p+2i;q;\sigma)(p-q)^2 + \phi(-p+2i;q;\sigma)(p-q)^2]dpdq. \tag{3.16}$$

Solving this equation for σ (as required by Eq. (3.12)) was not possible analytically. We found through trial and error that the following is a good approximation for σ in terms of σ_0:

$$\sigma \approx \tan(3.5\sigma_0)/3.5, \tag{3.17}$$

where the input to the tangent function is in radians. This approximation has errors of less than $3.4\,\%$ for σ_0 values up to 0.3 (Figure 3.4).

A summary of the process of updating a probabilistic forecast goes as follows: the variance of past PIT transition distances (σ_0) is computed by Eq. (3.15), which is used to approximate σ in Eq. (3.17); σ is then used in Eq. (3.12) to compute the transition function (S_n); The transition function, combined with the latest available verifying PIT value is used to calculate the PIT distribution (Ψ_n) by Eq. (3.14), which is used to update the original probabilistic forecast through Eq. (3.5).

3.3 Operational test case

3.3.1 Model data and configuration

Hourly surface temperature forecasts from the Mesoscale Compressible Community [MC2, Benoit et al. (1997)] model, the Penn State/NCAR Mesoscale Model [MM5, Grell et al. (1994)], and the Weather Research and Forecasting [WRF, Skamarock et al. (2005)] model were used for the case study period: 0000 UTC 1 Sep. 2005 to 2300 UTC 1 Feb. 2008. Two runs for the WRF model were used: one using GFS initialization (WRFG) and the other using NAM initialization (WRFN), while MC2 and MM5 both used NAM initialization. The MC2 and MM5 runs had outer domains with 108-km grid spacing, and inner 36-, 12-, and 4-km nested domains. The WRF runs were similar, but did not contain the 4-km nested domain. These domains comprised our 14-member ensemble.

The models were initialized once per day at 0000 UTC, and hourly forecast output to 60 hours was available. Probabilistic forecasts were generated for the same time period.

The model runs and probabilistic forecasts were generally completed by 0600 UTC, after which we used the latest observation to update the probabilistic forecasts valid for the subsequent 24 hours. The update process was repeated each hour as a new observation became available. This was done until 0600 UTC the next day, when the probabilistic forecasts from the next forecast cycle were used. This means that for each forecast cycle 24 24-h updated forecasts were produced, yielding 576 forecasts per day.

We tested the method on temperature probabilistic forecasts and observations for the following five airport stations in British Columbia, Canada: Vancouver International Airport station (CYVR), Abbotsford International Airport (CYXX), Victoria International Airport (CYYJ), Kamloops Airport (CYKA), and Kelowna Airport (CYLW), which provided a geographically diverse sample from within our smallest model domain.

3.3.2 Original probabilistic forecasts

We used the method of moments to produce the original probabilistic forecast from the forecast ensemble. The PDF using this method is computed by:

$$f_t(x) = \phi(x; \xi_t + \mu; s^2), \tag{3.18}$$

where again ϕ is a Gaussian PDF, x is a temperature value, ξ_t is the ensemble mean at time t, μ is a bias-correction term for the centre of the distribution, and s^2 is the variance of the distribution.

The last two parameters are determined by the forecast errors during the training period $\mathcal{T}$:

$$\mu = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_i - \xi_i \tag{3.19}$$

$$s^2 = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (x_i - \mu - \xi_i)^2, \tag{3.20}$$

Note that the spread in this case is independent of the ensemble spread.

The parameters μ and s were computed separately for each station and separately for each of the 24 forecast hours. They were computed from a 40-day sliding window that ended the day before the forecast was initialized. A training period of 40 days is a compromise between the need to use statistics that adapt quickly to seasonal changes and the requirement to have enough data to robustly estimate the parameters. Similar training lengths have been used to produce probabilistic forecasts using Bayesian Model Averaging (Raftery et al., 2005; Sloughter et al., 2007).

The spread parameter σ_0 (and consequently σ) was also computed separately for each station using a 40-day sliding window, however all 24 forecast offsets for a given station were pooled together to give a more robust estimate.

3.4 Analysis

3.4.1 Ignorance score

We use the logarithmic score of Good (1952), which has gained popularity over the last decade and has been referred to as "Ignorance" score owing to its ties with information theory (Roulston and Smith, 2002). It is defined as follows:

$$\text{IGN}(f) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} -log_2\big(f_t(x_t)\big). \tag{3.21}$$

IGN rewards forecasts that place high confidence in the value where the observation falls. Low ignorance scores are desired.

The total ignorance scores of the original probabilistic forecasts were computed by averaging ignorance scores over all forecast cycles, and forecast hours, but separately for each station and each value of n in order to see how far into the future a recent observation can improve the ignorance score.

Figure 3.5a shows the *improvement* in ignorance score provided by the updated probabilistic forecast as a function of time from the most recent observation. The updated forecasts at $0\,\mathrm{h}$ after an observation has been made has an ignorance score of $-\infty$ since the true state is fully known. However, this update forecast is of no value since it is only available after the observation has been made. As time since most recent observation increases, the improvement in the ignorance score reduces down toward 0.

3.4.2 Continuous ranked probability score

We also computed the continuous ranked probability score (CRPS) to further evaluate the quality of the probabilistic forecasts. It is defined as:

$$\mathrm{CRPS}(F) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \int_{-\infty}^{+\infty} [F_t(x) - H(x - x_t)]^2 dx, \tag{3.22}$$

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

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases} . \tag{3.23}$$

Low values of CRPS are preferred.

Figure 3.5b shows the percentage improvement due to the updated forecast relative to the original raw forecast. This is defined as:

$$\%\ \text{improvement} = \frac{\mathrm{CRPS}(F_{\text{raw}}) - \mathrm{CRPS}(F_{\text{updated}})}{\mathrm{CRPS}(F_{\text{raw}})} \times 100\%. \tag{3.24}$$

Results for CRPS show a similar pattern as for the ignorance score, with the update method providing less improvement as the time since the most recent observation increases. The average CRPS of the 5 stations was $1.50\,^{\circ}\mathrm{C}$ and the update method brought the value down to $1.06\,^{\circ}\mathrm{C}$ and $1.27\,^{\circ}\mathrm{C}$ at 3 and 6 hours respectively.

3.4.3 Reliability

A probabilistic forecast is reliable (or calibrated) when the PIT values are uniformly distributed between 0 and 1 (Gneiting et al., 2007). This can be diagnosed by a simple histogram of verifying PIT values as reliable forecasts will give a flat histogram.

Figure 3.5c shows the histogram of PIT values from all forecast hours, forecast cycles, stations, and values of n. The update method does not appear to degrade or improve the reliability of the original forecasts in any significant way.

3.4.4 Mean absolute error

A probabilistic forecast can also provide a best deterministic estimate, by using the median of the probability distribution (as shown by the $50\,\%$ lines in Figure 3.1a and Figure 3.1b). We used the mean absolute error (MAE) as a verification measure of this deterministic forecast:

$$\mathrm{MAE}(f) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left| x_t - F_t^{-1}(0.5) \right|, \tag{3.25}$$

where F_t^{-1} is the inverse of F_t giving the temperature value corresponding to a nominal proportion of 0.5.

The MAE of the deterministic forecast (Figure 3.5d) showed a similar pattern to the ignorance score and CRPS, with the update method improving MAE from $2.07\,^{\mathrm{o}}\mathrm{C}$ down to $1.42\,^{\mathrm{o}}\mathrm{C}$ and $1.73\,^{\mathrm{o}}\mathrm{C}$ at 3 and 6 hours respectively. Improvements in MAE suggest that the update method improves the central tendency of the probabilistic forecasts.

3.5 Conclusions

We have presented a method to update probabilistic forecasts of continuous variables based on recent observations, which should prove useful for a variety of nowcasting purposes. An alternative to this is to use data assimilation after new observations are available in order to create new initializations for the ensemble, followed by a complete rerun of the ensemble. This is considerably more expensive from a computational point of view, and may be infeasible for many operational systems.

The method improves the ignorance score and CRPS of the probabilistic forecasts, and improves the MAE of the median of the distribution significantly for forecasts up to six hours after a recent observation, while not affecting reliability negatively.

Future work includes investigating the benefits of using a higher-order Markov model for modeling PIT transitions. In addition to accounting for hour by hour correlation of PIT values, a higher-order Markov model can also incorporate any diurnal correlation of PIT values that may

exist thereby allowing for the potential to improve forecasts for $24\,\mathrm{h}$ after a recent observation.

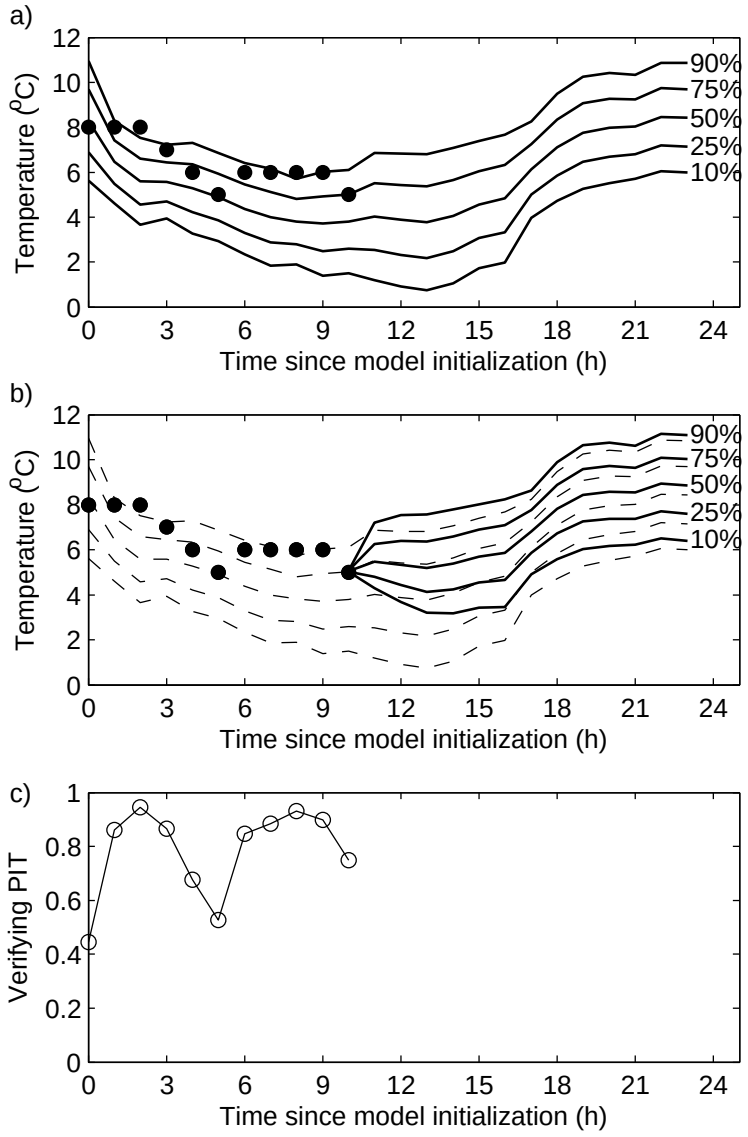

Figure 3.1: (a) A sample probabilistic temperature forecast initialized at 0000 UTC. Forecasted cumulative probability values are shown by lines. Observations are shown by solid dots. (b) The updated probabilistic forecast (solid lines) based on the most recent observation. The original forecast is shown by dashed lines. (c) The probability integral transform values of the original forecast corresponding to the observations.

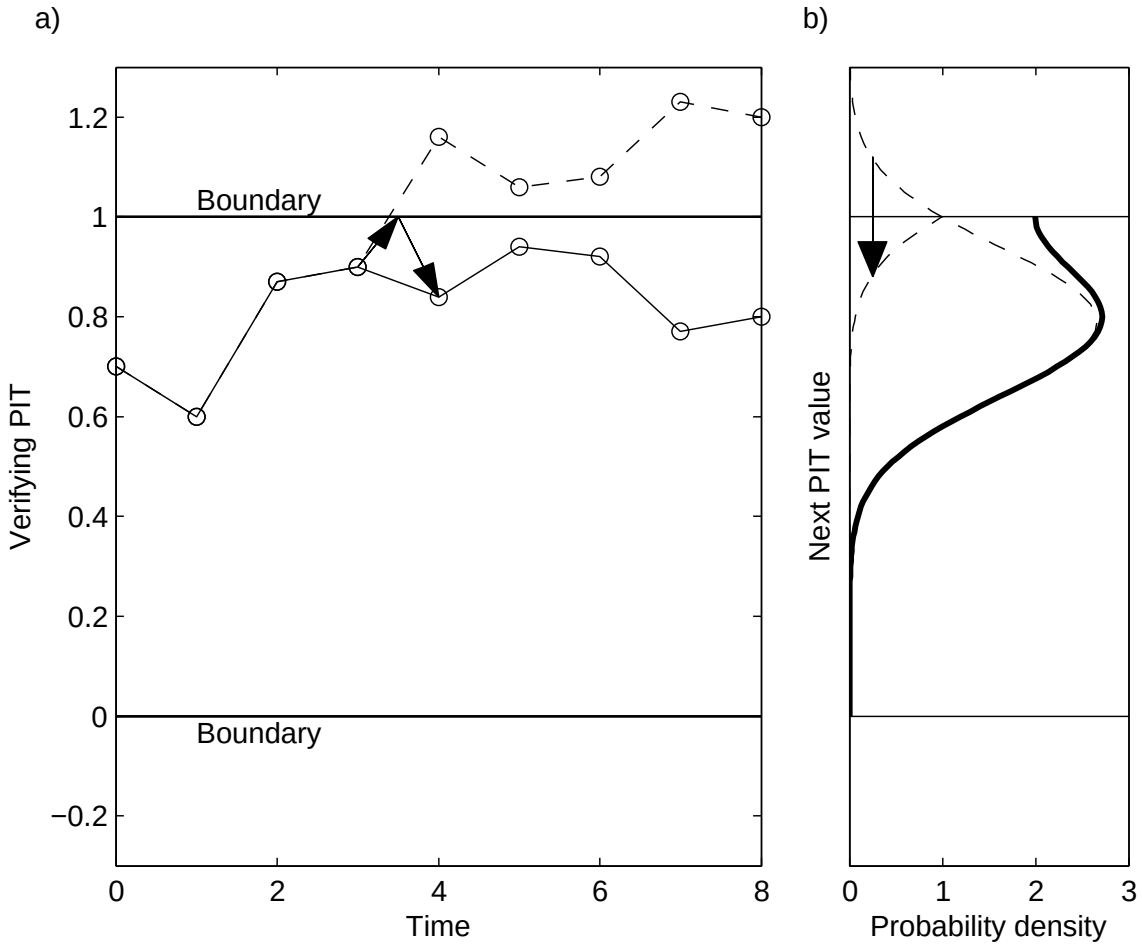

Figure 3.2: (a) A hypothetical time-series of verifying PIT values (solid line). Mirror barriers at 0 and 1 reflect any steps back into the domain. The dashed line shows the PIT time-series without reflections. The transition from time 3 to 4 involves a reflection across 1 as shown by the arrows. (b) The probability density function (thick solid line) of the PIT value for time 9, given that the PIT value at time 8 was 0.80. The dashed line shows the probability of the Gaussian distribution that has been reflected back into the domain.

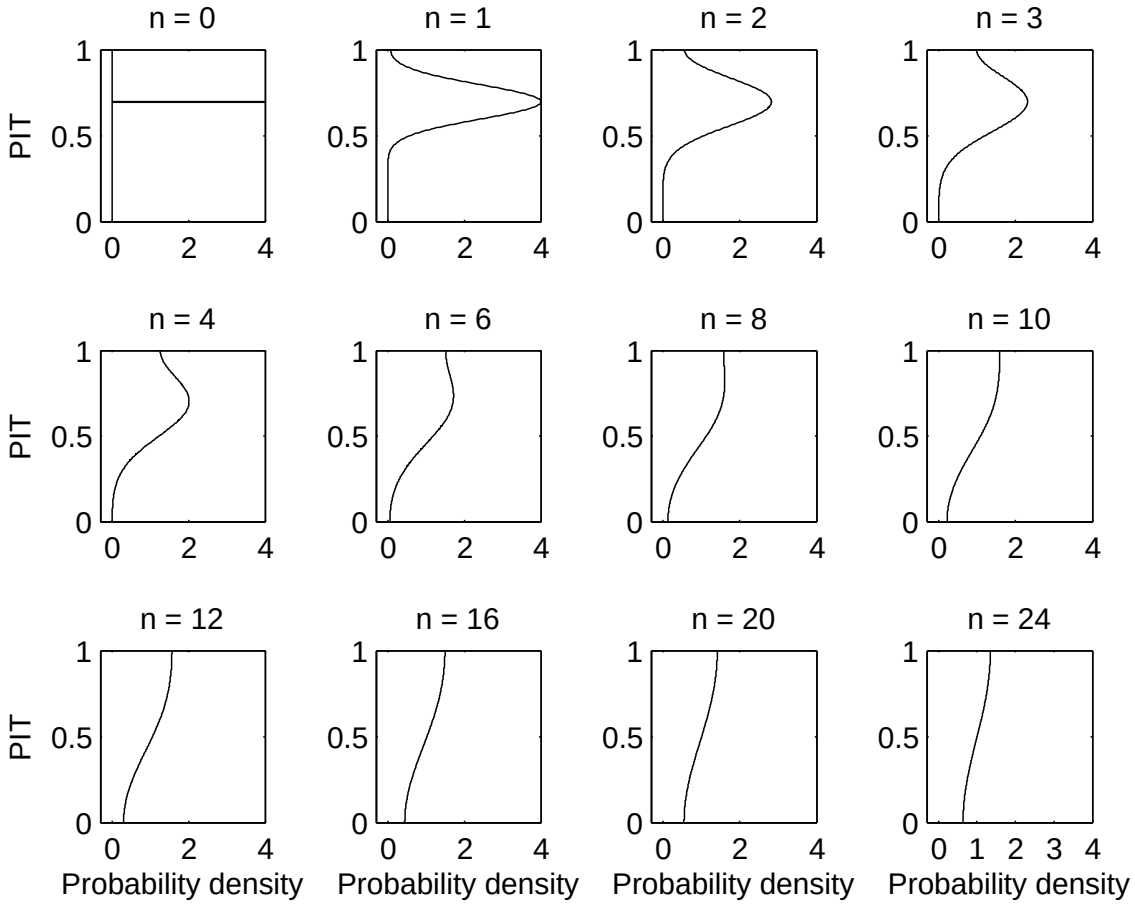

Figure 3.3: An example sequence of probability density functions of PIT values for different number of hours (n) after an observation has been made. In this case at $n = 0$, the PIT value is fully known to be 0.7.

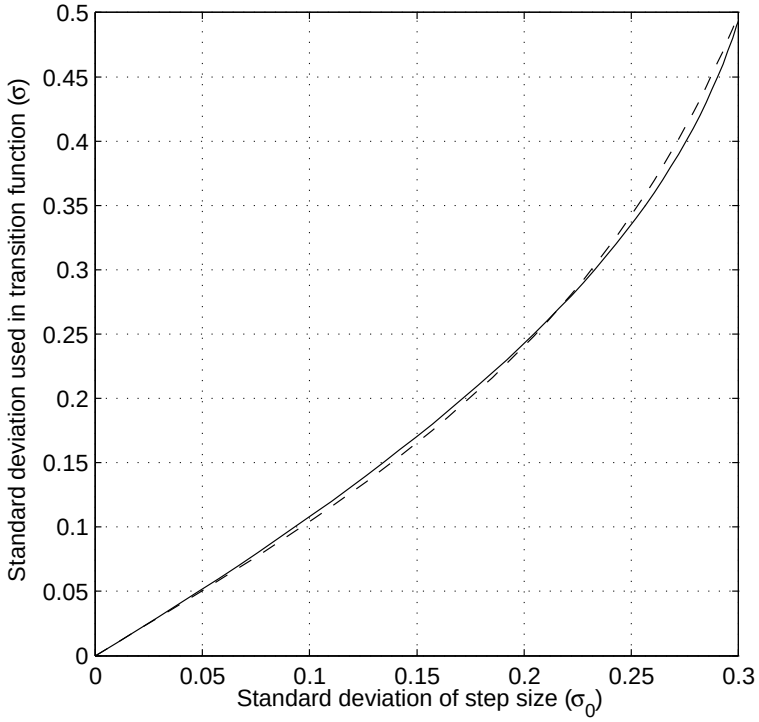

Figure 3.4: Standard deviation of PIT step sizes used in the transition function as a function of the measured standard deviation of step sizes of past PIT values (solid line) and the approximation $\sigma = \tan(3.5\sigma_0)/3.5$ (dashed line)

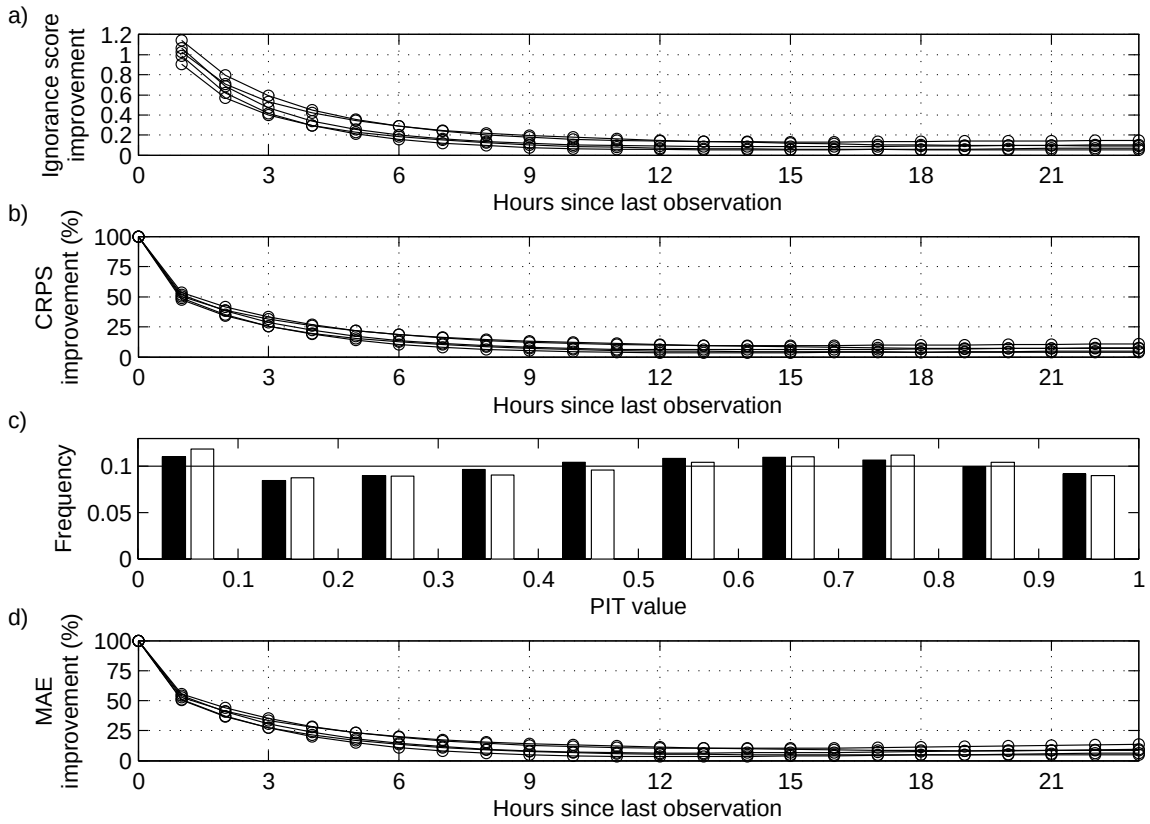

Figure 3.5: Verification statistics for the probabilistic forecasts used in the study. (a) Reduction (improvement) of the ignorance score by the updated probabilistic forecast relative to the original probabilistic forecast. Each of the five lines represents the score for a different station. (b) Percentage improvement in the continuous ranked probability score by the updated probabilistic forecast. (c) PIT histogram of the updated forecasts (black bars) and the original forecasts (white bars), indicating the reliability of the forecasts. (d) Percentage improvement in mean absolute error of the median of the updated probability distributions relative to the median of the original distribution.

Chapter 4

A modular operational probabilistic weather forecasting system

4.1 Introduction

Weather-forecast providers are increasingly being requested to provide forecasts in probabilistic form (AMS, 2008). A deterministic forecast provides a single estimate of a weather variable in the future. Probabilistic forecasts complement this information with an estimate of the prediction's uncertainty by indicating the probability of occurrence of all weather outcomes. Uncertainty information is especially useful for forecast users who make decisions based on balancing the risks and costs associated with weather outcomes (Murphy, 1977; Richardson, 2000; Palmer, 2000; Zhu et al., 2002). By knowing the likelihood of the occurrence of disastrous events, these users can adequately protect their weather-affected operations (see McCollor and Stull 2008b for an application in hydroelectric power management). Improving the quality of probabilistic forecasts is currently an active area of research.

Ensemble methods (Leith, 1974) are typically used as the basis for generating probabilistic forecasts. Ensembles aim to sample the probability density function (PDF) of the true error distribution, but are frequently found to be underdispersive (Hamill and Colucci, 1998; Buizza et al., 2005; Raftery et al., 2005). Also, the spread-skill relationship of ensembles is often found to be non-existent or weak (Hamill and Colucci, 1998; Stensrud et al., 1999), although stronger relationships have been found in some cases (Grimit and Mass, 2002; Stensrud and Yussouf, 2003; Scherrer et al., 2004). This has led to the development of statistical methods, which instead of assuming that the ensemble perfectly samples the true error PDF, assigns a probability distribution based on various attributes of the ensemble. Methods such as ensemble model output statistics (EMOS; Gneiting et al., 2005) and Bayesian model averaging (BMA; Hoeting et al., 1999; Raftery et al., 2005) are frequently used. The focus of our work is on such statistical methods for generating and improving probabilistic forecasts.

4.1.1 Notation

We focus our study on continuous weather variables (such as temperature) and mixed discrete-continuous variables (such as precipitation amount). Given a weather variable X, a probabilistic forecast for time t can be given by a cumulative probability distribution (CDF) $F_t(x)$, giving the probability that X takes on a value less than x. This distribution can be created in many ways, but will generally be based on output from numerical weather prediction (NWP) model runs. Also useful is the PDF given by:

$$f_t(x) = \frac{dF_t(x)}{dx}, \tag{4.1}$$

which indicates the relative likelihoods of various values of X.

4.1.2 Verification

The observed state of X at time t is denoted by x_t. These observations are used to determine the quality of the probabilistic forecasts. The continuous ranked probability score (CRPS; Hersbach, 2000) is a metric commonly used to evaluate the performance of probabilistic forecasts and is defined as:

$$\text{CRPS} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \int_{-\infty}^{\infty} \big[F_t(x) - H(x - x_t)\big]^2 dx, \tag{4.2}$$

where $\mathcal{T}$ represents a set of time points used for evaluation, $|\mathcal{T}|$ is the size of this set, and $H(s)$ is the Heaviside function defined by:

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases}. \tag{4.3}$$

The CRPS is the integral of the Brier Score (BS; Brier, 1950) over all values of X. The CRPS rewards probabilistic forecast distributions that are narrow and centred around the observed state. A low CRPS value is preferred.

The logarithm score (Good, 1952), often known as the ignorance score (Roulston and Smith, 2002), is also commonly used and is defined by:

$$\text{IGN} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} -log_2\big[f_t(x_t)\big]. \tag{4.4}$$

The ignorance score rewards forecasts that prescribe high probability density at the variable value that is verified. Low ignorance scores are preferred.

The CDF value corresponding to the verifying observation is commonly referred to as the prob-

ability integral transform (PIT) value and is denoted by p_t:

$$p_t = F_t(x_t). \tag{4.5}$$

A set of probabilistic forecasts are probabilistically calibrated if p_t values are uniformly distributed on the interval $[0, 1]$ (Gneiting et al., 2007).

4.1.3 Mixed discrete-continuous variables

An added complication occurs for any variable that has discrete probability mass in parts of the variable's domain. Quantitative precipitation rate, for example, can have a finite probability mass at $X = 0\,\mathrm{mm}$ corresponding to the probability of no precipitation. Following Sloughter et al. (2007), we separate the distribution into a discrete and a continuous part:

$$F(x) = \begin{cases} P & x = 0\,\mathrm{mm} \\ P + F_c(x)(1 - P) & x > 0\,\mathrm{mm}, \end{cases} \tag{4.6}$$

where P is the probability mass of the discrete part, and $F_c(x)$ is the probability distribution for the continuous part. The normalization by $(1 - P)$ allows us to define the range of $F_c(x)$ to be the interval $[0, 1]$.

When computing the CRPS for this mixed distribution, the lower bound of integration in Eq. (4.2) becomes $0\,\mathrm{mm}$. When computing the ignorance score for precipitation, we use the probability mass P_t for those cases when no precipitation was observed:

$$\mathrm{IGN} = \frac{1}{|\mathcal{T}| + |\mathcal{T}_0|}\left(\sum_{t\in\mathcal{T}} -log_2\big[f_t(x_t)\big] + \sum_{t\in\mathcal{T}_0} -log_2\big(P_t\big)\right), \tag{4.7}$$

where $\mathcal{T}$ represents cases where precipitation was observed and $\mathcal{T}_0$ represents cases where no precipitation was observed.

Evaluating the performance of the discrete part by itself can also be useful and for this we use the Brier Score with a threshold of $0\,\mathrm{mm}$:

$$BS(0\,\mathrm{mm}) = \begin{cases} (1 - P_t)^2 & x_t = 0\,\mathrm{mm} \\ P_t^2 & x_t > 0\,\mathrm{mm} \end{cases}. \tag{4.8}$$

We define an analogous metric for the ignorance score of the discrete part:

$$\text{IGN}(0\,\text{mm}) = \begin{cases} -log_2(P_t) & x_t = 0\,\text{mm} \\ -log_2(1 - P_t) & x_t > 0\,\text{mm} \end{cases}, \tag{4.9}$$

which is the ignorance score for a binary random variable.

4.1.4 Goals

Generating probabilistic weather forecasts has been well studied and a large number of methods and models have been developed to improve verification scores. The goal of this chapter is to present a probabilistic forecast system that can combine these techniques to find combinations that work particularly well. This is done by separating the process of probabilistic forecast generation into a series of sequential components, each of which has a specific role. This separation allows each component to be researched and improved independently of the other components.

The forecast system and its components are described in Section 4.2, a software approach for implementing the system is presented in Section 4.3, the functionality of the system is tested on a temperature and precipitation case study in Section 4.4, and conclusions are drawn in Section 4.5.

4.2 System description

We decompose the process of generating probabilistic forecasts into a series of steps originating with a set of input predictors (Figure 4.1). Before a probabilistic forecast can be disseminated, the predictors pass through each of these steps (termed *components*). The proposed system contains correction, uncertainty, calibration, and update components, each of which serve a very special purpose in the overall aim of producing high-quality probabilistic forecasts.

Each component can be defined and implemented in a number of different ways, and we call a specific implementation a *scheme*. This idea is analogous to the modular approach taken by some community-developed NWP models, such as the Weather Research and Forecasting (WRF; Skamarock et al., 2005) model, where the development of different microphysics, radiation, surface, and boundary layer schemes can be done relatively independently of the rest of the model provided the scheme conforms to the requirements imposed by the software framework. Instead of having physically-based components, the proposed probabilistic system has a number of statistically-based components. A user of the system selects one scheme for each component, which defines a *configuration* of components.

The system defines a set of input and output requirements for each component. Provided that a scheme conforms to these requirements, the scheme can be used in conjunction with any com-

bination of schemes for the other components. This setup easily allows improvement efforts to be focused into specific areas, and allows the research to be done independent of the other components. The components included in the decomposition were selected to allow us to include the majority of commonly used probabilistic methods found in NWP. We are unaware of any other studies that use such a decomposition approach.

Each component uses the output from the previous component as well as a set of parameters $\boldsymbol{\theta}_t$ to generate output for the next component. In an operational setting, these scheme-specific parameters can vary with time because they can be continuously trained using new observations as they become available.

The components of the system are discussed next. An illustration of the schemes that we have implemented for each component for our case study are summarized in Figure 4.2.

4.2.1 Predictors

The output from NWP model runs usually provides the basis for probabilistic weather forecasts. In many cases the output will be in the form of an ensemble of forecasts, although a single deterministic forecast could also be used. We term these forecasts *predictors* and the N predictors are denoted by $\xi_1, \xi_2, ..., \xi_N$.

The above specification does not prevent us from using an ensemble of past verifying observations as our predictors. These "predictors" form the basis for climatological forecasts, which can be used as a baseline against which to evaluate probabilistic-forecast skill based on NWP runs.

Although not investigated in this chapter, an ensemble of past analogs (Hamill and Whitaker, 2006; Delle Monache et al., 2011) or any other variables that have predictive capabilities could also be used.

4.2.2 Correction

The output from NWP models often exhibit biases that can be removed by post-processing. Post-processing uses the past behaviour of the predictors in order to correct for any systematic errors. A large variety of post-processing methods used for NWP exists, such as model output statistics (MOS; Glahn and Lowry, 1972), Kalman filtering (Homleid, 1995), neural networks (Yuval and Hsieh, 2002; Marzban, 2003), analog methods (Delle Monache et al., 2011), and gene-expression programming (Bakhshaii and Stull, 2009).

The requirement of the correction component is that it takes the set of predictors $\{\xi_1, \xi_2, ..., \xi_N\}$ and produces a corrected set $\{\hat{\xi}_1, \hat{\xi}_2, ..., \hat{\xi}_N\}$. An implementation of a correction scheme needs not concern itself with how the inputs were created, only that there are N available members. Similarly,

how the corrected members will be used later on is also not of concern.

As an example for our case study, we use two simple algorithms for correcting the input predictors for temperature. The first method removes a common bias term μ from each ensemble member:

$$\hat{\xi}_i = \xi_i - \mu, \tag{4.10}$$

where μ is determined from past performance of the ensemble. A second method removes a separate bias term from each ensemble member:

$$\hat{\xi}_i = \xi_i - \mu_i. \tag{4.11}$$

The two above correction methods cause problems for correcting precipitation variables as it can generate negative precipitation amounts. Instead of improving the amount of precipitation, for the purposes of this chapter we focus on improving the ensemble's ability to distinguish between the occurrence and non-occurrence of precipitation. The number of ensemble members predicting an occurrence of precipitation can be useful for gauging the probability of precipitation P. However, we found from the data set used in our case study that the ensemble often produces too many small, but non-zero ensemble members when evaluated against the observations. This had the effect that the ensemble could not resolve days with low chance of rain from those with a high chance since for the most part the number of non-zero ensemble members was always very high or very low. To correct this problem, a cut-off value ϵ can be used to define the minimum accumulation amount needed for a forecast to be considered an occurrence of precipitation:

$$\hat{\xi}_i = \begin{cases} 0\,\mathrm{mm} & x \leq \epsilon \\ \xi_i & x > \epsilon \end{cases}. \tag{4.12}$$

This allows the ensemble to better differentiate between low and high probability days.

4.2.3 Uncertainty model

An ensemble of corrected predictors does not by itself comprise a probabilistic forecast. Before a full probability distribution $F(x)$ can be constructed, a suitable interpretation of how the set of predictors represents forecast uncertainty must be chosen. We term this an uncertainty model. The uncertainty model prescribes a probability distribution based on the arrangement of the (corrected) input predictors.

There are several common ways to prescribe probability given an ensemble of forecasts (see discussion in Chapter 2), such as the binned probability ensemble (BPE; Anderson, 1996) technique,

BMA, moment-based methods (Jewson et al., 2005), and EMOS. Common to these models is that they prescribe probability based on both the central tendency of the predictors as well as the level of disagreement among the predictors.

Temperature variables

For temperature variables, we have restricted our study to models using Gaussian distributions. The probability distribution is created by centering a Gaussian distribution on the corrected ensemble mean, and adjusting its variance appropriately. That is:

$$F(x) = \phi\big(x; \hat{\xi}; \sigma^2\big), \tag{4.13}$$

where ϕ is a Gaussian CDF for some variable value x, $\hat{\xi}$ is the corrected ensemble mean, and σ^2 is the distribution variance. The variance is the one free parameter and effectively determines the uncertainty of the forecast.

There are a number of ways to fix the variance of this distribution (Jewson et al., 2005; Gneiting et al., 2005):

$$\sigma^2 = a \tag{4.14}$$

$$\sigma^2 = b\sigma^2_{\hat{\xi}} \tag{4.15}$$

$$\sigma^2 = a + b\sigma^2_{\hat{\xi}}, \tag{4.16}$$

where $\sigma^2_{\hat{\xi}}$ is the ensemble variance and a and b are constants. Eq. (4.14) refers to a model where the spread of the distribution is independent of the ensemble spread (e.g. Gaussian constant spread model in Figure 4.2). The ensemble spread can be used as a gauge for uncertainty, since the disagreement among the ensemble members suggests a general difficulty in determining the future state. Eq. (4.15) uses only the ensemble spread, with a scaling factor to account for under or over-dispersion of the ensemble (e.g. ensemble spread model) and Eq. (4.16) combines both (e.g. full regression model). Linear regression between the squared ensemble mean error $(\hat{\xi}_t - x_t)^2$ and the ensemble spread $\sigma^2_{\hat{\xi}}$ can be used to determine parameters a and b.

Precipitation

For precipitation, we need a separate model for each of P and $F_c(x)$. Logistic regression is commonly used for P (Sloughter et al., 2007) as this forces P to be restricted to the interval $(0, 1)$. Logistic regression in this case uses linear regression to fit the logarithm of the ratio of odds of no

precipitation to precipitation:

$$\text{logit}P = log\left(\frac{P}{1-P}\right). \tag{4.17}$$

We investigate three regression equations, which use the cubic root of precipitation (Sloughter et al., 2007) as a variate:

$$\begin{aligned} \text{logit}P &= c_0 + c_1\sqrt[3]{\hat{\xi}} && (4.18)\\ \text{logit}P &= c_0 + c_1\delta && (4.19)\\ \text{logit}P &= c_0 + c_1\sqrt[3]{\hat{\xi}} + c_2\delta, && (4.20) \end{aligned}$$

where δ is the fraction of ensemble members predicting no precipitation. Eq. (4.18) is referred to as the ensemble mean model in Figure 4.2, Eq. (4.19) as the ensemble fraction model, and Eq. (4.20) as the combined model. For climatological predictors, the empirical fraction of ensemble members predicting no precipitation can be used:

$$P = \delta. \tag{4.21}$$

The continuous part of the precipitation distribution is often modeled using a gamma distribution (Hamill and Colucci, 1998; Sloughter et al., 2007):

$$F_c(x) = \Gamma\big(x;\alpha;\beta\big), \tag{4.22}$$

where Γ is a gamma CDF, α is its shape parameter and β is its scale parameter. The shape and scale parameters as a function of the distribution's mean μ and variance σ^2 are:

$$\begin{aligned} \alpha &= \frac{\mu^2}{\sigma^2} && (4.23)\\ \beta &= \frac{\sigma^2}{\mu}. && (4.24) \end{aligned}$$

Sloughter et al. (2007) used

$$\begin{aligned} \mu &= c_0 + c_1\sqrt[3]{\hat{\xi}} && (4.25)\\ \sigma^2 &= c_2 + c_3\hat{\xi}, && (4.26) \end{aligned}$$

where c_0, c_1, c_2, and c_3 are constants. We had difficulty getting stable solutions using the adaptive method described in Section 4.3.3, since this model has four free parameters. We therefore use the

following model as it had one fewer free parameter:

$$\mu = c_0 + c_1 \sqrt[3]{\hat{\xi}} \tag{4.27}$$

$$\sigma^2 = c_2 \mu, \tag{4.28}$$

where c_0, c_1, and c_2 are constants. Contrary to temperature variables, precipitation uncertainty has been found to be better explained by the ensemble mean than the ensemble spread (Hamill and Colucci, 1998), which is why ensemble spread does not appear in these equations.

We found that when using climatological predictors, the empirical moments of the ensemble gave a good fit:

$$\mu = \hat{\xi} \tag{4.29}$$

$$\sigma^2 = \sigma^2_{\hat{\xi}}, \tag{4.30}$$

where $\sigma^2_{\hat{\xi}}$ is the variance of the climatological predictors.

4.2.4 Calibration

In some cases, the probability distribution produced by the uncertainty component may exhibit calibration deficiencies when the subsequent observations do not fit the assumptions used. For example, if a Gaussian distribution is used for cases where the distribution of observations is non-Gaussian, the resulting forecasts will exhibit distributional bias. The calibration component is the distributional analogy of the bias-correction performed by the correction component on the deterministic predictors.

Calibration deficiencies can be corrected by a calibration method that maps probability values F to calibrated probabilities $\hat{F}$ by using a calibration function Φ as follows:

$$\hat{F}(x) = \Phi[F(x)]. \tag{4.31}$$

Calibration in the form of Eq. (4.31) can be implemented in a number of ways (see for example Bremnes, 2007, or Chapter 2). For precipitation, we use separate calibration functions for the discrete and continuous parts:

$$\hat{F}(x) = \begin{cases} \Phi_0\big(P_0\big) & x = 0\,\mathrm{mm} \\ \Phi_0\big(P_0\big) + \Phi\big[F_c(x)\big]\big[1 - \Phi_0\big(P_0\big)\big] & x > 0\,\mathrm{mm} \end{cases}. \tag{4.32}$$

4.2.5 Updating

In the time between when a forecast is first produced and when the next forecast cycle starts, new observations may become available. The original probabilistic forecast can in general be improved by statistically assimilating this observation (Chapter 3). This component uses the fact that the PIT values are highly correlated in time. For example, when a forecast verifies in the 20th percentile, it will likely continue to do so for the next few hours. A probabilistic forecast $\hat{F}_t(x)$ valid for time t can be updated by a recent PIT value p_{t-n} n hours prior (i.e. $p_{t-n} = \hat{F}_{t-n}(x_{t-n})$). The updating is performed by an update function U as follows:

$$\hat{F}_{t|t-n}(x) = U\big[\hat{F}_t(x), p_{t-n}\big], \tag{4.33}$$

where time indices refer to time points within the same forecast run.

Chapter 3 used a reflected Gaussian distribution for U with a single fitting parameter σ determined by how correlated in time the PIT values are. Updating in this form can be a useful alternative to a computationally expensive full data assimilation followed by a re-initialization of the models driving the ensemble. Since the correlation of a recent PIT value vanishes quickly with time, this component is most useful for probabilistic nowcasting purposes.

At this point, the initial predictors have undergone several transformations and improvements, and are now ready to be disseminated and verified.

4.3 Implementation

The software strategy used to implement the system must aim to achieve the goal of modularity described earlier. Not only must schemes be interchangeable, but adding a new scheme should only require the developer to write code that directly defines the scheme, with a minimal amount of additional code to be added elsewhere in the framework. Finally, computationally efficiency is also important for operational purposes, as potentially thousands of locations and many weather variables must be processed every day.

4.3.1 Approach

An object oriented (OO) software approach is used here, as this allows the components to be easily modularized by exploiting polymorphism and function inheritance features of OO. Each component is defined by an abstract class that specifies what functionality must be implemented by a candidate scheme. Provided that a scheme implements all functions of its parent class, this scheme can be used by the system in combination with any scheme of any other component.

Code abstraction is important in this system. A developer of a calibration scheme, for example, should be insulated from code elsewhere in the framework. Any calibration implementation relies only on a CDF value provided by the uncertainty model, and need not concern itself with how that output was created. This simplifies the task of writing a scheme since only the input and output specifications must be dealt with.

We have made extensive use of function inheritance. For example for our case study, the three logistic regression models need only to specify what regression variables to use and how these are combined to predict the log odds ratio. These three classes only involve computer code that directly implements Eq. (4.18)-Eq. (4.20). Adding a new regression model is as simple as creating a new class that specifies the variables needed and implementing a log odds ratio function. Functionality to estimate the parameters of the model is inherited from the maximum-likelihood class (Figure 4.2). The maximum-likelihood class adaptively finds the optimal parameters for the model by relying on its subclasses to implement the likelihood function corresponding to those parameters.

Even though the goal of the discrete uncertainty class is to output a value for P, none of the three logistic regression schemes directly need to implement code that computes P. Thus, function inheritance greatly reduces duplication of code, and the adding of new schemes generally only requires the implementation of code that directly defines its core behaviour.

4.3.2 System outputs

Users of the system are interested in CDFs, PDFs, and in some cases in inverse CDFs (e.g. to compute a confidence interval). Instead of requiring all components to be able to provide functionality to compute all three types of output, we focus on producing a CDF. That is, from the uncertainty component and onward, a probability distribution described by a CDF is passed between the components. The PDF can then be computed by polling the CDF and computing the derivative numerically. To compute the inverse CDF, the system uses a simple iterative approach by polling the CDF for different values of x until one that gives an $\hat{F}(x)$ that is close to the desired inverse value.

4.3.3 Adaptive parameter estimation

Each scheme of the system can make use of a set of stored parameters θ. To achieve the goal of computational efficiency for operational (real-time) forecasts, we require all scheme parameters to be computed adaptively. This reduces the computational requirements since only the previous estimate of the parameter must be retrieved, instead of a long history of past data.

The best estimate of a parameter θ at time t is denoted by θ_t. Let θ_t^* represent the parameter computed solely by the information provided by the observation at time t. The parameter can then

be updated to a new value θ_{t+1} based on the recursive equation:

$$\theta_{t+1} = \frac{\tau - 1}{\tau}\theta_t + \frac{1}{\tau}\theta_t^*, \tag{4.34}$$

where $\tau > 0$ is a unitless time scale corresponding to how quickly the effect of new information vanishes over time. Large values of τ will cause θ to adapt slowly to new information. θ_{t+1} is a weighted average of the previously best estimate and the current evidence.

To update the parameters, each scheme must define how θ_t^* is determined. For example, the parameter μ in Eq. (4.10) can be updated as follows:

$$\mu_{t+1} = \frac{\tau - 1}{\tau}\mu_t + \frac{1}{\tau}\big(\hat{\xi}_t - x_t\big). \tag{4.35}$$

The parameter ϵ in Eq. (4.12) can be updated as follows:

$$\epsilon_{t+1} = \begin{cases} \frac{\tau-1}{\tau}\epsilon_t + \frac{1}{\tau}\hat{\xi}_t & x = 0\,\mathrm{mm} \\ \epsilon_t & x > 0\,\mathrm{mm} \end{cases}. \tag{4.36}$$

That is, over time, ϵ approaches the average value of the ensemble mean when there is no observed precipitation.

The regression parameters used to compute parameters a and b in Eq. (4.14)-Eq. (4.16) can be updated in a similar fashion.

For maximum-likelihood methods in the uncertainty component, we use the parameter estimation technique employed by Pinson and Madsen (2009). Here a vector of parameters $\boldsymbol{\theta}$ is determined simultaneously using:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \frac{1}{\tau}\mathbf{R}_t^{-1}\frac{\nabla L(\boldsymbol{\theta}_t, x_t)}{L(\boldsymbol{\theta}_t, x_t)}, \tag{4.37}$$

where $L(\boldsymbol{\theta}_t, x_t)$ is the likelihood function for the parameters $\boldsymbol{\theta}_t$ with verifying observation x_t, and where the covariance matrix $\mathbf{R}_t$ is defined as:

$$\mathbf{R}_{t+1} = \frac{\tau - 1}{\tau}\mathbf{R}_t + \frac{1}{\tau}\left(\frac{\nabla L(\boldsymbol{\theta}_t, x_t)}{L(\boldsymbol{\theta}_t, x_t)}\right)\left(\frac{\nabla L(\boldsymbol{\theta}_t, x_t)}{L(\boldsymbol{\theta}_t, x_t)}\right)^T. \tag{4.38}$$

The estimation method requires any scheme inheriting this class to define the likelihood function L, which for logistic regression is the solution to P in Eq. (4.17), and for the gamma model is the gamma PDF corresponding to Eq. (4.22).

The calibration method defined in Chapter 2 uses a large collection of past PIT values to calibrate

the probability distribution. Instead of storing individual PIT values, an adaptive alternative to this is to only use a few calibration points Φ_p corresponding to several values of p and adaptively move these calibration points to approximately match the distribution of past PIT values. These points can then be updated as follows:

$$\Phi_{p,t+1} = \frac{\tau - 1}{\tau}\Phi_{p,t} + \frac{1}{\tau}H(p - p_t), \tag{4.39}$$

where p_t is the PIT value corresponding to the verifying observation. We used 9 evenly distributed calibration points on the interval $[0, 1]$ for our case study as this balances the need to resolve patterns and smoothing out noise in the data (Chapter 2).

We chose to not create a calibration function Φ_0 for the discrete part, by reasoning that the probability model for the discrete part has several parameters to determine a single point, whereas the model used for the continuous part must fit to the entire distribution. Thus, the model for the discrete part should already create probabilities that do not have an overall tendency to under or overpredict probability of no precipitation.

For the update scheme, the single parameter σ can be updated adaptively as follows:

$$\sigma^2_{t+1,\hat{t}} = \frac{\tau - 1}{\tau}\sigma^2_{t,\hat{t}} + \frac{1}{\tau}(p_{t,\hat{t}+1} - p_{t,\hat{t}})^2, \tag{4.40}$$

where t represents different days the input predictors are initialized and $\hat{t}$ represents two different time points within the same forecast run. That is, $p_{t,\hat{t}}$ and $p_{t,\hat{t}+1}$ represent two consecutive PIT values for a certain forecast run.

For all corrector schemes, the update scheme, and uncertainty models (other than maximum-likelihood methods) we used a dimensionless time-scale of $\tau = 30$ iterations, as similar time-scales (in the form of window lengths) have been determined to be suitable for such methods (Raftery et al., 2005; Sloughter et al., 2007; McCollor and Stull, 2008a). The maximum-likelihood method required a longer time-scale since several parameters are estimated simultaneously. For these we used $\tau = 60$ iterations. PIT-based calibration requires on the order of 100 data-points to be effective (Chapter 2) and therefore we used $\tau = 90$ iterations for the calibration component.

4.3.4 Bypass schemes

When using the system, the correction, calibration, and update components are optional. For each of these we create a scheme that bypasses the component for cases where these methods are not required or wanted. The bypass scheme simply provides an output that is an unaltered version of its input.

The correction, calibration, and update bypass schemes implement respectively:

$$\hat{\xi}_i = \xi_i, \tag{4.41}$$

$$\hat{F}_t(x) = F_t(x), \text{ and} \tag{4.42}$$

$$\hat{F}_{t|t-n}(x) = \hat{F}_t(x). \tag{4.43}$$

Note that the predictor component is a required component because without it the forecast chain cannot be started and the uncertainty model is required because without it the predictors cannot be converted to probabilistic form.

4.3.5 Verification

Two verification metrics were mentioned in Section 4.1. Although not part of producing a probabilistic forecast, the same modular approach used for the probabilistic forecast system can be used to implement the verification of the forecasts. We again define an interface for verification that takes as input a probabilistic forecast and a corresponding observation, and outputs a verification score. The verification scheme can rely on the probabilistic forecast providing a CDF, PDF, and inverse CDF. If desired, an adaptive verification can be utilized that does not require the saving of a large array of past historical values, although we did not do this for our case study.

A side-effect of restricting the verification component to only take as input the forecast and corresponding observation, and not any information about the strategy taken by to produce the probabilistic forecast, is that all scores will be in accordance with Dawid's prequential principle (Dawid, 1984).

4.4 Case study

4.4.1 Data set

The goal of this section is to show how the proposed system can be used to gain insight into which combination of components provide the best probabilistic forecasts of surface weather. We tested the system on two ensemble prediction systems (EPS). Medium-range forecasts from the North American Ensemble Forecast System (NAEFS; Toth et al., 2006) were used for 24-h minimum temperature (MINT), 24-h maximum temperature (MAXT), and total 24-h precipitation (PCP). We used 42 members of the NAEFS ensemble: 21 members produced by the U.S. National Weather Service and 21 members produced by the Meteorological Service of Canada. Two of these members were control runs. MINT was taken from the 12UTC model output and MAXT was from 00UTC,

corresponding to night and day times in the domain of interest. PCP was computed by the total precipitation accumulation from 00UTC to 00UTC the next day. Data from 1 Sep 2010 to 31 Aug 2011 were used, with lead times between 1 and 15 days.

The UBC short-range ensemble forecasts (UBC-SREF) were also tested. This 20-member multi-model, multi-resolution ensemble consists of hourly forecasts from four NWP models including the Mesoscale Compressible Community (MC2; Benoit et al., 1997) model, the Penn State/NCAR Mesoscale Model (MM5; Grell et al., 1994), and versions 2 and 3 of the Weather Research and Forecasting (WRF; Skamarock et al., 2005) model. All models used initializations from the North American mesoscale (NAM) model and in addition the global forecast system (GFS) was used to initialize a second run of WRF version 2. Horizontal grid spacing ranging from 108-km to 1.3-km were used. The models were initialized once per day at 0000 UTC. We looked at hourly surface temperatures (THOUR) with lead-times up to 60 hours for the time period from 1 Jan. 2009 to 31 Dec. 2010. The first 10 forecast hours for this variable were discarded as this was generally the time required for the forecasts to complete after model initialization.

A large number of configurations of components were tested, as summarized in Table 4.1. These were based on schemes that were appropriate for each variable. The configurations used for the climatological baseline forecasts are summarized in Table 4.2.

Observations from 15 stations located in southern British Columbia, Canada (Figure 4.3) were used to train and evaluate the proposed probabilistic system. Probabilistic forecasts were constructed for each of the configurations of components, and were computed separately for each station and forecast offset. An example probabilistic forecast for THOUR as it passes through the components of the system is shown in Figure 4.4.

In the next subsections, we highlight how each component contributes to overall forecast quality.

4.4.2 Comparison of uncertainty models

We first look at how the choice of uncertainty model affects the overall quality of the resulting probabilistic forecasts. We found that this was dependent on which ensemble system used. What worked with one EPS did not necessarily work with the other.

For the temperature variables of the NAEFS ensemble (i.e. MINT and MAXT), the ensemble spread was found to be a useful variable in determining uncertainty. As seen in Figure 4.5a,c, the full-spread regression model [Eq. (4.16)] gave ignorance scores that were better than using a constant spread model [Eq. (4.14)]. This was especially true for lead-times greater than day 5. This model remained skillful (i.e. beating climatology) 3 days longer for MINT and 2 days longer for MAXT.

To determine the cause of this, we show ignorance scores as a function of the error of the median

of the forecast distribution (Figure 4.6a,c). The improvements in ignorance score arise due to poorer ignorance scores for forecasts with large errors. This is because the full-spread regression model is able to discriminate between days with low and high uncertainty. The lowest possible ignorance score for a Gaussian distribution with a fixed absolute error is one where the standard deviation matches the absolute error (dashed lines). The performance of the full-spread regression model is much closer to this minimum than the constant spread model.

The same corresponding improvement is much less evident in the CRPS (Figure 4.5b,d and Figure 4.6b,d). This is because CRPS is not sensitive to small probability values in the tails of the distribution. The difference of integrating $F(x)$ values of 0.001 or 0.01 is negligible [Eq. (4.2)]. However, the difference between ignorance scores for $f(x)$ values that are different by a factor of two is much larger. In fact, CRPS does not greatly penalize the occurrence of events deemed impossible by the forecast, whereas this would result in an infinitely large ignorance score. Accurate probability estimates in the tails of the distribution are important to users adverse to extreme weather conditions, and we therefore focus more on the conclusions provided by the ignorance score.

The ensemble spread model [Eq. (4.15)] proved to be less useful for lead-times of less than 5 days when evaluated by the ignorance score. However, for longer lead-times, the performance matched that of the full regression. This is likely because the ensemble spread has no skill as an uncertainty predictor for short lead times because an artificial spread is often imposed to create the initial ensemble states. This model then incorrectly responds to a noisy signal.

For THOUR, the full-spread regression model did not lead to any gains over the constant spread model (Figure 4.5e,f and Figure 4.6e,f). There are two possible explanations for this: 1) the UBC-SREF is a short-range system and ensemble spread is less useful at these lead-times, which also was the case for the NAEFS ensemble. 2) The UBC-SREF is a multi-model and multi-resolution ensemble, thereby being more heterogeneous than NAEFS, which only has two models each of which have constant grid resolution. This may mean that the spread-skill relationship of the UBC-SREF cannot be captured by a simple linear relationship.

The choice of discrete uncertainty model had an effect on PCP (Figure 4.5g,h). The full logistic regression model [Eq. (4.20)] performed best, however the ensemble mean model [Eq. (4.18)] performed only slightly worse. These two models outperformed the ensemble fraction model [Eq. (4.19)] mainly because of their ability to differentiate days with extreme precipitation events (Figure 4.7). These results suggest that the ensemble mean of precipitation was the strongest predictor of P and that the fraction of ensemble members predicting the non-occurrence of precipitation was less significant.

Another overall finding is that precipitation-probability forecasts are skillful out to only 5 to 6 day lead time, while temperature-probability forecasts are skillful out to 11 to 12 day lead times, for

these stations in mountainous western Canada.

4.4.3 Effect of correction and calibration

Temperature variables

We found that for the three temperature variables, applying a correction method significantly reduced the ignorance score (Figure 4.8) and CRPS (Figure 4.9) of the resulting probabilistic forecasts. The effect was investigated by running a configuration with different correction schemes for particular uncertainty models without using calibration and updating. For NAEFS (i.e. for MINT and MAXT), whether a common bias term [Eq. (4.10)] or a member-specific term [Eq. (4.11)] was used made little difference. This was also true for THOUR for the constant spread model [Eq. (4.15)], since a common and member-specific correction will give identical results, as the ensemble mean of the corrected predictors are identical in both cases. However, the choice of correction method did influence the performance of the two ensemble spread based methods [Eq. (4.15) and Eq. (4.16)], as seen in Figure 4.8h,i. This is likely due to the fact that each NWP model and grid resolution in the UBC-SREF has different biases. The member-specific correction method affected the ensemble spread in a way that benefited the full regression model [Eq. (4.16)], but caused lower performance for the ensemble spread model [Eq. (4.15)] when compared to the ensemble mean correction method.

We generally found that using a Gaussian model for temperature produced calibrated or near-calibrated forecasts provided the input predictors were corrected. However, for some stations, the PIT-histogram indicated skewness. This is an effect that cannot be corrected by increasing or decreasing the spread of the probability model, but rather is evidence that a Gaussian distribution is not a perfect fit. Calibration can correct for this distributional bias.

To investigate the effect of correction and calibration, we use the PIT-based decomposition of the ignorance score (Chapter 2). The ignorance score can be decomposed into an uncalibration component IGN_{uncal} and a base potential ignorance component IGN_{pot}:

$$\text{IGN} = \text{IGN}_{pot} + \text{IGN}_{uncal}. \tag{4.44}$$

IGN_{pot} is the lowest ignorance score possible if all calibration deficiencies are removed and IGN_{uncal} represents the added ignorance score due to deviations from a calibrated (i.e. flat) PIT-histogram. We used PIT-histograms with 10 equally sizes bins to compute the decomposition.

The effect of correction and calibration schemes on the components of ignorance can be shown schematically as in Figure 4.10. Two sample stations requiring calibration (panels a and b) and two

stations that do not (panels c and d) are shown. The forecasts without correction and calibration have high ignorance scores due to high values of IGN_{uncal}. The correction component removes most of IGN_{uncal}, as the correction centres the Gaussian distribution properly by removing bias. This is seen by the improvement moving along lines of constant potential ignorance, approaching the calibration deviation expected by a perfectly calibrated forecast. Since PIT-histograms are based on a limited sample of PIT values, even a perfectly calibrated set of forecasts is expected to have some small amount of calibration deviation. This amount is indicated by the gray vertical lines in Figure 4.10.

For YXJ and YXS, the correction did not remove all of IGN_{uncal} because it could not correct for skewness, as seen by the characteristic deviations in the PIT-histogram (Figure 4.10e-f). The calibration component completes the improvement. As noted in Chapter 2, calibrating models that are already nearly calibrated increases ignorance very slightly due to overfitting, which is seen for stations COQ and YVR. The corresponding PIT-histograms become flatter (i.e. better) after each type of correction.

Precipitation

For PCP, the rounding correction method [Eq. (4.12)] affects the logistic regression part of the uncertainty model, since it changes the number of predictors that forecast no precipitation. The correction had a favourable effect on the ensemble fraction discrete model [Eq. (4.19)] in both ignorance scores and CRPS, but a much smaller effect on the full regression [Eq. (4.20)] and ensemble mean models [Eq. (4.18)], as seen in Figure 4.8j-l and Figure 4.9j-l. The correction allowed the ensemble fraction model to better resolve cases of high probability of precipitation from cases with low probability.

Despite correction, the gamma model defined by Eq. (4.27) and Eq. (4.28), created forecasts with slight calibration deficiencies, regardless of the discrete model used (Figure 4.11). The non-flat PIT-histograms (Figure 4.11d-f) suggest that this model places too much confidence in lower precipitation values and too little in higher values. To alleviate this problem, a better gamma model could be devised using a different set of parameters, or the calibration method could be applied. The calibration method removed the remaining calibration deviation.

The ignorance decomposition also confirms that the correction improved the resolving abilities of the ensemble fraction model as the improvement in the ignorance score came solely from improved potential ignorance and not from reduced calibration deviation (Figure 4.11c).

4.4.4 Updating

As the observation reporting frequency of the UBC-SREF is higher than the frequency of NWP model initialization, statistical data assimilation of recent observations could be used to improve the THOUR forecasts (Figure 4.12). On an ignorance decomposition diagram, the updating improves the potential ignorance while remaining calibrated. It is interesting to note that even after applying the updating method, the ensemble spread model [Eq. (4.15)] still produced worse ignorance scores than the two other Gaussian models [Eq. (4.14) and Eq. (4.16)].

4.5 Conclusions

We have presented a system for creating and improving probabilistic forecasts. The modular system separates the major aspect of probabilistic forecast generation allowing research efforts to be focused into independent areas. The system is extendible such that new probabilistic forecasting models and methods can easily be added without affecting the rest of the system.

Each component serves a specific purpose to help improve forecast metrics. The correction and calibration schemes reduced the calibration component of ignorance, whereas better predictors, uncertainty models, and update schemes reduce potential ignorance. The implemented schemes are all adaptive, which is efficient for operational daily forecasts.

We anticipate that further improvements are possible by including more advanced correction schemes such as MOS, neural networks, analog methods, and genetic schemes, and more advanced uncertainty models such BMA. Also, other parameter estimation techniques commonly used for probabilistic weather forecasts such as expectation maximization (Dempster et al., 1977; Raftery et al., 2005), or CRPS minimization (Gneiting et al., 2005) could be incorporated.

In addition, the decomposition used in this study uses four components, however there are likely other components that could be added. As an example, a selection component could be added between predictors and corrector, which serves to select a set of the ensemble predictors (see for example Garaud and Mallet, 2011). This way, potentially low-quality ensemble members can be removed before the next stage, potentially resulting in better probabilistic forecasts.

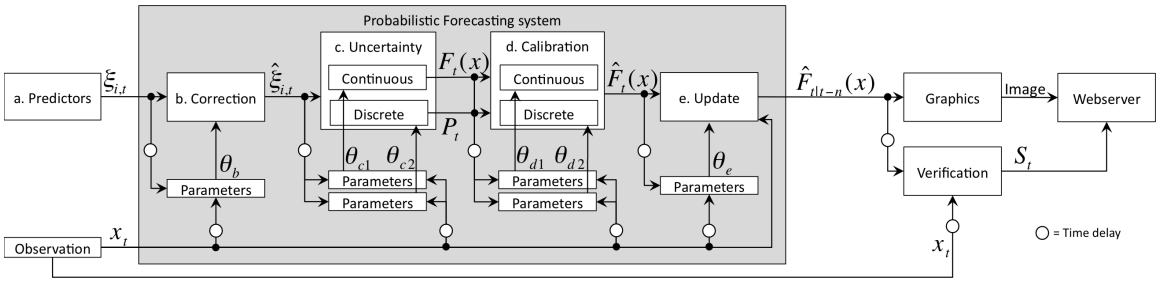

Figure 4.1: Schematic diagram of the components of the forecasting system. Input and outputs are shown by mathematical symbols and a delay in signals reaching components are shown by circles. Namely, parameter values used in the current time step are calculated using observations and predictors from the previous step.

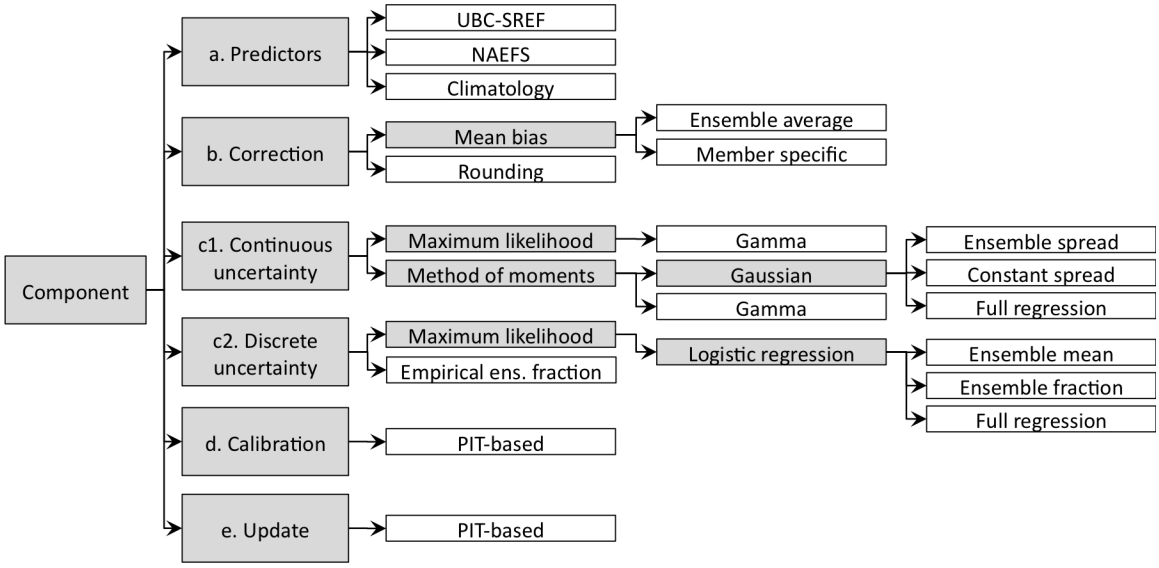

Figure 4.2: Inheritance hierarchy of the implemented schemes in the system as illustrated for our case study. Instantiable classes are shown in white and abstract classes (i.e. not instantiable) are shown in gray.

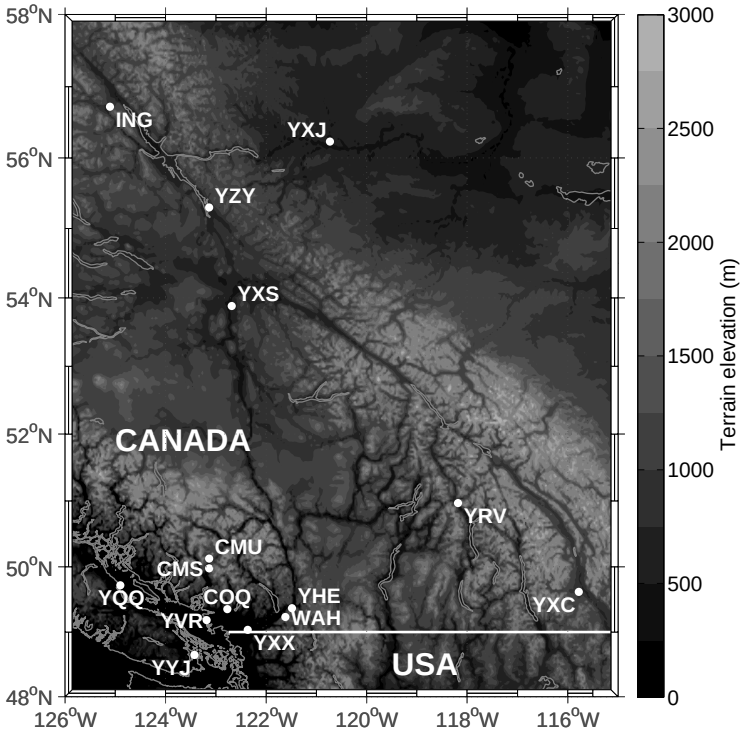

Figure 4.3: Case study observation stations in Southern British Columbia, Canada with their corresponding station code. Station codes starting with "Y" are for airport weather stations with ICAO designations that imply a prefixed "C" (e.g., YVR = CYVR = Vancouver International Airport). Station codes not starting with "Y" are operated by BC Hydro.

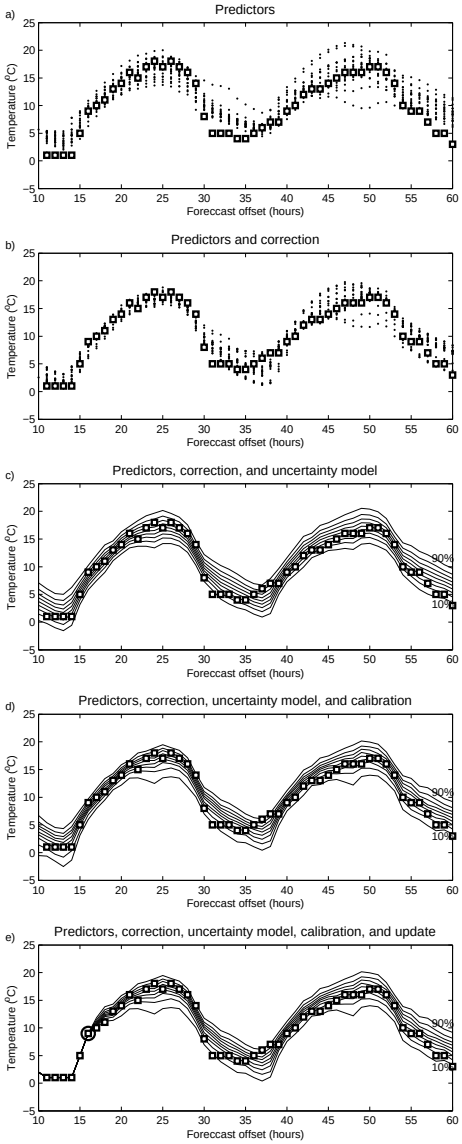

Figure 4.4: The state of the forecasts after passing different stages in the system for hourly temperature for station YXS (Prince George) for the forecast runs initialized on 00 UTC Jul 1, 2010. Ensemble members (dots), probabilistic forecasts for cumulative probabilities 10% to 90% with 10% increments (solid lines), and verifying observations (squares) are shown. a) Raw predictors; b) After the member-specific correction has been applied to the predictors; c) Forecast in probabilistic form after the constant spread uncertainty model has been applied; d) After calibration has been applied; e) After updating the forecast with the observation from 18 UTC Dec 31, 2010 (circled square).

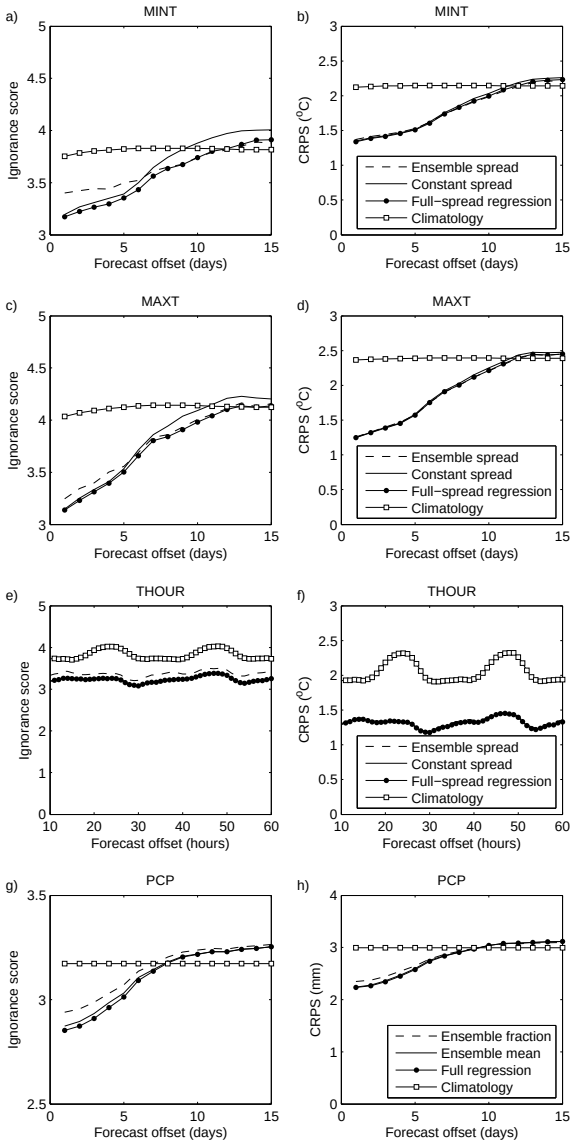

Figure 4.5: Overall verification scores for the forecast variables in the case study. Each row shows the scores for a particular variable, and each column shows a different a verification statistic. Scores for different uncertainty models are shown using rounding correction for PCP and member specific correction for temperature variables, except for THOUR Ensemble spread, which uses ensemble average correction. The PIT-based calibration scheme and no updating scheme were used. Smaller scores are better.

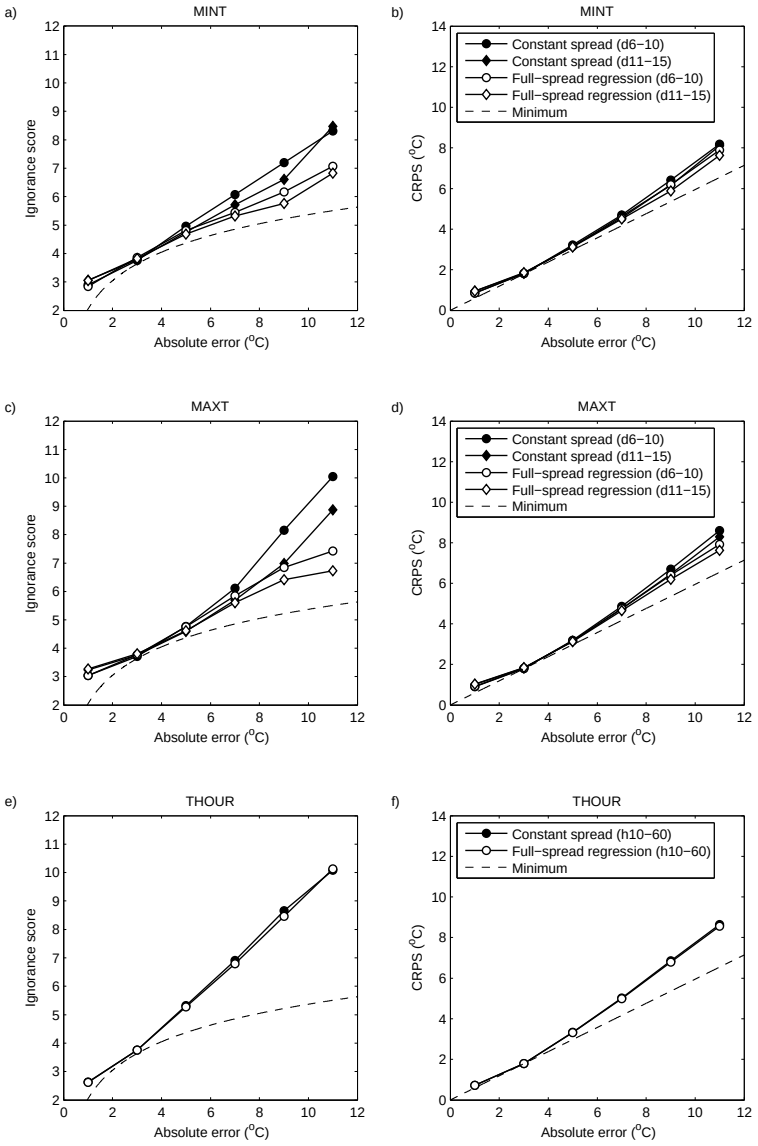

Figure 4.6: Ignorance and CRPS as a function of the absolute error of the median of the probability distributions. Each row represents one forecast variable and each column represents one verification statistic. For MINT and MAXT, averages for day 6-10 (circles) and day 11-15 (diamonds) are shown. For THOUR, averages over forecast hours 10-60 (circles) are shown. Smaller scores are better. Member-specific correction but no calibration and update scheme were used.

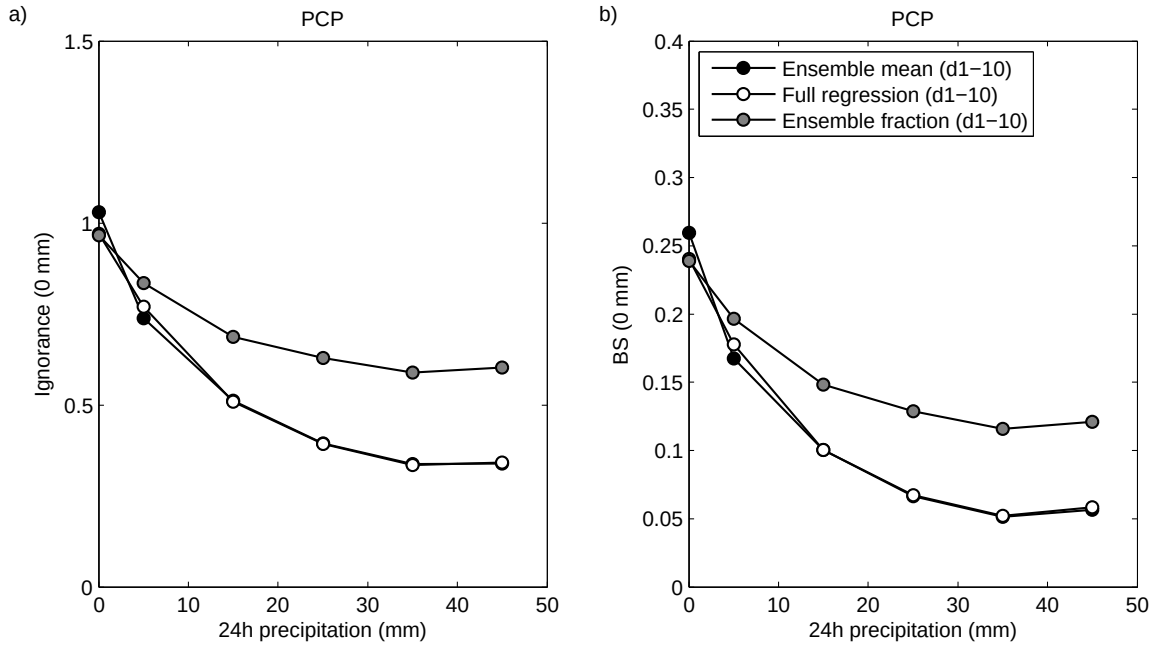

Figure 4.7: Similar to Figure 4.6 but for PCP and for the ignorance and Brier scores with thresholds of $0\,\mathrm{mm}$. Smaller scores are better. Rounding correction but no calibration and update scheme were applied.

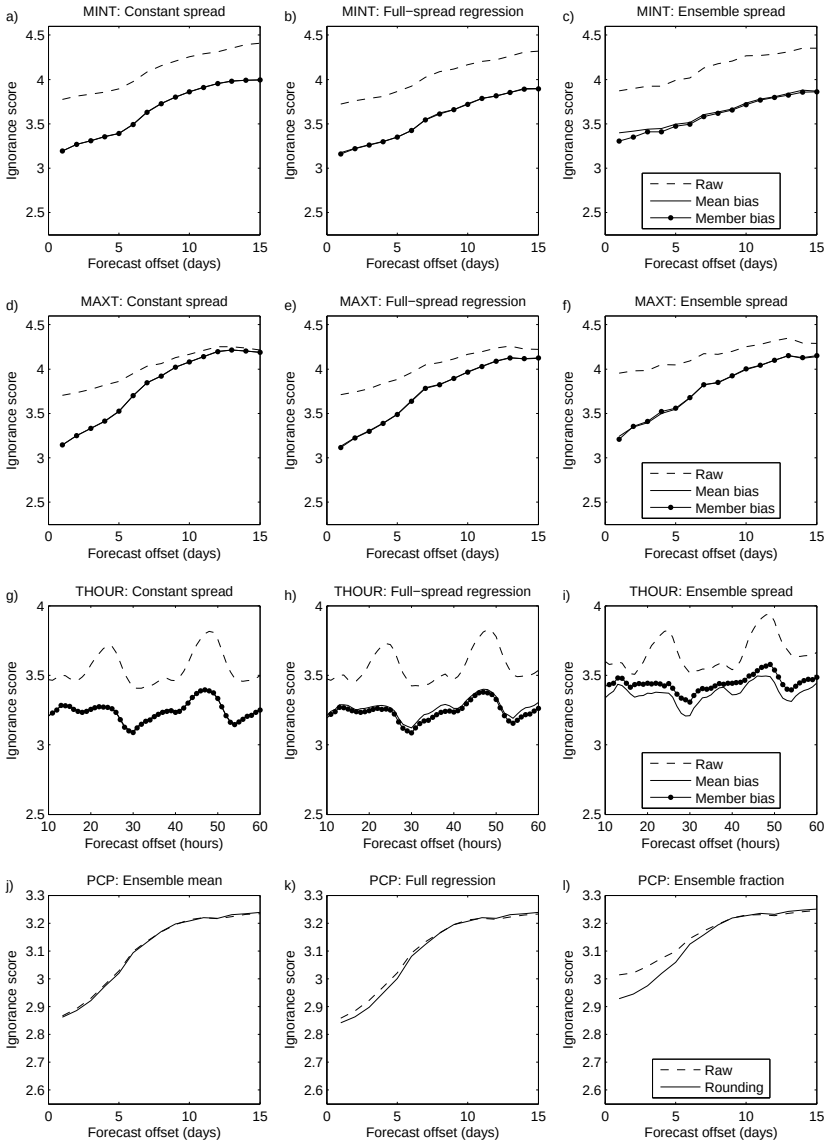

Figure 4.8: The effect of correction methods on the ignorance score of different variables (each row) and different uncertainty models (each column). No calibration and update schemes were used. For figures where the mean bias (solid line) is not visible, it coincides with the line for member bias. Smaller scores are better.

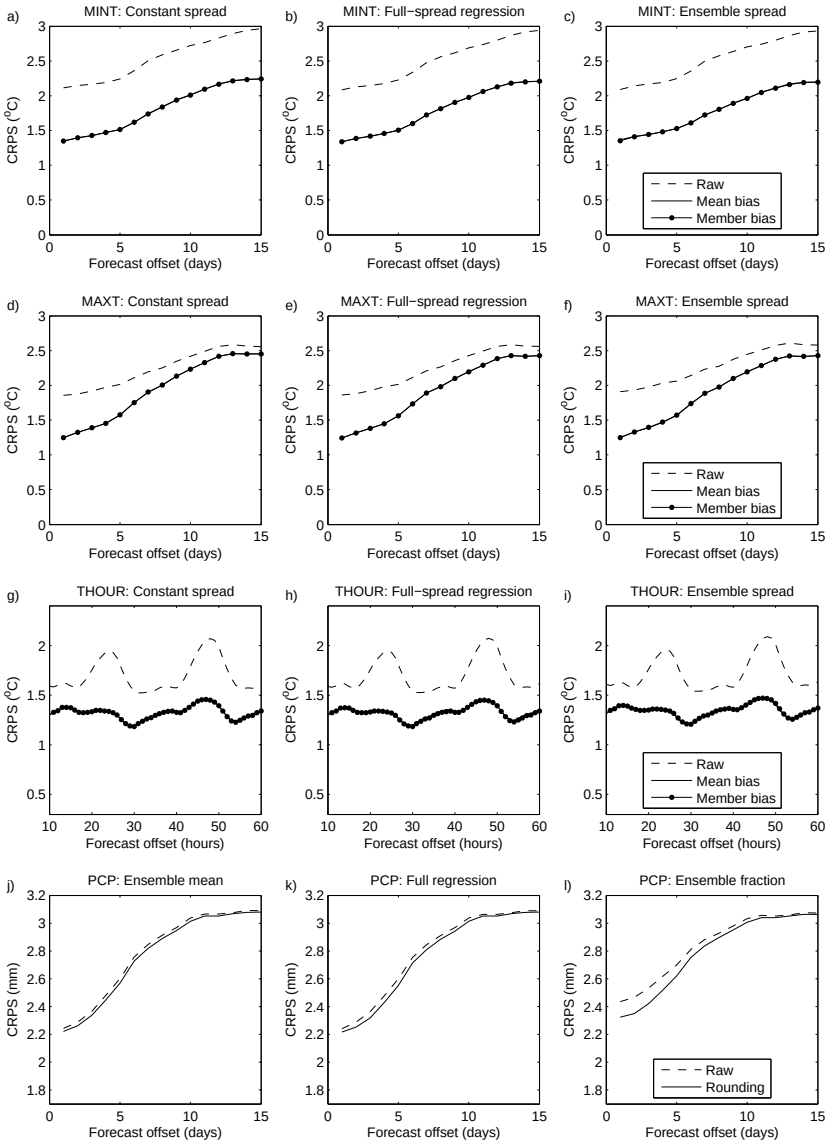

Figure 4.9: Similar to Figure 4.8 but for CRPS.

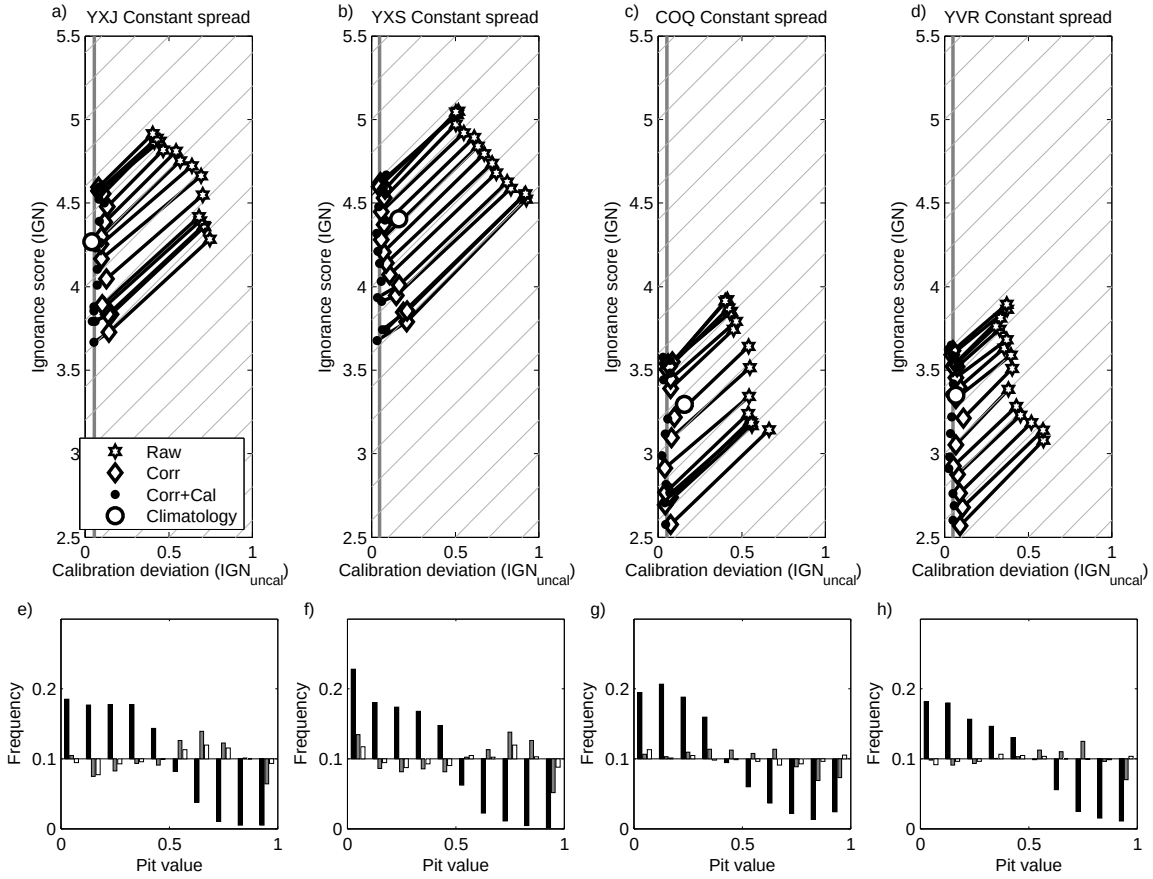

Figure 4.10: a-d) Ignorance decomposition graph for 4 stations for 24-h minimum temperature. Ignorance scores (IGN) and ignorance due to calibration deviation (IGN_{uncal}) is shown for the forecasts with different schemes enabled for the Constant spread uncertainty model. Each line represents a different lead-time, with short lead-times generally having lower ignorance scores. The expected calibration deviation of perfectly calibrated forecasts are shown by vertical gray lines. Diagonal lines represent lines of constant potential ignorance (IGN_{pot}). Corr refers to the Member specific correction technique. Smaller IGN and IGN_{uncal} values are better. e-h) PIT-histograms for the raw (black), corrected (gray), and corrected/calibrated (white) are shown. YXJ = Fort St. John, BC; YXS = Prince George, BC; COQ = Coquitlam, BC; YVR = Vancouver.

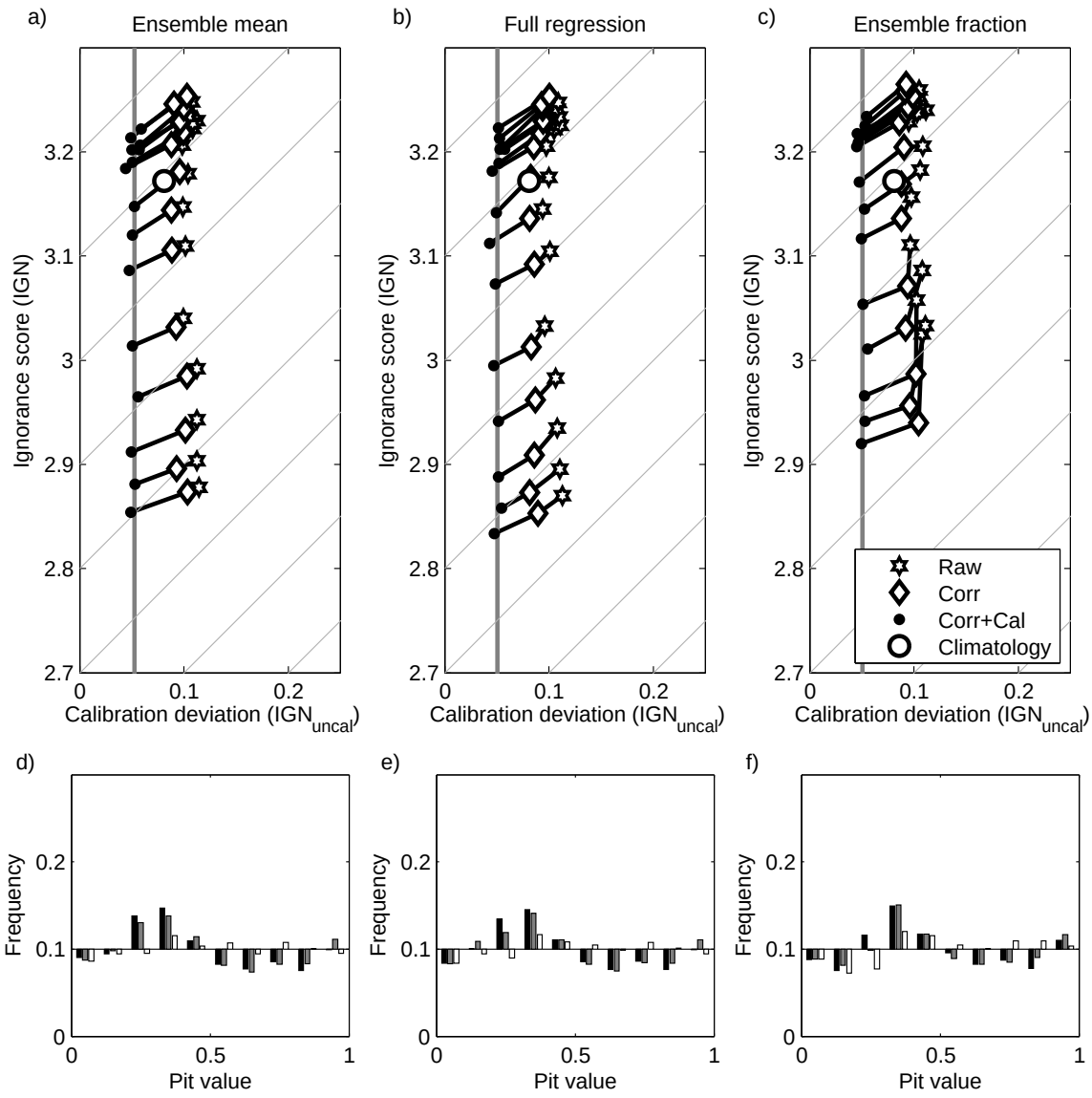

Figure 4.11: Similar to Figure 4.10 but for 24-h accumulated precipitation. Three discrete uncertainty models are shown: a) Ensemble mean, b) full-spread regression, and c) ensemble fraction. The gamma model was used for the continuous uncertainty model. The points are averages over all 15 stations and each line represents a different lead-time, with short lead-times generally having lower ignorance scores.

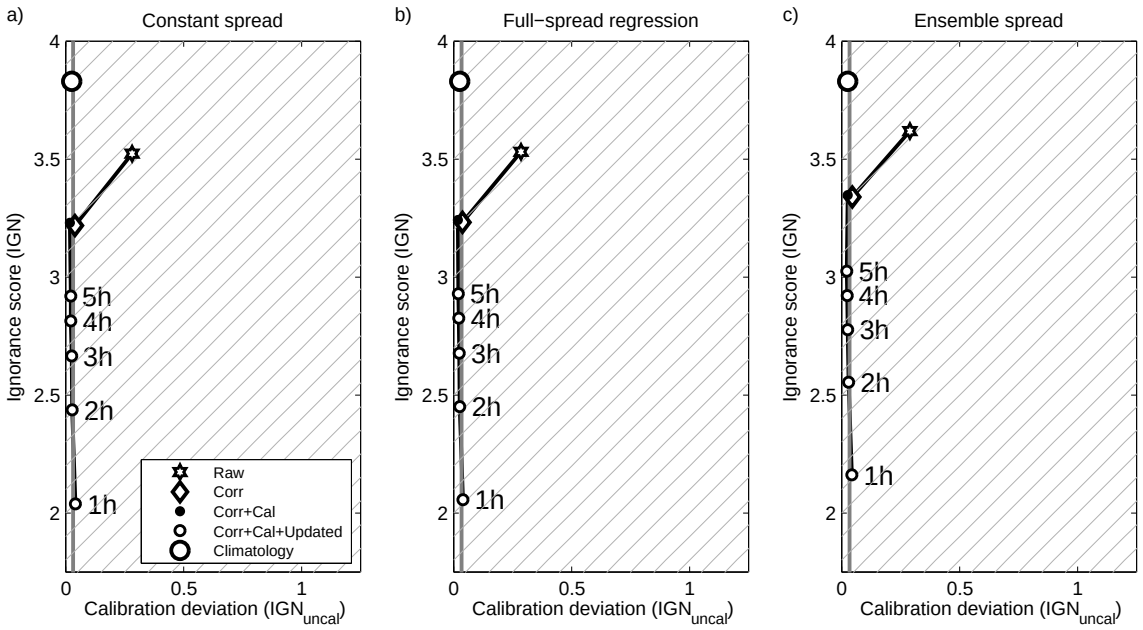

Figure 4.12: Similar to Figure 4.10 but for THOUR. Three uncertainty models are shown: a) Constant spread, b) Full regression, c) and Ensemble spread. The points have been averaged over all 15 stations and over forecast offsets 10h-60h. Smaller ignorance and calibration deviation are better.

Component	THOUR	MINT	MAXT	PCP
Predictors	UBC-SREF	NAEFS	NAEFS	NAEFS
Correction	Ensemble average Member specific	Ensemble average Member specific	Ensemble average Member specific	Rounding
Continuous Uncertainty	Ensemble spread Constant spread Full regression	Ensemble spread Constant spread Full regression	Ensemble spread Constant spread Full regression	Gamma (ML)
Discrete Uncertainty				Ensemble mean Ensemble fraction Full regression
Calibration	PIT-based	PIT-based	PIT-based	PIT-based
Updating	PIT-based			

Table 4.1: Combinations of schemes from Figure 4.2 used in the case study for hourly temperature (THOUR), 24-h minimum temperature (MINT), 24-h maximum temperature (MAXT), and 24-h accumulated precipitation (PCP). ML = Maximum likelihood.

Component	THOUR	MINT	MAXT	PCP
Predictors	Climatology	Climatology	Climatology	Climatology
Correction				
Continuous Uncertainty	Ensemble spread	Ensemble spread	Ensemble spread	Gamma (MM)
Discrete Uncertainty				Empirical ens. frac.
Calibration				
Updating				

Table 4.2: Similar to Table 4.1, but for climatological baseline forecasts. MM = Method of moments.

Chapter 5

Conclusions

The goal of this dissertation was to improve probabilistic forecasts for operational use. This has been achieved through the development of new methods and approaches to probabilistic forecasting.

5.1 Summary of methods and procedures

The dissertation proposed a four-stage decomposition process for generating probabilistic weather forecasts, as was presented in Chapter 4. This decomposition contains the following components: 1) correction; 2) uncertainty model; 3) calibration; and 4) updating. This allows research to be focused into specific areas, each of which help improve the overall quality of the resulting probabilistic forecast.

The decomposition was implemented in a probabilistic forecasting system using an object-oriented software strategy. This strategy insulates a developer of a particular scheme from the implementation details of the other components, and also allows for the interchangeability of schemes.

The decomposition resulted in the development of a number of other new methods:

- A new calibration method for probabilistic forecasts was presented in Chapter 2. This method takes as input a full probability distribution, and removes any distributional bias the distribution may have. The calibration is based on the distribution of past verifying probability integral transform (PIT) values.

- A new statistical updating scheme for probabilistic forecasts that incorporates recently made observations was presented in Chapter 3. The method improves probabilistic forecasts in the short-term by relying on verifying PIT values being correlated in time. The method models the sequence of PIT values as a first-order Markov process, using a reflected Gaussian transition function.

- A new decomposition of the ignorance score was presented in Eq. (2.27). This decomposition separates the ignorance score into a component related to the amount of distributional bias and a remaining component related to the base resolving ability of the forecast. The ignorance

decomposition identifies how a certain statistical method affects the overall ignorance score, as shown in Chapter 4.

5.2 Summary of findings

A number of findings were made from the evaluation of the probabilistic forecast system and its components:

- The schemes that work well for one set of input predictors did not necessarily work well for another. The performance of various methods depended in part on how the ensemble of inputs was constructed.
  - For temperature forecasts, the spread of an ensemble of predictors was found to be a useful predictor of forecast uncertainty in the medium-range. This was seen for days 6-15 for the North American Ensemble Forecasting System (NAEFS). Using the ensemble spread produced lower ignorance scores than using a constant spread mainly due to lower ignorance scores for events with large forecast errors.
  - For temperature forecasts in the short-term, little or no improvement was seen when using the ensemble spread. The effect was weak for days 1-5 for the NAEFS ensemble and non-existent for all forecast offsets for the University of British Columbia short-range ensemble forecasts (UBC-SREF). For the 48-h forecasts from the reforecast dataset (Chapter 2), Bayesian model averaging (BMA), which accounts for spread-skill relationships, did not provide any benefit over a constant spread model suggesting again that for the short-term the disagreement between members did not provide any skill in predicting uncertainty.
  - For the NAEFS precipitation forecasts, the strongest factor for determining probability of precipitation was the ensemble mean. A small overall improvement was found when the fraction of ensemble members forecasting the non-occurrence of precipitation was also included.
  - The binned probability ensemble technique generally produced probability forecasts with high ignorance scores (Section 2.6.2). This was attributed to the nature in which the method distributes probability mass between consecutive ensemble members and also due to its assumption of a perfect spread-skill relationship (Figure 2.10).
- Gaussian distributions (used for either the method of moments or BMA) generally produced calibrated forecasts for temperature variables. When Gaussian distributions were used in

cases where a skewed distribution was more appropriate, the calibration method improved the resulting forecast (Section 2.6.1).

- The correlation in time of PIT values is strong enough that it can be exploited to produce improved probabilistic forecasts for surface temperature in the short-term (Chapter 3). Modeling PIT values as a first-order Markov process resulted in improved ignorance scores, continuous ranked probability scores (CRPS), and mean absolute error, and did not further degrade reliability (Section 3.4).

- The CRPS and ignorance score will in some cases yield different conclusions about which probabilistic methods are best. This is due to their different treatment of probabilities in the tail of the forecast distribution (Chapter 4). For example, when the ensemble spread was found to be a useful predictor of uncertainty by evaluating with the ignorance score, the utility was found to be much lower when evaluated by the CRPS.

5.3 Potential applications

Parts of the probabilistic forecasting system presented are currently used in real-time by the Weather Forecast Research Team at UBC. Products in the form of cumulative probability plots are being used by BC Hydro to aid in medium-range planning of weather-affected activities.

The system presented here could be used by any forecasting centre interested in improved probabilistic forecasts. It allows centres to further implement new methods and determine what combinations of methods yield the highest quality probabilistic forecasts. These improved forecasts can form the basis for better decision making by any business, organization, or individual with weather-affected operations.

5.4 Limitations and recommendations for further work

I have analysed the performance of a small set of relatively simple methods within the proposed system. Although more advanced methods, such as BMA, was tested within the context of calibration in Chapter 2, it was not tested within the framework of the proposed system. It would be of great interest if such methods were added to the system, especially for evaluating the performance of medium-range forecasts, where spread-skill relationship may be stronger.

Also, the evaluation of more advanced correction schemes such as model output statistics, Kalman filtering, neural networks, analog methods, and gene-expression programming would be beneficial, provided adaptive parameter estimation algorithms can be devised for these methods.

The system itself could also be extended in many ways. Firstly, the decomposition used in this dissertation uses four components. There are likely other components that can be added to this decomposition. As an example, a selection component could be added between predictors and corrector, which serves to select a set of the ensemble predictors. This way, potentially low-quality ensemble members can be removed before the next stage, potentially resulting in better probabilistic forecasts.

Secondly, the system treats each forecast location independently. An alternative is to pool parameters between stations for potentially more robust results (see for example Raftery et al., 2005). This affects only the parameter estimation part of the system and does not affect the decomposition itself.

Thirdly, some applications require joint probability distributions. An example is winter road maintenance (Berrocal et al., 2010) where the joint distribution of precipitation and temperature is required. These forecasts consider multiple variables, which is currently not considered by the decomposition. A solution to this would be relax the output requirement of the uncertainty model, thereby requiring multivariate calibration and update methods.

Lastly, for computational efficiency, an adaptive approach was used to update the parameters of the schemes. This may limit which candidate schemes can be implemented, since it may be difficult to cast the parameter estimation part of a scheme into adaptive form. A solution to this would be to sacrifice computational efficiency by allowing methods to retrieve large sets of historical performance statistics.

The main limitation of the calibration method presented in Chapter 2 is the slight increase in ignorance score seen when the input probability distributions are already calibrated, or nearly calibrated. This was attributed to overfitting, as the calibration method attempts to calibrate based on a noisy distribution of verifying PIT values. Even perfectly calibrated forecasts are expected to exhibit some noise in the distribution of PIT values due to sampling errors. Smoothing was applied to the PIT values to reduce noise, but overfitting still occurred.

One solution to this problem is to use a method that detects when a set of forecasts are already nearly calibrated. When the uniformity of past PIT values are above a certain level, the calibration method would not attempt to calibrate, but instead would use the original probability distribution. Thus, calibration would only be attempted for variables that exhibit sufficient distributional bias. This could vary from stations to station, or even from season to season.

The updating method of Chapter 3 was found to work well for temperature variables, as a Gaussian distribution described the PIT transitions well. However, such a transition function may be inappropriate for other variables, in particular precipitation. Precipitation changes much more abruptly and irregularly than temperature and may therefore require an alternative model.

One solution to this would be to investigate other, more complex transition functions. Another improvement could result from the use of a higher-order Markov model. Instead of only using the most recent verifying PIT value, the method could use several recent PIT values.

Bibliography

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

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

Anthes, R. A., 1974: Data assimilation and initialization of hurricane prediction models. *J. Atmos. Sci.*, **31**, 702–719.

Bakhshaii, A. and R. Stull, 2009: Deterministic ensemble forecasts using gene-expression programming. *Wea. Forecasting*, **24**, 1431–1451.

Benjamin, S. G., et al., 2004: An hourly assimilationforecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518.

Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wideband atmospheric model suited for finescale process studies and simulation. *Mon. Wea. Rev.*, **125**, 2382–2415.

Berrocal, V. J., A. E. Raftery, T. Gneiting, and R. C. Steed, 2010: Probabilistic weather forecasting for winter road maintenance. *J. of the American Statistical Association*, **105**, 522–537.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347.

Bremnes, J. B., 2007: Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods. Tech. rep., Norwegian Meteorological Institute, 1–34 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

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

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097.

Chmielecki, R. M. and A. E. Raftery, 2010: Probabilistic visibility forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **139**, 1626–1636.

Dawid, A. P., 1984: Statistical theory: The prequential approach (with discussion). *J. of the Royal Statistical Society, Series A*, **147**, 278–292.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to post-process numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554–3570.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B*, **39**, 1–38.

Eckel, F. A. and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

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

Ehrendorfer, M., 1994: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.*, **122**, 703–713.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Evensen, G., 1994: Sequential data assimilation with nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162.

Garaud, D. and V. Mallet, 2011: Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast: Application to air quality. *J. Geophys. Res.*, **116**, doi:10.1029/2011JD015780.

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

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. of the Royal Statistical Society, Series B*, **69**, 243–268.

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. of the American Statistical Association*, **102**, 359–378.

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

Good, I. J., 1952: Rational decisions. *J. of the Royal Statistical Society, Series B*, **14**, 107–114.

Grell, G. J., J. Dudhia, and D. R. Stauffer, 1994: A description of the 5th generation Penn State/NCAR mesoscale model (MM5). Tech. Rep. TN-398+STR, National Centre for Atmospheric Research, 122 pp.

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

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

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 S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble foecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

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

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

Hart, K. A., J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic Winter Games. *Wea. Forecasting*, **19**, 200–218.

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

Hirschberg, P. A., et al., 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.*, **92**, 1651–1666.

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

Homleid, M., 1995: Diurnal correction of short-term surface temperature forecasts using the Kalman filter. *Wea. Forecasting*, **10**, 689–707.

Hopson, T. M. and P. J. Webster, 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., A. Brix, and C. Ziehmann, 2005: *Weather Derivative Valuation*. Cambridge University Press, 373 pp.

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

Karlin, S. and H. Taylor, 1981: *A second course in stochastic processes*. Academic Press, 582 pp.

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

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Lewis, J. M. and J. C. Derber, 1985: The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37**, 130–141.

Lorenz, E. N., 1963: Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130–141.

Marzban, C., 2003: Neural networks for postprocessing model output: ARPS. *Mon. Wea. Rev.*, **131**, 1103–1111.

McCollor, D. and R. Stull, 2008a: Hydrometeorological accuracy enhancement via postprocessing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.

McCollor, D. and R. Stull, 2008b: Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. *Wea. Forecasting*, **23**, 557–574.

Michalakes, J., J. Dudhia, D. Gill, J. Klemp, and W. Skamarock, 1999: Design of a next-generation weather research and forecasting model. *Towards teracomputing*, World Scientific, 117–124.

Molteni, F. and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.*, **119**, 269–298.

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

Murphy, A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816.

Nielsen, H. A., et al., 2006: From wind ensembles to probabilistic information about future wind power production – results from an actual application. *9th International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden.

Nipen, T. and R. Stull, 2011: Calibrating probabilistic forecasts from an NWP ensemble. *Tellus*, **63**, 858–875.

Nipen, T. N., G. West, and R. B. Stull, 2011: Updating short-term probabilistic weather forecasts of continuous variables using recent observations. *Wea. Forecasting*, **24**, 564–571.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**, 71–116.

Pinson, P. and H. Madsen, 2009: Ensemble-based probabilistic forecasting at Horns Rev. *Wind Energy*, **12**, 137–155.

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

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

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667.

Rose, C., 1995: A statistical identity linking folded and censored distributions. *Journal of Economic Dynamics and Control*, **19**, 1391–1403.

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

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spreadskill relations using the ECMWF ensemble prediction system over Europe. *Wea. Forecasting*, **19**, 552–565.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. Tech. Rep. TN-468+STR, National Centre for Atmospheric Research, 88 pp.

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

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Stensrud, D. J. and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

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

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Toth, Z., et al., 2006: The North American Ensemble Forecast System (NAEFS). *18th Conference on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc.

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

Yuval and W. W. Hsieh, 2002: An adaptive nonlinear MOS scheme for precipitation forecasts using neural networks. *Wea. Forecasting*, **18**, 303–310.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.