Ruprecht-Karls-Universität Heidelberg
Fakultät für Mathematik und Informatik

Probabilistic Forecasts of Precipitation Using Quantiles

Diplomarbeit

von

Corinna Maria Frei

Betreuer: Prof. Dr. Tilmann Gneiting
Dr. Thordis L. Thorarinsdottir

Mai 2012

Abstract

Probabilistic forecasts of precipitation are of high socio-economic interest, with applications in industries ranging from agriculture to tourism and logistics. Nowadays, such forecasts are usually based on forecast ensembles from numerical weather prediction models. However, even state-of-the-art ensemble prediction systems are uncalibrated and biased, and a variety of post-processing techniques aiming to calibrate the ensemble output have been proposed. In this project, we propose a novel way of post-processing forecast ensembles for precipitation based on quantile regression. Although conventional quantile regression has been found to perfom well, seperately fitted regressions are not constraint to be mutually consistent and the method does not yield predictive densities. In order to solve these issues, we propose a two-step approach which models the probability of precipitation occurrence using logistic regression and the precipitation amounts using a log-normal distribution. The method is applied to 48-h forecasts of 24-h precipitation accumulation over the North American Pacific Northwest, using the University of Washington mesoscale ensemble. The resulting probabilistic forecasts turn out to be much better calibrated than the unprocessed ensemble and a climatological reference forecast.

Zusammenfassung

Probabilistische Niederschlagsvorhersagen haben einen hohen sozio-ökonomischen Wert und finden Anwendung in den unterschiedlichsten Wirtschaftszweigen, beispielsweise in der Landwirtschaft, im Tourismus und in der Logistik. Heutzutage werden solche Vorhersagen oft mithilfe von Vorhersageensemblen numerischer Wettervorhersagemodelle erstellt. Allerdings sind selbst Ensemblevorhersagemodelle, die dem neuesten Stand der Forschung entsprechen, nicht kalibriert, sondern oft mit systematischen und Dispersionsfehlern behaftet. Um diese Fehler zu korrigieren wurde eine Vielzahl sogenannter Postprocessingmethoden entwickelt. In der vorliegenden Arbeit schlagen wir eine neue Postprocessingmethode vor, die mithilfe von Quantilregression Vorhersageensembles für Niederschlag nachbearbeitet. Reguläre Quantilregression liefert zwar gute Ergebnisse, bisher aber nicht in Form von Vorhersagedichten. Zudem sind separat geschätzte Regressionen nicht zwingenderweise konsistent. Um diese Probleme zu lösen, schlagen wir einen zweistufigen Ansatz vor, der zuerst die Niederschlagswahrscheinlichkeit mit logistischer Regression, und dann die Niederschlagsmenge mit einer Log-Normalverteilung modelliert. Diese Methode wurde im Anschluss in einer Fallstudie auf Ensemble-Niederschlagsvorhersagen über dem nordamerikanischen Pazifischen Nordwesten angewandt. Die resultierenden probabilistischen Vorhersagen waren wesentlich besser kalibriert als die Vorhersagen des Ensembles ohne Korrektur und klimatologische Referenzvorhersagen.

Contents

Chapter 1

Forecasting with Uncertainty

Whether it is forecasting the weather, the economy or the course of a disease – anticipating things to come is arguably one of the biggest human desires. By knowing what to expect, we feel better prepared to make decisions.

Henri Poincare said that "it is far better to foresee even without certainty than not to foresee at all". Still, how valuable can a forecast possibly be that, in the best case, fails to provide key information on how certain predictions are, and, in the worst case, is simply wrong? Besides, what if one is not interested in what is most likely to happen, but rather in minimizing the risks of unlikely but nontheless possible events?

One of the most substantial applications of forecasting is the weather. Apart from being convenient in everyday life, sophisticated weather forecasts are used for decision making in such diverse fields as electrical power generation, ship routing, pollution management, risk management in insurance and reinsurance, disease prediction, crop yield modeling, and many more (Palmer, 2002).

For millennia, humankind has tried to understand the complex dynamics of climate and weather. Nevertheless, personal experiences and common knowledge in form of weather lores were all sailors, farmers, and everyone else had to rely on until more advanced weather forecasts became possible. In the 21st century, weather modeling and forecasting have become a global multi-billion dollar endeavor involving thousands of scientists, countless weather stations, trillions of arithmetic operations per second, and some of the largest supercomputers in the world.

Weather forecasting has come a long way since scientists first made use of computational models for predictive purposes in the mid-20th century. In recent years, exponentially growing computer capacities and a comprehensive stream of live weather

data have lifted some of the restrictions on modern weather forecasting. However, even though today's 7-day-forecasts are likely to outperform the 24h-forecasts from 30 years ago, uncertainty still remains. Consequently, in spite of seeing uncertainty as a flaw on our ability to forecast, one should embrace uncertainty as the essence of forecasting.

To increase both accuracy and usefulness of weather forecasts, methods have to be developed to characterize, evaluate and, where possible, quantify uncertainty. The goal is not to foresee with certainty what is going to happen, but to better understand and represent uncertainty. Forecasts need to offer the necessary means to tackle concrete questions in the decision making process, i.e. they have to enable risk-based rational decision making. As a consequence, forecasts should be probabilistic rather than deterministic. Wherever possible they should be expressed as probability distributions over future weather quantities or events (Dawid, 1984; Gneiting, 2008).

Probabilistic versus Deterministic Forecasting

Probabilistic forecasts are scientifically more "honest" than deterministic forecasts, allowing the forecaster to admit and express uncertainty. In this way awareness for uncertainty is raised, as users are provided with the necessary means to quantify and predict weather related risk. Deterministic forecasts on the other hand are incapable of providing the necessary uncertainty information. Even more so, they possibly create the illusion of certainty in a user's mind, as they hide the predictive uncertainty behind the facade of a precise "best" estimate.

Ranging from weather and climate prediction to economic and financial risk management, the probabilistic approach is very effective in a variety of applications. Often, ranges or thresholds are crucial: farmers may be interested in the risk of freezing temperatures for efficient agricultural activities, or the chance of winds being low enough to spray pesticides safely, rather than the most likely estimates of temperature or wind speed. Also, for issuing severe weather warnings, possible extreme values are of particular interest. Still, regardless of the many situations in which probabilistic information could be of value, for reasons of communication and simplified decision making, the prevailing format of forecasts remains deterministic (Gneiting, 2011).

Krzysztofowicz (2001) illustrates suboptimal actions and fatal consequences, including massive economic and social opportunity losses, that can arise from forecasts suppressing information and judgment about uncertainty. In 1997, an estimate of 49 ft for the flood crest on the Red River close to Grand Forks, North Dakota was issued. City officials and

residents took the forecast literally and prepared themselves as if the estimate of 49 ft was a perfect forecast. Weeks later the actual crest of 54 ft (being 26 ft above the flood stage) was overtopping dikes, forcing not only evacuations, but also devastating the city. Bearing in mind that the conventional deterministic prediction of 49 ft for the flood crest did not imply that the risk of a higher flood crest was zero, suppose what difference a probabilistic forecast indicating a 30% risk of exceeding 50 ft could have made: the risks of the river crest overtopping dikes and causing heavy losses could have been traded off against the costs of additional precautionary actions and might have led to a different scenario.

In situations calling for risk-based decision making or cost/loss analysis, knowing the probabilities of different outcomes can make a huge difference. These varying probabilistic point predictions are often quantiles of the underlying predictive probability distribution. Opposed to probabilistic point predictions, which represent probabilities for certain specific events only, predictive probability density functions (PDFs) assign probabilities to all possible future outcomes at once. These distribution functions tend to not only be more convenient to manipulate than a set of point values (Bröcker and Smith, 2008), they also provide a better insight into the current circumstances as they look at the system as a whole.

Probabilistic Weather Forecasting

Probabilistic weather forecasts are often based on so called ensemble prediction systems. Numerical weather prediction (NWP) models are run several times with different initial conditions or model physics, addressing the inherent uncertainties in atmospheric prediction. However, even state-of-the-art ensemble systems lack calibration and are contaminated by systematic biases. Moreover, they do not generate probabilistic forecasts in terms of full predictive distributions per se, as the raw ensemble output only consists of a finite set of deterministic forecasts. In view of these limitations, ensembles call for some sort of post-processing in order to make use of all information available in an ensemble forecast.

In concert with statistical post-processing, ensemble prediction systems offer the possibility of well-calibrated probabilistic forecasts in form of predictive probability density functions (PDFs) over future weather quantities or events. Still, what exactly constitutes a good probabilistic forecast? And how can we evaluate and compare the performance of competing forecasts? According to the diagnostic paradigm of Gneiting

et al. (2007), the goal is to maximize the sharpness of a probabilistic forecast subject to its calibration. Calibration refers to the reliability of the forecast, that is, the statistical consistency between the probabilistic forecast and the actually occuring observations. Sharpness refers to the concentration of the predictive distribution; under the condition that all forecasts are calibrated, we define the sharpest to be the best. In other words, the sharper a calibrated predictive distribution, the fewer uncertainty and, ultimately, the better its performance.

Here we are concerned with probabilistic forecasts of precipitation. Precipitation poses a particular challenge, as for its mixed discrete-continuous probability distribution. Ranging from regression techniques such as linear (Glahn and Lowry, 1972), logistic (Wilks, 2009), and quantile regression (Bremnes, 2004) to binning techniques (Gahrs et al., 2003), and neural networks (Koizumi, 1999), a variety of methods has been developed to statistically post-process numerical model output and generate probabilistic precipitation forecasts. To date, however, the only statistical post-processing techniques transforming raw ensemble output into fully specified, calibrated predictive distributions are Bayesian Model Averaging (BMA; Sloughter et al., 2007) and, to some extend, logistic regression (Wilks 2009).

In this project, we take a closer look at techniques based on quantiles, i.e. logistic regression and quantile regression. So far, these approaches only give probabilistic point predictions, in form of probability forecasts at a given threshold or quantile forecasts at a given level. The problem then is to ensure consistency, as both, threshold non-exceedance probabilities and quantiles, are not automatically constrained to be monotonically increasing. Nevertheless, even though they are associated with additional obstacles, these conventional regression techniques were found to perform comparatively well in the statistical post-processing of ensemble forecasts. In the study of Sloughter et al. (2007), in which multi-analysis ensemble forecasts of precipitation were calibrated, conventional logistic regression and BMA showed comparable Brier skill scores for a large range of precipitation thresholds.

Wilks (2009) proposed to extend the logistic regression framework and fit logistic regressions for all thresholds simultaneously, thereby not only resolving the issues mentioned above, but also providing fully specified predictive densities. After studying Wilks' method from a probabilistic perspective, i.e. deriving a closed form expression for the predictive distribution, we intended to develop an analogous method for quantile regression. However, this turned out to be a more demanding endeavor than first

expected, as it was not possible to transfer his ideas to the quantile regression framework directly.

Instead, we developed a completely new method based on quantile regression, which also both resolves the issues mentioned above and yields fully specified predictive densities. Our approach proceeds in two steps: First, the probability of precipitation occurrence (PoP) is modeled using logistic regression. Second, the precipitation amounts are modeled using a log-normal distribution. Thus, the predictive PDF is a mixture of a point mass at zero and a skewed continuous distribution.

The thesis at hand is organized as follows:

In Chapter 2, we provide some background information on numerical weather prediction and ensemble forecasting. Chapter 3 is concerned with probabilistic forecasts of precipitation. Specifically, we discuss three statistical post-processing techniques that generate probabilistic forecasts of precipitation based on ensemble forecasts: We first portray BMA and logistic regression, including Wilks' extension, before we review conventional quantile regression and discuss our new approach in detail. Chapter 4 containts a case study, i.e. results for daily 48-h forecasts of 24-h accumulated precipitation over the North American Pacific Northwest in 2008, based on the eight-member University of Washington mesoscale ensemble (Grimit and Mass, 2002) and associated verifying observations. Throughout the thesis we use illustrative examples drawn from these data. Finally, in Chapter 5, we give a summary and discuss the results of the case study, leading to the question of how the method could be improved even further.

Chapter 2

Numerical Weather Prediction and Ensemble Forecasting

Weather forecasting has come a long way since the possibility of numerical weather prediction has first been suggested by Lewis Fry Richardson in 1922. He proposed to start by modeling the laws of physics governing the behavior of the atmosphere as a system of mathematical equations (Bjerknes, 1904) and then apply his finite difference method to solve the system. To demonstrate how this method could be put to use, he calculated the changes in surface pressure at two points in central Europe by hand, using the most complete set of observations available to him. Needless to say, that, in absence of computational powers, it took him more than six weeks to complete the tremendous number of required equations. In addition, both due to incomplete and imperfect initial data and serious deficiencies in his approach, the first results were very poor (Lynch, 2008).

Nevertheless, Richardson's work turned out to be more than visionary. Even though it took another 30 years until computation time dropped below the forecast period itself, his prospective ideas are now not only universally acknowledged among meteorologists, but constitute the foundation of modern weather forecasting. For years to come, weather forecasting was considered an intrinsically deterministic endeavor: For one set of "best" input data, deemed to represent the current weather conditions, one "best" weather prediction was to be generated. Alongside a comprehensive stream of live weather data, model complexity and the size of initial data sets have gradually been increased to take advantage of exponentially growing computer capacities, making numerical prediction models not only faster, but also more accurate.

Today, weather forecasts worldwide rely primarily on numerical weather prediction models very similar to the ones Richardson thought of almost 100 years ago. Alternative forecasting models based on purely statistical methods, using auto-regressive time series techniques (Brown et al., 1984) or neural network methods (Kretzschmar et al., 2004), have been implemented successfully at prediction horizons of a few hours. In the medium range, however, they are outperformed by numerical weather prediction models (Campbell and Diebold, 2005).

Still, weather forecasting is not exact science. In the early 1990s researchers began to realize that inherent limitations in atmospheric predictability restrict the value added by further increasing model complexity. The belief in one near-perfect model with complexity close to the real world faded, resulting in a shift of paradigm within the meteorological community. Even with ever increasing computational resources at hand, computer-generated forecasts will always stay flawed, containing a mix of errors due to incomplete initial estimates of atmospheric conditions and the chaotic nature of the partial differential equations used to simulate the atmosphere. As a consequence, alternative ways of using the available computational resources to improve numerical weather forecasts were explored, ensemble forecasting being a primary candidate (Leith, 1974).

Ensemble prediction systems utilize a varied set of initial conditions and/or numerical models to compute a set of separate deterministic forecasts (Gneiting and Raftery, 2005). Whereas no single one of these forecasts might be ultimately correct, ensembles can be exploited to better define the most likely weather outcome while also more accurately assessing the risks of rare and possibly dangerous events. However, even though ensembles do provide additional information compared to a single numerical forecast, they are often associated with an unsatisfactory degree of calibration (Hamill and Colucci, 1998).

The remainder of this chapter is organized as follows. First, numerical weather prediction techniques, including the inherent limitations to numerical predictability, are discussed. In the second section, ensemble forecasting is elaborately portrayed including the different types, advantages, and disadvantages of ensemble prediction systems and how they can be further improved.

2.1 Numerical Weather Prediction Models

Numerical weather prediction (NWP) models are based on the idea that the laws of physics will determine future atmospheric states, given that the current state of the atmosphere is known. These grid-based atmospheric simulations rely on a system of differential equations derived from the laws of fluid- and thermodynamics. Those equations are discretized and integrated forward from initial states based on observational data assimilated from a variety of sources.

Ever since the first operational real-time NWP forecasts in the 1950s, the enterprise was driven by the objective of improving the quality of these models, i.e. reducing model forecast errors under the constraint of model efficiency. Since then, our understanding of earth's atmosphere has improved significantly, providing a strong theoretical foundation for weather forecasting. On top of that, technology spurred the NWP development in two major ways: by providing ever-increasing computer capacities and the steady introduction of additional instruments for data acquisition.

Consequently, NWP forecasts have become more and more accurate over time. Data assimilation methodologies have become increasingly sophisticated to make best use of the ever increasing amount of observational data. Model complexity has been increased, facilitating more sophisticated numerical schemes and model parametrization along with the use of progressively finer resolutions in order to capture small scale processes. Having said that, why will even complexer models not automatically lead to better forecasts?

The errors in numerical weather prediction are essentially of two classes: incomplete or inadequate initial estimates of atmospheric conditions (analysis error), along with deficiencies within the numerical models itself (model error).

Differences between the true atmospheric conditions and the analyzed initial state serving as input to NWP models arise from instrument errors, as well as from imperfect data assimilation techniques. Irregularly spaced observations and areas with sparse data add further difficulties to the interpolation to the grid structure. In case of limited area models, additional difficulties arise from both uncertainties and errors due to artificial lateral boundary conditions. Errors introduced by the analyzed initial state therefore pose the first set of limitations to atmospheric predictability, no matter how skillful the model might be (Grimit and Mass, 2002).

Deficiencies within the numerical models itself further limit atmospheric predictions. Even increasingly complex numerical models are subject to imprecision, especially within the physical parametrisation, as these parametrisations are merely simplifications of

complex atmospheric processes. Clouds, for example, are very difficult to model as they encompass many physical processes on a vast range of scales. Still, they play a crucial role in transporting water and momentum throughout the atmosphere. In order to incorporate clouds into NWP models, they are related to variables on the scales the model resolves, i.e. they are parametrized by processes of various precision.

However, the most fundamental problem within the numerical weather prediction framework is the chaotic nature of the partial differential equations used to simulate the atmosphere (Lorenz, 1963). It is impossible to solve these equations exactly through analytical methods. Any so called 'solutions' of NWP models are therefore only approximations and even small errors, whether they result from inaccurate initial conditions or imperfect numerical models, grow with time and may lead to significant forecast errors.

As a result, although the benefits of complexer models and higher resolution forecasts are numerous, recognition of these limitations to atmospheric predictability has led to increased interest in alternative ways of using the available computational resources to improve numerical weather forecasts. Research efforts have shifted from minimizing the sources of uncertainty even further by incremental improvements to a better representation of uncertainty itself.

2.2 Ensemble Prediction Systems

As the belief in one near-perfect model with complexity close to the real world faded, alternatives to single-integration forecasts were explored. Ensemble prediction systems (EPS) allocate the available computational resources to compute a series, or ensemble, of reduced-resolution dynamical NWP model forecasts with varying initial conditions and/or model physics. Thereby ensemble forecasting offers the possibility to address the uncertainties within the initial state estimates and numerical models, i.e. the inherent uncertainties in atmospheric prediction. Whereas no single one of these forecasts might be ultimately correct, multiple results of the ensemble output both suggest the possibility and provide the means of probabilistic forecasts. Moreover, the ensemble mean forecast typically outperforms all or most of the individual forecasts comprising the ensemble. In doing so, it tends to be the best estimate for the verifying state of the atmosphere (Grimit and Mass, 2002).

Ensembles can not only be exploited to better define the most likely weather outcome, but also to more accurately assess the risks of rare and unlikely, but nonetheless possible

events. Thus, ensembles manage to reflect the predictability of a weather system, as well as identify predictability limits and reduce forecast surprises. This is particularly relevant as the most dangerous weather conditions are not likely to be the most probable.

In late December 1999, two storms of extreme force, subsequently named Lothar and Martin, swept across Central Europe, causing major damage in France, Southern Germany and Switzerland. Over 100 people were killed, and the storm caused extensive damage to buildings, infrastructure and forests resulting in substantial economic loss. Figure 2.1 shows the surface pressure 'stamp maps' for the 42-hour ensemble forecast generated at the European Centre for Medium-Range Weather Forecasts (ECMWF) just days before winter storm Lothar hit Northern France. The top left shows the best-guidance deterministic forecast for December 26th and the verifying analysis. Given the available data on December 24th, ECMWF's operational high-resolution deterministic forecast predicted a rather typical winter's day, and missed the storm completely. Even though a number of ensemble members support this forecast, it can be seen that several members predict some sort of a storm. In other words, the ensemble did display a significant risk of a severe event, while the single best deterministic forecast depicting the most-likely outcome failed to even indicate such a risk. In this particular case, the ensemble spread was enormous, i.e. the development of atmospheric flow was highly unpredictable. Differences between ensemble members predicting storms and those that did not, turned out to correspond to small perturbations in temperature and wind over the West Atlantic (Palmer, 2002).

Ensemble prediction systems can be designed in several ways, including a varied set of initial conditions or numerical models. In the former case, a collection of forecasts is generated running one numerical model with varying initial conditions, which are chosen to be consistent with current observations and the typical analysis error. These so called multi-analysis ensembles address forecast uncertainties due to imperfect representations of the atmosphere. Thus, they may diagnose sensitivity to the initial conditions (ICs). The example above conveys how big the impact of analysis errors, even at shorter lead times, can be. Today, the selection of initial conditions is a science in itself and a number of methods, including random perturbations and breeding methods, have been developed to generate initial conditions reflecting this uncertainty. Multi-model ensembles, on the other hand, include a number of independently derived models within the ensemble in order to stress model uncertainty. These numerical models are generally run using a single set of initial conditions. Alternative representations of forecast uncertainty due

Figure 2.1: Surface pressure stamp maps for a 42-h ensemble forecast for the infamous storm Lothar, which ended up devastating parts of mainland Europe in December 1999. Given the available data on December 24th, the best single deterministic forecast predicted a not-atypical winter's day, while the EPS indicated the risk of a severe storm (Palmer, 2002; Fig. 3).

to imperfect models include stochastic representations of the parameterized physical processes (Buizza et al., 2005) and a perturbated parameter approach (Murphy et al., 2004). Figure 2.2 shows a schematic representation of the combined approach, namely a multimodel multianalysis ensemble, where an ensemble forecast is generated running several numerical models with a collection of initial conditions.

In an ideal ensemble forecast setting, where a particular ensemble prediction system samples all sources of forecast error appropriately, ensembles are supposed to provide a flow-dependent sample of the probability distribution of possible future atmospheric states. The dispersion of a forecast ensemble then corresponds to the uncertainty in the forecast: a small ensemble spread indicates low uncertainty and vice versa. In addition, the probability of any event can be estimated directly from the relative event frequency in the ensemble (Wilks, 2006). For example, if 15 out of 50 ensemble members indicate the risk of a severe storm, the probability of the actual occurrence of such a storm should be 30%. The forecast would be considered calibrated, if in 30% of the cases when a 30% probability for a storm was issued, such an event did actually occur.

In practice, however, ensemble prediction systems are far from perfect. Initial condition selection procedures fail to randomly sample the current atmospheric state, and numerical models are still deterministic simplifications of atmospheric reality. In other words, ensembles capture only some of the uncertainties involved in numerical weather prediction – and even those only partially. As a result, ensemble forecasts are often uncalibrated, in that they are contaminated by systematic biases and dispersion errors. Nevertheless, ensembles can give us an indication of uncertainty, and relationships between forecast error and the a priori known ensemble spread have been established for several ensemble prediction systems, even when the ensemble is uncalibrated (Buizza et al., 2005). Especially for extreme spread events, ensemble spread and ensemble mean error tend to be highly correlated, i.e. they are a good estimator of the eventual forecast skill (Whitaker and Loughe, 1998). However, it has also been shown that ensemble forecasts typically turn out to be underdispersive, in that they fail to cover the full range of possibilities of the verifying weather. The ensemble spread, then, is on average too small and the observation lies outside of the ensemble range too often. Moreover, if ensemble relative event frequencies are used to estimate event probabilities, underdispersive ensemble outputs tend to lead to overconfidence in probability assessment. Consequently, in order to produce reliable probability forecasts, numerous methods for calibrating the ensemble output have been proposed, see e.g. Wilks (2011).

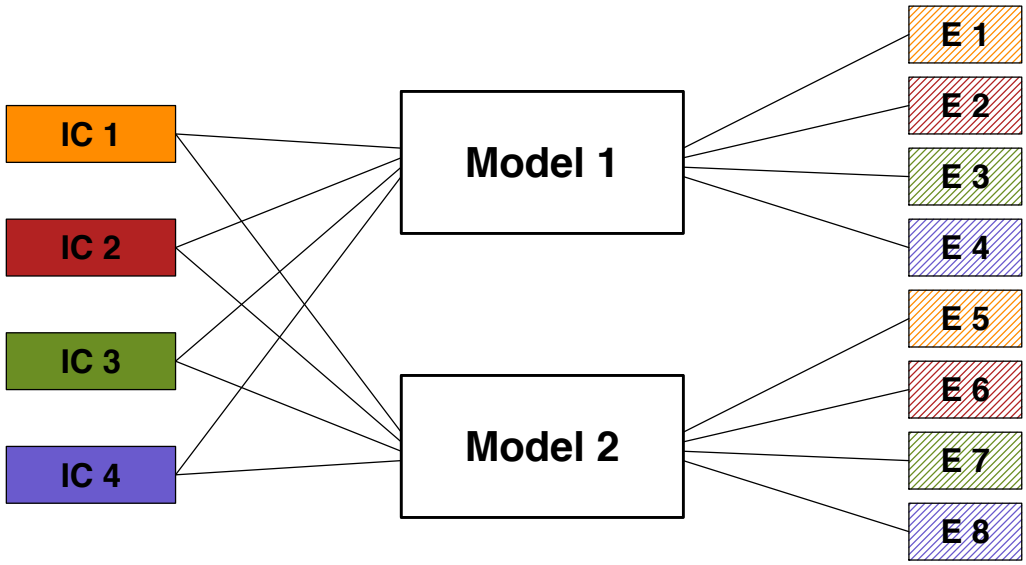

Figure 2.2: Diagram of an eight member multimodel multianalysis ensemble. Two numerical models are discretized and integrated forward from four sets of initial conditions (IC 1 - IC 4) resulting in eight seperate deterministic forecasts (E 1 - E 8).

In concert with statistical post-processing, ensemble prediction systems offer the possibility of well-calibrated, flow-dependent probability forecasts in form of predictive probability density functions (PDFs) over future weather quantities or events. Over the past years, several statistical post-processing methods for weather variables such as temperature, air pressure and quantitative precipitation have been developed. State-of-the-art-approaches include Bayesian model averaging (BMA; Raftery et al., 2005) and ensemble model output statistics (EMOS; Gneiting et al., 2005). BMA represents the predictive PDF of any future weather quantity of interest as a weighted average of PDFs centered on the individual bias-corrected ensemble member forecasts. EMOS is based on multiple linear regression, and e.g. fits a normal distribution to the ensemble member forecasts for temperature and pressure. The following chapter extensively portrays several post-processing techniques that have been proposed to generate probabilistic forecasts of precipitation based on ensemble forecasts.

Chapter 3

Probabilistic Forecasts of Precipitation

Probabilistic forecasts of precipitation are of high socio-economic interest. Agriculture, forestry, landscaping, hydraulic engineering, tourism, transportation and logistics are just few of the industries that depend on precise precipitation forecasts and particularly benefit from probabilistic forecasts for optimal decision making. The advantages of a probabilistic point of view become visible when looking at precipitation-related extreme events such as floods, flash floods, landslides or avalanches. When applied in such circumstances, probabilistic forecasts may not only help to minimize economic damages, but also potentially save human lives.

However, probabilistic forecasts of precipitation turn out to be especially challenging as precipitation has a mixed discrete-continuous probability distribution, i.e. the predictive distribution of precipitation is skewed, non-negative and has a positive probability at zero. Despite the fact that the U.S. National Weather Service already began to produce and disseminate the first probabilistic forecasts in terms of probability of precipitation occurrence (PoP) in the 1960s, the transition to fully specified probability distributions is still in progress.

As mentioned in the introduction, several statistical techniques have been proposed to generate probabilistic forecasts of precipitation based on ensemble forecasts. Below, we discuss three such statistical post-processing techniques: First, we portray Bayesian Model Averaging, which was the first method proposed in the literature to provide fully specified predictive densities for precipitation. In the second section, we outline the logistic regression approach of Wilks (2009) and investigate his method from a

probabilistic perspective. In the third and final section of this chapter, we propose a new method based on quantile regression, which also yields fully specified predictive densities.

3.1 Bayesian Model Averaging

Bayesian model averaging (BMA) was originally developed to combine predictions from different competing statistical models, in order to account for uncertainties within the model selection process (Hoeting et al., 1999). It has been successfully applied to several statistical model classes including linear regression and related models in the health and social sciences, improving predictive performance in all cases. Raftery et al. (2005) extended BMA from statistical to dynamical models, such as numerical weather prediction models. Moreover, they demonstrated how BMA can be used to statistically post-process forecast ensembles, that is, transforming ensemble output into well-calibrated probabilistic forecasts in form of predictive PDFs of future weather quantities.

In BMA for forecast ensembles, Y denotes the weather quantity of interest and $X_1, \ldots, X_K$ denote K ensemble member forecasts. In addition, every forecast X_k is associated with a component PDF, $h_k(y \mid x_k)$, which can be interpreted as the conditional PDF of the weather quantity Y given $X_k = x_k$, conditional on X_k providing the best forecast in the ensemble. The BMA predictive PDF for Y can then be expressed as a mixture of the component PDFs,

$$p\left(y \mid x_1, \ldots, x_K\right) = \sum_{k=1}^{K} w_k h_k(y \mid x_k), \tag{3.1}$$

where the weight w_k is based on ensemble member k's relative performance or skill in the training period. These weights can thus be interpreted as the posterior probability of forecast X_k being the best forecast in the ensemble. As the w_k's are probabilities, they are non-negative and sum to one, that is $\sum_{k=1}^{K} w_k = 1$.

The distribution of the component PDFs depends on the weather variable of interest. For weather variables whose predictive PDFs are approximately normal, such as temperature and sea level pressure, the component PDFs can be taken to be normal distributions centered at the bias-corrected ensemble member forecasts (Raftery et al., 2005).

In order to apply BMA to quantitative precipitation forecasts, Sloughter et al. (2007) propose to model the component PDF of precipitation, $h_k(y \mid x_k)$, as a mixture of a point mass at zero and a gamma distribution. More specifically, the probability of precipitation (PoP) is modeled as a function of the forecast X_k, using logistic regression with a power transformation of the forecast as the predictor variable, and the predictive PDF of the amount of precipitation is specified as a gamma distribution, given the amount of precipitation being greater than zero. The BMA predictive PDF is then a weighted average or a mixture of such distributions,

$$p\left(y \mid x_1, \ldots, x_K\right) = \sum_{k=1}^{K} w_k \left\{ \mathbb{P}\left[y = 0 \mid x_k\right] \mathbb{I}_{\{y=0\}} + \mathbb{P}\left[y > 0 \mid x_k\right] g_k(y \mid x_k) \mathbb{I}_{\{y>0\}} \right\}, \quad (3.2)$$

where y is the cube root of precipitation accumulation, $\mathbb{I}$ is the general indicator function, and w_k is the BMA weight.

The logistic regression model is

$$\text{logit}\mathbb{P}\left(y = 0 \mid x_k\right) = \log \frac{\mathbb{P}\left(y = 0 \mid x_k\right)}{\mathbb{P}\left(y > 0 \mid x_k\right)} = a_{0k} + a_{1k} x_k^{1/3} + a_{2k} \delta_k, \quad (3.3)$$

where $x_k^{1/3}$ is the power-transformed forecast x_k, and δ_k is an indicator variable, that is equal to one if $x_k = 0$ and equal to zero otherwise. The probability $\mathbb{P}\left(y = 0 \mid x_k\right)$ specifies the probability of zero precipitation given x_k, and $\mathbb{P}\left(y > 0 \mid y_k\right)$ its complement, the probability of non-zero precipitation given x_k, conditional on x_k providing the best forecast in the ensemble.

The predictive PDF $g_k(y \mid x_k)$ of the cube root of the precipitation amount y, given that it is positive, is specified as a gamma distribution

$$g_k(y \mid x_k) = \frac{1}{\beta_k^{\alpha_k} \Gamma\left(\alpha_k\right)} y^{\alpha_k - 1} \exp\left(-\frac{y}{\beta}\right),$$

where $\alpha_k = \frac{\mu_k^2}{\sigma_k^2}$ is the shape parameter and $\beta_k = \frac{\sigma_k^2}{\mu_k}$ is the scale parameter of the gamma distribution. The mean, μ_k, and the variance, σ_k^2, of the gamma distribution are modeled as

$$\mu_k = b_{0k} + b_{1k} x_k^{1/3} \qquad \text{and} \qquad \sigma_k^2 = c_{0k} + c_{1k} x_k.$$

As the variance parameters c_{0k} and c_{1k} do not vary much from one model to another, Sloughter et al. (2007) restrict them to be constant across all ensemble members. The c_{0k} and c_{1k} terms are replaced with c_0 and c_1, reducing not only the number of parameters to be estimated, but also the risk of overfitting.

Parameter estimation is based on training data, i.e. forecast-observation pairs from a training period. The parameters a_{0k}, a_{1k}, a_{2k} are member-specific and estimated separately for each ensemble member, using the logistic regression in (3.3) with precipitation/no precipitation as the dependent variable, and $x_k^{1/3}$ and δ_k as the independent variables. The mean parameters b_{0k} and b_{1k} are estimated using linear regression over all the cases where precipitation occurred. These parameters are also member-specific, and are estimated using the cube root of precipitation as the dependent variable, and the cube root of the forecasted accumulation amount as the independent variable. The remaining parameters c_0 and c_1 and the BMA weights $w_1, \ldots, w_K$ are estimated using the maximum likelihood technique and the EM algorithm.

3.2 Logistic Regression

Logistic regression is a nonlinear regression technique that has been implemented in a variety of fields ranging from statistics and epidemiology to the social sciences and econometrics. It is particularly suited to probability forecasting, as it predicts the occurrence of an event by fitting data to a logistic function, yielding 'S-shaped' prediction functions that are, like probabilities, strictly bounded to the unit interval.

Denoting p as the probability of a particular outcome, the logistic regression model can be expressed as

$$p = \frac{\exp\left[-f\left(\mathbf{x}\right)\right]}{1 + \exp\left[-f\left(\mathbf{x}\right)\right]} \tag{3.4}$$

or, equivalently

$$\log\left[\frac{p}{1-p}\right] = -f\left(\mathbf{x}\right), \tag{3.5}$$

where $f\left(\mathbf{x}\right)$ is a linear function of the predictor variables $\mathbf{x}$, say

$$f\left(\mathbf{x}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K. \tag{3.6}$$

Note that even though the log-odds ratio of p is related linearly to both the predictor variables $\mathbf{x}$ and the parameters β_i, the probabilities themselves are not. Thus, the

regression parameters β_i cannot be estimated directly using standard linear regression techniques; rather they are generally estimated using an iterative maximum likelihood procedure.

Post-Processing Ensemble Forecasts with Logistic Regression

An important recent implementation of logistic regression is in the statistical post-processing of ensemble forecasts of continuous weather variables such as temperature or precipitation, for which the forecast probabilities p pertain to the occurrence of the verification, Y, above or below a prediction threshold q:

$$p = \mathbb{P}\left(Y \leq q\right). \tag{3.7}$$

Furthermore, when the predictor variables $\mathbf{x}$ are given by a K member ensemble forecast, the predictors in (3.6) are usually simple functions of these forecasts, such as the mean value function, the variance function, or the square root of a single entry.

Although logistic regression has been found to perform comparably well for the statistical post-processing of ensemble forecasts, notable difficulties arise when it is used in its classical form. Specifically, separate regression equations are usually fitted for a finite number of predictand thresholds, yielding a collection of threshold probabilities rather than full forecast probability distributions. As a consequence, probabilities for intermediate predictand thresholds must be interpolated from the finite collection of regression equations, the number of which is limited, as the training sample size is limited.

However, the most problematic consequence of the conventional logistic regression framework is that threshold non-exceedance probabilities derived from the different equations are not constrained to be mutually consistent per se. As an example, consider probability forecasts for the lower tercile, $q_{1/3}$, and the upper tercile, $q_{2/3}$, of the climatological distribution of a predictand. According to (3.5), the two threshold non-exceedance probabilities, $p_{1/3} = \mathbb{P}\left(V \leq q_{1/3}\right)$ and $p_{2/3} = \mathbb{P}\left(V \leq q_{2/3}\right)$, are specified by $\log\left[\frac{p_{1/3}}{1-p_{1/3}}\right] = f_{1/3}\left(\mathbf{x}\right)$ and $\log\left[\frac{p_{2/3}}{1-p_{2/3}}\right] = f_{2/3}\left(\mathbf{x}\right)$, respectively. Then, unless the regression functions $f_{1/3}\left(\mathbf{x}\right)$ and $f_{2/3}\left(\mathbf{x}\right)$ are exactly parallel, i.e. they only differ with respect to their intercept parameters β_0, they will cross for some values of the predictors $\mathbf{x}$, leading to the nonsense result of $p_{1/3} > p_{2/3}$, which is clearly impossible.

Extending the Logistic Regression Structure

Wilks (2009) proposes to circumvent the problem of potentially incoherent forecast probabilities by fitting logistic regressions for all thresholds simultaneously. Specifically, he suggests to extend the logistic regression structure, i.e. equations (3.4) and (3.5), to include an additional predictor that is a non-decreasing function, $g(q)$, of the predictand threshold itself, yielding a unified logistic regression model that pertains to any and all predictand thresholds:

$$F(q \mid \mathbf{x}) = \mathbb{P}(Y \leq q \mid X = \mathbf{x}) = \frac{\exp[g(q) - f(\mathbf{x})]}{1 + \exp[g(q) - f(\mathbf{x})]}, \tag{3.8}$$

or,

$$\log\left[\frac{F(q \mid \mathbf{x})}{1 - F(q \mid \mathbf{x})}\right] = g(q) - f(\mathbf{x}) \ . \tag{3.9}$$

The unified logistic regression model offers several benefits. In addition to providing smoothly-varying forecast probabilities for all predictand thresholds, $F(q \mid \mathbf{x})$ can be explicitly computed for each fixed set of predictor variables $\mathbf{x}$. In this case, $F(q \mid \mathbf{x})$ is a function of $q \in \mathbb{R}$ and therefore a forecast CDF. Moreover, as different logistic regressions only differ in respect to the non-decreasing function $g(q)$, the number of parameters to be estimated is significantly reduced, and the resulting forecast probabilities are necessarily mutually consistent.

For the function $g(q)$, Wilks (2009) considers

$$g(q) = \alpha_1 \sqrt{q},$$

and (3.6) is given by

$$f(x) = \beta_0 + \beta_1 \sqrt{\overline{x}},$$

where $\overline{x}$ denotes the ensemble mean. Furthermore, this approach has been shown to perform slightly better than the classical logistic regression approach, in particular for small data sets.

A natural question arising in this context is whether the expression in (3.8) leads to a known parametric family of distributions, i.e. if we can choose $g(q)$ in such a way that (3.8) is a standard CDF. We found that if $g(q) = \alpha_0 + \alpha_1 q$ with $\alpha_1 > 0$, then $F(q \mid \mathbf{x})$ is a logistic distribution

$$F(q \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-\frac{q-\mu}{s}\right)}$$

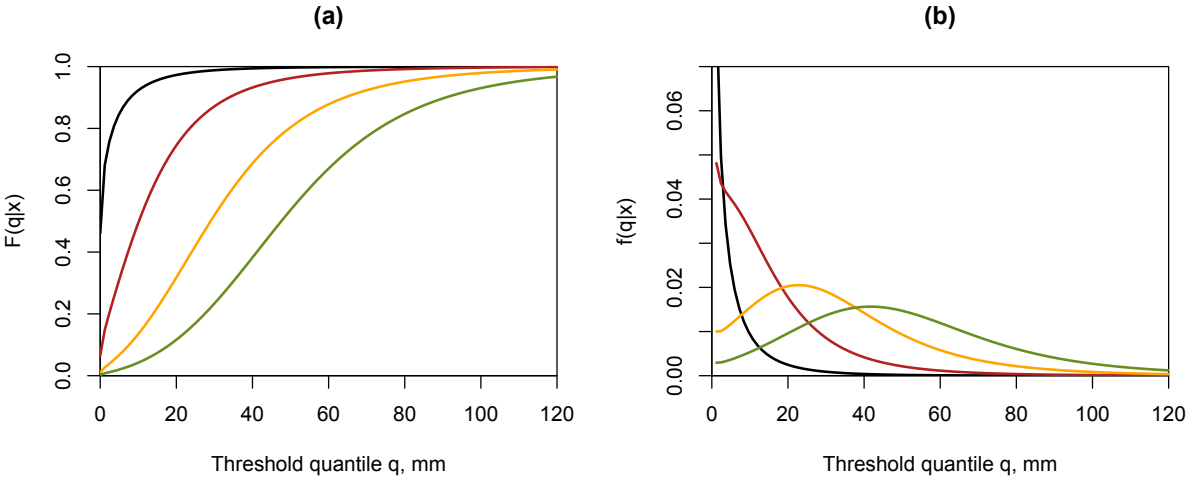

Figure 3.1: (a) CDFs and (b) PDFs for accumulated precipitation, evaluated at selected values of the ensemble mean (black corresponds to an ensemble mean of 0 mm, red to 5 mm, orange to 15 mm, and green to 25 mm). Parameter values as in Wilks (2009).

with location parameter, the mean, $\mu = \frac{f(x)-\alpha_0}{\alpha_1}$ and scale parameter $s = \frac{1}{\alpha_1}$. The variance of the logistic distribution is modeled as $\sigma^2 = \frac{\pi^2}{3\alpha_1^2}$. A result can be seen in Figure 3.1.

3.3 Quantile Regression

Opposed to conventional regression techniques, which focus on the conditional averages, quantile regression offers the possibility to estimate any and all conditional quantiles of a response variable distribution, thereby providing a more complete view of possible causal relationships.

Let Y denote the response variable, say the precipitation amount, and $X = x$ a forecast for Y. Assume that some $b \in \mathbb{R}_+$ exists such that Y, $X \in [0, b]$. Then

$$Q_Y(\tau \mid x) := \inf\{q : \mathbb{P}(Y \leq q \mid X = x) \geq \tau\}$$

denotes the τth conditional quantile $(0 \leq \tau \leq 1)$ of Y given X. A linear quantile regression model for $Q_Y(\tau \mid x)$, at a given τ, specifies

$$Q_Y(\tau \mid x) = \beta_0(\tau) + x\beta_1(\tau). \tag{3.10}$$

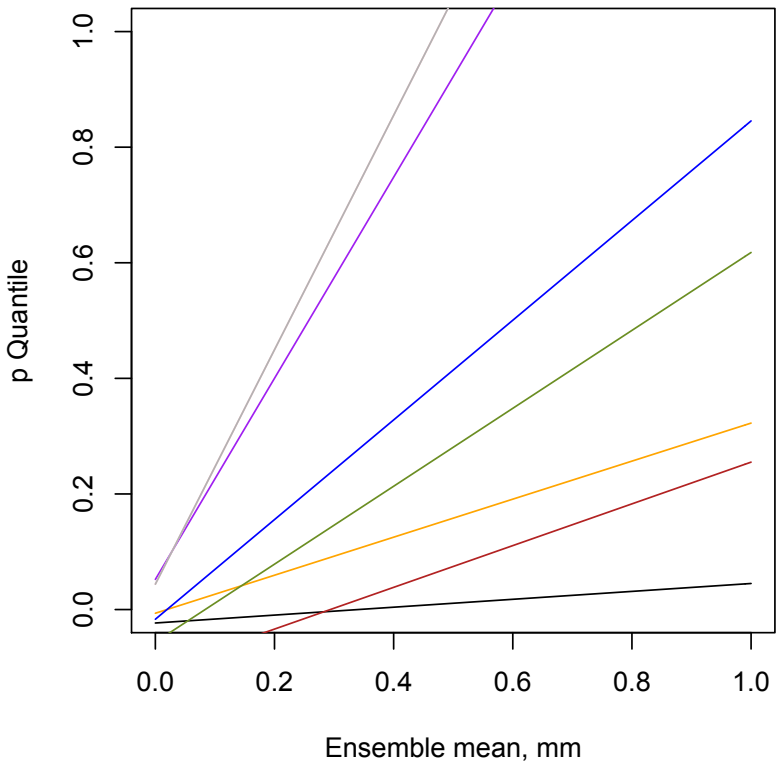

Figure 3.2: Quantile regression curves valid on 8 January 2008 in Astoria, Oregon. As these regressions are fitted separately, they are not constrained to be parallel, and logically inconsistent forecasts are inevitable for sufficiently extreme values of the predictor (black corresponds to the 1%, red to the 10%, orange to the 15%, green to the 30%, blue to the 50 %, purple to the 90% and grey to the 99% quantile).

By choosing τ appropriately, one can focus on any quantile of interest, such as the tails and other non-central parts of a response variable distribution. However, the conventional quantile regression framework can only give quantile forecasts for certain specific probabilities, rather than yielding full predictive PDFs. As a consequence, one is faced with similar problems as with the logistic regression framework, and the separately derived quantile forecasts are not automatically constrained to be mutually consistent.

Figure 3.2 shows an example for separately fitted quantile regression curves, based on data provided by the UWME (see Section 4.1 for a description). To estimate the regression coefficients, we use the R package `quantreg` with the ensemble mean as the only predictor. As the regressions were fitted separately, they are not constrained to be parallel. Thus, logically inconsistent forecasts are possible, especially for small predictor values.

3.3.1 Probabilistic Forecasts in Terms of Quantiles

Bremnes (2004) proposes a two-step approach for making reliable probability forecasts of precipitation in terms of quantiles. He suggests to first model the probability of precipitation using probit regression, and then to estimate selected quantiles in the distribution of precipitation amounts, given the occurrence of precipitation, by means of local quantile regression. By applying the laws of probability, these steps are combined in order to make unconditional quantile forecasts.

However, Bremnes (2004) does not intend to make probabilistic forecasts in terms of full predictive distributions; rather, he avoids distributional assumptions, and estimates the quantiles of interest directly. He reasons that the distribution for precipitation amounts varies highly with the predictors, and is difficult to model using standard parametric distributions whose parameters all depend on these predictors.

In order to circumvent consistency issues, the use of local quantile regression is suggested. Bremnes (2004) points out that it might be unrealistic to restrict quantiles to be linear, or difficult to find appropriate transformations for the predictors, and that it might be troublesome to put constraints on all the βs to avoid crossing quantile curves.

In contrast to Bremnes' approach, we propose a method based on quantile regression which resolves both the consistency issues and yields fully specified predictive densities. Our approach is based on a result of Tokdar and Kadane (2011).

3.3.2 New Approach Based on Tokdar and Kadane (2011)

Tokdar and Kadane (2011) introduce a semi-parametric Bayesian framework for a simultaneous analysis of linear quantile regression models. For a one-dimensional covariate, they present a simpler equivalent characterization of the monotonicity constraint through an interpolation of two monotone curves. In addition, they make use of the observation that equation (3.10) automatically lends itself to a conditional density for Y, which can be written as

$$f\left(y \mid x\right) = \frac{1}{\frac{\partial}{\partial\tau} Q_Y\left(\tau \mid x\right)}\Big|_{\tau=\tau_x(y)} = \frac{1}{\beta_0'\left(\tau\right) + x\beta'\left(\tau\right)}\Big|_{\tau=\tau_x(y)}, \tag{3.11}$$

where $\tau_x(y)$ solves $y = Q_Y\left(\tau \mid x\right)$ in τ, and β' denotes the derivative of the function β.

Based on a similar result in Tokdar and Kadane (2011), we derive the following theorem and corollary.

Theorem 1. *A linear specification of* $Q_Y(\tau \mid x)$ *as in (3.10) for* $\tau \in [0,1]$ *is monotonically increasing in* τ *for every* $x \in [0,b]$ *if and only if*

$$Q_Y(\tau \mid x) = \alpha_1 + \alpha_2 x + \left(1 - \frac{1}{b}x\right)\eta_1(\tau) + \frac{1}{b}x\eta_2(\tau) \tag{3.12}$$

where α_1, α_2 *are constants and* η_1, η_2 *are monotonically increasing in* τ.

Proof. If $Q_Y(\tau \mid x)$ is given by (3.12), then it must be monotonically increasing in τ for every $x \in [0,b]$, for which both $1 - \frac{1}{b}x$ and $\frac{1}{b}x$ are non-negative. One can express such $Q_Y(\tau \mid x)$ as in (3.10) by defining $\beta_0(\tau) = \alpha_1 + \eta_1(\tau)$ and $\beta(\tau) = \alpha_2 + \frac{1}{b}(-\eta_1(\tau) + \eta_2(\tau))$.

For the converse, every monotonicity-obeying $Q_Y(\tau \mid x)$ of the form (3.10) can be expressed as (3.12) by taking $\alpha_1 = 0$, $\alpha_2 = 0$, $\eta_1 = Q_Y(\tau \mid 0)$, $\eta_2 = Q_Y(\tau \mid b)$. □

Corollary 2. *It follows that*

$$\beta_0'(\tau) = \eta_1'(\tau) \tag{3.13}$$

and

$$\beta_1'(\tau) = \frac{1}{b}\left[\eta_2'(\tau) - \eta_1'(\tau)\right]. \tag{3.14}$$

Corollary 2 implies that the density in (3.11) is of the form

$$f(y \mid x) = \frac{1}{\left(1 - \frac{x}{b}\right)\eta_1'(\tau) + \frac{x}{b}\eta_2'(\tau)}\Big|_{\tau=\tau_x(y)}. \tag{3.15}$$

Let us assume for a moment that $\eta_1 = \eta_2$. Under this assumption, we can easily write the conditional density f_Y in terms of η_1. In order to find $\tau_x(y)$, we need to solve

$$y = \alpha_1 + \alpha_2 x + \eta_1(\tau) \tag{3.16}$$

for τ. This gives

$$\tau = \eta_1^{-1}(y - [\alpha_1 + \alpha_2 x]). \tag{3.17}$$

If we combine (3.15), (3.13) and (3.17), we get

$$f(y \mid x) = \frac{1}{\eta_1'\left(\eta_1^{-1}(y - [\alpha_1 + \alpha_2 x])\right)}. \tag{3.18}$$

Recall that

$$\frac{1}{\eta_1'\left(\eta_1^{-1}(z)\right)} = \frac{\partial}{\partial z}\eta_1^{-1}(z). \tag{3.19}$$

When now comparing (3.18) and (3.19), keeping in mind that f_Y is a density, we see that η_1 must be something like the quantile function of the forecast error $z = y - [\alpha_1 + \alpha_2 x]$. However, this approach will not work if $y \geq 0$: Without loss of generality, assume $\alpha_1 = 0$ and $\alpha_2 \geq 1$. Then it follows from (3.16) that $\eta_1(\tau) \geq 0$ for all τ. Hence, we obtain $y \geq x$ for all values of x.

In order to resolve these problems, we log-transform the support of f_Y and model the predictive density for Y as a mixture of a point mass at zero and a log-normal distribution.

3.3.3 Discrete-Continuous Model

Let Y denote precipitation, $\mathbf{x} = (x_1, \ldots, x_k)$ an ensemble forecast for Y, $\overline{x}$ the ensemble mean, and s^2 the ensemble variance. We model the predictive density for Y as

$$p(y \mid \mathbf{x}) = \mathbb{P}[y = 0 \mid \mathbf{x}]\, \mathbb{I}_{\{y=0\}} + \mathbb{P}[y > 0 \mid \mathbf{x}]\, g(y \mid \mathbf{x}) \mathbb{I}_{\{y>0\}}. \tag{3.20}$$

For precipitation occurrence, we employ logistic regression with a power transformation of the ensemble mean as a first predictor, an indicator function $\delta_{\overline{x}}$ as a second predictor, and the ensemble variance as a third predictor:

$$\text{logit}\mathbb{P}(y > 0 \mid \mathbf{x}) = \log\frac{\mathbb{P}(y > 0 \mid \mathbf{x})}{\mathbb{P}(y = 0 \mid \mathbf{x})} = \nu_0 + \nu_1 \overline{x}^{1/3} + \nu_2 \delta_{\overline{x}} + \nu_3 s^2.$$

The predictive distribution of the precipitation amount y, given that it is positive, is specified as a log-normal distribution with density function

$$g(y \mid \mathbf{x}) = \frac{1}{y\sigma_{\mathbf{x}}\sqrt{2\pi}} \exp\left(-\frac{(\log y - \mu_{\mathbf{x}})^2}{2\sigma_{\mathbf{x}}^2}\right). \tag{3.21}$$

This distribution can be derived from the linear quantile regression model in (3.10), where the parameters $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ may depend on the ensemble forecast $\mathbf{x}$ in various ways.

Homoskedastic Model Depending on the Ensemble Mean

In this context we employ a homoskedastic model depending on the ensemble mean. However, alternative models are possible, see Appendix B for examples.

Let $z = \log y$ denote the precipitation amount on the log-scale. Assume that the covariate $\mathbf{x}$ from the previous section is the ensemble mean $\overline{x}$, and that

$$\eta_1(\tau) = \eta_2(\tau) = F^{-1}\left(\tau; 0, \omega^2\right) = \omega\Phi^{-1}(\tau),$$

where $F(\cdot; 0, \omega^2)$ is the CDF of a normal distribution with mean 0 and variance ω^2, and Φ is the CDF of a standard normal distribution. From (3.18) and (3.19) it follows that

$$f(z \mid \overline{x}) = \frac{1}{\omega}\varphi\left(\frac{z - (\alpha_1 + \alpha_2\overline{x})}{\omega}\right),$$

where φ is the PDF of a standard normal distribution. That is, the parameters of the predictive density g in (3.21) are given by

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2\overline{x} \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \omega. \tag{3.22}$$

3.3.4 Parameter Estimation

For forecasts on any given day, parameter estimation is based on forecast and observation data from a rolling training period, which consists of the N most recent days available. In Section 4.2, we give details for the choice of N.

Logistic Regression Model: Parameters ν_0, ν_1, ν_2, and ν_3

The parameters ν_0, ν_1, ν_2, and ν_3 are estimated using the maximum likelihood technique for the logistic regression model:
Let W be a Bernoulli random variable, indicating whether precipitation occurs. We then obtain

$$p(w) = p^w \cdot (1-p)^{1-w}$$

as the density for w, with the respective probability of precipitation p.

The likelihood function $L(\nu_0, \nu_1, \nu_2, \nu_3) = \prod_{i=1}^{N} p(w_i)$ can be interpreted as the probability of the training data being observed, viewed as a function of the parameters.

As it is customary, we maximize the log-likelihood function rather than the likelihood function itself, in order to obtain the optimal parameters based on the training data:

$$\begin{aligned} l\left(\nu_0, \nu_1, \nu_2, \nu_3\right) = \log\left(\prod_{i=1}^{N} p\left(w_i\right)\right) &= \sum_{i=1}^{N} \log\left(p_i^{w_i} \cdot \left(1-p_i\right)^{1-w_i}\right) \\ &= \sum_{i=1}^{N} \left[w_i \log\left(p_i\right) - \left(1-w_i\right) \log\left(1-p_i\right)\right] \end{aligned}$$

As $\log\left(\frac{p}{1-p}\right) = \nu_0 + \nu_1 \overline{x}^{1/3} + \nu_2 \delta_{\overline{x}} + \nu_3 s^2$ denotes the logit for p,

$$\begin{aligned} l\left(\nu_0, \nu_1, \nu_2, \nu_3\right) &= \sum_{i=1}^{N} \left[w_i \left(\nu_0 + \nu_1 \overline{x}_i^{1/3} + \nu_2 \delta_{\overline{x}_i} + \nu_3 s_i^2\right)\right] - \\ &\quad \sum_{i=1}^{N} \left[\log\left(1 + \exp\left(\nu_0 + \nu_1 \overline{x}^{1/3} + \nu_2 \delta_{\overline{x}} + \nu_3 s^2\right)\right)\right]. \end{aligned}$$

Log-Normal Distribution: Parameters α_1, α_2, and ω

The parameters α_1, α_2, and ω are estimated employing minimum continuous ranked probability score (CRPS) estimation, as proposed in Gneiting et al. (2005). As the predictive density is modeled with a log-transformed normal distribution, we use the CRPS for normal distributions (see Appendix A.2) with the log-transformed observation set z.

When expressing the CRPS in terms of the parameters, the average score over all N pairs of forecasts and observations contained in the training data set is

$$\Gamma\left(\alpha_1, \alpha_2; \omega\right) = \frac{1}{N} \sum_{i=1}^{N} \omega \left\{ Z_i \left[2\,\Phi\left(Z_i\right) - 1\right] + 2\,\varphi\left(Z_i\right) - \frac{1}{\sqrt{\pi}} \right\},$$

where

$$Z_i = \frac{z_i - \left(\alpha_1 + \alpha_2 \overline{x_i}\right)}{\omega},$$

and $\Phi\left(\cdot\right)$ and $\varphi\left(\cdot\right)$ denote the CDF and the PDF of the standard normal distribution, respectively.

Both optimization processes are performed with the `optim` function in `R` and the Broyden-Fletcher-Goldfarb-Shanno algorithm (R Development Core Team, 2012). As initial values for the algorithm, we use the results of the previous day's estimation. For the forecast errors, we assume independence of time and space, which is reasonable as we

only make forecasts for one time and one location simultaneously (Raftery et al., 2005; Gneiting et al., 2005).

Furthermore, because of the similarities to the family of EMOS post-processing techniques, we refer to our technique in the following as EMOS.

Chapter 4

Case Study

We apply our EMOS method in a case study to 48-h precipitation forecasts of 24-h precipitation accumulation over the North American Pacific Northwest. We describe the forecast and observation data used, and the choice of training data in the subsequent Sections 4.1 and 4.2. In Section 4.3, we illustrate how the method works in practice. In the final section, Section 4.4, we evaluate the performance of our new approach, by employing the assessment tools presented in Appendix A and comparing its performance to the unprocessed raw ensemble and climatology.

4.1 Forecast and Observation Data

This case study is based on forecast data provided by the University of Washington mesoscale ensemble (UWME; Eckel and Mass, 2005), for the period between 1 January 2008 and 31 December 2008. During this time, the eight-member multi-analysis ensemble consisted of multiple runs of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5) with initial and lateral boundary conditions from eight different operational centers around the world, see Table 4.1.

The region covered by the UWME is the North American Pacific Northwest. Specifically, an inner nest, of 12 km grid spacing, roughly consists of the states Washington, Oregon and Idaho, as well as the southern part of the Canadian province British Columbia. The outer domain covers the western part of North America and large parts of the eastern Pacific Ocean, having a horizontal grid spacing of 36 km.

We use 48-h ahead forecasts of daily (24 hour) precipitation accumulation generated on the 12 km grid and bilinearly interpolated to observation locations from the forecast

Table 4.1: The UWME: initial and lateral boundary conditions from eight different operational centers around the world.

No	IC / LBC Source	Operational Center
1	Global Forecast System	USA National Centers for Environmental Prediction
2	Global-Environmental Multi-Scale Model	Canadian Meteorological Center
3	ETA Limited-Area Mesoscale Model	USA National Centers for Environmental Prediction
4	Global Analysis and Prediction Model	Australian Bureau of Meteorology
5	Global Spectral Model	Japanese Meteorological Agency
6	Navy Operational Global Atmospheric Prediction System	Fleet Numerical Meteorological and Oceanographic Center
7	Global Forecast System	Taiwan Central Weather Bureau
8	Unified Model	UK Met Office

grid points. The observation data come from meteorological stations located in the Pacific Northwest, and were subject to quality control procedures, described in Bahrs (2005), that is, dates and locations with missing forecasts or observations were removed from the data set.

For the time period considered, the data set contains 17,270 pairs of ensemble forecasts and observations, corresponding to 297 forecast days. There are 69 observation locations, whose distribution is pictured in Figure 4.1. In order to provide an appropriate training period for all days of 2008, additional data from the year 2007 is used. Further information about the UWME, now using the WRF mesoscale model, as well as real time forecasts and observations can be found at `http://www.atmos.washington.edu/~ens/uwme.cgi`.

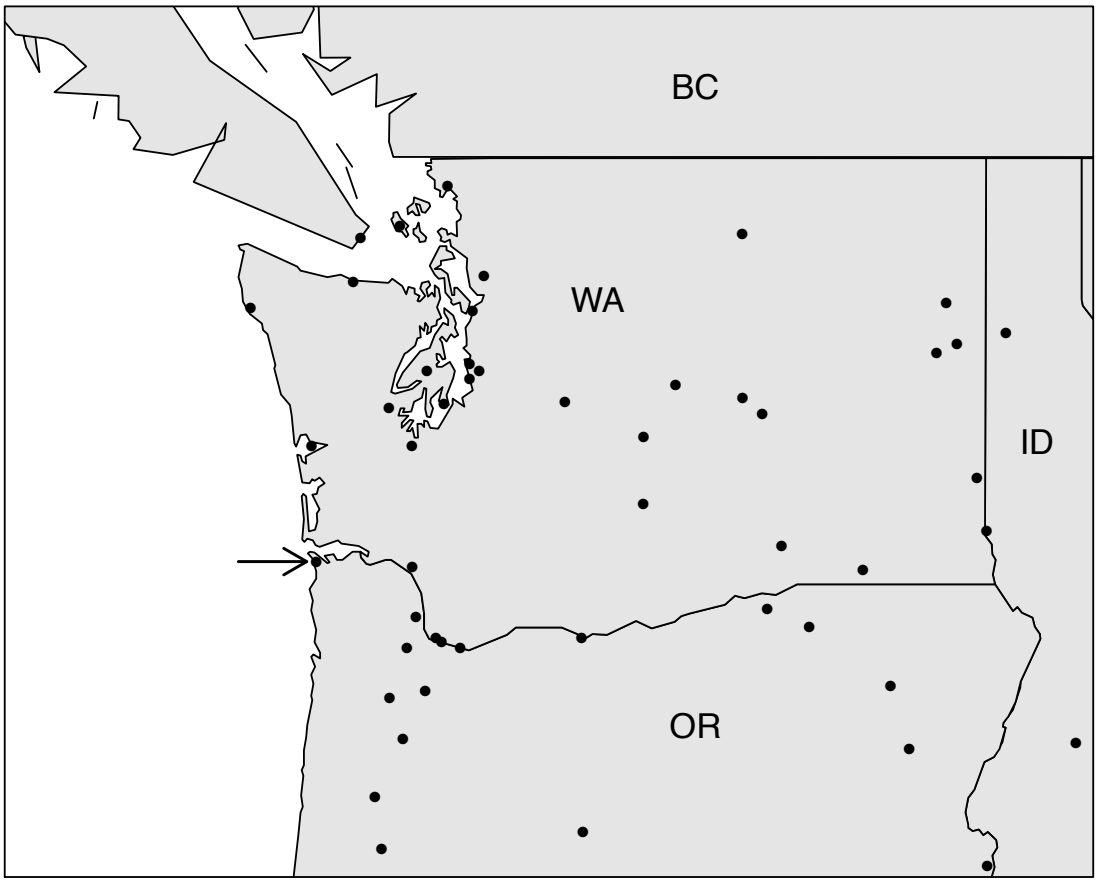

Figure 4.1: The 69 observation locations contained in the 2008 data set are located in the Pacific Northwest, including the Canadian province British Columbia (BC), and the US states of Washington (WA), Oregon (OR) and Idaho (ID). The arrow indicates the city of Astoria, Oregon (see Figure 3.2 and Section 4.3).

4.2 Choice of Training Data

As noted, we fit the parameters of our EMOS model using forecast and observation data from a rolling training period. That is, on any given day, we use training data from the N most recent days available. In principle, a longer training period reduces the statistical variability in the parameter estimation. However, if a long training period is chosen, the model is also less adaptive to e.g. changes in atmospheric regimes. For each day, we use training data from all 69 stations to estimate a single set of parameters across the Pacific Northwest, comparable to the regional EMOS method (Thorarinsdottir

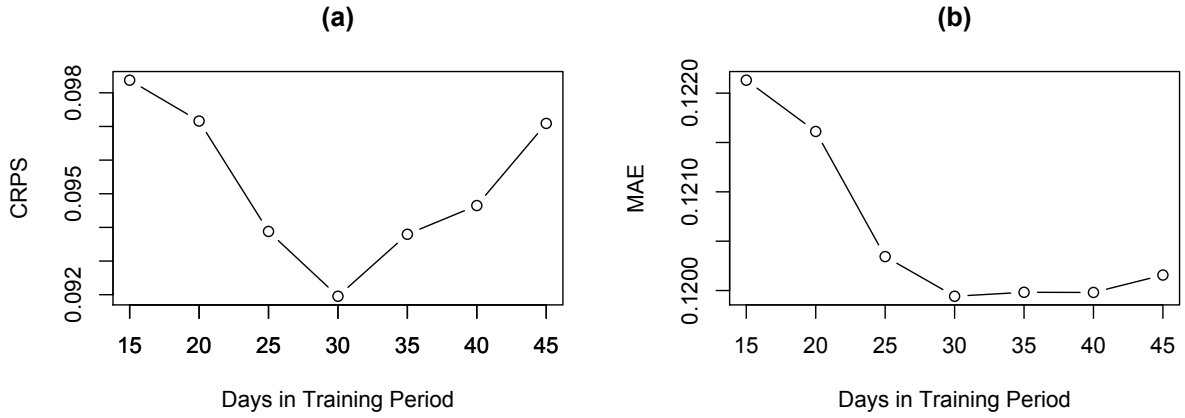

Figure 4.2: Comparison of training period lengths: (a) CRPS of EMOS forecasts and (b) MAE of the EMOS median forecasts.

and Gneiting, 2010). These parameters are then used to create EMOS forecasts at each of the 69 stations.

To make an informed decision about the length of the training period, we computed the average continuous ranked probability score (CRPS; see Appendix A) for the probabilistic forecasts and the mean absolute error (MAE) of the resulting deterministic forecasts, considering various training period lengths. The results are shown in Figure 4.2. Both the CRPS and the MAE improve (decrease) as the length of the training period N increases to 30 days, and thereafter they deteriorate (increase). We therefore use a training period of $N = 30$ days.

Note that it is however possible that other forecast lead times and other geographic regions require other choices for the training period.

4.3 Example: Forecast for Astoria, 8 January 2008

In order to illustrate how the new EMOS method works, we show an example of a post-processed 48-hour-ahead precipitation forecast valid on 8 January 2008 at station KAST in Astoria, Oregon.

The parameters estimated for this particular date are shown in Table 4.2. The logistic regression model assigns the highest coefficient, ν_1, to the ensemble mean, indicating a high predictive skill for this predictor. Note that the coefficient, ν_3, for the ensemble variance is negative, which implies an inversely proportional relationship between the

Table 4.2: Parameter estimates for the predictive model in (3.20), valid on 8 January 2008.

	ν_0	ν_1	ν_2	ν_3	α_1	α_2	ω
EMOS	−3.12	5.87	0.16	−0.99	−3.84	3.19	0.96

Table 4.3: Ensemble and EMOS forecast characteristics and the verifying observation for station KAST, valid on 8 January 2008.

	PoP	mean	variance	median	observation
Ensemble	1.00	0.17	0.01	0.13	0.03
EMOS	0.54	0.19	0.05	0.03	

variance and the probability of precipitation. Furthermore, both intercept parameters, ν_0 and α_1, have a rather high negative value, suggesting a high bias in the ensemble.

Table 4.3 shows the ensemble mean, median and variance, the EMOS results, and the verifying observation. By applying the logistic regression model in order to derive a calibrated PoP forecast, the unrealistically high probability denoted by the raw ensemble is reduced to 54% of the original value. Considering the quantative forecast, we find that EMOS adjusts the ensemle spread, thereby giving a more realistic estimate of forecast uncertainty. As for the deterministic forecasts, we can see that the observation lies fairly close to the EMOS median, which is substantially lower than the ensemble median forecast.

The corresponding predictive PDF at this date and location is displayed in Figure 4.3. The probability of exceeding a particular threshold can be derived by the proportion of the area under the solid curve to the right of the threshold, multiplied by the probability of precipitation. Although the observation lies far outside the ensemble range, it is contained in the region of high probability of the EMOS predictive PDF.

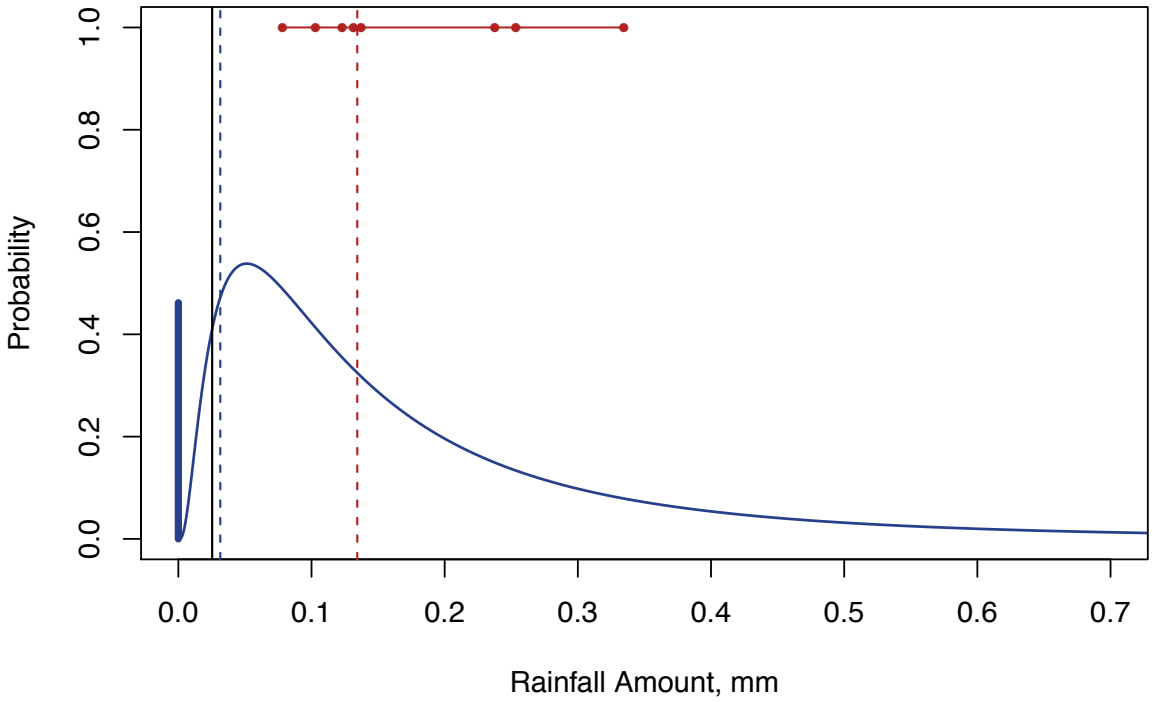

Figure 4.3: 48-hour-ahead EMOS density forecast of precipitation valid on 8 January 2008, in Astoria, Oregon. The blue vertical line at zero represents the EMOS estimate for the probability of no precipitation, and the solid curve the EMOS PDF of the precipitation amount given that it is non-zero. The dashed red line represents the ensemble median forecast, the solid red line the ensemble range, and the red dots the ensemble member forecasts. The dashed blue line indicates the EMOS median forecast, and the black vertical line the verifying observation.

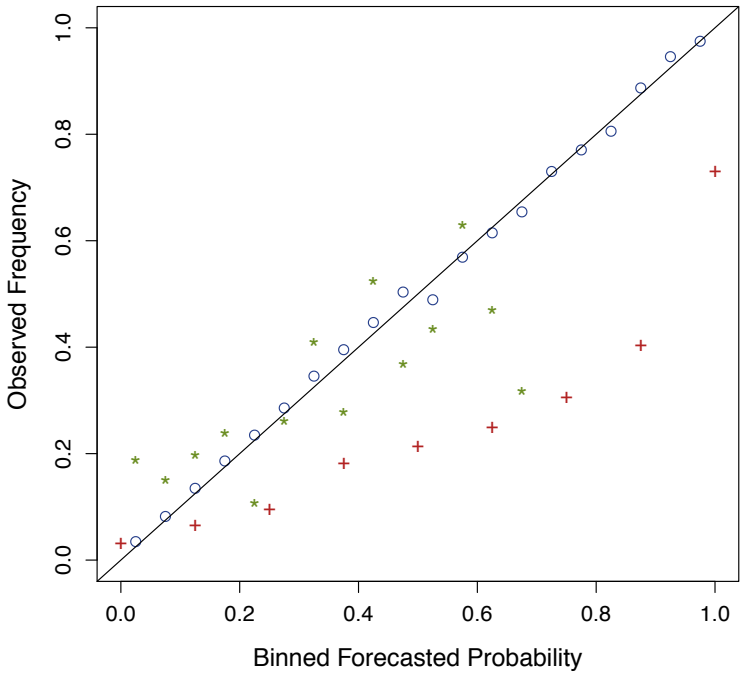

Figure 4.4: Reliability diagram of binned PoP forecast versus observed relative frequency of precipitation, for climatology (stars), consensus voting of the raw ensemble (crosses) and EMOS (circles).

4.4 Results for the Pacific Northwest, 2008

In this section, we present the results for both the probability of precipitation (PoP) forecasts and the precipitation amount forecasts, aggregated over all locations for the full calendar year 2008. We compare our new method to the raw ensemble forecast and a climatological forecast, which was produced for each day by taking all available observations from the training data as an ensemble forecast.

4.4.1 Probability of Precipitation Forecasts

We begin with a discussion of the PoP forecasts. Figure 4.4 shows the reliability diagram. Our new approach produces well-calibrated results, while both the consensus vote from the raw ensemble and the climatological ensemble produce severely uncalibrated forecasts.

Note that the climatologal ensemble failed to produce any PoP forecasts greater that 0.7, see Figure 4.5. This is due to both the large number of ensemble members and the high variability within the observations, as we use a regional approach. Furthermore we notice that EMOS calibrates the raw ensemble, by reducing the amount of probability

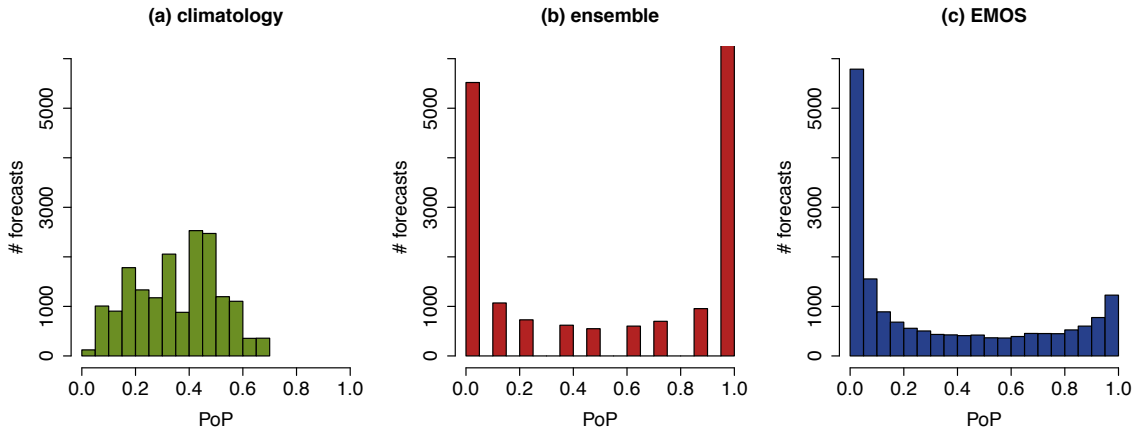

Figure 4.5: Histograms of forecast frequency for probability of precipitation, over all locations and available test dates in 2008.

forecasts equal to one. The Brier scores for PoP forecasts, see Table 4.4, reflect these results, in that the Brier scores based on our new approach were significantly better than both the raw ensemble and climatology.

4.4.2 Precipitation Amount Forecasts

In assessing probabilistic forecasts of quantitative precipitation, we follow Gneiting et al. (2005) and aim to maximize the sharpness of the predictive PDFs, subject to calibration. In order to assess calibration, we consider Figure 4.6, which shows the verification rank histogram (VRH) for the raw ensemble forecast and climatology, and the probability integral transform (PIT) histogram for the EMOS forecast distributions, see Appendix A for a general description of these methods.

For the verification rank histogram, there were incidences where the observed value was zero (no precipitation), and one or more forecasts were also zero. To obtain a rank in these situations, the observation rank was randomly chosen between zero and the number of forecasts equal to zero. In order to calculate the values for the PIT histogram, the EMOS cumulative distribution function was evaluated at its corresponding observation. In the case of an observation of zero, a value was randomly drawn between zero and the probability of no precipitation.

The histogram for the raw ensemble forecast is far from flat, indicating a lack of calibration. Specifically, it shows that the raw ensemble is underdispersed, that is, too

Table 4.4: Brier scores (BS) for probability of precipitation forecasts, and MAE and CRPS for quantitative precipitation forecasts over the Pacific Northwest in 2008. MAE and CRPS are given in mm, and the MAE refers to the deterministic forecast given by the median of the respective forecast distribution.

	BS	MAE	CRPS
Climatology	0.22	0.18	0.15
Ensemble	0.19	0.14	0.11
EMOS	**0.11**	**0.12**	**0.09**

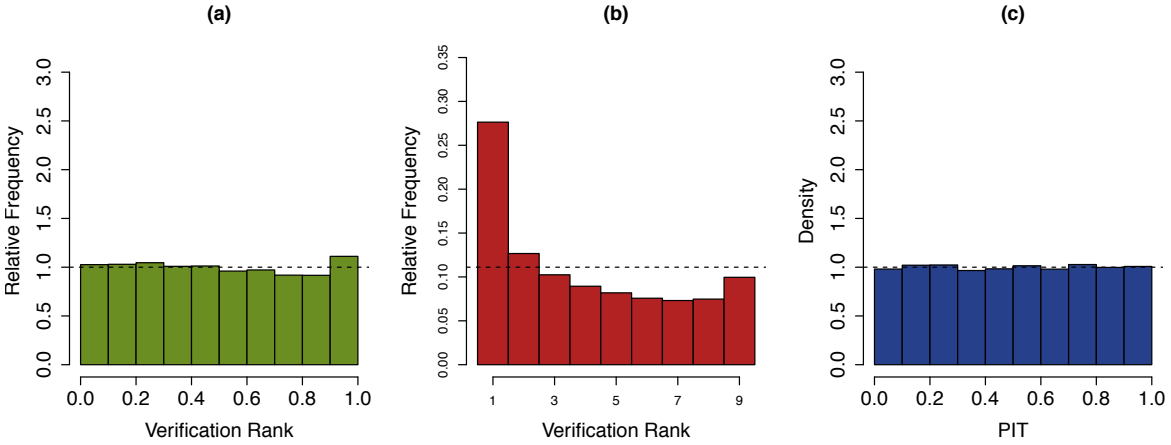

Figure 4.6: Verification rank histogram for (a) the climatological forecast and (b) the raw ensemble forecast, and (c) PIT for EMOS forecast distributions of precipitation accumulation.

many observations fall out of the ensemble range. The PIT histogram for the post-processed EMOS forecast on the other hand shows substantially better calibrated results, reflected by a uniformly distributed histogram. As can be expected, the climatological forecast seems to be fairly well calibrated as well.

To further assess the predictive performance of the three forecasting methods, we employ the MAE and the CRPS, see Appendix A. Table 4.4 shows the MAE and CRPS values for climatology, raw ensemble forecasts, and EMOS forecasts, all measured in millimeters. Deterministic forecasts can be created from all forecasts by finding the median of the predictive PDF, or the forecast ensemble, respectively; the MAE refers to the error of this desterministic forecast. EMOS outperfomed both climatology and the raw ensemble. The results for the CRPS were even stronger, in that the post-processed EMOS forecast improved by 18% compared to the raw ensemble forecast.

Chapter 5

Discussion

We proposed a novel way of statistically post-processing ensemble forecasts of precipitation based on quantile regression. The resulting predictive distribution is a mixture of a point mass at zero and a log-normal distribution. That is, it has two components: the probability of zero precipitation occurrence, and a predictive distribution for the precipitation accumulation given that it is greater than zero; it thus provides both probability of precipitation forecasts and probabilistic quantitative precipitation forecasts in a unified form.

In our experiments with the University of Washington mesoscale ensemble (UWME; Eckel and Mass, 2005), we applied the new technique to 48-h forecasts of 24-h precipitation over the North American Pacific Northwest. The EMOS probabilistic forecasts turned out to be much better calibrated than the unprocessed ensemble and a climatological reference forecast. Specifically, the EMOS forecasts outperformed both reference forecasts in terms of their respective Brier scores. Moreover, the EMOS median forecast had a lower MAE, and the EMOS forecast PDFs had substantially lower CRPS than both reference forecasts.

In addition, by providing a full predictive PDF, EMOS offers the advantage of giving quantile forecasts for any and all precipitation thresholds of interest, and the resulting quantile forecasts are necessarily mutually consistent. This provides a substantial improvement over the conventional quantile regression framework, where separate quantile regressions have to be fitted for each threshold of interest, and logically inconsistent forecasts are possible, as separately fitted regressions are not constrained to be parallel.

Still, various improvements to our method are possible. We estimated only a single set of parameters using data from the entire Pacific Northwest, and a more local approach

might perform better, see e.g. the local EMOS method (Thorarinsdottir and Gneiting, 2010). Moreover, our choice of training period is specific to our data set and region, where it rains relatively often. Other forecast lead times and other geographic regions are likely to require a different training period. Furthermore, we assumed independence of forecast errors in space and time, which is reasonable as we only make forecasts for one time and one location simultaneously (Raftery et al., 2005; Gneiting et al., 2005). However, methods for probabilistic forecasts at multiple locations have been developed for precipitation, see e.g. Berrocal et al. (2008). Finally, besides the used homoskedastic model for quantitative precipitation, several other models, which e.g. include heteroskedasticity or depend on additional predictors, are possible.

Appendix A

Performance Measures for Probabilistic Forecasts

In concert with statistical post-processing, ensemble prediction systems offer the possibility of well-calibrated probabilistic forecasts in form of predictive probability density functions (PDFs) over future weather quantities or events. Often it is critical to assess the predictive ability of these forecasts, or to compare and rank competing forecasting methods.

In the introduction we stated that, according to the diagnostic paradigm of Gneiting et al. (2007), the goal is to maximize the sharpness of a probabilistic forecast subject to its calibration. Calibration refers to the reliability of the forecast, that is the statistical consistency between the probabilistic forecast and the actually occurring observations. Sharpness refers to the concentration of the predictive distribution; it is a property of the forecasts only. Under the condition that all forecasts are calibrated, we define the sharpest to be the best.

Here we present several methods to assess the predictive performance of probabilistic forecasts for both dichotomous events, such as probability of precipitation, and probabilistic density forecasts, such as probabilistic forecasts of quantitative precipitation. For further reference see e.g. Wilks (2011).

A.1 Assessing Dichotomous Events

Brier Score (Accuracy Measure)

The most common accuracy measure for verification of probabilistic forecasts of dichotomous events, such as the probability of precipitation, is the *Brier score* (BS). It measures the total probability error, considering that the observation is 1 if the event occurs, and 0 if the event does not occur. The Brier score averages the squared differences between pairs of forecast probabilities p_i and the subsequent binary observations x_i,

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - x_i)^2 ,$$

where the index i denotes the numbering of the N forecast-event pairs. The Brier score can take values in the range $0 \leq BS \leq 1$, and is negatively oriented, i.e. perfect forecasts exhibit $BS = 0$. Less accurate forecasts receive higher Brier scores, note however that it weights large errors more than small ones.

Reliability Diagram (Calibration)

The *reliability diagram* is a graphical device that shows the full joint distribution of forecasts and observations for probability forecasts of a binary predictand, such as the probability of precipitation occurrence. It measures the agreement between predictand probabilities and observed frequencies, i.e. if the forecasts are calibrated. In order to obtain a reliability diagram, the binned forecast probabilities are plotted against the observed relative frequencies. For perfect calibration, the binned forecast probabilities and the observed frequencies should be equal, and the plotted points should lie on the diagonal.

A.2 Assessing Probabilistic Forecasts

PIT / VRH (Calibration)

The prefered tool to assess calibration of density forecasts is the *probability integral transform histogram* (PIT histogram; Dawid 1984, Diebold et al. 1998). If the predicted

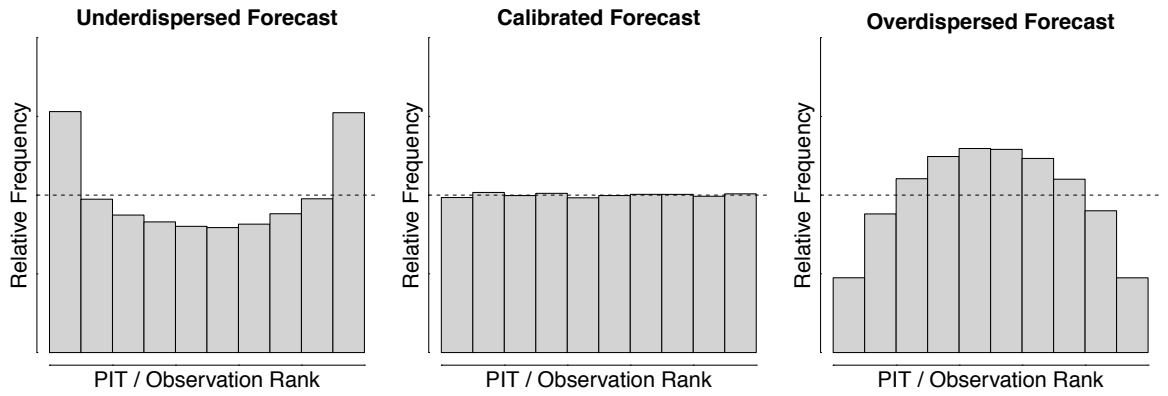

Figure A.1: Example VRHs /PITs, illustrating characteristic dispersion errors. Perfect uniformity is indicated by the horizontal dashed lines.

distribution F is equal to a hypothetical "true" distribution, the value of the predictive cumulative distribution function (CDF) at the observation x has a uniform distribution

$$p = F(x) \sim \mathcal{U}[0, 1].$$

For perfect calibration, the histogram of all PIT values p, computed for all forecasts available, should show a flat shape. Deviations from uniformity can be used to diagnose aggregate deficiencies, as they are both easy to recognize and to interpret. A U-shaped histogram indicates overconfidence, in that the spread of the underlying predictive distribution is too small, and the predictive distribution is under-dispersed. An over-dispersed distribution reveals itself in a hump-shaped diagram, with too many observations in the center of the distribution.

The discrete analog of the PIT histogram is the *verification rank histogram* or *Talagrand diagram* (VRH; Anderson 1996, Hamill and Colucci 1998, Talagrand et al. 1997), which can be used to evaluate ensemble forecasts. It plots the frequency of the observation ranks within the forecast ensemble. As before, a U-like shape implies under-dispersion, while the opposite holds for a hump-like shape. Skewed histograms indicate a certain bias and calibrated ensembles produce a flat histogram. The VRH can be interpreted in the same manner as the PIT histogram, see Figure A.1 for characteristic shapes of these histograms.

Note however that a flat histogram does not necessarily indicate a skilled forecast, i.e. rank uniformity is a necessary but not sufficient criterion for determining that an ensemble is reliable.

Proper Scoring Rules (Accuracy Measures)

In order to evaluate calibration and sharpness simultaneously, we employ so called proper scoring rules. These accuracy measures assign penalties depending on forecast accuracy, and can be used to compare competing forecasting methods.

Continuous Ranked Probability Score (CRPS)

A widely used proper scoring rule for probabilistic forecasts is the *continuous ranked probability score* (CRPS). It is defined as

$$\operatorname{crps}(P, x) = \int_{-\infty}^{\infty} \left(F(y) - \mathbb{I}\{y \geq x\}\right)^2 \mathrm{d}y,$$

where F denotes the CDF associated with the predictive distribution P, and x denotes the observation.

For ensemble forecasts, where the predictive distribution P_{ens} places a point mass of $\frac{1}{N}$ on the ensemble members $x_1, \ldots, x_N \in \mathbb{R}$, the CRPS can be evaluated as

$$\operatorname{crps}(P_{\text{ens}}, x) = \frac{1}{N} \sum_{j=1}^{N} |x_j - x_i| - \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_j - x_i|.$$

For a normal distribution, $\mathcal{N}(\mu, \sigma^2)$, there exists a closed form of the CRPS,

$$\operatorname{crps}\left(\mathcal{N}\left(\mu, \sigma^2\right), x\right) = \sigma \left\{ \frac{x-\mu}{\sigma} \left[2\,\Phi\left(\frac{x-\mu}{\sigma}\right) - 1 \right] + 2\,\varphi\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\},$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ denote the CDF and the PDF of the standard normal distribution, respectively (Gneiting et al., 2005).

Mean Absolute Error (MAE)

The CRPS generalizes the *mean absolute error* (MAE) and reduces to it for point forecasts; the MAE can thus be viewed as a special case of the CRPS. The MAE is specified as the mean absolute difference between the predictive median and the realizing observations:

$$\mathrm{mae}\,(P, x) = \frac{1}{N} \sum_{i=1}^{N} |\mu_i - x_i| \,,$$

where μ_i is the median of the predictive distribution and the sum extends over all forecast cases. It is used to determine the skill of the distribution median as a point forecast.

Appendix B

Alternative Models

In addition to the previously used model for quantitative precipitation, see Section B.1 and 3.3.3, we propose several other possible models, described in Section B.2 and Section B.3.

B.1 Homoskedastic Model Depending on the Ensemble Mean

Let $z = \log y$ denote the precipitation amount on the log-scale. Assume that

$$\eta_1(\tau) = \eta_2(\tau) = F^{-1}\left(\tau; 0, \omega^2\right) = \omega\Phi^{-1}(\tau),$$

where F is the CDF of a normal distribution with mean 0 and variance ω^2, and Φ is the CDF of a standard normal distribution. From (3.18) and (3.19) it follows that

$$f(z \mid \overline{x}) = \frac{1}{\omega}\varphi\left(\frac{z - (\alpha_1 + \alpha_2\overline{x})}{\omega}\right),$$

where φ is the PDF of a standard normal distribution. That is, the parameters of the predictive density g in (3.21) are given by

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2\overline{x} \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \omega. \tag{B.1}$$

B.2 Heteroskedastic Model Depending on the Ensemble Mean

We can include heteroskedasticity in our model by allowing η_1 and η_2 to differ. For instance, let

$$\eta_1(\tau) = \omega_1 \Phi^{-1}(\tau), \ \eta_2(\tau) = \omega_2 \Phi^{-1}(\tau).$$

Equation (5.16) then becomes

$$z = \alpha_1 + \alpha_2 x + \left[\left(1 - \frac{\overline{x}}{b}\right)\omega_1 - + \frac{\overline{x}}{b}\omega_2\right] \Phi^{-1}(\tau)$$

which leads to a predictive density g with

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2 \overline{x} \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \left(1 - \frac{\overline{x}}{b}\right)\omega_1 - + \frac{\overline{x}}{b}\omega_2. \tag{B.2}$$

B.3 Models Depending on the Ensemble Mean and Variance

From Equation (12) in Tokdar and Kadane (2011) it follows that it should be easy to extend the models discussed above to any of the following,

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2 \overline{x} + \alpha_3 s^2 \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \omega, \tag{B.3}$$

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2 \overline{x} + \alpha_3 s^2 \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \left(1 - \frac{\overline{x}}{b}\right)\omega_1 - + \frac{\overline{x}}{b}\omega_2, \tag{B.4}$$

$$\mu_{\mathbf{x}} = \alpha_1 + \alpha_2 \overline{x} + \alpha_3 s^2 \qquad \text{and} \qquad \sigma_{\mathbf{x}} = \left(1 - \frac{\overline{x}}{b'}\right)\omega_1 - + \frac{\overline{x}}{b'}\omega_2, \tag{B.5}$$

where b' is the upper bound for s^2.

Note that in all the models (B.1) - (B.5), the forecasts are used on their original scale. This is due to the fact that Theorem 1 only holds for bounded positive variables.

Bibliography

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

Bahrs, J. (2005) Observations QC documentation. Available at `http://www.atmos.washington.edu/mm5rt/qc_obs/qc_doc.html`.

Berrocal, V. J., A. E. R. and T. Gneiting (2008) Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals of Applied Statistics*, **2**, 1170–1193.

Bjerknes, V. (1904) Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorologische Zeitschrift*, **21**, 1–7.

Bremnes, J. B. (2004) Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Monthly Weather Review*, **132**, 338–347.

Bröcker, J. and L. a. Smith (2008) From ensemble forecasts to predictive distribution functions. *Tellus A*, **60**, 663–678.

Brown, B. G., R. W. Katz and A. H. Murphy (1984) Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology*, **23**, 1184–1195.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei and Y. Zhu (2005) A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems. *Monthly Weather Review*, **133**, 1076–1097.

Campbell, S. D. and F. X. Diebold (2005) Weather forecasting for weather derivatives. *Journal of the American Statistical Association*, **100**, 6–16.

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

Diebold, F. X., T. A. Gunther and A. S. Tay (1998) Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39**, 862–883.

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

Gahrs, G. E., S. Applequist, R. L. Pfeffer and X.-F. Niu (2003) Improved results for probabilistic quantitative precipitation forecasting. *Weather and Forecasting*, **18**, 879–890.

Glahn, H. R. and D. A. Lowry (1972) The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11**, 1203–1211.

Gneiting, T. (2008) Editorial: Probabilistic forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**, 319–321.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106**, 746–762.

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

Gneiting, T., A. Raftery, A. Westveld and T. Goldman (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098–1118.

Gneiting, T. and A. E. Raftery (2005) Weather forecasting with ensemble methods. *Science*, **310**, 248–9.

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

Hamill, T. M. and S. J. Colucci (1998) Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, **126**, 711–724.

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

Koizumi, K. (1999) An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Weather and Forecasting*, **14**, 109–118.

Kretzschmar, R., P. Eckert, D. Cattani and F. Eggimann (2004) Neural network classifiers for local wind prediction. *Journal of Applied Meteorology*, **43**, 727–738.

Krzysztofowicz, R. (2001) The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, **249**, 2–9.

Leith, C. E. (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, **102**, 409–418.

Lorenz, E. N. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20**, 130–141.

Lynch, P. (2008) The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, **227**, 3431–3444.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins and D. A. Stainforth (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772.

Palmer, T. N. (2002) The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, **128**, 747–774.

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. Available at `http://www.r-project.org`.

Raftery, A. E., T. Gneiting, F. Balabdaoui and M. Polakowski (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Sloughter, J. M., A. E. Raftery, T. Gneiting and C. Fraley (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135**, 3209–3220.

Talagrand, O., R. Vautard and B. Strauss (1997) Evaluation of probabilistic prediction systems. *Proc., ECMWF Workshop on Predictability*, 1–25, Reading, UK, European Centre for Medium-Range Weather Forecasts.

Thorarinsdottir, T. L. and T. Gneiting (2010) Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173**, 371–388.

Tokdar, S. and J. B. Kadane (2011) Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, **6**, 1–22.

Whitaker, J. S. and A. F. Loughe (1998) The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, **126**, 3292–3302.

Wilks, D. S. (2006) Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13**, 243.

Wilks, D. S. (2009) Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, **16**, 361–368.

Wilks, D. S. (2011) *Statistical Methods in the Atmospheric Sciences*. Academic Press, 3rd edition.

Erklärung

Hiermit versichere ich, dass ich meine Arbeit selbstständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

08. Mai 2012