BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 71 (2015) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sabrina Wahl

Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation

BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 71 (2015) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sabrina Wahl

Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation

Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Sabrina Wahl
aus Köln

Bonn, Mai 2015

Diese Arbeit ist die ungekürzte Fassung einer der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn im Jahr 2015 vorgelegten Dissertation von Sabrina Wahl aus Köln.

This paper is the unabridged version of a dissertation thesis submitted by Sabrina Wahl born in Köln to the Faculty of Mathematical and Natural Sciences of the Rheinische Friedrich-Wilhelms-Universität Bonn in 2015.

Anschrift des Verfassers: Address of the author:

Sabrina Wahl
Meteorologisches Institut der
Universität Bonn
Auf dem Hügel 20
D-53121 Bonn

1. Gutachter: PD Dr. Petra Friederichs, Universität Bonn
2. Gutachter: Prof. Dr. Andreas Hense, Universität Bonn

Tag der Promotion: 01. Oktober 2015
Erscheinungsjahr: 2015

Abstract

Over the last decade, advances in numerical weather prediction (NWP) led to forecasts on even finer horizontal scales and a better representation of mesoscale processes. High-resolution models provide the user with realistic weather patterns on the km-scale. However, the evaluation of such small-scale model output remains still a challenge in forecast verification and the quantification of forecast uncertainty. Ensembles are the main tool to assess uncertainty from NWP models. The first operational mesoscale NWP ensemble was developed by the German Meteorological Service (DWD) in 2010. The German-focused COSMO-DE-EPS is especially designed to improve quantitative precipitation forecasts, which is still one of the most difficult weather variables to predict.

This study investigates the potential of mesoscale NWP ensembles to predict quantitative precipitation. To comprise the uncertainty inherent in NWP, precipitation forecasts should take the form of probabilistic predictions. Typical point forecasts for precipitation are the probability that a certain threshold will be exceeded as well as quantiles. Quantiles are very suitable to predict quantitative precipitation and do not depend an a priori defined thresholds, as is necessary for the probability forecasts. Various statistical methods are explored to transform the ensemble forecast into probabilistic predictions, either in terms of probabilities or quantiles. An enhanced framework for statistical postprocessing of quantitative precipitation quantile predictions is developed based on a Bayesian inference of quantile regression.

For a further investigation of the predictive performance of quantile forecasts, the pool of verification methods is expanded by the decomposition and graphical exploration of the quantile score. The decomposition allows to attribute changes in the predictive performance of quantile forecasts either to the reliability or the information content of a forecasting scheme. Together with the Bayesian quantile regression model, this study contributes to an enhanced framework of statistical postprocessing and probabilistic forecast verification of quantitative precipitation quantile predictions derived from mesoscale NWP ensembles.

Contents

1. Introduction

Since the beginning of numerical weather prediction (NWP), quantification of forecast uncertainty is a major desire. Uncertainty arise from the nature of numerical prediction: the assumptions about model physics, the discretization in space and time, the parameterization of subgrid-scale processes, and imperfect initial conditions. All this affects the accuracy of numerical forecasts of complex systems like the earth's atmosphere. On the other side, the chaotic nature of the atmosphere itself leads to an intrinsic uncertainty inherent in every weather forecasting system. Some weather situations (e.g. large scale flows) will always be more predictable than others, e.g. small-scale weather events like thunderstorms, hail, or wind gusts. Predictability is a measure of forecast error and defines a horizon for skillful predictions (Lorenz, 1963b). On the global scale, NWP gives skillful forecasts for about 10 days, while on the convective scale the weather is mainly predictable for several hours. However, much effort is put in the development of NWP models. The increase of computational power allows to calculate numerics on even finer spatial grids, which are capable to describe more and more detailed physical processes. Although NWP has seen great advances and has become more accurate during the last century, the quantification of forecast uncertainty is still a crucial task. More complex weather prediction models lead to more realistic weather forecasts, but do not have smaller uncertainties.

The focus of this study is on the assessment of forecast uncertainty from convective-scale NWP models. The small-scale nature of mesoscale processes leads to faster error growth and hence less predictability (Lorenz, 1969). Predictions of small-scale events therefore must be probabilistic in nature, accounting for the uncertainty which is inherent to those forecasts (Murphy, 1991). Convective-scale ensemble systems are used to obtain probabilistic guidance. The main objectives of this study are

- the evaluation of ensemble forecast performance,
- the verification of probabilistic forecasts derived from the ensemble,
- the development of ensemble postprocessing techniques in order to obtain skillful probabilistic predictions.

The evaluation is focused on precipitation, which is still one of the most difficult weather variables to predict (Ebert et al., 2003). Especially during summer, the skill of quantitative precipitation forecasts is very low (Fritsch and Carbone, 2004). Precipitation is a result of very complex, dynamical and microphysical processes and is often used to measure model performance of mesoscale NWP systems.

1.1. Convective-scale weather prediction

Convective-scale weather prediction yields a better representation of small-scale weather phenomena triggered by deep moist convection. Non-hydrostatic model dynamics and a horizontal resolution of just a few kilometers allow to simulate convective processes more explicitly. The benefit of convection-permitting NWP models is a better physical representation of mesoscale convective systems, more realistic looking weather patterns and localized intense events like heavy precipitation (Mass et al., 2002; Done et al., 2004; Schwartz et al., 2010). They do not necessary improve point specific forecasts and often suffer from positioning and timing errors. Convective-scale weather prediction models are combined with ensemble techniques in order to assess forecast uncertainty.

The assessment of forecast uncertainty does not necessarily focus on the forecast error at the end of forecast lead time. At first one is concerned about the forecast error at the beginning of the forecast, the initial time step. Forecast uncertainty starts with the definition of an initial atmospheric state, a 3-dimensional field around the globe which can never be known with certainty. In a second step, one is concerned about how these initial uncertainties will evolve during model integration using imperfect model physics. Instead of the trajectory of the deterministic atmospheric state in the phase space one is interested in the evolution of the multivariate probability distribution of the atmospheric state (Epstein, 1969). The time evolution of a probability function can be solved directly by the Liouville equation. However, solving the Liouville equation is not feasible for high-dimensional systems like the atmosphere. A pragmatic solution to the Liouville equation is the so called Monte Carlo ensemble (Leith, 1974). A Monte Carlo ensemble consists of several model integrations, starting from different initial conditions using different model physics. The ensemble of weather trajectories is an indicator of forecast uncertainty and predictability, and represents the probability of the atmosphere to be in a certain state.

Ensemble forecasts provide the user with additional information. An ensemble issues the most probable state of the atmosphere, e.g. the ensemble mean, together with its uncertainty, e.g. the ensemble spread. But ensemble predictions are only useful if they obey the principles of good forecasts (e.g. Murphy, 1993). Altogether we want to know how much confidence we can put into a forecast system. That leads us to the large field of forecast verification.

1.2. Verification and ensemble postprocessing

The verification of ensemble forecasts has mainly two branches. A verification based on the individual ensemble members specifies attributes like reliability, discriminative power, or information content. It answers the questions: Does the ensemble represent sufficient ensemble spread? Can the ensemble discriminate between different outcomes of the observations? This does not necessarily lead to a ranking of several competitive ensemble systems. The second branch is the verification of probabilistic products derived from the ensemble, like predictive distribution or density functions, but also functionals thereof (e.g. mean, quantiles, probabilities). The verification of probabilistic forecasts is based on proper scoring rules (Gneiting and Raftery, 2007; Bröcker, 2012), which can either be regarded as cost-functions which a forecaster wants to minimize, or as a reward which should be maximized. In both cases, score functions assign a value to a forecast system which allows to define a ”best” system or a ranking of systems. One has to keep in mind, that such score functions not only evaluate the ensemble but also the process used to derive the probabilistic forecast.

In this sense, statistical postprocessing is closely related to forecast verification. The translation of a set of realizations into e.g. an empirical distribution function, a mean value, or quantiles is a simple form of postprocessing. More advanced methods use a historic data set of ensemble forecasts and observations to define a statistical relationship. Regression techniques allow to link covariates from the ensemble to the expected outcome of an observed variable. Different covariates can increase the information content of an ensemble, while the statistical relationship can account for calibration and systematic biases. Statistical models are estimated such that the postprocessed forecasts optimize their respective score function, e.g. a score function which is consistent for the type of prediction (Gneiting, 2011a). The drawback of such a statistical postprocessing is that we often have to make assumptions e.g. about the distribution of a variable or about the form of the statistical relationship. The performance of statistical models strongly depends on how well these assumptions fit to the real data. However, if a suitable statistical relationship for the historic data set can be found, it can be used to make future predictions given that the forecasting system does not change. The added value of postprocessing can be expressed in terms of an improvement of the score function. A decomposition of score functions allows to attribute the improvement directly to forecast characteristics like reliability/calibration, resolution/information content, or discrimination.

1.3. Bayesian postprocessing

Statistical postprocessing is often limited in the dependence structure and complexity of statistical relationships. Bayesian models offer a more flexible and complex formulation. Fundamental in the Bayesian framework of statistical postprocessing is the treatment of unknown model parameters (e.g. regression coefficients) as random variables. Prior knowledge (i.e. expert opinion or external knowledge) about the parameters can be included into the postprocessing by appropriate prior distributions. Moreover, the hierarchical structure of a Bayesian model is suitable to describe complex structures, like spatial variations of model parameters. The drawback of Bayesian models are the high-computational costs. Numerical solutions often rely on iterative processes, which require a vast amount of computational capacities. Increasing technical resources have made Bayesian modeling more feasible during the last decades. However, the exploration of Bayesian models for numerical weather prediction application is still an active field of research.

1.4. Outline

This study was conducted in the framework of the research project ”Bayesian ensemble postprocessing”, funded by the German Meteorological Service (Deutscher Wetterdienst, DWD) within the extramural research program. The main tasks of the project was the development of ensemble postprocessing techniques tailored for precipitation forecasts derived from a convective-scale ensemble system. The project started in 2009 and used a skeleton EPS interim solution, based on the convective-scale NWP model COSMO-DE which is centered over Germany. A poor man's ensemble was constructed from the deterministic COSMO-DE model and time-shifted model runs. The objectives of the first project phase focused on different types of probabilistic predictions (e.g. predictive distributions, functionals), the translation of ensemble forecasts into probabilistic predictions, and the exploration of methods for statistical calibration. The main results are published in Bentzien and Friederichs (2012).

In the second phase of the project, the most promising methods were applied to the COSMO-DE-EPS, the first operational convective-scale ensemble prediction system. The German-focused COSMO-DE-EPS was implemented 2010 by DWD. In a pre-operational phase between December 2010 and May 2012, COSMO-DE-EPS run under operational conditions, and became operational on May 22, 2012. The data set used in this study holds forecasts from the pre-operational phase for the year 2011. The focus lies on probability and quantile forecasts derived from logistic and quantile regression. A Bayesian quantile regression model is developed and explored for a further enhancement of quantile forecasts derived from the ensemble.

Special focus was hold on the verification of the probabilistic forecasts. Both forecast types use a consistent scoring function. Probabilities are evaluated using the Brier score (Brier, 1950; Murphy, 1973). The well known decomposition into reliability, resolution and uncertainty gives more detailed insights in forecast performance than a single score value. The reliability diagram yields as a graphical representation of forecast calibration. Verification of quantile forecasts uses a score function based on the asymmetric check-loss function. Since the quantile score is a proper score function, an analog decomposition into reliability, resolution and uncertainty must exist. In Bentzien and Friederichs (2014), we have derived this decomposition in order to extend the verification framework for quantile forecasts. We now dispose over a decomposition which gives us detailed insights in the calibration of quantile forecasts, as well as a quantification of their information content. A graphical representation of reliability for quantile forecasts is explored.

Part I of this study gives a brief overview about numerical weather prediction and ensemble generation. Chapter 4 is dedicated to ensemble forecast verification, and introduces the newly developed extended framework for quantile verification. Part II comprises the statistical methods for ensemble postprocessing. The main results for the poor man's ensemble are given in chapter 7, which is a summary of the key findings of Bentzien and Friederichs (2012). Chapter 8 presents the results for COSMO-DE-EPS. In Part III the Bayesian quantile regression model is explored. This study is closed in Part IV by a summary and conclusion.

Part I.

Numerical weather prediction and verification

2. Mesoscale numerical weather prediction

Modern weather forecasting describes the atmospheric state and motion by a set of mathematical equations. The equations follow the physical laws of fluid dynamics and thermodynamics, e.g. the primitive equations. The initial atmospheric state is derived from irregular spaced observations on the one hand, as well as satellite or radar data on the other hand. Data assimilation methods are required to obtain the best available initial state to start the model integration. Numerical weather prediction (NWP) models solve the set of mathematical equations on a discrete 3-dimensional grid defined around the globe. The effect of subgrid-scale processes (e.g. clouds, precipitation, solar radiation, turbulence, soil and vegetation) on the atmospheric state must be incorporated by empirical parameterizations, which play an important role in the setup of a NWP model.

Since the beginning of operational weather forecasts in the 1950s, NWP models have seen great advances (Harper et al., 2007). With increasing computer powers, the horizontal resolution of global NWP models lies between 30-50 km. In contrast to global models, limited-area models cover only a limited part of the earth thereby allowing for even higher spatial and temporal resolutions. They account for more complex physical processes which are treated explicitly instead of parameterizations and represent surface conditions and orography in more detail. However, limited area models strongly depend on lateral boundary conditions which must be obtained from a driving host model (e.g. global model).

A major task of meteorological services is the prediction and warning of weather that has the potential for hazardous impacts, denoted as high-impact weather. High-impact weather in western Europe is related to strong mean winds, severe gusts, and heavy precipitation (Craig et al., 2010). Especially during summer, these weather situations are often related to moist convective processes. In order to resolve such mesoscale processes explicitly, high-resolution models (HRM) with a horizontal grid spacing of less then 10 km are developed. A prerequisite for NWP on these spatial scales is a non-hydrostatic formulation of the model dynamics. Today, many meteorological services use HRMs for operational forecasts and weather warnings for their specific area of responsibility (e.g. Skamarock and Klemp, 2008; Saito et al., 2006; Staniforth and Wood, 2008; Baldauf et al., 2011b; Seity et al., 2011).

Despite all advances in HRM, precipitation is still one of the major challenges in NWP. Due to its high temporal and spatial variability, it is one of the most difficult meteorological variables to predict (Ebert et al., 2003). Precipitation can be induced by many processes on larger and smaller scales (e. g. convection, convergence, orography), all of which have to be represented within the model. Moreover, a complex chain of microphysical processes is necessary to describe the building and life cycle of hydrometeors. Processes involved in precipitation range over all scales from microphysics to the mesoscale and the larger scale. The skill of precipitation forecasts critically depends on an accurate prediction of the whole atmospheric state, and thus is often used to measure model performance in NWP (Ebert et al., 2003).

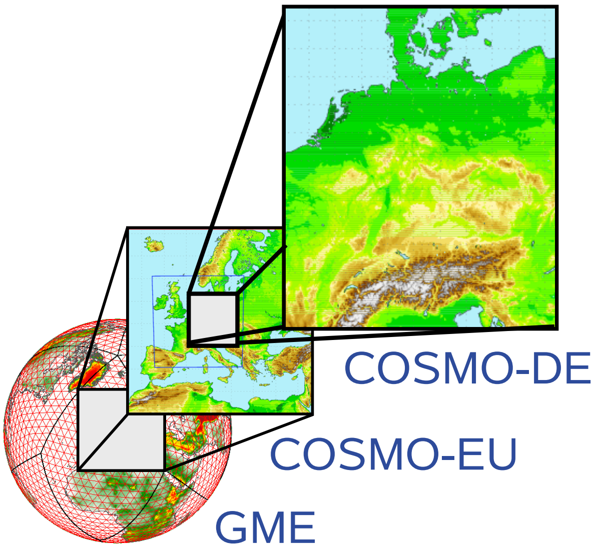


Figure 2.1.: Illustration of the operational model chain of DWD (Source: DWD).

The focus of this study is on precipitation forecasts for Germany, derived from the operational HRM of the German Meteorological Service (DWD). The operational model chain of DWD consists of the global model GME with a horizontal resolution of 30 km, the regional model COSMO-EU (7 km) which is centered over central Europe and is nested into the GME, and the high-resolution model COSMO-DE (2.8 km) which retrieves hourly boundary conditions from COSMO-EU. The model domain of COSMO-DE covers the area of Germany, parts of the neighboring countries and most of the Alps region. The model chain is illustrated in Fig. 2.1. COSMO-EU and COSMO-DE are both applications of the flexible COSMO model which is developed and maintained by the Consortium for Small-scale Modeling. The models are particularly designed to predict high-impact weather in Europe and Germany. The following section gives a general overview of the COSMO model. Section 2.2 describes the operational setup of COSMO-DE. Note that the forecast system is subject to steady changes which are documented on the webpage http://www.dwd.de/modellierung (see Changes in the NWP-system of DWD).

2.1. The COSMO model

The COSMO model is a non-hydrostatic limited-area NWP model for operational forecasts and research applications. It is developed and maintained by the members of the consortium, which comprises the national weather services of Germany, Swiss, Italy, Greece, Poland, Romania, and Russia. Other academic institutes and regional and military services are also participating. Detailed information about COSMO and its various applications, including a large number of documentations, can be found on the webpage http://www.cosmo-model.org. The following overview of the COSMO model is taken from Schättler et al. (2013).

The main features of the COSMO model are the non-hydrostatic model dynamics which are based on the primitive hydro-thermodynamical equations. They describe a full compressible flow in a moist atmosphere on a rotated latitude-longitude grid with generalized terrain-following vertical coordinates. The prognostic variables are wind, pressure disturbances, temperature, specific humidity, and cloud water content, with options for a prognostic treatment of cloud ice content and precipitation in form of rain, snow, and graupel. Numerical time integration is based on variants of two time-level Runge-Kutta or three time-level leapfrog schemes.

The non-hydrostatic model formulation allows for simulations on a broad range of spatial scales. The focus lies on the meso-β and meso-γ scale. A horizontal resolution of 10 km or less leads to a better representation of near-surface weather conditions like clouds, fog, frontal precipitation and orographically and thermally forced wind systems. On spatial scales of 1-3 km deep moist convection should be explicitly resolved by the model dynamics. That allows for a direct simulation of small-scale severe weather events like thunderstorms, squall-lines, mesoscale convective systems and winter storms.

The COSMO model provides a comprehensive package of physical parameterizations to cover different applications, spatial and temporal scales. The package includes parameterizations for moist convection (Tiedtke, 1989; Kain and Fritsch, 1993), radiation (δ two-stream radiation scheme after Ritter and Geleyn, 1992), subgrid-scale clouds, subgrid-scale turbulence, amongst others. Precipitation is parameterized by a Kessler-type bulk formulation with options for cloud ice and graupel. The microphysical scheme also allows for a prognostic treatment of precipitation in forms of rain, snow and graupel. COSMO includes variants of a multilayer soil model, a fresh-water lake parameterization and a sea ice scheme.

Initial and lateral boundary conditions are generally provided by coarser gridded models, like the global model GME or a COSMO model with lower resolution. COSMO uses a continuous 4-dimensional data assimilation scheme based on observation nudging (Newtonian relaxation). Observations are taken from radiosondes (wind, temperature, humidity), aircrafts (wind, temperature), wind profiler, and surface data from observational sites (SYNOP), ships, and buoys (pressure, wind, humidity). In order to provide a full data assimilation cycle, COSMO has an optional soil moisture analysis to improve the 2m-temperature, a sea surface temperature analysis, and a snow depth analysis.

The COSMO model is a very flexible model and the actual setup depends on the application and the availability of observational data. It can be used for short-range weather predictions (e.g. the operational COSMO-EU or COSMO-DE) as well as for long-term climate projections (COSMO-CLM; Rockel et al., 2008). Special versions of the COSMO model are developed by academic researches, e.g. for aerosols and reactive tracers (COSMO-ART; Vogel et al., 2009), or fog forecasting (COSMO-FOG; Masbou, 2008). Most recently, a regional reanalysis system for Europe based on the COSMO model has been setup by the Climate Monitoring Branch of the Hans Ertel Center for Weather Research (Bollmeyer et al., 2015).

2.2. The COSMO-DE forecasting system

COSMO-DE is at the high-resolution end of the DWD model chain and in operational use since April 2007. The model setup is described in Baldauf et al. (2011a,b). The model grid covers Germany and parts of the neighboring countries with a horizontal grid spacing of 0.025° ($\sim$ 2.8 km) and a total of 421 $\times$ 461 gridpoints ($\sim$ 1200 $\times$ 1300 km^2). COSMO-DE uses 50 vertical layers in generalized terrain-following height coordinates. The levels range between 10 m and 22 km above sea level. The dynamical core of COSMO-DE uses a two time-level split-explicit

Runge-Kutta variant. The advection of scalar fields is based on a three dimensional extension of the Bott scheme (Bott, 1989).

Due to the horizontal grid spacing of 2.8 km deep moist convection should be explicitly resolved by the model dynamics. Only shallow convection is parameterized by a reduced Tiedtke scheme. Prognostic precipitation in forms of rain, snow, and graupel is modeled within a three-category ice scheme described in Reinhardt and Seifert (2006). Subgrid-scale turbulence is parameterized according to the level-2.5 scheme of Mellor and Yamada (1974).

A key feature of COSMO-DE is the assimilation of radar derived rain rates through latent heat nudging (LHN). The 3-dimensional thermodynamical field is adjusted such that the modeled precipitation rates better match the observed radar field (Stephan et al., 2008). LHN initializes convective events at the beginning of the simulation thereby improving forecasts during the first forecast hours and leading to a short model spin-up time. Bierdel et al. (2012) showed, that COSMO-DE produces horizontal wind fields that represent a realistic energy spectrum on the atmospheric mesoscale down to 12-15 km which indicate an effective resolution of 4 to 5 of the horizontal grid spacing.

COSMO-DE retrieves hourly boundary conditions from the coarser gridded COSMO-EU. The model domain of COSMO-EU covers western Europe with a horizontal grid-spacing of 7 km. In COSMO-EU, deep moist convection is fully parameterized by the Tiedtke scheme. The microphysical scheme considers a prognostic treatment of cloud ice and precipitation in form of rain and snow. However, a LHN scheme is currently not applied to the operational COSMO-EU. COSMO-DE and COSMO-EU both use a multilayer soil model (TERRA-ML) and a freshwater lake parameterization scheme (FLake). A sea-ice scheme is only applied to COSMO-EU. While the update cycle for COSMO-EU starts every 6 hours for a forecast lead time of 2-3 days, COSMO-DE is initialized every 3 hours and produces forecasts for the next 21 hours.

3. Mesoscale ensemble prediction

Forecasts of deterministic NWP models as described in Section 2 start from a single set of initial conditions and predict the future state of the atmosphere. Such forecasts can never be certain. The initial state of the atmosphere is always known within a certain margin of error and hence affects forecast accuracy. Moreover, imperfect model dynamics and unresolved scales contribute to the forecast error. The demand for ensemble prediction and probabilistic forecasting arose already at the very beginning of numerical weather prediction by Eady (1949) and Thompson (1957). Due to the uncertain character of initial conditions, the ”answer” in terms of numerical forecasts must also be stated in terms of probabilities (Eady, 1951). The idea was further motivated by the research of Edward Lorenz in the 1960s. Predictability is a measure of forecast error at a certain time step and provides additional information about the confidence of a deterministic forecast (Lorenz, 1963b). It defines a horizon for skillful predictions from a NWP model. The quantification of model uncertainty and hence predictability is a central part in NWP.

3.1. Overview of operational ensemble prediction

The initial state of the atmospheric system can be considered as a single point in a phase space, where NWP describes the evolution of the system along a certain weather trajectory. However, small perturbations in the initial state lead to varying trajectories. Such forecast errors grow with forecast lead time, and the future state of the atmosphere becomes uncertain or unpredictable after some integration time (Lorenz, 1963a). In order to extend the range of skillful forecasts, Lorenz (1965) proposed to use an ensemble of possible initial states instead of a single estimate. Variations in the initial conditions should resemble the errors in observations. A model integration is started from each of the initial conditions, leading to an ensemble of future states. Probabilistic guidance in terms of the probability of an event or the mean and variance of a certain weather quantity can be achieved. The skill of probabilistic forecasts at longer time scales overcomes the limit of deterministic predictions. Ensembles of this kind are called Monte Carlo ensembles.

A theoretical concept of Monte Carlo ensembles is given by Epstein (1969). Instead of calculating several model runs as an approximation to the forecast distribution, the evolution of the probability density function of the atmospheric state in phase space can be predicted directly. This is done by solving the Liouville equation, the continuity equation for probabilities. However, for high-dimensional problems like NWP a solution of the Liouville equation is computational unattainable. Instead, Monte Carlo forecasts can be regarded as a feasible approximation to stochastic dynamic predictions (Leith, 1974), and became the common choice of operational ensemble forecasting. Moreover, Monte Carlo ensembles can easily be extended to represent model uncertainties, e.g. by combining different NWP models (multi-model ensembles; see also Palmer et al., 2005) or by using different setups of the same model (multi-physics ensemble). A historical review of ensemble methods is given in Lewis (2005, 2014).

The generation of meaningful initial condition perturbations is a complex task. Kalnay et al. (2006) show the close relation to data assimilation and give a comprehensive overview of the variety of methods which are developed. Following Buizza et al. (2005), the performance of ensemble forecasts strongly depends on the data assimilation scheme to create the initial conditions and the numerical model to generate the forecasts. Moreover, a successful ensemble should also represent model-related uncertainties. The generation of an appropriate ensemble design is still a field of active research, and there is no general solution to define a perfect ensemble setup.

3.1.1. Global ensemble prediction

After decades of active research, ensemble predictions on the global scale became routinely available in the mid-nineties by the European Center for Medium-Range Weather Forecasts (ECMWF; Molteni et al., 1996), the National Center for Environmental Predictions (NCEP; Tracton and Kalnay, 1993), and the Canadian Meteorological Center (CMC; Pellerin et al., 2003). Several competing schemes of initial perturbation generation were developed. The ECMWF EPS uses singular vectors (Buizza and Palmer, 1995; Barkmeijer et al., 1999) to create 32 ensemble members, and later 50 members (Buizza et al., 1998). Toth and Kalnay (1993) introduced the breeding vectors, which are used by the NCEP Global Ensemble Forecast System (GEFS). Since 2006, 20 perturbed initial conditions are created by an extended version of breeding vectors using the ensemble transform and rescaling (Wei et al., 2008). The CMC ensemble is based on perturbations from data assimilation cycles described in Houtekamer et al. (1996). Since 2005, the CMC EPS uses the ensemble Kalman filter (Houtekamer et al., 2009).

Model uncertainty was implemented into the ECMW EPS in 1998 by a stochastic parameterization scheme (Buizza et al., 1999; Palmer et al., 2005). The NCEP GEFS implemented a stochastic total tendency perturbation scheme in 2010 (Hou et al., 2010). A multi-model approach is used by the CMC EPS. Two different global models are used to drive 8 ensemble members, respectively. Meanwhile other meteorological services follow the ensemble approach, and some of these global ensemble systems are part of the THORPEX Interactive Grand Global Ensemble (TIGGE; Park et al., 2008).

3.1.2. Regional ensemble prediction

Additional challenges arise for regional ensembles based on limited area models. The generation of initial perturbations is not straight forward (e.g. nonlinear error growth, faster error growth on smaller scales). Model errors have a larger impact on regional ensembles. Moreover, the perturbation of lateral boundary conditions has to be considered. Eckel and Mass (2005) and their references give a comprehensive overview about the challenges of short-range ensemble forecasting. A pragmatic approach is the nesting of a limited area model into an ensemble or set of different global or coarser grid models. The first operational short-range ensemble forecasting systems became available in the first years of the 21th century, e.g. for

North-America (NCEP SREF), the Pacific North-West (UWME; Grimit and Mass, 2002), and Europe (COSMO-LEPS; Marsigli et al., 2005). A more detailed overview is given in Bowler et al. (2008).

3.1.3. Convective-scale ensemble prediction

The first mesoscale ensemble system with a convection-permitting NWP model was implemented by DWD in 2010. The COSMO-DE-EPS is a multi-analysis and multi-physics ensemble. Initial and boundary conditions are obtained from different global models, while model uncertainty is accounted by different formulations of model physics. A detailed description of COSMO-DE-EPS follows in Section 3.2.2. The UK MetOffice also implemented a convection permitting ensemble (MOGREPS UK), which became operational in 2012 (Golding et al., 2014). MOGREPS UK is a downscaling ensemble with a horizontal resolution of 2.2 km, covering the area of UK and surroundings. The 12 members of MOGREPS UK are driven by initial and lateral boundary conditions from the regional (and later from the global) ensemble MOGREPS R (MOGREPS G). Currently under development is the AROME EPS by Météo France (Vié et al., 2011). The generation of convective-permitting ensembles is still a field of active research, and a brief overview is given in Peralta et al. (2012) and Vié et al. (2011).

3.2. Ensembles based on the COSMO-DE forecasting system

3.2.1. COSMO-DE lagged average forecasts

Before Monte Carlo ensembles became routinely available for NWP, Hoffman and Kalnay (1983) proposed the method of lagged average forecasts (LAF) as pragmatic alternative to the computational expensive Monte Carlo ensemble. Forecasts from successive initialization times are combined to an ensemble forecast for a common verification period. The LAF ensemble comes at no additional costs, since the different members are already provided by the operational update cycle of NWP. Several studies show the benefit of LAF in short-range weather prediction, e.g. Lu et al. (2007); Mittermaier (2007); Yuan et al. (2009). However, LAF is a pragmatic approach to ensemble generation. It ignores model errors and therefore does not represent all sources of uncertainty.

In Bentzien and Friederichs (2012) we construct a LAF ensemble from the rapidly updated COSMO-DE forecasting system. COSMO-DE is initialized every three hours and simulates a period of 21 hours ahead. Four successively started forecasts describe a joint verification period of at most 12 hours. The combination of model runs is illustrated in Fig. 3.1. Each forecast is initialized with different initial and boundary conditions. Thus the LAF can be considered as an multi-analysis ensemble. Note that the different initial conditions derived from the time-lagged members are not independent since they are obtained from the previous forecast cycle, modified by observations. However, COSMO-DE-LAF serves as a benchmark for the more sophisticated ensemble prediction system COSMO-DE-EPS.

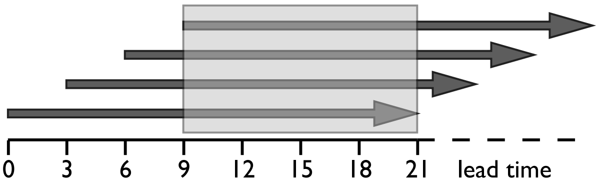
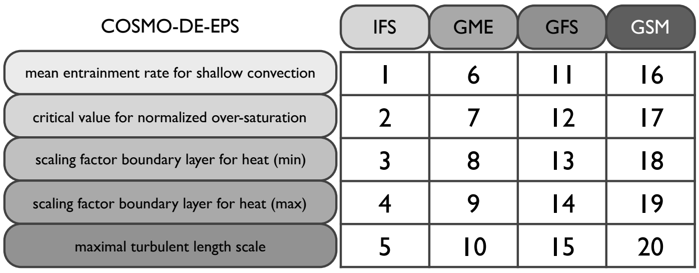


Figure 3.1.: Illustration of the COSMO-DE-LAF forecast for a common verification period of at most 12 hours. Forecasts are initialized every 3 hours. The lead time of the forecasts is 0 - 12h, 3 - 15h, 6 - 18h, and 9 - 21h, respectively.

3.2.2. COSMO-DE ensemble prediction system

COSMO-DE-EPS is developed by DWD as a multi-physics and multi-analysis ensemble based on the COSMO-DE forecasting system (Gebhardt et al., 2011; Peralta et al., 2012). It runs pre-operational at DWD since December 2010, and got operational on May 22nd, 2012. However, the ensemble design and the operational setup did not change, and in this study we focus on pre-operational forecasts for the year 2011. The 20 members of COSMO-DE-EPS differ from COSMO-DE with respect to initial and boundary conditions and physical parameterizations. Boundary conditions are provided by four global models (IFS from ECMWF, GME from DWD, GFS from NCEP, and GSM from the Japanese Meteorological Agency). A so-called boundary conditions EPS (BCEPS) is constructed using the coarser-grid model COSMO-EU. The ensemble model chain emphasizes forecasts from the four global models, which are used to drive four different members from BCEPS. Every group of 5 members from COSMO-DE-EPS are nested into one member of COSMO-BCEPS. In order to preserve the benefit of latent heat nudging, initial conditions for the different COSMO-DE-EPS members are obtained by slightly modifying the original COSMO-DE analyses with differences from the respective COSMO-BCEPS member and COSMO-EU (Theis et al., 2012). Therefore, each member of COSMO-DE-EPS receives boundary and initial conditions from one of the four BCEPS members. Model physics are disturbed by parameter variation of parameterizations for microphysics, turbulence, and shallow convection. Perturbations are applied to the turbulent length scale, the scaling factor for the thickness of the laminar boundary layer for heat, the critical value for normalized over-saturation, and the mean entrainment rate for shallow convection. More information about the meaning of the perturbed parameters can be found in Baldauf et al. (2011a) and Schättler et al. (2013). The variation of parameters is kept constant over time and for each ensemble member. Altogether we have five different model physics configuration, each of them driven by four different initial and boundary conditions from the BCEPS. The ensemble setup for COSMO-DE-EPS is illustrated in Fig. 3.2.

While the LAF method was originally developed as pragmatic approach to ensemble prediction, it nowadays becomes popular again as a useful tool to extend existing ensemble systems which are often restricted in member size due to computational limits. In this study, the LAF approach is used to create a time-lagged ensemble from COSMO-DE-EPS. Analogously to COSMO-DE, the COSMO-DE-EPS has a rapid update cycle of 3 hours and simulates 21-hour forecasts. For each of the 20 members, four time-lagged forecasts are derived according to the scheme in Fig. 3.1. All forecasts are equally weighted. This is in accordance to Ben Bouallègue et al. (2013), who used a combination of three time-lagged model runs to enlarge COSMO-DE-EPS.



COSMO-DE-EPS	IFS	GME	GFS	GSM
mean entrainment rate for shallow convection	1	6	11	16
critical value for normalized over-saturation	2	7	12	17
scaling factor boundary layer for heat (min)	3	8	13	18
scaling factor boundary layer for heat (max)	4	9	14	19
maximal turbulent length scale	5	10	15	20

Figure 3.2.: Illustration of the COSMO-DE-EPS setup. The 20 members are driven by different global models (initial and boundary conditions) and perturbed physics.

The time-lagged COSMO-DE (COSMO-DE-TLE in the following) consists of 80 members, and allows inference of the ensemble size and the generation of ensemble spread. To this end, another 20-member ensemble COSMO-DE-TLE$_{sub}$ is constructed, which consists of 5 members of the COSMO-DE-EPS and their respective time-lagged forecasts[1]. Comparison of these ensembles will show the contribution of the time-lagged members to the ensemble spread.

[1] The members 1,7,13,15,19 from Fig. 3.2 are chosen in order to have one member from each physical perturbation and each global model.

4. Verification of ensemble forecasts

Forecast verification in general is based on the inference of the joint distribution of forecasts and observations. The joint distribution describes the degree of association between predictions for future quantities and the events that have materialized. It is an a posteriori assessment of forecast performance. In the simple case of binary forecasts and observations, the joint distribution can be represented by a contingency table. It shows the relative frequencies of possible combinations of predicted and observed events. The factorization of the joint distribution into a conditional and marginal distribution allows to assess different attributes of forecast performance. This is known as calibration-refinement or likelihood-base rate factorization and described in detail by Murphy and Winkler (1987). Table 4.1 provides a list with certain characteristics of forecast performance which might be of interest for users. A comprehensive overview of traditional forecast verification methods based on this distribution-oriented approach is given in Wilks (2006b), Chapter 7, and Jolliffe and Stephenson (2012). However, most of the traditional methods focus on the verification of deterministic forecasts, thereby comparing a single-valued forecast to a single-valued observation.

The verification of ensemble forecasts faces new challenges. We have multiple forecasts on the one side, which, in the ideal case, represent independent realizations from the distribution of the observations. On the other side we still have a single-valued observation. Hence we cannot observe what we want to predict: the distribution of future weather quantities. Forecast and verification strategies are manifold. Ensemble forecasts are at first finite sets of deterministic forecast realizations. An evaluation based on the individual members measures attributes of forecast performance. Typical methods are the rank histogram to check ensemble consistency, the discrimination score or the spread-skill relationship to assess the information content of the ensemble. A brief overview of such methods is given in Weigel (2012). However, a set of realizations is in general not a useful forecast for potential users, e.g. decision makers or economists. A typical forecast strategy is to transform the ensemble into a probabilistic prediction, e.g. a predictive distribution or statistical functionals (e.g. moments, quantiles, probabilities). The verification of probabilistic forecasts relies on proper score functions, e.g. cost functions which a forecaster aims to minimize. Probabilistic forecast verification is described in detail by Gneiting and Raftery (2007) and Bröcker (2012), amongst others. One has to keep in mind, that, in the case of ensemble forecasting, the score evaluates not only the ensemble system but also the process used to derive the probabilistic forecast. In this sense, verification is closely related to postprocessing of ensemble forecasts. Probabilistic forecasts derived from the ensemble can be optimized by minimizing the corresponding score function. More details will be given in Chapter 5.

Proper scores are a quantitative measure of forecast accuracy. They assign a single value to a forecast system which allows to define a ”best” system or a ranking of systems. To get more detailed insights, decompositions of proper scores are proposed. Of particular interest are

Table 4.1.: Glossary of forecast attributes which are of interest in evaluating forecast performance. Descriptions are taken from Murphy (1993), Table 2, and Wilks (2006b), Section 7.1.3.

The joint distribution of an observation y and a forecast f can be factorized into a conditional and marginal distribution following Murphy and Winkler (1987)

$$\underbrace{p(y,f)}_{\text{joint distribution}} = \underbrace{p(y|f)\,p(f)}_{\text{calibration-refinement}} = \underbrace{p(f|y)\,p(y)}_{\text{likelihood-base rate}}$$

Association	$p(y,f)$	(linear) relationship between individual pairs of forecasts and observation
Accuracy	$p(y,f)$	correspondence between individual pairs of forecasts and observation; generally assessed by score functions
Skill	$p(y,f)$	accuracy of forecasts relative to a reference forecast; generally measured by skill scores
Bias	$p(f)$, $p(y)$	*unconditional bias* or *systematic bias* correspondence between the mean of the forecasts and the mean of the observations
Reliability	$p(y|f)\,p(f)$	*calibration* or *conditional bias* correspondence between conditional mean observations and conditioning forecasts
Resolution	$p(y|f)\,p(f)$	difference between conditional mean observations (conditional on the forecasts) and the unconditional mean of the observations
Discrimination	$p(f|y)\,p(y)$	converse of resolution; difference between conditional mean forecasts (conditional on the observations) and the unconditional mean of the forecasts
Sharpness	$p(f)$	*refinement* variability of forecasts; sharpness and resolution become identical if forecasts are completely reliable
Uncertainty	$p(y)$	variability of observations

the forecast attributes reliability and resolution. Their estimation is related to the calibration-refinement factorization proposed by Murphy and Winkler (1987). A decomposition has already been derived for several scores, e.g. the continuous ranked probability score and the Brier score. In Bentzien and Friederichs (2014), we derive a similar decomposition of the quantile score and explore a graphical representation of quantile reliability. With this decomposition, we contribute to an extended framework for quantile forecasts.

The remainder of this chapter is organized as follows: Section 4.1 focus on ensemble verification using the rank histogram. An introduction to probabilistic forecast verification is given in Section 4.2. The section describes the concept of proper scores and elucidates the general decomposition of score functions to assess different attributes of forecast performance. Section 4.3 presents methods for the estimation of scores, with a special focus on the calibration-refinement factorization.

4.1. Rank statistics and the beta score

A first evaluation of statistical consistency between the ensemble and the verifying observations is commonly done by the analysis rank histogram (e.g. Anderson, 1996; Hamill and Colucci, 1997). If the ensemble members represent mutual independent realizations from a perfect predictive distribution (i.e. a distribution that corresponds to the best forecaster's estimate), then the ranks of the observations within the ensemble are uniformly distributed. A generalization of the rank histogram which applies to predictive distribution functions either empirical or parametric is the probability integral transforms (PIT, Gneiting et al., 2007). Consider an ensemble of forecasts $E_1, ..., E_M$ when y is the event that materializes. If F_P is the predictive distribution function based on $E_1, ..., E_M$, the probability integral transform is given by PIT $= F_P(y)$. For a perfect or ideal ensemble, the PIT values are uniformly distributed. Deviations from the uniform distribution can be used to identify deficiencies of the ensemble forecasting system. They are usually displayed graphically by a histogram of the PIT values. A flat histogram indicates statistical consistency between the ensemble and the verifying observations. A skewed distribution of PIT values indicates a bias in the ensemble mean. If the histogram exhibits a bulb (u-shaped) form, this points to an over (under) representation of ensemble spread. The observations are too frequently in the middle (outside) of the ensemble forecast range. Note that if the verifying dataset contains aggregations over a large spatial or temporal domain, deficiencies can be averaged out (Hamill, 2001).

Keller and Hense (2011) propose the beta score (β_S) and beta bias (β_B) to quantitatively evaluate the PIT histograms. A beta distribution which is determined by two parameters α, β is fitted to the histogram of PIT values. Beta score and beta bias are then calculated as

$$\beta_S = 1 - \sqrt{\frac{1}{\alpha \cdot \beta}}\,,$$
$$\beta_B = \beta - \alpha\,.$$

For a perfectly flat histogram, the beta score equals zero. The ensemble spread is underestimated (overestimated) for a negative (positive) β_S. A beta bias greater (smaller) than zero indicates a bias towards higher (lower) values (L- or J-shaped histogram).

4.2. Probabilistic forecast verification

We now turn to probabilistic forecast verification. Consider again an ensemble forecast with a finite set of realizations, and a probabilistic forecast f derived from the ensemble, e.g. a predictive distribution F_P. Here, f can also take the form of statistical functionals $T[F_P]$ which can be understood as point forecasts of the predictive distribution (Gneiting, 2011a). Typical functionals are the mean $E[F_P]$, the variance, or quantiles. A *score function* assigns a real value to individual pairs of forecast and observation $S(f, y)$. Table 4.2 shows some score functions applicable to probabilistic predictions. The *expected score* is now the expectation of $S(f, y)$ with respect to the joint distribution $p(f, y)$. Thus, the expected score is a measure of forecast accuracy. Smaller scores indicate a better agreement between the probabilistic predictions and the events that materializes.

Table 4.2.: Consistent score functions for probabilistic forecasts. The predictive distribution or density is denoted by $F(t)$ or $f(t)$. Forecasts in terms of statistical functionals are denoted by x. Observations are continuous $y \in \mathcal{R}$, while $\mathcal{R}$ can be the real line or any interval on the real line, e.g. the positive half axis. Probability forecasts are taken as probabilities for the excess of a certain threshold u. The abbreviations are: CRPS – continuous ranked probability score, LS – logarithmic score, MSE – mean squared error, MAE – mean absolute error, QS – quantile score, BS – Brier score.

forecast type		score function	
cumulative distribution	F(t)	CRPS	$\int(F(t) - H(t-y))^2 dt$
probability density	f(t)	LS	$-\log(f(y))$
mean	$x = E[F(t)]$	MSE	$(y-x)^2$
median	$x = F^{-1}(0.5)$	MAE	$\|y-x\|^2$
τ-quantile, $\tau \in [0,1]$	$x = F^{-1}(\tau)$	QS	$\rho_\tau(y-x)$
probability, $u \in \mathcal{R}$	$x = 1 - F(t=u)$	BS	$(x - \mathbb{I}(y>t))^2$

An important property of probabilistic forecast verification is the propriety of the score function (Murphy, 1973; Gneiting and Raftery, 2007). A score function is (strictly) proper if the expected score is minimized if (and only if) the forecaster's best judgment is issued as forecast. Only proper score functions guarantee honesty and prevent hedging. There exists a wide range of proper score functions, and their application depends on the kind of probabilistic forecast that is issued. In this sense, Gneiting (2011a) demands that score functions must be carefully matched with the type of probabilistic prediction. All score functions listed in Table 4.2 are proper and consistent for the given functional. We will concentrate in the following on the continuous ranked probability score, the Brier score and the quantile score, which all have a close relationship.

4.2.1. Proper score functions

We consider in the following continuous observations $y \in \mathcal{R}$, while $\mathcal{R}$ can be the real line or any interval on the real line, e.g. the positive half axis. Forecasts issued in terms of a predictive distribution function F_P are commonly verified by the continuous ranked probability score (CRPS, Matheson and Winkler, 1976; Hersbach, 2000)

$$S_{CRP}(F_P, y) = \int_{\mathcal{R}} \left[F_P(t) - H(t-y)\right]^2 dt\,. \tag{4.1}$$

Here, $H(t-y)$ denotes the Heaviside step function. The predictive distribution F_P can either be obtained as empirical distribution of the ensemble members, or from statistical postprocessing with the ensemble forecasts as covariates. For a deterministic forecast, the CRPS reduces to the mean absolute error.

The integral in eq. (4.1) averages the quadratic loss $(F_P(t) - H(t-y))^2$ over the whole range of forecast values $t \in \mathcal{R}$. Deficiencies in different parts of the distribution function may remain undetected by the CRPS. An evaluation with respect to certain thresholds or probability levels is highly recommended (Gneiting and Ranjan, 2011). In this sense, we focus here on two other proper scoring rules which are widely used in probabilistic forecast verification and are closely related to the CRPS, namely the Brier score (BS) and the quantile score (QS).

The BS is used to assess the predictive performance of probability forecasts for a dichotomous event. In the context of a continuous predictand, a probability forecast is defined as the probability that a certain threshold $u \in \mathcal{R}$ will be exceeded. In terms of a predictive distribution this probability is given by $p_u = 1 - F_P(u)$. However, p_u can also be estimated as the expectation of a Bernoulli distribution derived from postprocessing. The BS is the squared difference between the forecasts $p_u \in [0,1]$ and observations $\{0,1\}$ and is given by (Brier, 1950)

$$S_B(p_u, y) = (p_u - \mathbb{I}(y > u))^2 \,. \tag{4.2}$$

Here, $\mathbb{I}$ is an indicator function which is set to 1 if the condition in brackets is true and zero otherwise.

Another representation of the predictive distribution is given by its inverse, the quantile function. Quantile forecasts are derived from the predictive distribution as $q_\tau = F_P^{-1}(\tau)$ for the probability levels $\tau \in [0,1]$. However, q_τ can also be estimated via postprocessing, e.g. quantile regression. The verification of quantile forecasts is done using the QS (e.g. Koenker and Machado, 1999; Gneiting and Raftery, 2007; Friederichs and Hense, 2007)

$$S_Q(q_\tau, y) = \rho_\tau(y - q_\tau) = \begin{cases} \mid y - q_\tau \mid \tau & \text{if } y \geq q_\tau \,, \\ \mid y - q_\tau \mid (1-\tau) & \text{if } y < q_\tau \,. \end{cases} \tag{4.3}$$

Here, $\rho_\tau(.)$ is the so called check loss function. The check loss is the absolute error between observations and quantile forecasts, weighted with τ if the quantile forecast does not exceed the observations and weighted with $(1-\tau)$ otherwise. The QS is minimized if q_τ is the "true" quantile of y. For more information about the check loss function and its relation to quantiles, the reader is referred to Koenker (2005) (pp. 5-7) and Gneiting (2011b).

Both the BS and QS generalize to the CRPS by the integral of the BS over all thresholds u or the integral of the QS over all probability levels τ

$$S_{CRP} = \int_{\mathcal{R}} S_B(p_u, y) du = 2 \int_0^1 S_Q(q_\tau, y) d\tau \,.$$

The second equality is based on the work of Laio and Tamea (2007). The three representations of the CRPS are illustrated graphically in Fig. 4.1. In its original representation (eq. 4.1), the CRPS is the square of the gray shaded error in the left panel, which is the difference between the predictive distribution F_P and the Heaviside function evaluated at y. The BS for a certain threshold is the square of the distance between a point of the curve $1 - F_P$ and 0 for $y \leq t$ and 1 for $y > t$. The distances are shown by the vertical blue lines in the middle panel. Integrated over all possible thresholds, this results in the same representation as the left panel. In the QS representation (right panel), the CRPS is obtained as overlapping squares. Each square

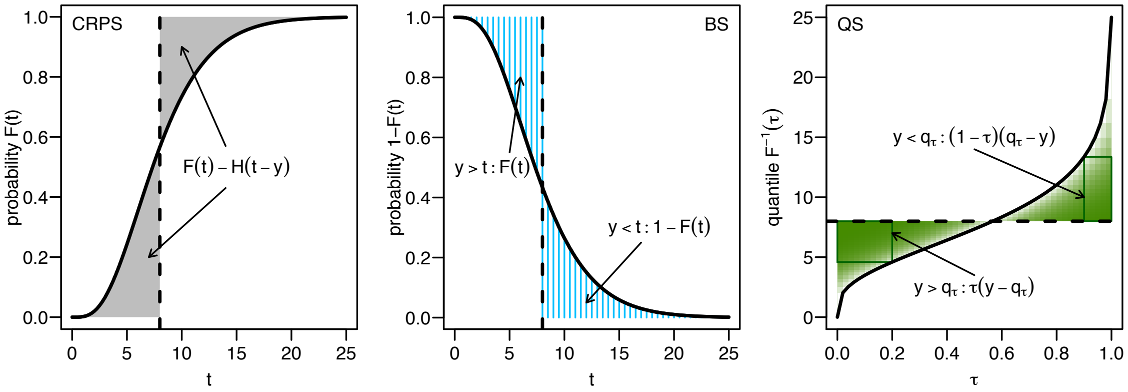


Figure 4.1.: Illustration of the CRPS and its relation to the Brier score and quantile score. The solid line shows the predictive distribution (or a transformation thereof), and the dashed vertical line the verifying observation.

represents the QS for a certain τ, and is given by the distance between the observation and the quantile curve $|y - F^{-1}(\tau)|$ times τ if $y > F^{-1}(\tau)$ and times $1 - \tau$ if $y < F^{-1}(\tau)$. The integration over all probability levels results in half of the squared area of the left panel.

4.2.2. Decomposition of proper scores

For simplification, F_P is used in the following for any type of probabilistic prediction, either in terms of a predictive distribution or as a statistical functional $T[F_P]$, which can either be a quantile $T[F_P] = F_P^{-1}$ or a probability $T[F_P] = 1 - F_P$. Proper scoring rules can generally be decomposed into the three main characteristics uncertainty, reliability, and resolution (Gneiting and Raftery, 2007; Bröcker, 2009). The decomposition is related to the calibration-refinement factorization proposed by Murphy and Winkler (1987).

The *uncertainty* is obtained from the climatological forecast $F_{\bar{Y}}$ (i.e. the marginal distribution of the verifying observations), and is given by the score function $S(F_{\bar{Y}}, y)$. It describes the variability of observations and hence is a property of the observations alone.

The *reliability*, also known as calibration, describes the statistical consistency between forecasts and observations. A forecast system is reliable, if the forecast distribution is equal to the conditional probability of the verifying observation $p(y \mid f) = p(f)$. In terms of the score function, the reliability is given by the positive score difference

$$D(F_P, F_{Y|P}) = S(F_P, y) - S(F_{Y|P}, y)\,, \tag{4.4}$$

where $F_{Y|P}$ is the conditional distribution of the observations given the forecasts. A small reliability term indicates a good agreement between F_P and $F_{Y|P}$. Note that the reliability is also denoted as divergence of the score function (e.g. Thorarinsdottir et al., 2013).

The *resolution* is related to the information content of a forecasting scheme. It describes the ability of a forecasting system to a priori distinguish between different outcomes of the observations (with respect to the climatology $F_{\bar{Y}}$). The resolution is given by the positive score difference

$$D(F_{\bar{Y}}, F_{Y|P}) = S(F_{\bar{Y}}, y) - S(F_{Y|P}, y)\,. \tag{4.5}$$

A larger resolution indicates a better discrimination of events with respect to climatology. Given the divergences (4.4) and (4.5), the score function $S(F_P, y)$ can be expressed as

$$\begin{aligned} S(F_P, y) &= S(F_{Y|P}, y) + D(F_P, F_{Y|P}) \\ &= \underbrace{S(F_{\bar{Y}}, y)}_{\text{uncertainty}} - \underbrace{D(F_{\bar{Y}}, F_{Y|P})}_{\text{resolution}} + \underbrace{D(F_P, F_{Y|P})}_{\text{reliability}}\,. \end{aligned} \tag{4.6}$$

Since the uncertainty solely depends on the verifying observations, changes in the predictive forecasting scheme will only affect the resolution and reliability part of the score.

Decompositions as in (4.6) have been derived for the CRPS (Hersbach, 2000; Candille and Talagrand, 2005) and the BS (Murphy, 1973). As part of this dissertation, the decomposition of the QS was recently developed by Bentzien and Friederichs (2014). Software routines for the calculation and decomposition of the CRPS, BS, and now also the QS, are freely available for the R statistical language (R Core Team, 2014) within the ”verification” package (Gilleland, 2014). However, the calculation of the score decomposition requires an estimation of the conditional distribution $F_{Y|P}$, and will be discussed in the next section.

4.3. Score estimation

Typically scores are calculated as the average value of a score function within a sufficiently large data set of forecast-observation pairs $\{(F_P, y)_i\}$, with $i = 1, ..., N$ the sample size. Hence, verification strongly depends on the size of the data set, spatial and temporal coverage, amongst others. Following Gneiting and Raftery (2007), the expected score is estimated empirically by the *average score* which is given by

$$\mathcal{S}(F_P) = \frac{1}{N}\sum_{i=1}^{N} S(F_{P_i}, y_i)\,. \tag{4.7}$$

A single score value is assigned to a forecasting system and can be used to compare different forecasting schemes on the same verifying data set $\{y_i\}$. A smaller score denotes a system with better predictive performance. Often it is more intuitive to compare skill scores, which measure the relative gain of a forecast system with respect to a reference forecast (e.g. climatology)

$$Skill = 1 - \frac{\mathcal{S}(F_P)}{\mathcal{S}(F_{ref})}\,.$$

Skill scores are positively oriented, where negative values indicate no predictive skill. Positive values show the percentage of improvement with respect to the reference forecasts and are bounded by 1 (100% improvement).

The evaluation of resolution and reliability requires an estimation of the conditional distribution function $F_{Y|P}$, which is also denoted as calibration function. The estimation relies on a categorization of forecast values. The data set is divided into groups or subsamples of similar forecast values. Each subsample is described by a discretized forecast value $F_P^{(k)}$, with $k = 1, ..., K$ the number of subsamples. The conditional probability $F_{Y|P}^{(k)} = F(y \mid F_P = F_P^{(k)})$ (or the respective statistical functional $T[F_{Y|P}^{(k)}]$) is calculated from the respective observations in subsample k. Note that for a statistic meaningful evaluation, each subsample must be sufficiently represented by the data set.

Given the values for $F_{Y|P}^{(k)}$ and $F_P^{(k)}$, the reliability is calculated from the divergence

$$REL = \frac{1}{N} \sum_{k=1}^{K} N_k \, D(F_P^{(k)}, F_{Y|P}^{(k)}) \,, \tag{4.8}$$

where N_k is the number of values in the subsample k and $N = \sum_k N_k$. For a perfect calibrated forecast, the reliability part is zero. Calibrated forecasts can be obtained from every forecasting system using the calibration function $F_{Y|P}$ instead of the predictive distribution F_P. But calibration alone is not a sufficient criterion of predictive performance. For example, a forecast system issuing always the climatology frequency of an event is perfectly calibrated, but cannot distinguish between different observations. The information content of a forecasting scheme is assessed by the resolution part. The resolution is calculated from the divergence

$$RES = \frac{1}{N} \sum_{k=1}^{K} N_k \, D(F_{\bar{Y}}, F_{Y|P}^{(k)}) \,. \tag{4.9}$$

A larger resolution indicates a better discrimination of different observations conditional on the forecasts. The uncertainty is calculated as the average score for the sample climatology $F_{\bar{Y}}$

$$UNC = \frac{1}{N} \sum_{i=1}^{N} S(F_{\bar{Y}}, y_i) \,. \tag{4.10}$$

4.3.1. Decomposition of the Brier score

Let p_u be the probability functional $T[F_P] = 1 - F_P(u)$. The discretized forecast values are given by $p_u^{(k)}$. The conditional observed frequencies are estimated from the observations y_i which belong to the k-th subsample $i \in \mathcal{I}_k$, as $\bar{y}_u^{(k)} = N_k^{-1} \sum_{i \in \mathcal{I}_k} \mathbb{I}_{y>u}$. The climatological forecast is given by the unconditional mean $\bar{y}_u = N^{-1} \sum_{i=1}^{N} \mathbb{I}_y > u$. Using the Brier score function (4.2) and the expressions (4.8)-(4.10), the decomposition of the BS is given by (Wilks, 2006b)

$$\begin{aligned} &\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (p_u^{(k)} - \mathbb{I}(y_i > u))^2 \\ &\quad = \underbrace{\frac{1}{N} \sum_{k=1}^{K} N_k \, (p_u^{(k)} - \bar{y}_u^{(k)})^2}_{\text{reliability}} - \underbrace{\frac{1}{N} \sum_{k=1}^{K} N_k \, (\bar{y}_u - \bar{y}_u^{(k)})^2}_{\text{resolution}} + \underbrace{\bar{y}_u(1 - \bar{y}_u)}_{\text{uncertainty}} \,. \end{aligned}$$

4.3.2. Decomposition of the quantile score

Let q_τ be the quantile functional $T[F_P] = F_P^{-1}(\tau)$. The discretized forecast values are given by $q_\tau^{(k)}$. Conditional observed quantiles $y_\tau^{(k)}$ are estimated as sample quantiles from the observations y_i, with $i \in \mathcal{I}_k$, which belong to the k-th subsample. The climatological quantile is estimated from all observations and denoted by $\bar{y}_\tau$. Given the representation of (4.8)-(4.10) and using the quantile score function in (4.3), the decomposition of the QS is given by (Bentzien and Friederichs, 2014)

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\rho_\tau(y_i - q_\tau^{(k)})$$

$$= \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left[\rho_\tau(y_i - q_\tau^{(k)}) - \rho_\tau(y_i - y_\tau^{(k)})\right] \tag{4.11}$$

$$- \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left[\rho_\tau(y_i - \bar{y}_\tau) - \rho_\tau(y_i - y_\tau^{(k)})\right] \tag{4.12}$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\rho_\tau(y_i - \bar{y}_\tau)\,. \tag{4.13}$$

The r.h.s. of the equation describes the reliability (4.11), the resolution (4.12), and the uncertainty (4.13) of the quantile score.

4.3.3. Graphical representation of reliability

The reliability can be displayed graphically within a reliability diagram. This is well known for probability forecasts and the Brier score (e.g. Hsu and Murphy, 1986), and can also be adopted for a graphical exploration of quantile reliability (Bentzien and Friederichs, 2014). A reliability diagram shows the values of the calibration function $F_{Y|P}^{(k)}$, plotted against the discretized forecast values $F_P^{(k)}$. For a well calibrated forecast, i.e. forecasts which are realizations from the same data generating underlying distribution function as the observations, the points should lie close to the diagonal line. Deviations of the diagonal line reveal deficiencies of forecast performance, like constant over- or underforecasting. A comprehensive discussion of the reliability diagram can be found in Wilks (2006b), Section 7.4.4.

4.3.4. Discretization error

The discretization procedure described in Sec. 4.3 automatically leads to a bias in score estimates. The average score of the discretized forecasts will differ from the score of the original forecast values. Moreover, the discretization will also affect the estimation of the resolution and the reliability part of the score. The intervals for the discretization have to be chosen carefully to keep the biases as small as possible. The uncertainty is estimated from the observations alone and is not affected by the discretization.

The discretization is determined by the number of intervals, the interval width, and the representation of discretized forecast values. Several studies investigate the discretization error of the BS (e.g. Atger, 2003, 2004; Bröcker and Smith, 2007; Bröcker, 2008). Probability forecasts are bounded by 0 and 1. A categorization is often based on 10 intervals of equal width, and the discretized forecast values are set to the mid of the interval. A sharpness diagram shows the number of forecast-observation pairs in each interval. However, this might not be an optimal binning strategy. If the intervals are not sufficiently represented, a robust estimate of the calibration function cannot be guaranteed. The undersampling will result in strong biases in the decomposition. The bias can be reduced, if intervals are chosen such that they all contain an equal number of forecast-observation pairs (Atger, 2004). Moreover, the number of categories should be adjusted with regard to the sample size. The discretized forecast values may also be better represented by the mean or median of forecast values within an interval. This will also affect the graphical representation of reliability (Bröcker and Smith, 2007).

A comprehensive study about an optimal binning procedure for quantile forecasts can be found in Bentzien and Friederichs (2014). In general, the quantile forecast range should be split into non-overlapping intervals which are equally populated with forecast-observation pairs. The intervals are thus defined by the $1/K$-percentiles of the forecast values, with K the number of intervals. We have shown that an equal-distributed binning will largely reduce the discretization error compared to an equi-distant binning procedure. Moreover, we investigated the influence of the number of intervals onto the bias of reliability and resolution. For small K, the resolution is largely underestimated due to less variability between the forecast values. A large number of intervals will lead to a better representation of the resolution, but strongly affects the reliability. There has to be a trade-off between the gain in resolution and the loss in reliability to determine the optimal value for K (see Fig. 2 in Bentzien and Friederichs, 2014).

Part II.

Probabilistic forecasting and statistical postprocessing

5. Methodology

Probabilistic forecasting requires the transformation (postprocessing) of the ensemble into probabilistic predictions, e.g. an empirical distribution function, statistical moments, or quantiles. Simple postprocessing methods use solely the ensemble forecast to derive probabilistic products and will be explained in Sec. 5.1. More advanced postprocessing methods require a sufficiently large historic data set of forecasts and observations. A statistical relationship is defined which accounts for biases and systematic errors. Wilks (2006a) gives an overview of state-of-the-art ensemble postprocessing techniques.

We distinguish between point forecasts like the mean, quantiles, or probabilities on the one hand, and distributional forecasts in terms of a probability density function or cumulative distribution function on the other hand. Typical point forecasts for precipitation are the probability that a certain threshold will be exceeded, as well as quantiles. Probability forecasts are important for weather services to issue warnings for severe weather events. However, quantile forecasts gain more and more importance in probabilistic forecasting of quantitative precipitation. Quantiles need no prior knowledge about the range of data, as is necessary for probability forecasts to define meaningful thresholds. Boxplots are a graphical representation of a distribution in terms of its quantiles, and are a very intuitive tool to communicate uncertainty.

Regression techniques can be used to calibrate point forecasts. Logistic regression (Hamill et al., 2004) and quantile regression (Bremnes, 2004) directly estimate conditional probabilities or quantiles of the variable of interest. These semi-parametric techniques do not require an a priori distributional assumption. However, both techniques are limited to values that are sufficiently sampled and are generally applied for each quantile or probability separately. Methods for distributional forecasts often rely on parametric distribution functions which require the estimation of only a few distribution parameters. Typical methods are e.g. non-homogeneous Gaussian regression (EMOS, Gneiting et al., 2005), Bayesian model averaging (BMA, Raftery et al., 2005), or kernel dressing (Bröcker and Smith, 2008). A parametric distribution function allows to calculate probabilities and quantiles directly from the distribution parameters. However, the performance of such forecasts strongly depends on how suitable the a priori selected distribution fits the data. Sloughter et al. (2007) use BMA for precipitation forecasts, assuming a mixture of a point mass at zero and a gamma distribution. Scheuerer (2014) utilizes a generalized extreme value distribution censored at zero using EMOS. In Bentzien and Friederichs (2012), we utilize generalized linear models (GLM) for the estimation of various parametric distribution functions for precipitation. A generalized Pareto distribution is used for a better representation of the tail of the distribution.

This chapter is organized as follows: Section 5.1 starts with the translation of ensemble forecasts into probabilistic products. Regression techniques for point forecasts in terms of probabilities and quantiles are described in Section 5.2. A parametric mixture model for a full predictive distribution as introduced by Bentzien and Friederichs (2012) is given in Section 5.3.

5.1. From ensemble to probabilistic forecasts

Consider ensemble forecasts E_i with $i = 1, ..., M$ members. A probabilistic forecast can be derived by calculating statistical functionals directly from the M realizations at each grid point. These functionals, e.g. mean, standard-deviation or quantiles, are point-estimates and each of them represents a property of the underlying predictive distribution. The mean $\bar{E}$ and variance var(E) represent statistical moments, estimated by the arithmetic averages given by

$$\bar{E} = \frac{1}{M}\sum_{i=1}^{M} E_i \, , \quad \text{var}(E) = \frac{1}{M-1}\sum_{i=1}^{M}(E_i - \bar{E})^2 \, .$$

Threshold exceedance probabilities or quantiles are specific points of the predictive distribution. Threshold exceedance probabilities are estimated by the fraction of ensemble members which exceed a certain threshold u. The probabilities are given by the arithmetic average

$$Pr(u|E_1, ..., E_M) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(E_i > u) \, ,$$

where $\mathbb{I}(.)$ is the indicator function which is 1 if the condition in brackets is true and zero otherwise. With regard to precipitation forecasts, we consider two types of precipitation events: the occurrence of precipitation (i.e. precipitation above zero) using the probability of precipitation (PoP) and precipitation above a threshold u using the probability of threshold exceedance (PoT).

Quantiles are estimated from the order statistics of the ensemble member. We assume that the ensemble members are already ordered such that $E_1 \leq, ..., \leq E_M$. There is no unique solution for the estimation of sample quantiles, and various software packages will handle the calculation differently. A comprehensive overview about the most common sample quantile definitions is given in Hyndman and Fan (1996). They generalize sample quantiles to a weighted mean of the form

$$Q(\tau|E_1, ..., E_M) = (1-\gamma)E_j + \gamma E_{j+1} \, ,$$

where the index j is given by $j = \lfloor \tau M + r \rfloor$ for some $r \in \mathbb{R}$. The floor function $\lfloor . \rfloor$ denotes the largest integer not greater than the value in brackets. This study uses the type 8 estimator, which is given by $r = 1/3(\tau + 1)$ and the weight function $\gamma = \tau M + r - j$.

5.1.1. Neighborhood method and first-guess forecasts

Operational ensemble systems are often limited in ensemble size due to the high computational costs. Time-lagging (Sec. 3.2.2) is one inexpensive way to increase the ensemble size and hence to improve the representation of ensemble spread. Another way is the so-called neighborhood method introduced by Theis et al. (2005). It was originally introduced to obtain probabilistic guidance from a deterministic model forecast as illustrated in Fig. 5.1. A 12-hour precipitation forecast from one ensemble member (deterministic forecast) is shown in Fig. 5.1(a). Statistical functionals are calculated from a spatial neighborhood of 5×5 grid points. In case of the mean functional, the neighborhood method leads to a smoothed forecast field as illustrated in Fig.

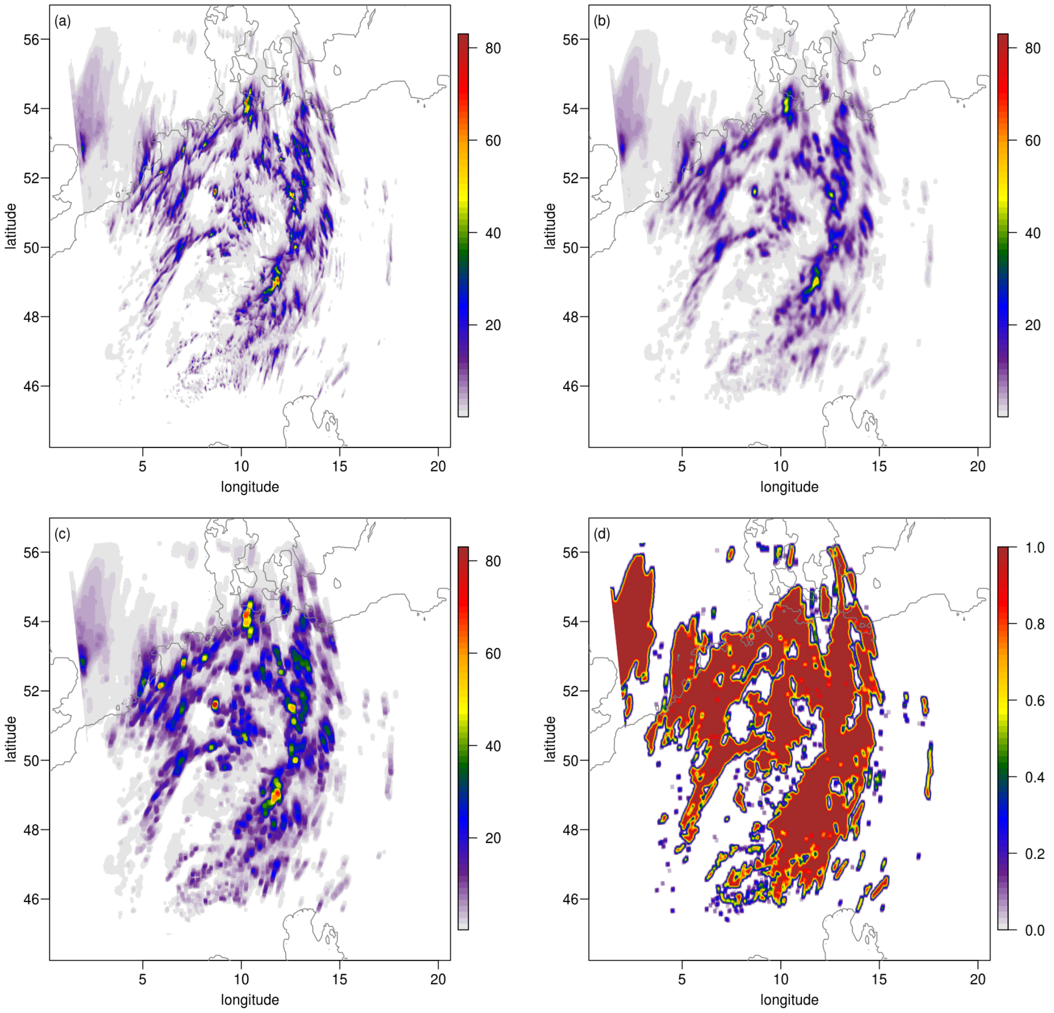


Figure 5.1.: Precipitation forecast in mm/12h valid for June 6, 2011 (12-24 UTC). The deterministic forecast in (a) is one realization of the ensemble forecast. Probabilistic products are derived from the deterministic forecast using a 5 × 5 neighborhood. Shown here is the mean (b), the 0.9-quantile (c), and the PoT for a threshold of 1 mm/12h (d).

5.1(b). The fine scale structure is smoothed through the averaging process, which also removes extreme precipitation values. The 0.9-quantile, estimated from the 25 gridboxes within each neighborhood, is displayed in Fig. 5.1(c). A smoother structure of the precipitation pattern is obtained compared to the deterministic forecasts, but extreme values are more pronounced than in the mean forecast. The PoT for a threshold of 1 mm/12h is shown in Fig. 5.1(d).

The neighborhood-method can be applied to ensemble forecasts, which was first introduced by Schwartz et al. (2010). Quantiles, probabilities or mean values are estimated from an enlarged ensemble E_{ij}, where $i = 1, ..., M$ are the ensemble member and $j = 1, ..., S^2$ are the numerated grid points within a spatial neighborhood of $S \times S$ grid points. This method is especially suitable for precipitation forecasts, which show small-scale features which are well represented in high-resolution numerical weather prediction forecasts, but suffer strongly from displacement errors. Fig. 5.2 shows probabilistic products derived from the 20 member COSMO-DE-EPS (left

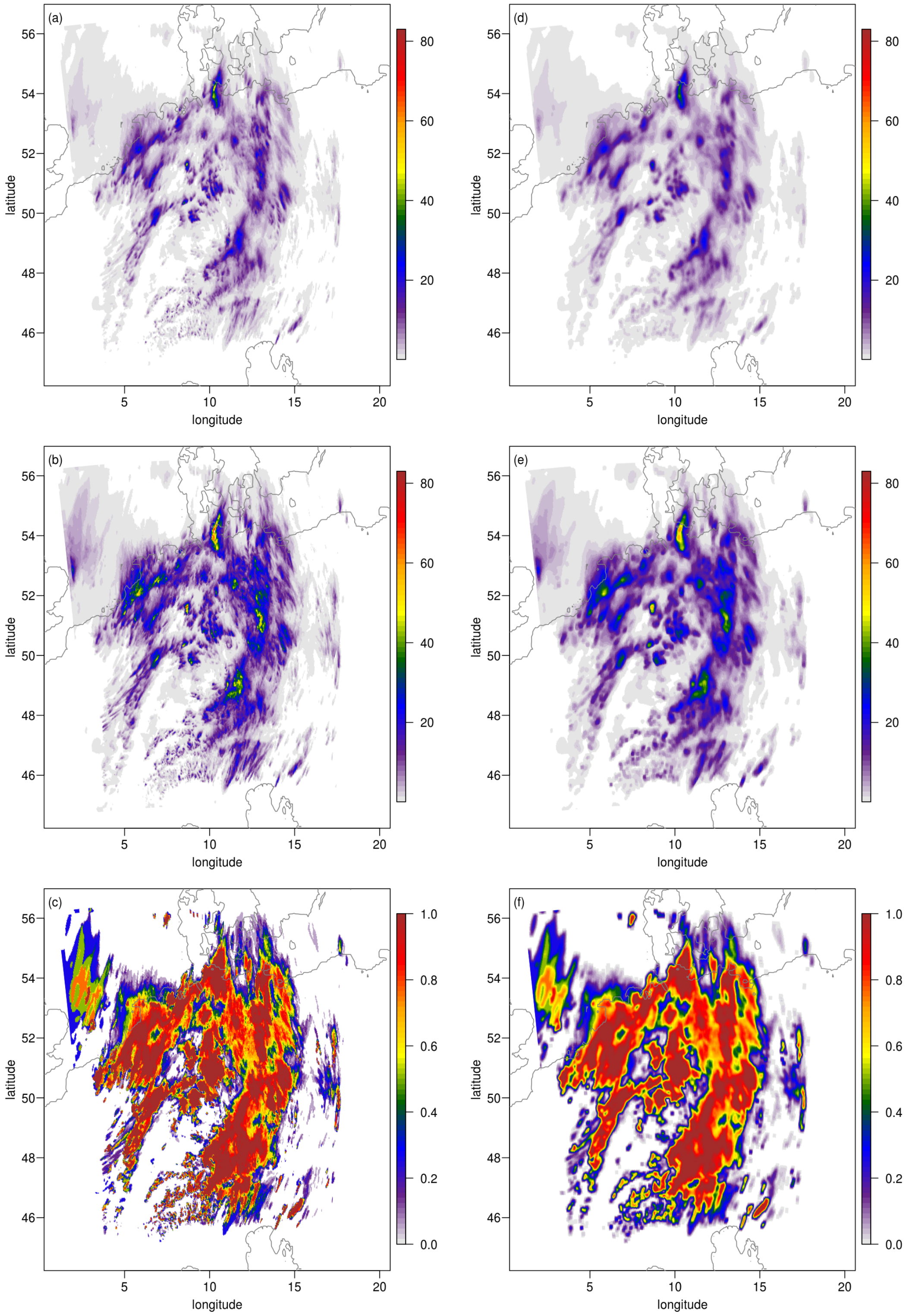


Figure 5.2.: Probabilistic products derived from the ensemble forecast valid for June 6, 2011 (12-24 UTC). Shown here is the mean (a,d), the 0.9-quantile (b,e), and the PoT for a threshold of 1 mm/12h (c,f). The left column (a-c) is derived from the 20 members of the ensemble alone, while the right column (d-f) uses additionally a 5 × 5 neighborhood.

column) and under consideration of a 5 × 5 neighborhood (right column). The latter is built from 20 members times 25 neighbors which lead to an enlarged ensemble of 500 values for each gridpoint. Especially the PoT forecasts in Fig. 5.2(c+f) show larger variability compared to the forecast system in Fig. 5.1(d).

The benefit of the neighborhood method to probabilistic quantitative precipitation forecasts from an ensemble has been shown recently by Schaffer et al. (2011), Bentzien and Friederichs (2012), Ben Bouallègue et al. (2013), and Scheuerer (2014). What remains difficult is the determination of an appropriate size of the spatial neighborhood. Increasing the spatial scale does improve the predictive performance of probabilistic forecasts in terms of reliability, but also reduces the sharpness (Ben Bouallègue, 2011). The choice of spatial scale depends on the user. Roberts and Lean (2008) suggest that the spatial neighborhood should be at least of the size of the effective resolution or scale of predictability as estimated e.g. by Skamarock (2004) or Bousquet et al. (2006). For COSMO-DE-EPS, a comprehensive study based on kinetic energy spectra was made by Bierdel et al. (2012). They found that only processes of at least 4-5 times the horizontal grid spacing can appropriately resolved by the model dynamics. We therefore conclude that for COSMO-DE-EPS, the size of the neighborhood should at least be set to 5×5 grid boxes.

In the following, probabilistic products for precipitation derived from the ensemble under consideration of a 5×5 neighborhood are denoted as first-guess forecasts. They serve as a benchmark for probabilistic forecasts derived from a statistical model which is based on a historical data set. First-guess forecasts often show deficiencies with respect to calibration, especially if the raw ensemble does not show sufficient ensemble spread. These deficiencies can be overcome using one of the regression techniques explained in the next sections.

5.2. Logistic and quantile regression

More advanced methods for postprocessing use historic data of observations and ensemble forecasts to define a statistical relationship, e.g. in terms of a regression ansatz. Once a suitable regression model has been specified, it can be used to provide future predictions conditional on the ensemble forecasts for future time steps. In the following sections, the response variable of such a regression model is denoted by Y. The regression model depends on the ensemble by the definition of covariates, further denoted by X. The advantage of the regression ansatz is, that one can use more than one variable of the ensemble forecasts, and not necessarily the variable which is to be predicted. The following notation is used in the remainder of this chapter: Capital letters refer to random variables, while small letters denote a realization thereof. Bold letters indicate multivariate quantities. Vectors $\mathbf{x}$ are taken as column vectors, while the transposed vector is given by $\mathbf{x}'$.

5.2.1. Logistic regression

Logistic regression (LR) is used to derive calibrated forecasts for the PoP and PoT. LR is a generalized linear model, where the variable of interest follows a Bernoulli distribution (Fahrmeir and Tutz, 1994). Since we have a continuous response variable Y, a dichotomous event is defined by the exceedance of a threshold u. LR assumes that the probability that $Y > u$ depends on a linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}_u$ transformed by the inverse logit function

$$Pr(Y > u|\mathbf{X} = \mathbf{x}) = \text{logit}(\eta) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_u)}{1 + \exp(\mathbf{x}'\boldsymbol{\beta}_u)} .$$

Here, $\mathbf{x}$ is a vector with predictive covariates, starting with a 1 accounting for the intercept. The regression coefficients $\boldsymbol{\beta}_u$ are estimated by maximizing the likelihood Λ from a training sample $i = 1, ..., N$ of forecast-observation pairs

$$\hat{\boldsymbol{\beta}}_u = \arg \max_{\boldsymbol{\beta}_u} \Lambda(\boldsymbol{\beta}_u) , \text{ with}$$
$$\Lambda(\boldsymbol{\beta}_u) = \prod_{i=1}^{N} (\mathbf{x}_i'\boldsymbol{\beta}_u)^{\mathbb{I}(y_i > u)} (1 - \mathbf{x}_i'\boldsymbol{\beta}_u)^{\mathbb{I}(y_i \leq u)} .$$

Here and in all regression models that follow, the predictive covariates $\mathbf{x}$ are taken from the ensemble forecasts. We are not restricted to precipitation forecasts and the mean functional, respectively, as is the case for e.g. BMA or kernel dressing. The regression ansatz allows to incorporate any kind of information from the ensemble: categorical as well as continuous variables, probabilities, quantiles, et cetera. Note that covariates should be normalized before they enter the regression model.

5.2.2. Quantile regression

Quantile regression (QR) is a method to estimate conditional quantiles of the response variable Y given the covariates $\mathbf{X}$ by the use of a linear model and is described in detail by Koenker (2005). For a given probability level $\tau \in [0, 1]$, the quantile regression function is defined as

$$Q(\tau|\mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_\tau . \tag{5.1}$$

The regression coefficients $\boldsymbol{\beta}_\tau$ are estimated in order to minimize the quantile score function

$$\hat{\boldsymbol{\beta}}_\tau = \arg \min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{N} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau) . \tag{5.2}$$

Censored quantile regression

An important property of the quantile function is the equivariance to monotone transformations (Koenker, 2005, pp. 39). Quantiles are determined from the order statistics of the data, which is not destroyed by monotone transformations. The equivariance of the quantile functions allows the formulation of a censored quantile regression for variables which are bounded by zero. Analogously to (5.1) and (5.2), the censored QR is formulated as

$$Q(\tau|\mathbf{X} = \mathbf{x}) = \max\left(0, \mathbf{x}'\boldsymbol{\beta}_\tau\right) ,$$
$$\hat{\boldsymbol{\beta}}_\tau = \arg \min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{N} \rho_\tau\left(y_i - \max\left(0, \mathbf{x}_i'\boldsymbol{\beta}_\tau\right)\right) .$$

The coefficients $\hat{\boldsymbol{\beta}}_\tau$ of the censored QR are estimated in a three step procedure (according to Chernozhukov and Hong, 2002; Friederichs and Hense, 2007). In the first step, the probability of precipitation is estimated using logistic regression. Based on the estimated PoP, which we denote by π_0 in the following, a subsample is chosen with $J_0 = \{i : \pi_{0,i} > 1 - \tau\}$. The second step calculates an initial estimate of the regression coefficients with standard quantile regression based on the subsample J_0. Another subsample $J_1 = \{i : \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\tau > 0\}$ is selected. Step three estimates the regression coefficients $\hat{\boldsymbol{\beta}}_\tau$ based on the subsample J_1.

The advantage of this three step procedure is, that standard quantile regression can be calculated by existing software packages like the ”quantreg” package by Koenker (2013) for the R statistical language (R Core Team, 2014).

5.3. Mixture models

Probability and quantile forecasts can also be derived from a parametric distribution function. Following the approach of Sloughter et al. (2007), the conditional probability density function (PDF) of precipitation $f_Y(Y \mid \mathbf{X} = \mathbf{x})$ can be described using a two step model. In a first step the probability of precipitation $\pi_0 = Pr(Y > 0 \mid \mathbf{X} = \mathbf{x})$ is estimated via LR. In a second step the PDF of the amount of precipitation given that it is not zero is assumed to follow a parametric distribution f_*, defined on $\mathbb{R}^+$. This yields a predictive PDF of the form

$$f_Y\,(y \mid \mathbf{X} = \mathbf{x}) = \begin{cases} 1 - \pi_0 & \text{for } y = 0 \\ \pi_0\, f_*(Y \mid \mathbf{X} = \mathbf{x}) & \text{for } y > 0\,. \end{cases} \tag{5.3}$$

The cumulative distribution function (CDF) for precipitation of the PDF in (5.3) is given by

$$\begin{aligned} F_Y(y \mid \mathbf{X} = \mathbf{x}) - F_Y(0 \mid \mathbf{X} = \mathbf{x}) &= \int_0^y f_Y(t \mid \mathbf{X} = \mathbf{x}) dt \\ &= (1 - \pi_0) + \pi_0\, F_*(y \mid \mathbf{X} = \mathbf{x}, y > 0)\,. \end{aligned} \tag{5.4}$$

If the CDF in (5.4) is known, it is possible to calculate the PoT forecasts as well as conditional quantiles. The PoT’s for a threshold u are derived as

$$\begin{aligned} Pr(Y > u \mid \mathbf{X} = \mathbf{x}) &= 1 - F_Y(u \mid \mathbf{X} = \mathbf{x}) \\ &= \begin{cases} \pi_0 & \text{for } u = 0 \\ \pi_0(1 - F_*(u \mid \mathbf{X} = \mathbf{x})) & \text{for } u > 0\,. \end{cases} \end{aligned}$$

The quantile function $Q_Y(\tau) = F_Y^{-1}(\tau)$ is obtained by inverting the CDF

$$\begin{aligned} F_Y(y \mid \mathbf{X} = \mathbf{x}) = \tau &= 1 - \pi_0 + \pi_0\, F_*(y \mid \mathbf{X} = \mathbf{x}) \\ \tilde{\tau} &= \frac{\tau - 1 + \pi_0}{\pi_0} = F_*(y \mid \mathbf{X} = \mathbf{x})\,, \\ Q_Y(\tau \mid \mathbf{X} = \mathbf{x}) &= \begin{cases} 0 & \text{for } \tilde{\tau} \leq 0\,, \\ F_*^{-1}\,(\tilde{\tau} \mid \mathbf{X} = \mathbf{x}) & \text{for } \tilde{\tau} > 0\,. \end{cases} \end{aligned}$$

Since precipitation is a censored random variable $Y \in [0, \infty]$, the quantile function is censored, too. A quantile can take values greater than zero only if the probability of precipitation π_0 is greater than $1 - \tau$. This follows directly from the condition

$$\begin{aligned}\tilde{\tau} = \frac{\tau - 1 + \pi_0}{\pi_0} &> 0 \\ \tau - 1 + \pi_0 &> 0 \\ \pi_0 &> 1 - \tau\,.\end{aligned}$$

For f_* we assume a parametric distribution function defined on the positive real line which belongs to the exponential family. The expectation of f_* and the variance parameter can thus be estimated by generalized linear models (GLMs).

5.3.1. Generalized Linear Model

The theory of GLMs is well described in Fahrmeir and Tutz (1994) or McCullagh and Nelder (1989). A GLM is based on two assumptions. The *distributional assumption* expects that the response variable Y is conditionally independent given the covariates $\mathbf{X}$ and that their type of distribution belongs to the exponential family. The *structural assumption* implies a relation between the conditional expectation value μ and the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ of the form

$$\mu(Y \mid \mathbf{X} = \mathbf{x}) = h(\eta) = h(\mathbf{x}'\boldsymbol{\beta})\,.$$

Hereby, the function h is called the response function and the inverse is called the link function $h^{-1}(\mu) = g(\mu) = \eta$. Thus, the GLM is described by the type of the exponential family, and the response or link function. The conditional variance of Y is of the form

$$\mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x}) = \phi V(\mu)\,,$$

where ϕ is the dispersion parameter. The variance function $V(\mu)$ is a function of the mean and depends on the distributional assumption.

The parameters of a GLM are estimated from training data. At first, the regression coefficients for the mean are estimated using maximum-likelihood techniques

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \Lambda(\boldsymbol{\beta} | y_i, \mu_i)\,, \text{ with } \mu_i = \mathbf{x}_i'\boldsymbol{\beta}\,.$$

The likelihood function Λ is given by the selected distribution function. The dispersion parameter is estimated subsequently from the moment estimator

$$\hat{\phi} = \frac{1}{N - P} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{V(\mu_i)}\,,$$

where N is the number of observations and P is the number of predictors. If μ_i and ϕ are estimated, they can be related to the distribution parameters. In this study, we use the Gamma, log-normal, and inverse-Gaussian distribution.

Gamma distribution

The Gamma distribution is fully determined by a shape parameter α and a scale parameter θ. It has the form

$$f_\Gamma(y|\alpha,\theta) = \frac{1}{\theta\Gamma(\alpha)}\left(\frac{y}{\theta}\right)^{\alpha-1}\exp\left(-\frac{y}{\theta}\right) \quad \text{with } \alpha,\theta > 0\,,$$

with expectation $\mu_\Gamma = \alpha\theta$ and variance $\sigma_\Gamma^2 = \alpha\theta^2$ (Wilks, 2006b). Hereby, $\Gamma(\cdot)$ denotes the Gamma function. f_Γ can be reparameterized in terms of μ_Γ and α as

$$f_\Gamma(y|\mu_\Gamma,\alpha) = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu_\Gamma}\right)^{\alpha} y^{\alpha-1}\exp\left(-\frac{y\alpha}{\mu_\Gamma}\right)\,.$$

The response function determines how the linear predictor enters the gamma distribution. Typical response functions for a gamma distribution are the reciprocal $\mu_\Gamma = \eta^{-1}$, the identity $\mu_\Gamma = \eta$, or the exponential function $\mu_\Gamma = \exp(\eta)$. The variance is given by $\sigma_\Gamma^2 = \mu_\Gamma^2/\alpha$, with $V(\mu) = \mu^2$ and $\phi = 1/\alpha$ for the Gamma distribution. Hence, the shape parameter α is derived by

$$\hat{\phi} = \frac{1}{\hat{\alpha}} = \frac{1}{N-P}\sum_{i=1}^{N}\left[\frac{y_i - \hat{\mu}_{\Gamma,i}}{\hat{\mu}_{\Gamma,i}}\right]^2\,.$$

Log-normal distribution

The log-normal distribution with mean μ_{ln} and standard deviation σ_{ln} is given by

$$f_{ln}(y;\mu_{ln},\sigma_{ln}) = \frac{1}{y\sigma_{ln}\sqrt{2\pi}}\exp\left[-\frac{(\ln(y)-\mu_{ln})^2}{2\sigma_{ln}^2}\right]\,, \quad y > 0\,.$$

A log-normal GLM is estimated with a Gaussian linear model for log-transformed variables. The variance is related to the dispersion parameter using the variance function $V(\mu) = 1$

$$\hat{\phi} = \hat{\sigma}_{ln}^2 = \frac{1}{N-P}\sum_{i=1}^{N}(\ln(y_i) - \mu_i)^2\,.$$

Inverse-Gaussian distribution

The inverse-Gaussian distribution is described by a mean $\mu_{ig} > 0$ and shape parameter $k > 0$

$$f_{ig}(y;\mu_{ig},k) = \sqrt{\frac{k}{2\pi y^3}}\exp\left(\frac{-k(y-\mu_{ig})^2}{2\mu_{ig}^2 y}\right)\,, \quad y > 0\,.$$

The variance is given by $\sigma_{ig}^2 = \mu_{ig}^3/k$, with $V(\mu) = \mu^3$ and $\phi = 1/k$ for the inverse-Gaussian distribution. The shape parameter k is related to the dispersion parameter by

$$\hat{\phi} = \frac{1}{\hat{k}} = \frac{1}{N-P}\sum_{i=1}^{N}\frac{(y_i - \hat{\mu}_{\Gamma,i})^2}{\hat{\mu}_{\Gamma,i}^3}\,.$$

5.3.2. A mixture model with GPD tail

The parametric mixture might well represent the bulk of the distribution, but not necessarily captures the right tail behavior. All proposed distributions in previous section exhibit an exponential tail behavior, but several studies show evidence for a heavy tail behavior of precipitation (e.g. Friederichs, 2010). A misrepresentation of the tail behavior might lead to large prediction errors for extreme precipitation events. Extreme value theory provides an asymptotic theory for the tail of a distribution, and is particularly developed to make predictions beyond the range of the data. Hence, a very natural extension of the mixture model is to represent the tail of the conditional distribution using a generalized Pareto distribution (GPD, Coles, 2001).

The GPD is used to model excesses $Z = Y - u_\tau$ over large thresholds u_τ. Extreme value theory proves that under very general conditions Z asymptotically follows a GPD for large u_τ

$$F_{GPD}(z) = 1 - \left(1 + \frac{\xi z}{\sigma_u}\right)^{-1/\xi}, \quad z > 0\,.$$

Hereby, σ_u denotes the scale parameter and ξ the shape parameter.

A relatively simple formulation of a mixture with variable tail behavior is used in Bentzien and Friederichs (2012). The mixture model in Eq. (5.4) models the range below u_τ, which is taken as the conditional τ_u-quantile, and the GPD models the thresholds above. The probability τ_u is set to 0.95. The GPD is additionally conditioned on the covariates by assuming a linear model for the scale parameter with

$$\sigma_u = \mathbf{x}'\boldsymbol{\beta}_\sigma\,.$$

The shape parameter $\xi = \xi_0$ is kept constant. The complete CDF of the GPD mixture reads

$$F_Y(y \mid \mathbf{X}=\mathbf{x}) - F_Y(0 \mid \mathbf{X}=\mathbf{x}) = \begin{cases} (1-\pi_0) + \pi_0\, F_*(y \mid \mathbf{X}=\mathbf{x}) & \text{for } y \le u_\tau\,, \\ \tau_u + (1-\tau_u) F_{GPD}(y - u_\tau \mid \mathbf{X}=\mathbf{x}) & \text{for } y > u_\tau\,. \end{cases}$$

PoT forecasts for large thresholds $u > u_\tau$ are obtained by

$$Pr(Y > u \mid \mathbf{X}=\mathbf{x}) = (1-\tau_u)\left(1 + \frac{\xi(y - u_\tau)}{\sigma_u}\right)^{-1/\xi}.$$

Estimates of conditional quantiles for $\tau > \tau_u$ are obtained by the quantile function

$$Q_{GPD}(\tau \mid \mathbf{X}=\mathbf{x}) = \begin{cases} u_\tau + \frac{\sigma_u}{\xi}\left[(1-\tilde{\tau})^{-\xi} - 1\right] & \text{for } \xi \ne 0\,, \\ u_\tau + \sigma_u \log(1-\tilde{\tau}) & \text{for } \xi = 0\,, \end{cases}$$

where $\tilde{\tau}$ is defined as $\tilde{\tau} = \frac{\tau - \tau_u}{1-\tau_u}$ (Friederichs, 2010).

6. Precipitation: observations and model data

In this chapter we will shortly discuss the nature of observational data for precipitation and introduce the data sets used in this study. Precipitation results from complex micro-physical and dynamical processes on smaller and larger scales and exhibits a large temporal and spatial variability. Measurements have to capture the mixed discrete-continuous character of precipitation, the identification of rainfall occurrence and the amount of precipitation. There exists mainly two sources of observational data for precipitation:

- In-situ measurements are obtained from rain gauges located at observational sites. They are point measurements limited to single locations, but with high data quality.

- Remote sensing observations like radar or satellite have a good spatial and temporal coverage. However, they cannot directly measure the actually rainfall amount. Radar reflectivity has to be converted to precipitation rates by empirical relationships.

In this study, in-situ measurements from rain gauges are used as target for precipitation. The data is taken from the observational network of DWD[1]. DWD disposes over a dense network of rain gauges located all over Germany. This study uses data from about ~ 1000 observational sites with hourly measurements of precipitation. One has of course to keep in mind that the NWP model output represent area-mean values of precipitation for a single gridbox (here 2.8km×2.8km). The area-mean can naturally not capture heavy precipitation amounts measured by localized rain gauges. However, the ability of statistical postprocessing to link area-mean forecasts to localized observations will be explored. Throughout the study, forecast-observation pairs are build from the nearest gridpoint of COSMO-DE to the observational sites.

The following sections describe the two data sets for COSMO-DE-LAF and COSMO-DE-EPS which are used for the evaluation of precipitation forecasts derived from the COSMO-DE ensembles. Focus of the study are daily 12-hourly precipitation accumulations between 12 to 24 UTC.[2]

6.1. Data set I: COSMO-DE-LAF

The first part of the study considers the development and adaption of suitable postprocessing methods for precipitation from COSMO-DE based ensemble systems. Since data of COSMO-DE-EPS was only available at the end of 2011, this part was done using the COSMO-DE-LAF as skeleton EPS interim solution. Three years of data were collected from the DWD archive, namely the time period from July, 2008 until June, 2011. Statistical postprocessing relies on

[1] The rain gauge data was kindly provided by M. Göber, Deutscher Wetterdienst.

[2] Note that the construction of the 4 member LAF is restricted to a common forecast period of 12 hours, see section 3.2.1.

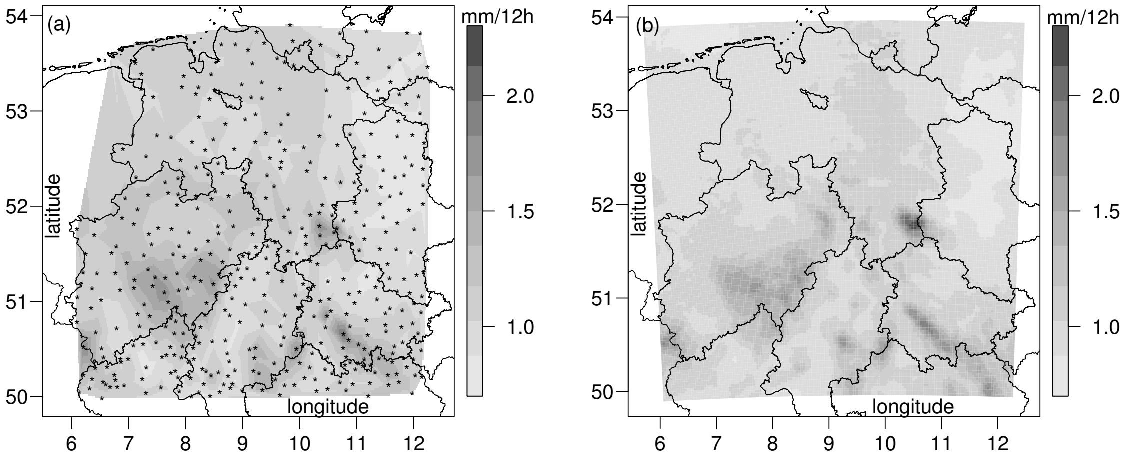


Figure 6.1.: Mean precipitation intensities in mm/12h for (a) station measurements (linearly interpolated) and (b) COSMO-DE forecasts for the period July 1, 2008 to June 30, 2011. The stars in (a) represent the locations of observational sites (Figure 1 from Bentzien and Friederichs, 2012).

an extensive training period, and robust estimates require a sufficiently long data set. However, if the training period is too long, changes in the operational forecast systems can deteriorate the predictive skill of the postprocessing. Reforecasts are useful to study the effect of long-time training periods (e.g. Hamill et al., 2008; Fundel et al., 2010) but are so far not available for COSMO-DE.

To construct 12-hourly precipitation forecasts from the LAF ensemble, we use daily model runs of COSMO-DE initialized at 03, 06, 09, and 12 UTC with forecast lead times 9-21h, 6-18h, 3-15h, and 0-12h, respectively. The analysis is restricted to a sub-domain of the original COSMO-DE model domain, which consists of 160×160 gridboxes and covers large parts of Northwest-Germany (see Fig. 6.1). The limited model domain reduces the computational costs and the amount of data remains considerable. Altogether, 445 observational sites are located in the subdomain. The data set thus contains values from 1095 days at 445 locations (see Tab. 6.1).

Figure 6.1 shows mean precipitation intensities averaged over the evaluation period. The gauge measurements are linearly interpolated only for enhanced visibility. Both, COSMO-DE and station measurements show enhanced precipitation over the mountain ranges in the southern parts. The highest precipitation intensities occur over Harz (51.8°N, 10.6°E) and Thüringer Wald (50.7°N, 10.8°E), followed by Vogelsberg (50.5°N, 9.2°E) and Sauerland/Rothaargebirge (51°N, 8°E). Lower precipitation intensities are observed and modeled over the northern parts with mostly flat topography.

6.2. Data set II: COSMO-DE-EPS

The evaluation of COSMO-DE-EPS is based on a one-year evaluation period from January to December, 2011. Ensemble forecasts are available for 357 days within this period. Daily 12-hourly precipitation forecasts are taken from the COSMO-DE-EPS initialized at 12 UTC with

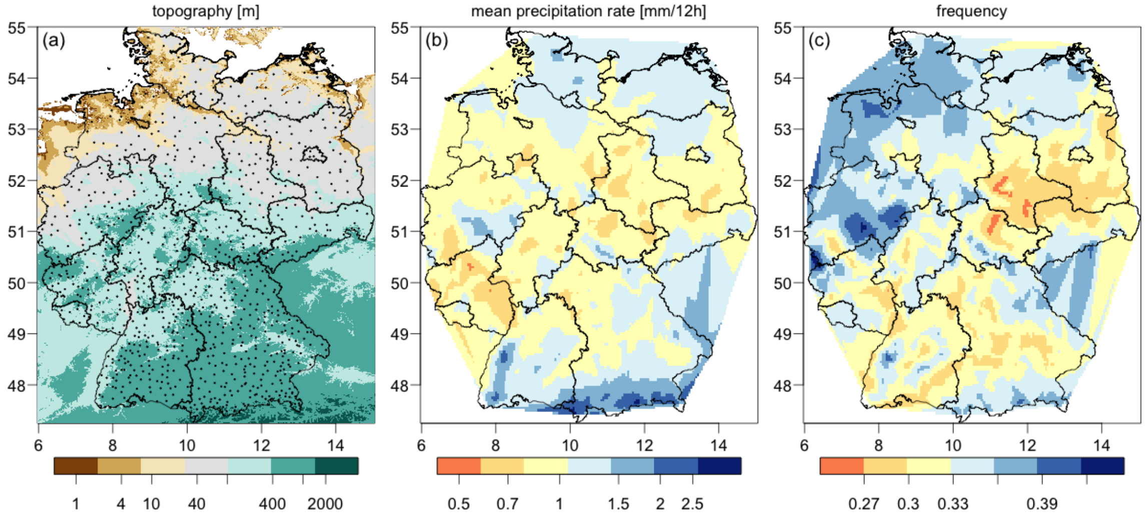


Figure 6.2.: Climatology of the observations for the year 2011: (a) The rain gauge network at DWD (color coding shows the topography of Germany). (b) Mean precipitation intensities in mm/12h at each station, linearly interpolated for enhanced visibility. (c) Frequency of days with precipitation above zero.

lead time 0-12h. The time-lagged ensemble COSMO-DE-TLE is constructed analogously to the LAF from the ensemble runs initialized at 03, 06, 09, and 12 UTC. The verification domain is extended to Germany, where we dispose of observational data from 1079 rain gauges. Thus, the data set contains values from 357 days at 1079 locations (see Tab. 6.1). Figure 6.2(a) shows the location of the rain gauges together with the topography. Germany is characterized by a flat topography in the northern parts and the coastal regions of Nordsee and Ostsee. The topography is continuously rising towards the southern parts of Germany, with the highest elevations in the alps.

Figure 6.2(b) shows the mean precipitation intensities for each station averaged over the year 2011. The data is linearly interpolated for enhanced visibility. The highest precipitation intensities are observed in the alps region and in the mountainous area of Schwarzwald (48°N, 8°E). Higher precipitation intensities are also observed in the German Mittelgebirge, especially over the Rheinisches Schiefergebirge (51°N, 8°E) and Thüringer Wald (50.7°N, 10.8°E). The north-eastern parts (coastal regions of Ostsee) also exhibit higher precipitation intensities. Lower precipitation intensities occured over the federal states Rheinland-Pfalz (50°N, 8°E), Sachsen-Anhalt (52°N, 12°E) and the northern part of Thüringen.

The frequency of rain days (days with precipitation above zero) as observed at each station in 2011 is shown in Fig. 6.2(c). The frequencies range between 25% and 45%. More rain days are observed over north-western Germany and the coastal regions, but also over mountainous area. The most rain days occur at the Rheinisches Schiefergebirge. The eastern part of Germany (Sachsen-Anhalt, Brandenburg) is very dry, with small precipitation intensities, but also the lowest number of rain days in 2011. Lower frequencies can also be found along the river valleys of Rhein and Donau.

Fig. 6.3(a) shows the seasonal variability of mean precipitation intensities. The year 2011 is characterized by lower precipitation intensities in spring, and higher intensities during summer

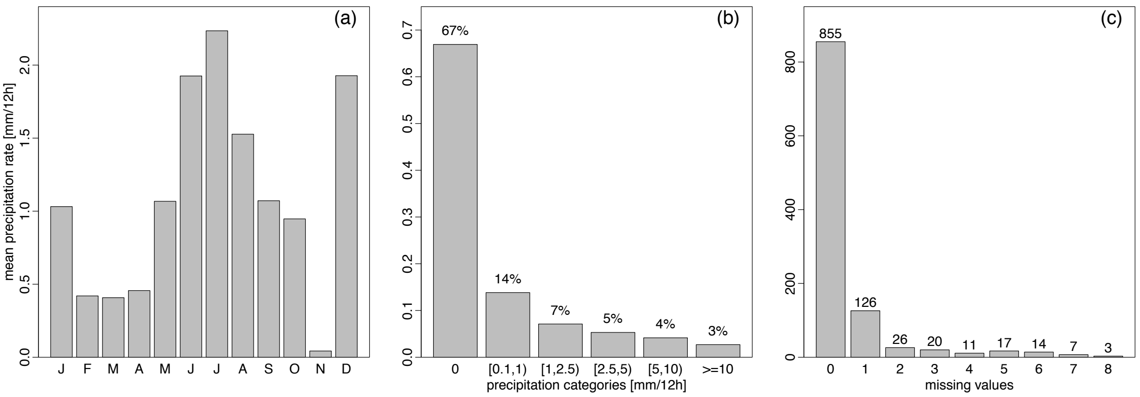


Figure 6.3.: Summary statistics of the observations for the year 2011: (a) Monthly average of precipitation between 12-24 UTC at each station. (b) Distribution of precipitation intensities. (c) Number of stations with missing data.

which are often related to convective events. The November 2011 was unusually dry, and nearly no precipitation was observed in Germany. The November was followed by a wet December with precipitation intensities nearly as high as in summer[3]. The distribution of 12-hourly rainfall events is shown in Fig. 6.3(b) for different categories. No rainfall is observed in 67% of the data. Higher rainfall events occur less frequently, and values of 10 mm/12h or more are only observed in 3% of the data (10.359 observations $\geq$ 10 mm/12h). However, the range of 12h rainfall amounts goes up to 90 mm/12h. The highest rainfall event of 93.2 mm/12h was observed in Kubschütz (51.2°N, 14.5°E) in the very eastern part of Germany on July 20, 2011 between 12 and 24 UTC. A low pressure system lead to strong precipitation and thunderstorms in most of the south-eastern parts of Germany.

The amount of missing data is small for the selected stations. An overview is given in Fig. 6.3(c). 79% of the stations have complete time series for 2011, and 12% have only one missing value. Other stations have 2 to 8 missing values. A summary of the number of observations within both data sets is given in Tab. 6.1.

[3] A very detailed analysis about the general precipitation climatology in Germany can be found on the webpage http://www.dwd.de/klimaatlas (Deutscher Klimaatlas).

Table 6.1.: Overview about the size of the data sets: model domain (NW-GER: north-western Germany, GER: Germany), time period of investigation, number of days, number of stations, number of total observations, number of missing values (NA; less than 1%).

	domain	time period	# days	# stations	# total	# NA
Data set I	NW-GER	01.07.2008 - 30.06.2011	1095	445	487 275	3 141
Data set II	GER	01.01.2011 - 31.12.2011	357	1079	385 203	524

7. Evaluation of COSMO-DE-LAF

The COSMO-DE-LAF serves as benchmark model to study the generation and calibration of probabilistic quantitative precipitation forecasts from a high-resolution, convection-permitting NWP model. The evaluation focus on the one hand on the general capability of the benchmark system to provide probabilistic guidance for precipitation. On the other hand, the evaluation determines a suitable setup for a statistical calibration model. This includes the setup of training- and forecast periods, the choice of predictive covariates and the comparison of various postprocessing techniques as described in Chapter 5. The main results of the generation and calibration of probabilistic quantitative precipitation forecasts from COSMO-DE-LAF are published in Bentzien and Friederichs (2012) using the data set I as described in Section 6.1. The key results are briefly summarized in the following sections.[1]

7.1. Statistical model setup

Statistical postprocessing relies on training data to determine the unknown parameters (i.e. regression coefficients) for the statistical model. An important aspect for the temporal setup is a strict separation of training and verification data. The observational data used for verification have to be independent from those used for the training. To this end, the data set is divided into blocks of 15 days which build the verification periods. For each verification period, the preceding time period is used for the training of the parameters of the postprocessing. Forecasts are then derived for the following 15 days. In this way we obtain a series of forecasts for the evaluation period of 3 years that are independently derived from the respective observations used for verification. Note that in this setup the data from all stations are pooled into one vector and the relation between the local predictors and the predictand is assumed to be spatially constant. An illustration of the gliding training and verification periods is given in Fig. 7.1.

In a first step, the length of training period has to be determined. The training period has to be sufficiently large to obtain stable parameter estimates. However, longer training periods are affected by inhomogeneities due to the seasonal cycle and changes in the operational model version of COSMO-DE. Training periods between 20 and 90 days previous to the 15 days of verification are tested. The predictive skill increases with the length of the training period up to 50 days. Longer training periods lead to inconsistent behavior of a variety of scores, and slowly deteriorates the predictive performance. Hence, the training length is set to 50 days in the following. To study the effect of sampling uncertainty, an analysis is included with a training sample that makes use of the complete data. Only the prediction period of 15 days extended by 15 days is withheld from the training. Hence, the training period encompasses 1065 days from the 3-year time period of investigation. In order to account for the seasonal changes within the

[1] Please note that the COSMO-DE-LAF was abbreviated by COSMO-DE-TLE in the original publication.

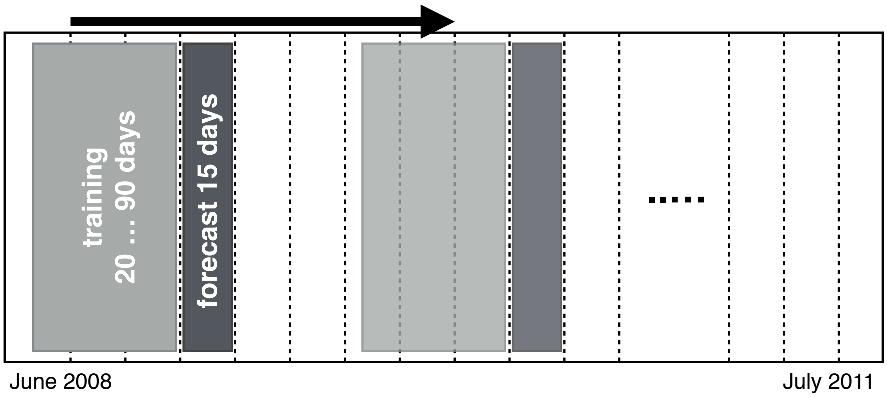


Figure 7.1.: Illustration of a gliding training and verification period. For each block of 15 days, forecast are obtained by a training period of the preceding 20 to 90 days.

training data, a sine and cosine with a period of one year is included into the postprocessing.

In a next step, a set of informative predictors has to be chosen. A typical predictor is the ensemble mean, as used by Wilks (2009) and Hamill et al. (2004), amongst others. Bremnes (2004) uses quantiles, minima and maxima from the ensemble, as well as relative frequencies of precipitation as predictive covariates. In this study, first-guess ensemble forecasts derived from the LAF under consideration of a spatial neighborhood of 5×5 gridboxes are taken as covariates. These forecasts are moreover considered as uncalibrated ensemble forecasts, and the value of the postproccesing is determined by the added value over the first-guess forecasts. The covariates comprise

- the ensemble mean, denoted as first-guess mean (fgM),
- first-guess quantiles (fgQ$_\tau$) ranging from $\tau = 0.25$ to $\tau = 0.999$,
- the first-guess probability of precipitation[2] (fgPoP), and
- first-guess probabilities of threshold exceedance (fgPoT$_u$) for thresholds of 5 and 10 mm/12h.

Several combinations of predictors are tested for each threshold u and probability level τ separately. The identification of the most informative predictors will be presented in the next section.

To further improve the predictive performance of statistical models, many studies apply power transformations to precipitation accumulations before they enter the postprocessing. Wilks (2009) used the square root, Sloughter et al. (2007) the 3rd root and Hamill et al. (2004, 2008) the 4th root of precipitation. In preliminary tests, largest improvements in terms of skill scores could be obtained with a 3rd root transformation. The power transformation is applied to the target precipitation, as well as to the predictor precipitation in terms of fgM and fgQ$_\tau$. The power transformation is not applied within the log-normal GLM.

[2] The fgPoP is determined from the ensemble as the probability that precipitation *is equal or greater* than 0.1 mm. The threshold of 0.1 mm is used since this is the smallest amount of precipitation which is measurable by rain gauges. The fgPoP shows a significantly better predictive performance than fgPoT$_{u=0}$. However, as covariate for statistical postprocessing, the fgPoT$_{u=0}$ is a more informative predictor and is used here synonymous for fgPoP.

7.2. Predictive covariates

The performance of statistical models strongly depends on the choice of predictive covariates. Several combinations of predictors have been tested for LR and PoP/PoT with a threshold of 5 and 10 mm/12h, for QR with τ ranging between 0.25 and 0.999, and the mixture models.

The fgM is the most informative predictor for LR and all thresholds. The combination of fgM and the respective fgPoP/fgPoT leads to a small amount of additional skill for PoP forecasts, and does not influence the predictive performance of PoT forecasts. For QR, the fgM is again a good predictor, especially for the lower quantiles. However, higher quantiles perform better if the respective fgQ$_\tau$ is used as covariate. The combination of fgM and fgQ$_\tau$ lead to a good performance for all quantiles.

The mixture models need a more complex setup of covariates. The LR part of the mixture model is used to derive the PoP. Here, a combination of fgM and fgPoP is used as predictors which performed best in the previous analysis. For the GLM part (Gamma, log-normal, inverse-Gaussian) a combination of fgQ$_{0.9}$ and fgPoP lead to the best performance. Note that the identity link function is used for all distributions. The estimation of the GPD part of the mixture model also relies on fgQ$_{0.9}$ and fgPoP as predictive covariates.

The regression coefficients for all statistical models (LR, QR, and the mixtures) exhibit large variations between the different training samples of 50 days, and often follow a distinct seasonal cycle. The size of the training sample may become an issue for the extremal quantiles and the GPD parameter. A longer training period of 1065 days (i.e. 3 years without the 15 days for verification extended by the following 15 days) is tested for the LR-Gamma-GPD model. A sine and cosine is added as covariate, as well as interaction terms with the predictors. Hence, the linear predictor for a certain day of the year $d = 1, ..., 365$ has the form

$$\begin{aligned}\boldsymbol{\eta} =& \beta_1 + \beta_2 \sin(\phi_d) + \beta_3 \cos(\phi_d) + (\beta_4 + \beta_5 \sin(\phi_d) + \beta_6 \cos(\phi_d))\mathbf{x}_1 \\ & + (\beta_7 + \beta_8 \sin(\phi_d) + \beta_9 \cos(\phi_d))\mathbf{x}_2\,. \end{aligned} \tag{7.1}$$

Hereby, $\mathbf{x}_1$ and $\mathbf{x}_2$ are the respective covariates and $\phi_d = 2\pi D/d$ is a function of the day of year with $D = 365$. The regression coefficients for the GPD scale parameter and the shape parameter are shown in Fig. 7.2 for a 50-day gliding training and a 1065-day training period. The variability within the seasonal cycle is largely reduced for the longer training period, leading to a stable estimation of the regression coefficients throughout the year. Moreover, the larger training data lead to a more robust and stable shape parameter estimate, which is nearly constant over the year. However, the extended training period provides much more stable parameter estimates, but only lead to a small increase in the predictive performance, which will be discussed in the next section.

7.3. Predictive performance

7.3.1. First-guess forecasts and calibration with LR/QR

The fgPoP and fgPoT forecasts for thresholds of 5 and 10 mm/12h show positive skill in terms of the Brier skill score (BSS). A stationwise climatology is used as reference forecast. The BSS lies

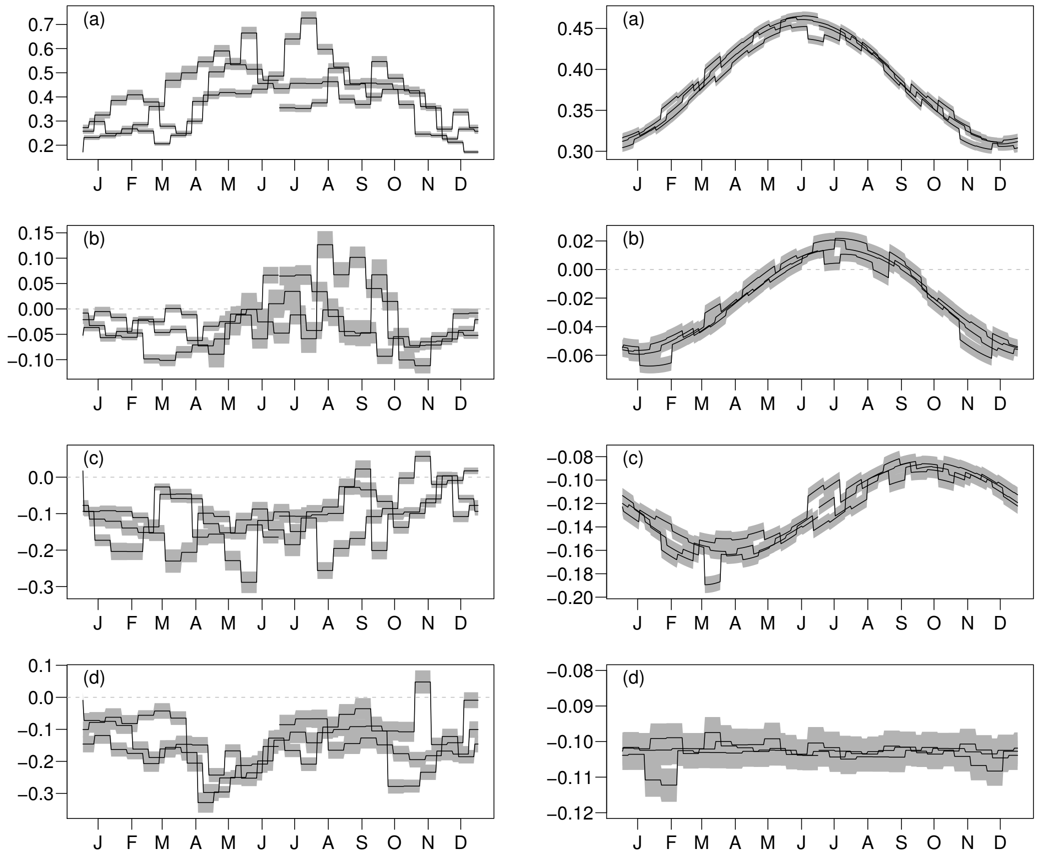


Figure 7.2.: *Left panel:* Temporal evolution of (a) the predictive intercept, the regression coefficients for (b) fgQ$_{0.9}$ and (c) fgPoP for the scale parameter, and (d) the shape parameter of the GPD with a 50-day training period. The gray shading indicates the standard error of the parameter estimates. The three lines refer to the three years of data used. *Right panel:* Same as left panel, but for the 1065-day training period. (Fig. 13 and 14 from Bentzien and Friederichs, 2012).

between 20-25% for fgPoP and fgPoT$_5$ and is about 10% for fgPoT$_{10}$ (see Fig. 2 from Bentzien and Friederichs, 2012). However, the variability between the 445 stations is large, ranging between 0% to 45%. The number of outliers (stations with negative skill) becomes larger for higher thresholds. Calibration with LR improves the predictive performance of first-guess forecasts. The benefit is largest for PoP, and the BSS of the calibrated forecasts ranges between 40% to 60% for all stations. The variability between stations is largely reduced. A strong improvement in the reliability as well as a significant increase in resolution is obtained from the postprocessing for PoP. For higher thresholds, the gain in predictive performance through LR mainly results from a better calibration, while the resolution of PoT$_5$ and PoT$_{10}$ is similar to the first-guesses.

First-guess quantile forecasts for τ ranging between 0.25 and 0.999 show very different forecast performances (see Fig. 6 from Bentzien and Friederichs, 2012). The quantile skill score (QSS) gives the percental improvement with respect to a stationwise climatology. Nearly no skill is obtained for the lower quantile fgQ$_{0.25}$. Since the mean PoP amounts to 39%, the 0.25-

quantiles is frequently censored. However, calibration with QR results in a QSS of about 10%, leading to skillful forecasts at nearly all stations. The QSS for $\mathrm{fgQ}_{0.5}$ ($\mathrm{fgQ}_{0.75}$/$\mathrm{fgQ}_{0.9}$) increases to 25% (45%/50%). Postprocessing only slightly affects the predictive performance for this range of quantiles. However, the skill of first-guess quantiles decreases for $\tau > 0.9$, and guidance for extreme quantiles ($\tau > 0.95$) cannot be given by the raw ensemble. QR is necessary to obtain skillful quantile forecasts for τ between 0.95 and 0.999. The QSS of the calibrated 0.999-quantile forecasts varies between 20% and 70%. The variations between the stations become larger for higher quantiles.

We summarize that the first-guess ensemble forecasts from COSMO-DE-LAF show a certain degree of skill and can be useful to obtain probabilistic guidance for precipitation. However, postprocessing largely improves the performance of PoP forecasts, and yields a better calibration of PoT forecasts. Postprocessing of quantile forecasts is indispensable for higher quantiles with $\tau > 0.9$, but also for the lower 0.25-quantile.

7.3.2. Parametric mixture models

A mixture model aims to provide the complete predictive distribution based on a small number of parameters. Quantiles and probability forecasts for all probability levels τ or thresholds u can be calculated directly from the distribution parameters. The CRPS is a general measure of forecast performance for distributional forecasts. Since the CRPS averages over the whole range of the distribution ($0 \leq \tau \leq 1$), differences within certain parts of the distribution remain undetected. A more complete picture for the performance of distributional forecasts is given by an evaluation with respect to various thresholds or probability levels.

Quantile forecasts from the different mixtures (Gamma, log-normal, inverse-Gaussian) are compared to those derived from QR. The most promising results are obtained for the Gamma mixture. The QSS is comparable to QR for a range of τ between 0.25 and 0.95. For higher τ, the QSS of the Gamma mixture decreases compared to QR, which is mainly due to a systematic overestimation of precipitation. Although the log-normal mixture has a similar QSS compared to the Gamma mixture for τ between 0.25 and 0.9, its predictive performance decreases more rapidly for higher probability levels. The inverse-Gaussian distribution completely fails to represent the bulk of the distribution (τ between 0.75 and 0.99), but recovers skill for the very high quantiles. Among the mixture models, the Gamma model provides the best performance in terms of the QSS. However, the QR approach is still superior to the mixture models, especially for higher quantiles. This can be seen from Fig. 7.3 (left panel), which shows the three-month moving average of the QSS of QR minus the QSS of the Gamma mixture. The differences are positive for all seasons, and vary over the year with the lowest predictive performance during spring and early summer. The bias increases with τ, which results from a general overestimation of the tail of the Gamma mixture.

The adaptive GPD tail largely improves the Gamma mixture. With a slightly negative shape parameter[3] (compare Fig. 7.2(d)), the GPD corrects for the overestimation of precipitation and lead to a predictive performance of higher quantiles which is similar to QR. The bias in QSS is largely reduced, as can be seen from Fig. 7.3 (right panel). The mean differences in QSS

[3] Although a negative shape parameter indicates an upper bound on extreme precipitation, an interpretation in terms of physical mechanisms should be avoided.

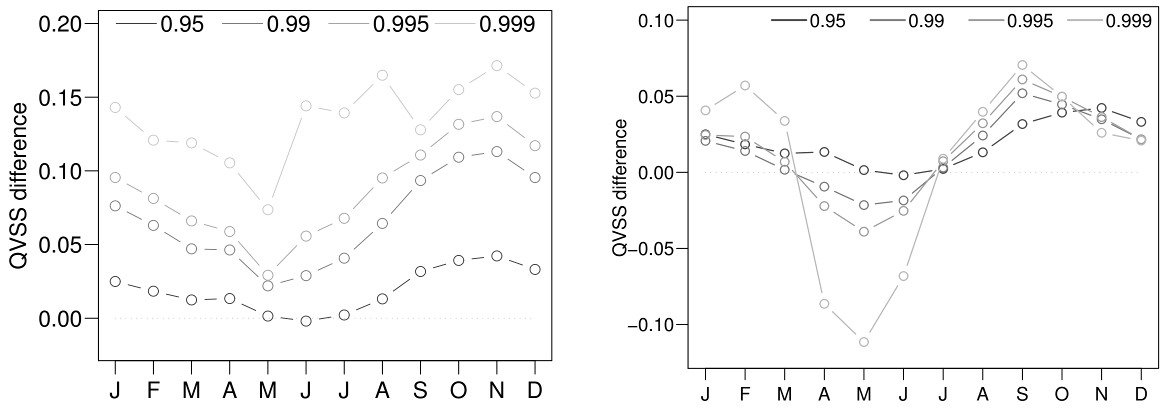


Figure 7.3.: *Left:* Three-month moving average of QSS for QR minus QSS for the LR-Gamma mixture. *Right:* Three-month moving average of QSS for QR minus QSS for the LR-Gamma-GPD mixture. (Fig. 11 and 15 from Bentzien and Friederichs, 2012)

between QR and the Gamma mixture range between zero and 15%. The differences for the GPD tail are largely reduced and are of the order of ±5%.

The mixture model using a gamma distribution with an adaptive GPD tail is an appropriate parametric alternative that allows for an extrapolation towards high quantiles. However, none of the parametric mixtures outperforms LR or QR. It depends on the user's needs which choice of postprocessing is most appropriate. If a complete predictive distribution is required, e.g. in order to sample from this distribution, the LR-Gamma-GPD is an appropriate alternative to LR and QR.

8. Evaluation of COSMO-DE-EPS

We now turn to the evaluation of the COSMO-DE ensemble prediction system using the data set II described in Sec. 6.2. This chapter starts with an evaluation of the raw ensemble system using rank statistics and the beta score in Sec. 8.1. Probabilistic forecasting of precipitation focuses on predictions in terms of probabilities and quantiles. Based on the results in Chapter 7, calibrated forecasts are obtained from logistic and quantile regression models. The performance of probability and quantile forecasts from COSMO-DE-EPS compared to the benchmark system COSMO-DE-LAF are presented in Sec. 8.2 and Sec. 8.3. This chapter closes with a summary and conclusion in Sec. 8.4.

8.1. Ensemble consistency

Probability integral transform histograms are constructed for the COSMO-DE-EPS, the COSMO-DE-TLE$_{sub}$, and the COSMO-DE-TLE. The PIT values are determined from the empirical distribution function for each day and site separately during the one-year time period of investigation. Hence, the histograms are build from a total of $\sim 385\,000$ forecast-observation pairs and are displayed in Fig. 8.1.

The COSMO-DE-EPS reveals a largely U-shaped histogram (Fig. 8.1(a)). The observations are lying too often outside the ensemble forecast range, indicating a general underestimation of ensemble spread. Moreover, observations are lying more often below the ensemble forecasts, leading to a small positive bias. The PIT histogram shows several peaks, indicating that observations are ranked frequently between different groups of members. This groups are build by ensemble members which are driven by the same boundary conditions. The different driving models contribute more to the ensemble spread (on average) than the different physical parameterizations.[1] A much flatter PIT histogram is obtained for the 20-member TLE$_{sub}$ in Fig. 8.1(b). Here, all members have different initial and boundary conditions either from a different driving model or a different initialization time. The underestimation of ensemble spread is largely reduced, and the beta scores decreases from -0.669 for the EPS to -0.358 for the TLE$_{sub}$. However, the TLE$_{sub}$ reveals a stronger positive bias than the EPS. The histogram for the 80-member TLE in Fig. 8.1(c) shows less overpopulation of the outer values, and therefore a further increase in ensemble spread due to the additional members. Moreover, the bias is largely reduced, which means that precipitation is less overestimated. The beta bias reduces from 0.041 for the EPS to 0.021 for the TLE. The time-lagged ensemble thus has a strong impact on ensemble consistency. It leads to a better representation of ensemble spread and also reduces the bias in precipitation forecasts compared to COSMO-DE-EPS.

[1] Note that this conclusion is only valid for the current setup of COSMO-DE-EPS as illustrated in Fig. 3.2 and as an average over one year of data. For a specific event, the physical disturbances may have a larger contribution to the ensemble spread than the driving models.

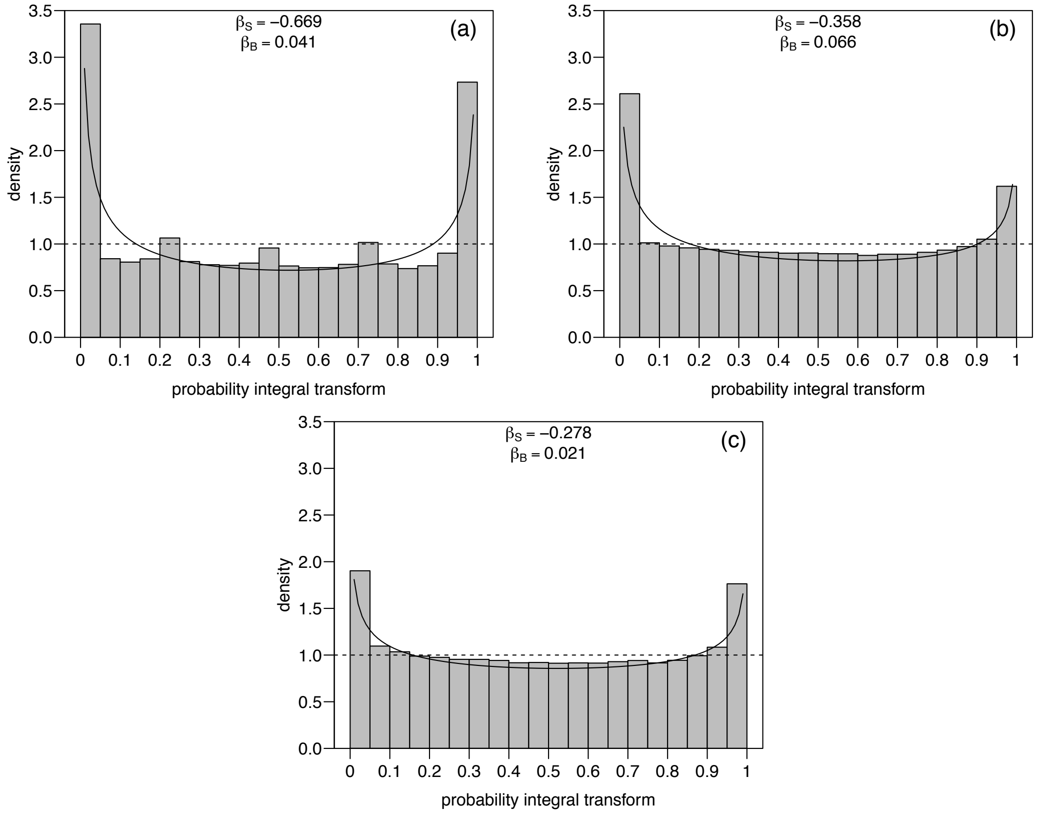


Figure 8.1.: Probability integral transform histogram and beta score for (a) COSMO-DE-EPS, (b) 20-member COSMO-DE-TLE$_{sub}$ and (c) 80-member COSMO-DE-TLE. The solid line shows the beta distribution fitted to the histogram values. The dashed line shows the uniform distribution.

To further investigate the influence of the different ensemble members, various sub-ensembles are constructed. The three ensemble categories are given by members with the same initialization time, with the same driving model, or the same model physics. An overview is given in Table 8.1, together with the respective beta scores. All ensembles reveal a negative beta score and a positive beta bias. Hence, all ensembles underestimate the ensemble spread and generally overestimate precipitation. Category I comprises the COSMO-DE-EPS for different forecast lead times (0-12h, 3-15h, 6-18h, 9-21h). The 20 members of each ensemble have the same initialization time, but differ with respect to the driving global model and model physics. Longer forecast lead times show a better representation of ensemble spread as the beta score increases (i.e. becomes less negative) from -0.669 for EPS$_{12UTC}$ to -0.476 for EPS$_{3UTC}$. In contrast, shorter lead times have smaller biases and hence show less overestimation of precipitation.[2]

Ensembles in category II are time-lagged ensembles, each build from 5 members of the EPS which have the same driving model. The 20 members differ with respect to model physics and initialization time. The four ensembles TLE$_{IFS}$, TLE$_{GME}$, TLE$_{GFS}$, and TLE$_{GSM}$ perform similar in terms of the beta score, which lies between -0.54 and -0.57. A positive bias of ~ 0.06

[2] Note that forecast quality in general might depend on the time of the day when forecasts are initialized, which is not analyzed here.

Table 8.1.: Overview of the categories (I-III) of the various sub-ensembles and their respective beta score and beta bias. The total number of members in each ensemble is given by M.

Cat.	COSMO-DE	beta score	beta bias	member	init time	M
	TLE	-0.278	0.021	1-20	12,09,06,03	80
	TLE$_{sub}$	-0.358	0.066	1,7,13,15,19	12,09,06,03	20
I	EPS$_{\mathrm{12UTC}}$	-0.669	0.041	1-20	12	20
	EPS$_{\mathrm{9UTC}}$	-0.552	0.053	1-20	09	20
	EPS$_{\mathrm{6UTC}}$	-0.501	0.057	1-20	06	20
	EPS$_{\mathrm{3UTC}}$	-0.476	0.061	1-20	03	20
II	TLE$_{\mathrm{IFS}}$	-0.561	0.033	1-5	12,09,06,03	20
	TLE$_{\mathrm{GME}}$	-0.572	0.060	6-10	12,09,06,03	20
	TLE$_{\mathrm{GFS}}$	-0.544	0.058	11-15	12,09,06,03	20
	TLE$_{\mathrm{GSM}}$	-0.551	0.058	16-20	12,09,06,03	20
III	TLE$_{\mathrm{P1}}$	-0.366	0.126	1,6,11,16	12,09,06,03	16
	TLE$_{\mathrm{P2}}$	-0.378	0.070	2,7,12,17	12,09,06,03	16
	TLE$_{\mathrm{P3}}$	-0.373	0.073	3,8,13,18	12,09,06,03	16
	TLE$_{\mathrm{P4}}$	-0.383	0.057	4,9,14,19	12,09,06,03	16
	TLE$_{\mathrm{P5}}$	-0.385	0.065	5,10,15,20	12,09,06,03	16

is obtained for nearly all ensembles of cat. II, with the exception of the TLE$_{\mathrm{IFS}}$. Boundary conditions obtained from the IFS model lead to a much smaller beta bias of 0.033. While cat. I and II ensembles show similar performances, they are still inferior compared to TLE and TLE$_{sub}$.

Category III comprises ensembles where all members share the same model physics, but use different driving models and initialization times. Note that ensembles of category III consist of only 16 members, where 4 members are taken from COSMO-DE-EPS, complemented by the time-lagged model runs. The ensembles TLE$_{\mathrm{P1}}$ to TLE$_{\mathrm{P5}}$ perform similar in terms of the beta score, which lies between -0.36 and -0.39. Hence, the performance is better than for the cat I and II ensembles, and is comparable to the TLE$_{sub}$. However, the physical representations used for the cat. III ensembles lead to different biases, ranging from 0.126 for TLE$_{\mathrm{P1}}$ to 0.057 for TLE$_{\mathrm{P4}}$.

We can summarize that the members of COSMO-DE-EPS show deficiencies in representing sufficient ensemble spread and generally overestimate precipitation. Time-lagging increases the ensemble spread, and different boundary conditions contribute more to the spread than variations of model physics. The impact on the bias is different for the various ensemble members. Members driven by the IFS model show general a smaller bias, while the representation of model physics in P1 leads to a larger bias. The bias is widely reduced for the COSMO-DE-TLE. However, as Hamill and Colucci (1997) already pointed out, there is potential for statistical postprocessing to generate calibrated forecasts, even if the ensemble is underdispersive and biased.

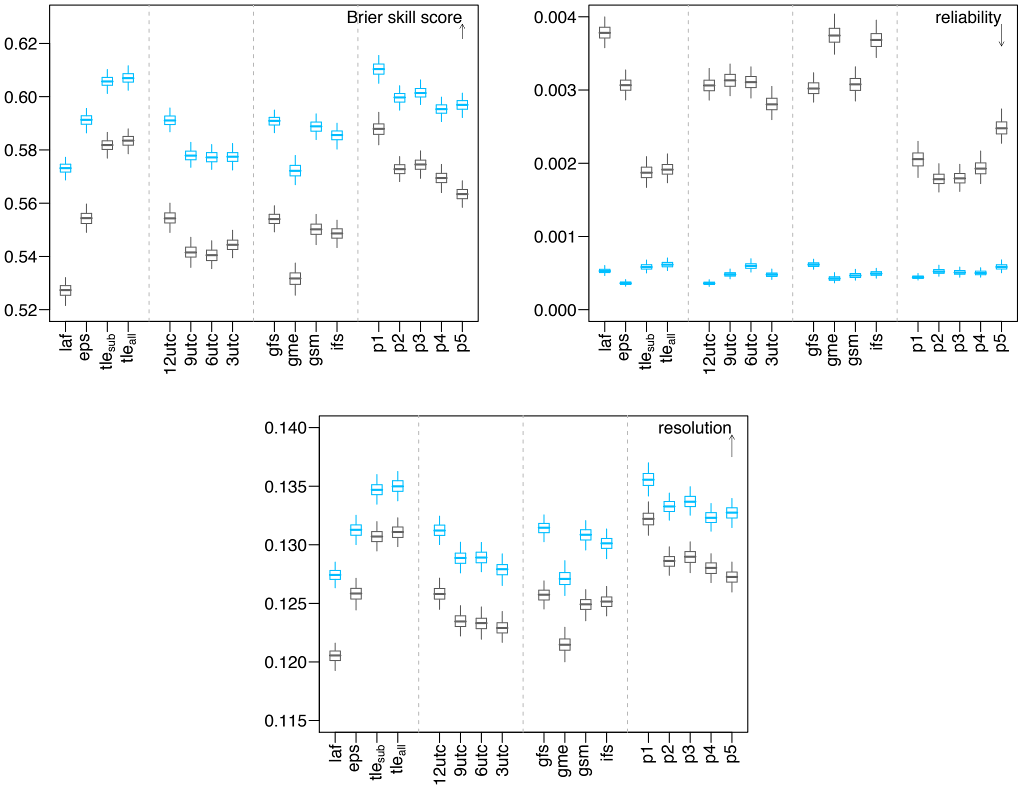


Figure 8.2.: BSS, reliability and resolution for fgPoP (gray boxplots) and calibrated PoP forecasts (blue boxplots) derived from various ensembles. The BSS is calculated with reference to a station-wise climatology. The boxplots show the 95% confidence interval of the scores estimated via 7-day block-bootstrapping.

8.2. Probability forecasts

Probabilistic forecasting in terms of probabilities is done using first-guess forecasts on the one hand and calibrated forecasts derived from LR on the other hand. First-guess forecasts for PoP and PoT are derived from the various ensembles in Tab. 8.1, as well as the COSMO-DE-LAF as benchmark ensemble. A LR model is set up using a cross-validation technique that makes use of the complete data. The one-year data set is divided into blocks of 21 days. Forecasts for each block are obtained from a training period which consists of all other blocks. The seasonal cycle is taken into account by a sine and cosine wave, as well as interaction terms with the covariates (see equation 7.1). Based on the results from Bentzien and Friederichs (2012), the covariates are taken by the fgM and and the respective fgPoP/fgPoT. The postprocessing is now based on observational sites from a larger geographical region (Germany, data set II). Therefore the elevation of the observational sites is included as a third covariate. The elevation is a stationary covariate which accounts for spatial inhomogeneities within the extended model domain.

Figure 8.2 shows the BSS, the reliability, and the resolution component for fgPoP and calibrated PoP forecasts derived from the different ensembles. The scores are averaged over all time steps and observational sites. The sampling uncertainty is estimated via 7-day block-bootstrapping (Efron and Tibshirani, 1993) with 1000 replicates simultaneously for all stations,

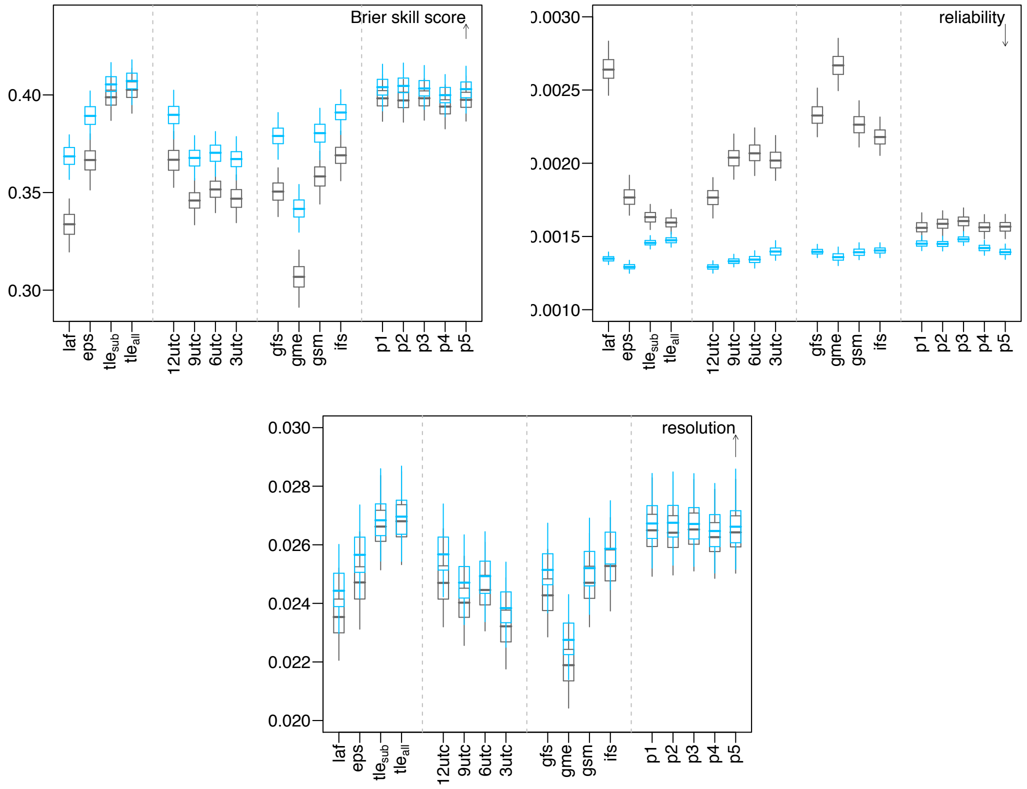


Figure 8.3.: Same as Fig. 8.2 but for PoT_5.

thereby preserving spatial and temporal correlations. The boxplots show the 95% confidence interval of the score. The arrows denote the orientation of the score. The reliability is negatively orientated (i.e. the smaller the better, with zero for a perfect forecast), whereas the resolution and the BSS are better the higher the values.

All forecasts in terms of fgPoP or calibrated PoP show positive skill with respect to a station-wise climatology. The BSS ranges between 52% and 62%. The calibrated PoP forecasts show a better predictive performance than the first-guess forecasts, and the benefit lies between 2% to 4%. The improvement is due to a better reliability as well as an increase in resolution. The EPS outperforms the benchmark ensemble LAF with a significant gain in reliability and resolution. The benefit becomes even larger for the time-lagged ensembles TLE and TLE_{sub}, which both show a similar performance despite their differences in ensemble size. After calibration with LR, the differences between the ensembles become smaller, indicating that statistical post-processing can account for a lack of calibration of the raw ensembles. LR also increases the resolution, since more predictive covariates (and hence more information) can be included into the statistical model. However, after calibration with LR, the EPS shows the best reliability, whereas both TLE and TLE_{sub} have the highest resolution.

We now analyze the performance of COSMO-DE-EPS for different forecast lead times (cat. I ensembles). Fig. 8.2 shows that the youngest forecast run EPS_{12UTC} has the highest BSS, mostly due to a much better resolution. That indicates that the youngest forecast run provides the most information. The older forecast runs have lower skill in terms of BSS, but perform similar

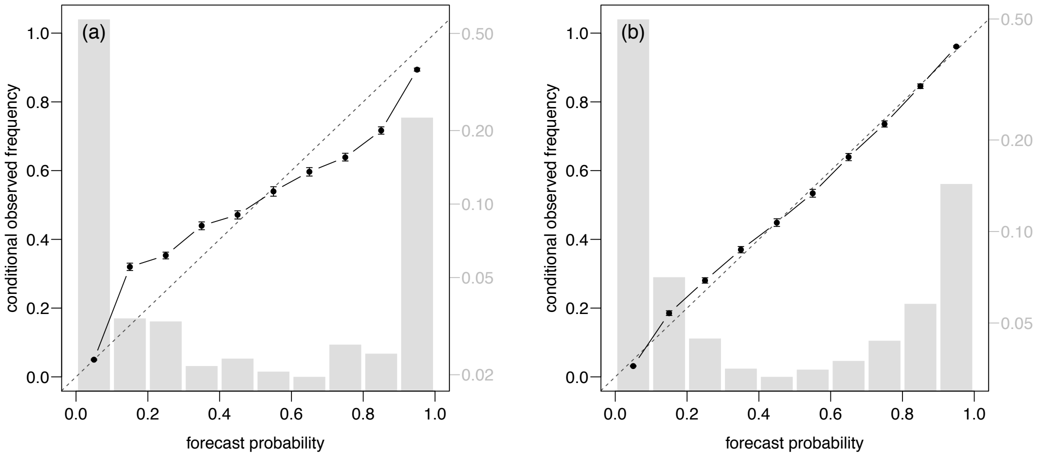


Figure 8.4.: Reliability diagram for (a) fgPoP and (b) calibrated PoP forecasts from COSMO-DE-EPS. The error bars show the 95% confidence interval of the observed frequencies conditional on each of the 10 forecast probabilities (estimated via 7-day block-bootstrapping). The barplot refers to the frequency of forecasts in each bin. A total of $\sim 385\,000$ pairs of observations and forecasts are used for each diagram.

compared to each other. However, the fgPoP of the oldest forecast run EPS$_{\mathrm{3UTC}}$ has a significant better reliability than the younger forecast runs. This might be due to a better representation of ensemble spread as can be seen from the beta score in Tab. 8.1. However, LR strongly affects the reliability part, and the youngest forecast run performs significant better after postprocessing. Note that even when the older forecast runs have a less good performance than EPS$_{\mathrm{12UTC}}$, their combination within a time-lagged ensemble largely improves the overall forecast skill.

The predictive skill of the category II ensembles varies with the driving model. The weakest performance in terms of the BSS is obtained for the TLE$_{\mathrm{GME}}$, with 2% less skill than the other three ensembles. The resolution is largely reduced. The differences in predictive performances between TLE$_{\mathrm{GFS}}$, TLE$_{\mathrm{GSM}}$, an TLE$_{\mathrm{IFS}}$ are much smaller. The reliability shows a weaker performance of both TLE$_{\mathrm{GME}}$ and TLE$_{\mathrm{IFS}}$. However, the differences in reliability are largely removed after postprocessing. Although LR does also slightly increase the resolution, the differences between the 4 global models remain the same.

We have already seen that different initial and boundary conditions from different driving models or time-lagged forecasts contribute more to the ensemble spread than different formulations of model physics. The category III ensembles thus show a better predictive performance for both first-guess and calibrated forecasts compared to cat. I and II ensembles. The BSS for fgPoP varies between 58% for TLE$_{\mathrm{P1}}$ to 56% for TLE$_{\mathrm{P5}}$. The TLE$_{\mathrm{P1}}$ shows the highest BSS, mostly due to a better resolution. The slightly worse reliability might be due to a larger bias as indicated by the beta bias in Tab. 8.1. LR increases the skill of PoP forecast, and differences in the reliability are removed. The resolution of all ensembles is increased after postprocessing, but the differences between the ensembles remain again the same.

The results for PoP can be generalized to PoT forecasts. First, the BSS for PoT$_u$ forecasts decreases with increasing threshold u, since higher thresholds are generally less predictable. The BSS is about 50% for $u = 1$ mm, between 30% and 40% for $u = 5$ mm, and about 20%

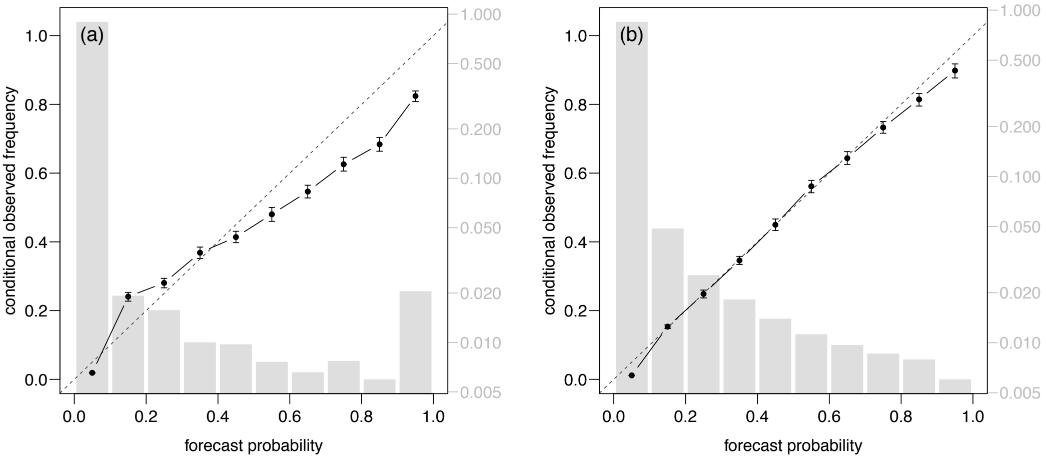


Figure 8.5.: Same as Fig. 8.4 but for PoT_5.

for $u = 10$ mm (not shown). Second, the differences between the various ensembles is similar to the PoP forecasts. The TLE and TLE$_{sub}$ ensemble are superior to the EPS and show the best predictive skill. The cat. III ensembles generally perform better than ensembles of cat. I and II. Third, the benefit of postprocessing (i.e. the increase in BSS) is smaller for PoT compared to PoP. The benefit mainly results from a better calibration, while the resolution component is less affected for higher thresholds. Fig. 8.3 shows the BSS, reliability and resolution for a threshold of 5 mm/12h as illustration.

Although the improvement in terms of the BSS seems to be small, the impact on the reliability through postprocessing becomes remarkable. The impact is stronger for the LAF and EPS than for the TLE, which is already better calibrated. Calibrated forecasts can be derived from all ensembles, regardless of the lack of reliability of the first-guess forecasts. Reliability diagrams for fgPoP and calibrated PoP from COSMO-DE-EPS are shown in Fig. 8.4. The reliability curve of fgPoP indicates an underforecasting of smaller forecast probabilities and an overforecasting of higher probabilities. These deficiencies are removed by LR, and good calibrated forecasts are obtained with a reliability curve which is close to the diagonal. The reliability diagram for a threshold of 5 mm/12h is shown in Fig. 8.5. The first-guess forecasts lack calibration for probability forecasts greater than 40%. Here, the ensemble overestimates the exceedance probability. LR reduces the overestimation of higher forecasts, and lead again to a reliability curve which is close to the diagonal.

Higher thresholds become less predictable, and often lack calibration for high probability values (not shown). The reliability diagram is strongly affected by the sample size. Since high probability forecasts of extreme thresholds are very rare, the estimation of conditional observed frequencies becomes uncertain. If the forecast intervals are not sufficiently represented (e.g. as revealed by the sharpness diagram), conclusions about the calibration cannot be made.

8.3. Quantile forecasts

We will now turn to quantile forecasts. First-guess quantiles from the ensembles are compared to calibrated quantile forecasts derived from QR. The QR model uses training and verification

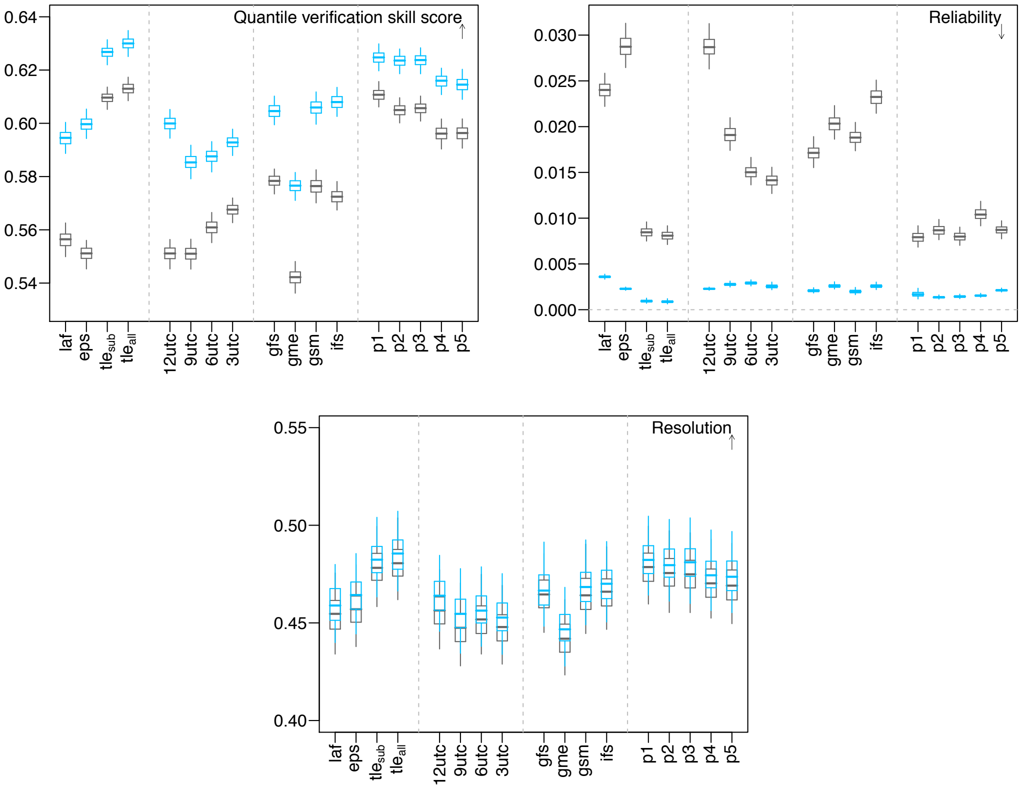


Figure 8.6.: QSS, reliability and resolution for fgQ$_{0.9}$ (gray boxplots) and calibrated 0.9-quantile forecasts (blue boxplots) derived from various ensembles. The QSS is calculated with reference to a stationwise climatology. The boxplots show the 95% confidence interval of the scores estimated via 7-day block-bootstrapping.

periods analog to the LR model. The predictive covariates are the fgM, the respective fgQ$_\tau$, and the elevation of observational sites, as well as the seasonal cycle. Fig. 8.6 shows the QSS for the 0.9-quantile forecasts derived from the different ensembles. The 0.9-quantile yields the highest predictive skill from a range of τ between 0.25 and 0.999, and is used by weather forecasters for an estimation of quantitative precipitation. The decomposition is estimated after Bentzien and Friederichs (2014) with an equi-distributed binning of forecasts into 30 bins. The boxplots show again the 95% confidence interval of the scores averaged over the time and observational sites, estimated via 7-day block-bootstrapping.

Both the TLE and TLE$_{sub}$ largely improve the predictive skill of fgQ$_{0.9}$. The QSS for LAF and EPS lies between 54% and 56%, and for both TLE between 60% and 62%. The TLE's show both an increase in resolution and a largely reduced reliability component compared to LAF and EPS. Although calibration improves the reliability for all ensembles, there are still significant differences between the ensembles. Reliability curves are therefore presented in Fig. 8.7 for the LAF, EPS, TLE$_{sub}$ and TLE. The deficiencies in calibration of fgQ$_{0.9}$ result from a general underestimation of quantiles, which becomes quiet visible in the double-logarithmic representation. The underestimation is significantly reduced for the TLE compared to LAF and EPS, especially for the lower range of forecast values. QR compensates for the underestimation, but lead to

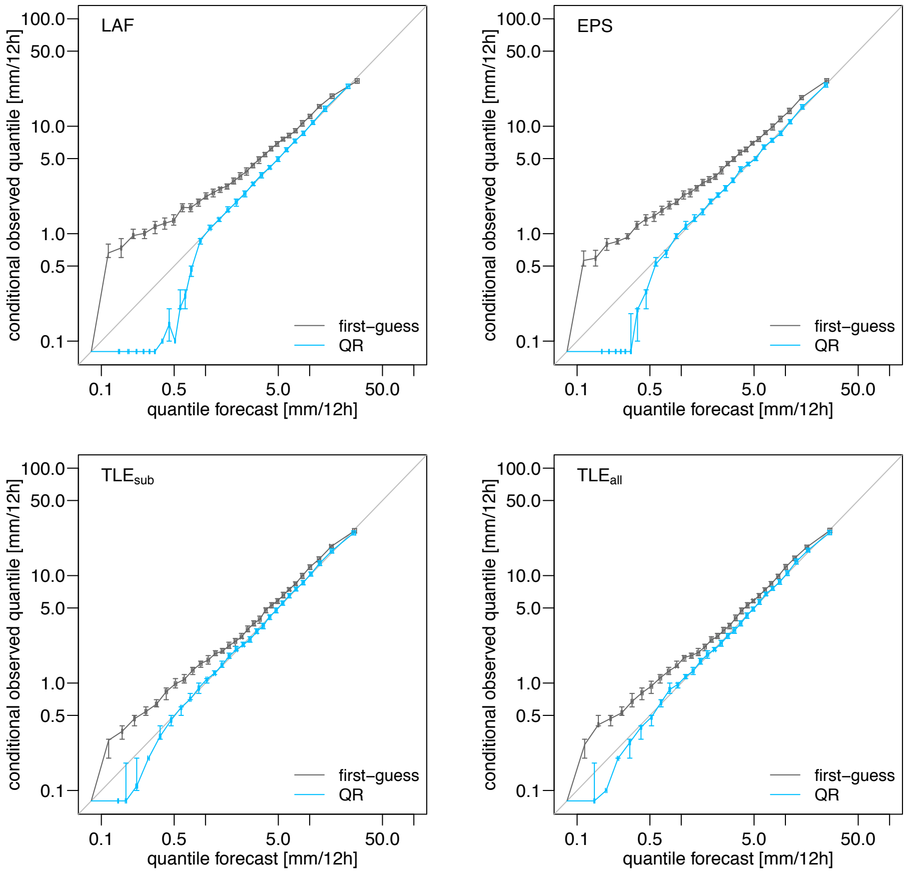


Figure 8.7.: Reliability diagram for fg$Q_{0.9}$ (gray lines) and calibrated 0.9-quantile forecasts (blue lines) derived from COSMO-DE-LAF, COSMO-DE-EPS, COSMO-DE-TLEPS$_{sub}$, and COSMO-DE-TLE. The error bars show the 95% confidence interval of the observed conditional quantiles on each of the 30 discretized forecast values (estimated via 7-day block-bootstrapping). A total of $\sim 385\,000$ pairs of observations and forecasts are used for each diagram.

a miscalibration of very small forecast values. After postprocessing, quantile forecasts are well calibrated for values above 1 mm for LAF and EPS, and values above 0.5 mm for TLE. Smaller values are slightly overestimated and less frequently censored.

We turn back to Fig. 8.6 and briefly discuss the predictive performance of the cat. I-III ensembles. The predictive skill of fg$Q_{0.9}$ varies with forecast lead time. Longer forecast lead times yield a better predictive performance which is mainly due to a better reliability. Postprocessing again accounts for the lack of calibration, and has a strong impact especially on the youngest forecast run EPS$_{12\text{UTC}}$. The QSS of the different driving models is very similar for TLE$_{\text{GFS}}$,TLE$_{\text{GSM}}$,TLE$_{\text{IFS}}$, while the TLE$_{\text{GME}}$ shows again a significant lower predictive performance. This is mainly due to a much lower resolution. The ensembles show some differences in reliability, which are widely removed after postprocessing. The cat. III ensembles show again a general better predictive performance than the cat. I and II ensembles. The QSS ranges between 59% and 61% for the first-guess, and between 61% and 63% for the calibrated quantile forecasts. Hence, the skill of TLE$_{\text{P1}}$ to TLE$_{\text{P5}}$ is much better than for LAF and EPS. While QR mainly affects the calibration of quantile forecasts, the impact on the resolution component is small for all ensembles. QR leads to a small increase of resolution, but differences between the ensembles remain the same.

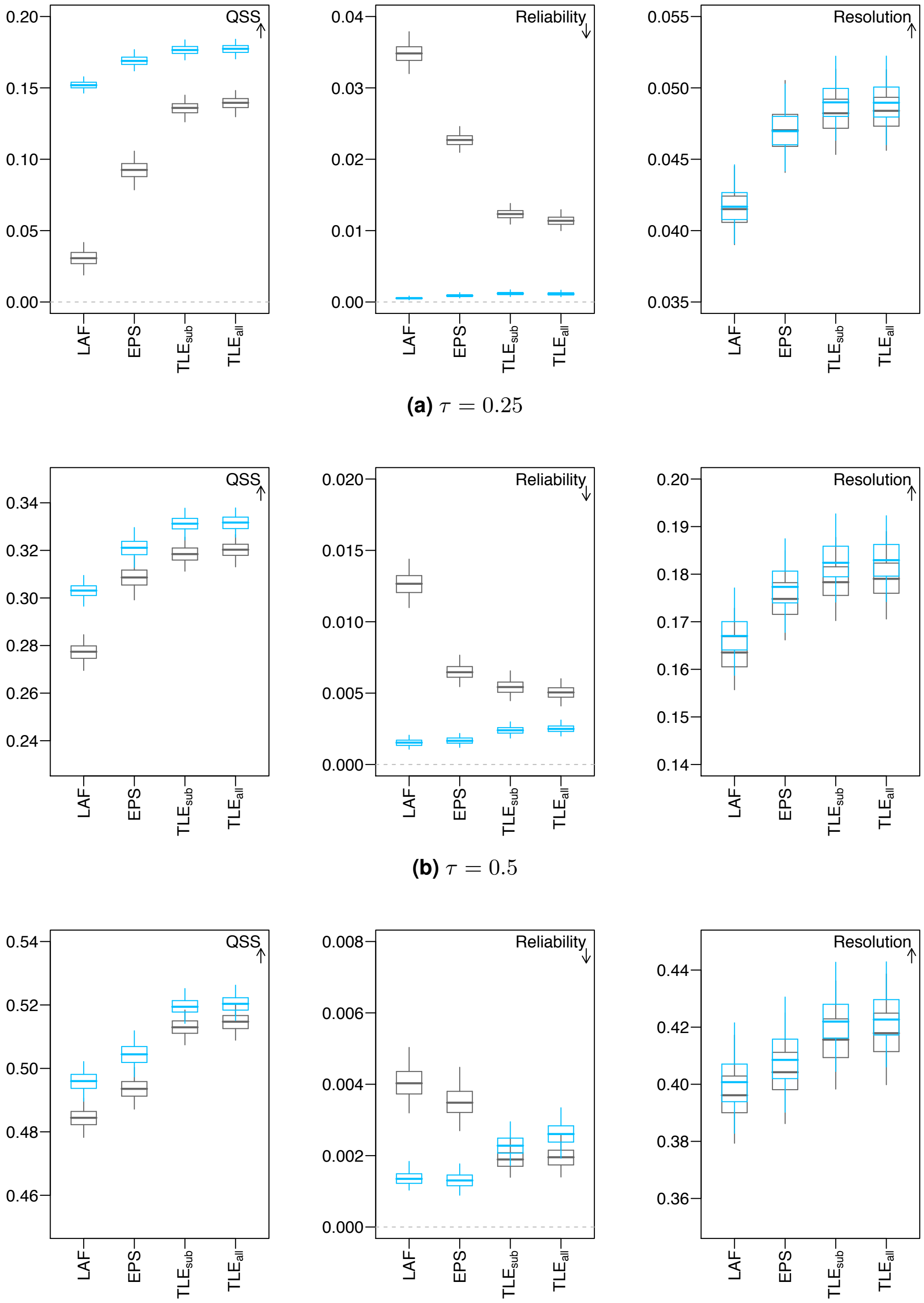
In the following, we will concentrate on the LAF, EPS and TLE ensembles. Fig. 8.8 shows the predictive skill of quantile forecasts for τ between 0.25 and 0.75. Note that the mean PoP of all observations is about 33%, and lower quantiles are frequently censored. Generally, the QSS increases for higher probability level, ranging between 3% to 18% for the 0.25-quantile, about 30% for the median, and up to 50% for the 0.75-quantile. The skill of fgQ$_{0.25}$ is quite low for the LAF, with a significantly lower resolution. The EPS outperforms the benchmark system, with both a better calibration and higher resolution. The predictive performance of fgQ$_{0.25}$ is further improved by the TLE and TLE$_{sub}$, with a large gain in QSS compared to the EPS. QR largely improves the performance of 0.25-quantile forecasts mainly due to a better reliability. The differences between the ensembles become smaller for higher τ. The impact of QR is smaller for the median and 0.75-quantile which are already well represented by the first-guesses. However, the TLE and TLE$_{sub}$ are still superior in predicting quantiles. Although the TLE and TLE$_{sub}$ show the best predictive skill and the highest resolution, they have a slightly worse reliability compared to calibrated forecasts from LAF and EPS. Again, postprocessing has only a slight effect on the resolution component for all quantile forecasts.

The predictive skill of extreme quantile forecasts for $\tau = 0.99$ and $\tau = 0.999$ is shown in Fig. 8.9. Note that first-guess quantiles are estimated from a sample of 100 (500/2000) values[3] for the LAF (EPS and TLE$_{sub}$ / TLE). Sampling uncertainty will become an issue for the predictive performance of first-guess forecasts. The QSS for the fgQ$_{0.99}$ amounts to 10% for LAF and EPS, and up to 50% for TLE and TLE$_{sub}$, which both have better reliability and higher resolution. Postprocessing is indispensable for LAF and EPS, and also improves forecasts for both TLE. Calibrated forecasts yield a QSS of over 60%. The first-guess forecasts fail to predict the 0.999-quantile, and skillful forecasts can only be obtained by QR. The TLE and TLE$_{sub}$ are again superior in predicting extreme quantiles. They show 10% more skill than forecasts from LAF or EPS.

Quantile reliability curves for the TLE are discussed in Bentzien and Friederichs (2014) and are shown here in Fig. 8.10. The double-logarithmic scale is chosen again for enhanced visibility. Since the fgQ$_{0.999}$ shows no predictive skill, it is omitted from the plot and results are only shown for the calibrated 0.999-quantile forecasts. The fgQ$_{0.75}$ is already well calibrated, and the reliability curve is close to the diagonal. For $\tau > 0.75$, the first-guess quantiles are significantly underestimated. The fgQ with $\tau < 0.75$ are in turn largely overestimated. Only the zero quantile forecasts are well calibrated for almost all probability levels. However, the miscalibration of fgQ forecasts is a consequence of the underrepresentation of ensemble spread.

The potential of calibration is larger for lower and higher τ, and only small improvements can be obtained for $\tau = 0.75$. QR compensates the overestimation of lower quantiles and the underestimation of higher quantiles. For τ between 0.25 and 0.75, the reliability curves are now close to the diagonal for forecast values above 0.3 mm. It remains a slight miscalibration of smaller quantile values, which was already discussed for the 0.9-quantile. QR widely reduces the underestimation of fgQ$_{0.99}$, but the reliability curve still shows small deviations from the diagonal. Forecast values up to 2 mm are still slightly overestimated. Higher forecasts values

[3] The number of values is given by the number of ensemble members multiplied by the size of the spatial neighborhood.

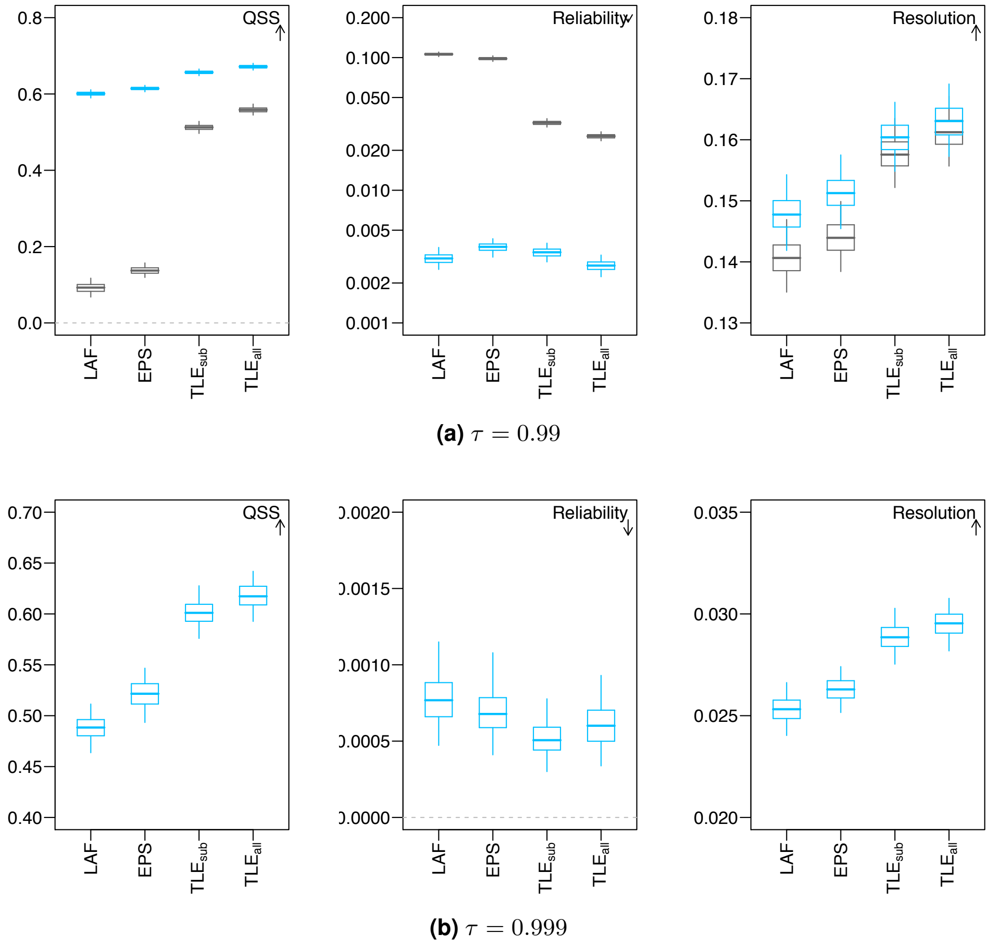


(a) $\tau = 0.25$

(b) $\tau = 0.5$

(c) $\tau = 0.75$

Figure 8.8.: QSS, reliability and resolution for fgQ_τ (gray boxplots) and calibrated quantile forecasts (blue boxplots) derived from COSMO-DE-LAF, COSMO-DE-EPS, COSMO-DE-TLEPS$_{sub}$, and COSMO-DE-TLE. The probability level τ is set to 0.25, 0.5, and 0.75.



(a) $\tau = 0.99$

(b) $\tau = 0.999$

Figure 8.9.: Same as Fig. 8.8, but for extreme quantiles with τ set to 0.99 and 0.999.

are well calibrated except for the highest category above 40 mm. These forecasts are now significantly overestimated, which might be an effect of the large compensation through QR. The 0.999-quantile shows moderate calibration, but the sampling uncertainty is very large. The average bin size for the estimation of the QS decomposition and hence for the reliability curve is shown in Fig. 8.11. The first bin contains all censored quantile forecasts, i.e. all quantile forecasts equal to zero. The number of censored forecasts N_0 is shown in the left panel for τ between 0.25 to 0.999. For $\tau > 0.5$, first-guess quantile forecasts are frequently more censored than the calibrated quantile forecasts. Quantile forecasts above zero are divided into 30 equi-distributed forecast bins. The mean size of each bin N_k is shown in the right panel of Fig. 8.11. Due to the censoring, the bin size is very mall for the 0.25-quantile (2000-3000 values), and increases with τ. Higher quantiles are thus estimated from a larger bin size. Although the 0.999-quantile is estimated from a sample of ~12000 values, this might still not large enough to obtain robust estimates for the extreme quantile.

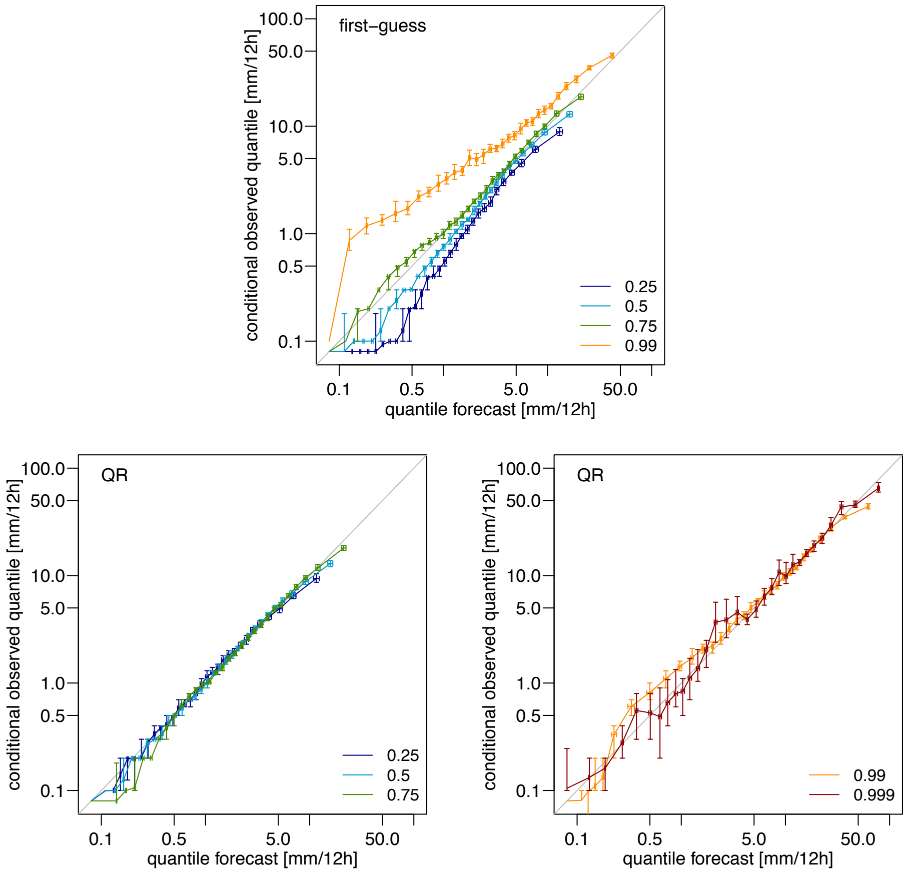


Figure 8.10.: Reliability diagram for first-guess (upper panel) and calibrated quantile forecasts (lower panel) derived from COSMO-DE-TLE. τ ranges between 0.25 and 0.999 (see legend). The error bars show the 95% confidence interval of the observed conditional quantiles on each of the 30 discretized forecast values (estimated via 7-day block-bootstrapping).

8.4. Conclusion

In this chapter, forecasts from the mesoscale ensemble system COSMO-DE-EPS are evaluated and compared to the benchmark system COSMO-DE-LAF. Moreover, the impact of an enlarged ensemble which includes forecast from longer lead times (time-lagged members) is assessed. The probabilistic forecasts emphasize probability forecasts for threshold excess as well as quantiles forecasts. First-guess forecasts are derived from the ensemble under consideration of a spatial neighborhood of 5×5 gridboxes.

The results show that the COSMO-DE-EPS generally outperforms the COSMO-DE-LAF. The higher predictive skill results from a significant gain in resolution and a better calibration for nearly all first-guess probability and quantile forecasts. The enlarged ensemble COSMO-DE-TLE yields to a further significant improvement of probabilistic forecasts. The time-lagged members largely increase the resolution of the EPS, and lead to even better calibrated forecasts. It is remarkable that most of the benefit of the TLE can also be obtained from a smaller ensemble (TLE$_{sub}$), where only 5 members of the EPS are chosen, complemented by their respective time-lagged members. Note that the 5 members are arbitrarily chosen, and the TLE$_{sub}$ might be

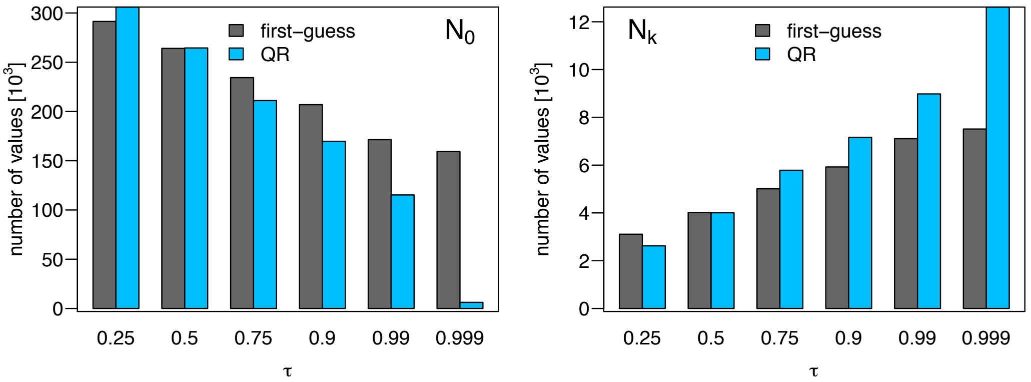


Figure 8.11.: Average bin size for the estimation of the QS decomposition from COSMO-DE-TLE. The first bin contains all censored quantile forecasts (N_0). The remaining forecasts are divided in $K = 30$ equi-distributed bins of size $\bar{N}_k$.

optimized by selecting another set of EPS members.

An investigation of various sub-ensembles reveals the contribution of the different ensemble members. Boundary conditions from various driving models contribute more to the ensemble spread on average than different model physics. Time-lagged members are also members with different boundary conditions, and thus largely contribute to the ensemble spread. The results for the various ensembles is in close connection with the investigations of the PIT histograms in Sec. 8.1. Ensembles with a better beta score (i.e. a beta score closer to zero), show a better predictive performance of probability and quantile forecasts. The improvement is mainly due to both a better calibration and an increase in resolution.

Postprocessing in terms of LR and QR largely improves the calibration of probabilistic predictions. The differences in the reliability of the various ensembles are largely reduced. Although first-guess forecasts show a moderate skill, postprocessing becomes indispensable for extreme quantile forecasts. While calibrated forecasts can be obtained for nearly all ensembles, the impact on the resolution is different. LR and QR generally increase the resolution part of the score. The effect is stronger for the occurrence of precipitation then for higher threshold exceedance, and only slightly affects the resolution of quantile forecasts. However, both LR and QR can not overcome structural deficiencies in resolution within the ensemble, and differences between the ensembles remain the same after postprocessing. One way of increasing the resolution might be to include other meteorological variables, and hence more information, into the statistical model.

Probability and quantile forecasts represent point estimates from the distribution, either with respect to different thresholds or to different probability levels. An advantage of the quantile view is that no prior knowledge of the data is needed to define the probability levels, whereas probability forecasts require the knowledge of the range of the data to define meaningful thresholds. In this sense, quantiles may be more useful particularly for extremal levels. The 0.99-quantile for example always lies in the tail of the distribution, but a threshold of 10 mm may be extreme in some regions and very normal in others. Quantiles can be displayed graphically in a boxplot and are a very intuitive way to communicate uncertainty to users.

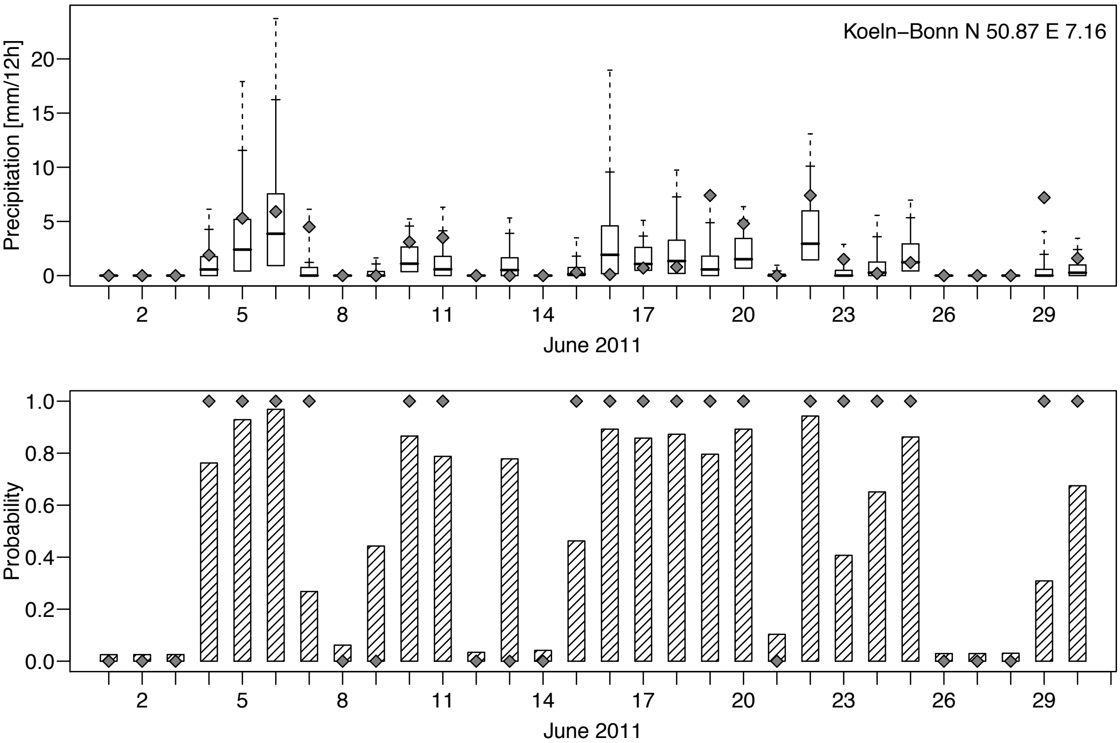


Figure 8.12.: Quantitative precipitation forecasts from COSMO-DE-TLE for Cologne-Bonn airport in June 2011. The boxplots in the upper panel represent the interquartile range (boxes), and the 0.9- and 0.95-quantiles (whiskers). The lower panel shows the probability of precipitation. The dots are the observations in mm/12h (upper panel) and as zero and one for precipitation below and above 0.1 mm.

Fig. 8.12 shows quantitative precipitation forecasts from COSMO-DE-TLE in terms of quantiles, complemented by the probability of precipitation. Precipitation at Cologne-Bonn airport in June 2011 is dominated by a series of convective events, especially around the 5th and 20th of June. The forecasting systems successfully distinguishes between days with the potential for high precipitation events and days without rainfall. Higher quantile forecasts characterize the risk of extreme precipitation events. Note that there is still a 5% chance that the 0.95-quantile will be exceeded, as happened here on the 29th of June.

A representation of probabilistic quantitative precipitation predictions in terms of the PoP and quantiles is highly recommended. A further investigation of quantile forecasts within a Bayesian framework will follow in the next chapters. This involves the exploration of more meteorological variables as predictive covariates, as well as the predictive performance of spatial quantile forecasts. So far a spatially constant relationship between covariates and point level measurements was assumed. Predictions can thus easily be interpolated to locations which are not included into the training data. The predictive performance of such spatial predictions will be investigated, and compared to a spatial modeling of regression coefficients.

Part III.

Bayesian postprocessing

9. Bayesian quantitative precipitation quantile prediction B(QP)^2

Quantile forecasts are a very useful and intuitive way to predict quantitative precipitation together with the uncertainty. QR largely improves the reliability of quantile forecasts, and thus has a strong positive impact on the predictive performance. However, small miscalibrations of quantile forecasts remain which cannot be captured by the classical QR model. This might partly be due to the formulation of the censored QR. The three step procedure described in Sec. 5.2.2 might lead to some biases in the regression parameters which affects the predictive performance. Miscalibrations might also result from the spatial homogeneous postprocessing. Moreover, the resolution of quantile forecasts depends on the set of covariates. Although the regression ansatz allows to include more predictive covariates, the selection of informative variables remains difficult.

All these issues (uncertainty, variable selection, spatial modeling) can be addressed within a Bayesian framework. Fundamental in Bayesian analysis is the treatment of model parameters (i.e. regression coefficients) as random variables. This allows for the incorporation of prior knowledge about the parameters into the statistical model. Moreover, the hierarchical structure of Bayesian models supports the formulation of complex data models, which is necessary for the spatial modeling. A comprehensive overview about Bayesian statistics and data analysis can be found in Banerjee et al. (2004); Gelman et al. (2004); Clark and Gelfand (2006), amongst others.

This chapter starts with a brief introduction to Bayesian modeling in Sec. 9.1. The formulation of a Bayesian QR model is provided in Sec. 9.2. Special prior distribution allow for the selection of informative covariates from a large number of variables. A spatial QR model is described in Sec. 9.3.

9.1. Bayesian inference

In Bayesian statistics, observations and unknown model parameters are treated as random variables. Let $\mathbf{y}$ denote the data vector of length N (number of observations) and let $\boldsymbol{\theta}$ be the vector of unobservable model parameters of length P (e.g. number of regression coefficients). The joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$ can be factorized by the conditional sampling distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$ (also denoted as likelihood), and the prior distribution of $\boldsymbol{\theta}$

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, .$$

Bayes theorem is applied to obtain the posterior distribution, which is given by the distribution of the unobservable parameter vector $\boldsymbol{\theta}$ conditional on the historic data $\mathbf{y}$

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} .$$

The marginal distribution of the observation $p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ can often not be calculated analytically. It is independent of the parameter vector $\boldsymbol{\theta}$ and only acts as a normalization constant. Inference about the posterior distribution can be done using the unnormalized posterior density (Gelman et al., 2004)

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{y})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y} \mid \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} . \tag{9.1}$$

The posterior distribution describes our knowledge about the model parameters given the historic data and hence also characterizes the uncertainty. External knowledge or expert opinion is incorporated by the specification of the prior distribution. The Bayesian model is specified by the likelihood of the data (e.g. a parametric distribution function) and the prior distribution of the parameters. It is a very flexible model which offers a common construction and analysis for a wide range of applications (Clark and Gelfand, 2006).

9.1.1. Hierarchical modeling

The factorization of the joint probability distribution enables the user to include more complex layers into the Bayesian model, also known as hierarchical modeling. Clark and Gelfand (2006) describe the hierarchical model in terms of three entities: The data level describes the underlying process from which observations are drawn (i.e. the likelihood). The process level specifies the model parameters $\boldsymbol{\theta}$ and typically involves unknown hyperparameters $\boldsymbol{\lambda}$. The hyperparameters are specified by a prior distribution. In this sense, the joint distribution is split into the hierarchical layers

$$\begin{aligned} p(\text{data}, \text{process}, \text{parameters}) \propto\, & p(\text{data} \mid \text{process}, \text{parameters}) \\ & \times p(\text{process} \mid \text{parameters}) \\ & \times p(\text{parameters}) . \end{aligned}$$

A common application of hierarchical models is the spatial modeling of the model parameters. $\boldsymbol{\theta}$ is described by a spatial latent processes which in turn depends on a parameter vector $\boldsymbol{\lambda}$. The conditional probability $p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ might be given by a multivariate normal distribution, and $\boldsymbol{\lambda}$ consists of parameters describing its expectation and covariance. The posterior of the hierarchical model is obtained as

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_\lambda p(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})p(\boldsymbol{\lambda})d\lambda . \tag{9.2}$$

The hyperprior $p(\boldsymbol{\lambda})$ must be defined and specifies the statistical properties of the parameters for the process level $p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$.

9.1.2. Markov Chain Monte Carlo

The Bayesian framework offers a very flexible construction and less limitations in the complexity of statistical models. However, the posterior in (9.1) or (9.2) can often not be calculated analytically. Numerical approximations are generally obtained by Markov Chain Monte Carlo (MCMC) techniques. A comprehensive overview of MCMC is given in e.g. Gilks et al. (1996). A Markov chain is a sequence of random variables $\{\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, ...\}$, such that the next state $\mathbf{z}_{r+1}$ is sampled from a distribution $q(\mathbf{z}_{r+1} \mid \mathbf{z}_r)$ which depends only on the current state of the chain. $q(. \mid .)$ is called the transition kernel of the chain.

The aim of MCMC is to construct a Markov chain such that its stationary distribution is precisely the posterior distribution of interest, i.e. $p(\boldsymbol{\theta} \mid \mathbf{y})$. This can be done using the Metropolis-Hastings algorithm, based on the work of Metropolis et al. (1953) and Hastings (1970). Starting from an initial value $\boldsymbol{\theta}_0$, the iterative process is described by repeating the following steps for $r = 1, 2, ..., R:$

- A candidate of $\boldsymbol{\theta}^\star$ is drawn from a proposal distribution $q(. \mid \boldsymbol{\theta}_{r-1})$, which only depends on the last value $\boldsymbol{\theta}_{r-1}$ of the chain.

- The acceptance probability is calculated from the ratio

$$\alpha(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_{r-1}) = \min\left(1, \frac{p(\boldsymbol{\theta}^\star \mid \mathbf{y}) q(\boldsymbol{\theta}_{r-1} \mid \boldsymbol{\theta}^\star)}{p(\boldsymbol{\theta}_{r-1} \mid \mathbf{y}) q(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}_{r-1})}\right) .$$

- For $\alpha = 1$, the candidate is accepted and $\boldsymbol{\theta}_r = \boldsymbol{\theta}^\star$. Otherwise, $\boldsymbol{\theta}^\star$ is only accepted with probability α, and $\boldsymbol{\theta}_r$ is set to

$$\boldsymbol{\theta}_r = \begin{cases} \boldsymbol{\theta}^\star & \text{with probability } \alpha\,, \\ \boldsymbol{\theta}_{r-1} & \text{with probability } 1 - \alpha\,. \end{cases}$$

After a sufficient number of iterations (burn-in period), the Markov chain will converge to the correct stationary distribution independent of the form of the proposal distribution $q(. \mid .)$. However, the choice of the proposal will affect the length of the burn-in period and the rate of convergence of the chain. An important tuning parameter is the proposal variance. A small variance lead only to small steps of the candidate, and the chain will converge very slowly. For large variances, the chain is characterized by jumps followed by long gaps where the chain does not move for several iterations. Several studies suggest to chose the proposal variance such that 30% to 50% of the candidates are accepted (e.g. Clark and Gelfand, 2006; Gelman et al., 2004).

A simplification of the Metropolis-Hastings algorithm is obtained if only symmetric proposals $q(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}_{r-1}) = q(\boldsymbol{\theta}_{r-1} \mid \boldsymbol{\theta}^\star)$ are considered. The acceptance probability is thus independent of $q(. \mid .)$ and given by the ratio of the posteriors

$$\alpha(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_{r-1}) = \min\left(1, \frac{p(\boldsymbol{\theta}^\star \mid \mathbf{y})}{p(\boldsymbol{\theta}_{r-1} \mid \mathbf{y})}\right) .$$

This is the original algorithm proposed by Metropolis et al. (1953).

9.2. Bayesian quantile regression

We will now turn to a Bayesian formulation of the quantile regression model. That requires the specification of an appropriate error distribution. Recall the classical regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \,, \quad i = 1, ..., N \,,$$

where the predictand y_i is explained by some covariates $\mathbf{x}_i \in \mathbb{R}^P$. For mean regression, the error distribution is assumed to be Gaussian with zero mean and some error variance $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. A generalization to quantile regression requires an error distribution which is defined such that its τ-quantile is equal to zero (c.f. Kneib, 2013)

$$\epsilon \sim f_\tau(.) \,, \quad \text{with} \int_{-\infty}^{0} f_\tau(\epsilon) d\epsilon = \tau \,.$$

An appropriate error distribution is the asymmetric Laplace distribution (ALP; Yu and Zhang, 2005), which is used by Yu and Moyeed (2001) to define a Bayesian quantile regression model. Other studies use mixtures of ALP and Dirichlet processes (e.g. Kottas and Krnjajic, 2009; Taddy and Kottas, 2010), or infinite mixtures of Gaussian distributions (e.g. Reich et al., 2010).

The Bayesian QR model of Yu and Moyeed (2001) can also be applied to censored variables as proposed by Yu and Stander (2007). The ALP is determined by three parameters, the location parameter μ, the scale parameter σ, and the asymmetry parameter τ which correspond to the probability level of the quantile

$$f_\tau(y; \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\} \,.$$

Yu and Stander (2007) use a simple form of the ALP with $\sigma = 1$. Inference with the scale parameter showed that the results do not vary with the choice of σ. Fig. 9.1(a) illustrates the ALP for different values of τ. The distribution is symmetric for $\tau = 0.5$ and shows a larger tail to higher (lower) values for $\tau < 0.5$ ($\tau > 0.5$). The likelihood function for the censored QR model is given by

$$L_\tau(\mathbf{y} \mid \boldsymbol{\beta}_\tau) = \tau^N (1-\tau)^N \exp\left\{-\sum_{i=1}^{N} \rho_\tau(y_i - \max(0, \mathbf{x}_i'\boldsymbol{\beta}_\tau))\right\} \,,$$

where the location parameter $\mu_i = \max(0, \mathbf{x}_i'\boldsymbol{\beta}_\tau)$ correspond to the censored quantile forecast. The likelihood depends on the unknown parameters $\boldsymbol{\theta} = \boldsymbol{\beta}_\tau$. The posterior distribution of the regression coefficients is build by the product of the likelihood and the prior distribution $p(\boldsymbol{\beta}_\tau)$

$$p(\boldsymbol{\beta}_\tau \mid \mathbf{y}) \propto L_\tau(\mathbf{y} \mid \boldsymbol{\beta}_\tau)\, p(\boldsymbol{\beta}_\tau) \,. \tag{9.3}$$

Since prior knowledge is generally not available, Yu and Stander (2007) use zero-mean Gaussian or Laplace distributions with large variances as prior distributions for $\boldsymbol{\beta}_\tau$. This corresponds to flat or uninformative priors and put little restrain on the regression coefficients.

We adopt the Bayesian QR model of Yu and Stander (2007), and use a MCMC procedure with a single-component Metropolis algorithm to obtain draws from the posterior in (9.3).

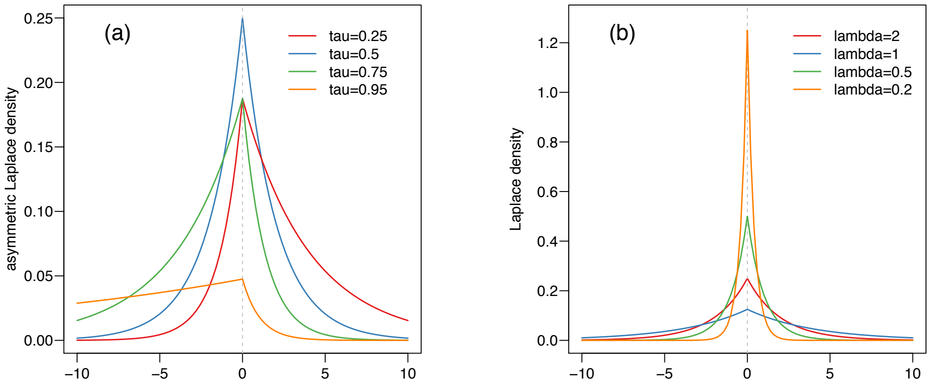


Figure 9.1.: Illustration of (a) the asymmetric Laplace density for different τ ($\mu = 0$, $\sigma = 1$) and (b) zero-mean Laplace density for different scale parameter λ.

Each regression coefficient is updated separately (single-component). The proposal density is a normal distribution with proposal variance υ_{β_j}. Within each MCMC iteration $r = 1, ..., R$, each regression coefficient $\beta_\tau^{(j)}$ with $j = 1, ..., P$ is updated as following:

- Draw a new candidate $\beta_\tau^{(j),*} \sim \mathcal{N}(\beta_\tau^{(j),r-1}, \upsilon_{\beta_j})$ from the proposal distribution, conditioning on the previous (initial) value $\beta_\tau^{(j),r-1}$.
- Calculate the acceptance probability $\alpha = \min\left(1, \frac{p(\boldsymbol{\beta}_\tau^* | \mathbf{y})}{p(\boldsymbol{\beta}_\tau^{r-1} | \mathbf{y})}\right)$ from the ratio of new and old posterior (Metropolis algorithm).
- Set $\beta_\tau^{(j),r} = \begin{cases} \beta_\tau^{(j),*} & \text{with probability } \alpha\,, \\ \beta_\tau^{(j),r-1} & \text{with probability } 1 - \alpha\,. \end{cases}$

After a sufficiently long burn-in period, the regression coefficients represent draws from the posterior $p(\boldsymbol{\beta}_\tau \mid \mathbf{y})$.

9.2.1. Variable selection

The selection of predictive covariates plays an important role for the performance of regression models. While regression models can improve the calibration of forecasts, the resolution strongly depends on the information content of the predictors. Several strategies have been developed to identify a set of informative covariates from a wide range of predictors. This includes stepwise backward or forward selection (see for example Fahrmeir and Tutz, 1994, chapter 4.1.2), or penalized regression techniques (Kyung et al., 2010). In the Bayesian framework, inference about variable selection can be done using appropriate prior distributions which put constrains onto the regression coefficients (analog to penalized regression). A Bayesian implementation of the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) is obtained by using independent zero-mean Laplacian priors for the regression coefficients

$$p(\boldsymbol{\beta}_\tau \mid \lambda) = \left(\frac{1}{2\lambda}\right)^P \exp\left\{-\frac{1}{\lambda}\sum_{j=1}^{P} \mid \beta_\tau^{(j)} \mid \right\}.$$

The zero-mean Laplace distribution only depends on the scale parameter λ and is illustrated in Fig. 9.1(b). The variance is given by $2\lambda^2$. Smaller values of λ correspond to less variance and a distribution which is sharper around zero. For a given λ, the posterior of the regression coefficients is given by

$$p(\boldsymbol{\beta}_\tau \mid \mathbf{y}) \propto L_\tau(\mathbf{y} \mid \boldsymbol{\beta}_\tau)\, p(\boldsymbol{\beta}_\tau \mid \lambda)\,.$$

A sharp prior distribution ($\lambda \ll 1$) forces the regression coefficients to be close to zero, unless they contribute significantly to the likelihood. Only the most informative covariates will have regression coefficients which are significantly unequal to zero. The scale parameter controls the degree of sparseness of the regression coefficients, and is also denoted as LASSO parameter. Li et al. (2010) studied Bayesian regularized quantile regression for non-censored data using the LASSO prior. An alternative approach is the Bayesian stochastic search variable selection framework, applied to censored quantile regression model by Ji et al. (2012). Most of the work concerning Bayesian quantile regression, including censored data or penalized regression, has been done using survival data in medical or biological applications. To the best of my knowledge, this is the first attempt to use a Bayesian quantile regression model for quantitative precipitation forecasts from short-range NWP.

9.3. Spatial quantile regression

So far, the regression coefficients $\beta_\tau^{(1)}, ..., \beta_\tau^{(P)}$ are spatially constant over the model domain. For the spatial model, each regression coefficient is now a function of the location vector $\mathbf{r} = (r_1, ..., r_S)'$, with S the number of observational sites. Hence, for each covariate with $j = 1, ..., P$, a vector of regression coefficients $\boldsymbol{\beta}_\tau^{(j)}(\mathbf{r})$ has to be modeled.

Reich et al. (2011) developed a Bayesian spatial quantile regression model for tropospheric ozone predictions under different climate scenarios. The regression coefficients are represented by a set of basis functions (Bernstein basis polynomials), with spatially varying basis function coefficients which are described by a latent Gaussian process. Another spatial quantile regression model was proposed by Lum and Gelfand (2012) introducing the asymmetric Laplace process.

This study proposes a spatial quantile regression model for quantitative precipitation which is based on the Bayesian hierarchical model by Cooley et al. (2007). In Cooley et al. (2007), return levels for extreme precipitation are estimated using a Generalized Pareto distribution. We adopt the hierarchical structure to our application using the Bayesian quantile regression model for censored variables. The observations $\mathbf{y}(\mathbf{r}, t)$ for each time step $t = 1, ..., T$ are considered as a partial realization of a spatial random process observed at fixed locations $\mathbf{r}$. The total number of forecasts and observations is given by $N = S \cdot T$. The predictive covariates are given by $\mathbf{x}_1(\mathbf{r}, t), ..., \mathbf{x}_P(\mathbf{r}, t)$. The spatial quantile forecast for location r_i and time step t is obtained by

$$q_\tau(r_i, t) = \max\left(0, \beta_\tau^{(1)}(r_i) x_1(r_i, t) + ... + \beta_\tau^{(P)}(r_i) x_p(r_i, t)\right)\,.$$

For simplification, we use the following notation:

- $\mathbf{Y} = (\mathbf{y}(\mathbf{r}, 1), ..., \mathbf{y}(\mathbf{r}, T))$: $S \times T$ matrix of observations;
- $\mathbf{B} = (\boldsymbol{\beta}_\tau^{(1)}(\mathbf{r}), ..., \boldsymbol{\beta}_\tau^{(P)}(\mathbf{r}))$: $S \times P$ matrix of spatially varying regression coefficients.

Data layer The likelihood of the data, given the regression coefficients, is described again by the asymmetric Laplace distribution

$$L_\tau(\mathbf{Y} \mid \mathbf{B}) = \tau^N (1-\tau)^N \exp\left\{ -\sum_{t=1}^{T} \sum_{s=1}^{S} \rho_\tau(y(r_s,t) - q_\tau(r_s,t)) \right\} .$$

Process layer Each of the regression coefficients $\boldsymbol{\beta}_\tau^{(1)}(\mathbf{r}), ..., \boldsymbol{\beta}_\tau^{(P)}(\mathbf{r})$ is represented by a Gaussian random field over the model domain

$$\boldsymbol{\beta}_\tau^{(j)}(\mathbf{r}) \sim \mathcal{MVN}(\mu_j \mathbf{I}_S, \boldsymbol{\Sigma}^{(j)}) , \quad j = 1, ..., P .$$

Here, $\mathbf{I}$ denotes the unit vector of length S. The multivariate Gaussian distribution is described by a spatially constant expectation μ_j and a parametric covariance function $\boldsymbol{\Sigma}^{(j)}$. We assume a stationary and isotropic covariance function which depends solely on the minimum distance between two locations $\mid r_r - r_s \mid$. The exponential covariance function is described by three parameters (e.g. Banerjee et al., 2004, Sec. 2.1)

$$(\boldsymbol{\Sigma}^{(j)})_{r,s} = \begin{cases} \sigma_j + \eta_j & \text{for } |r_r - r_s| = 0 \\ \sigma_j \exp(-\Phi_j |r_r - r_s|) & \text{otherwise} . \end{cases}$$

The parameters are the partial sill σ, the decay parameter Φ, and the nugget η. Note that the regression coefficients vary in space while the process parameters $\boldsymbol{\mu} = (\mu_1, ..., \mu_P)', \boldsymbol{\sigma} = (\sigma_1, ..., \sigma_P)', \boldsymbol{\eta} = (\eta_1, ..., \eta_P)'$ and $\boldsymbol{\Phi} = (\Phi_1, ..., \Phi_P)'$ are spatially constant.

Priors Prior distributions must be assigned to the parameter vectors $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\Phi}, \boldsymbol{\eta} \in \mathbb{R}^P$. The location parameters $\boldsymbol{\mu}$ are assumed to follow independent zero-mean normal distributions with high variance (flat prior)

$$p(\mu_j) \sim \mathcal{N}(0, \sigma_{\mu_j}^2) .$$

For the partial sill $\boldsymbol{\sigma}$, nugget $\boldsymbol{\eta}$ and decay $\boldsymbol{\Phi}$ parameters, independent inverse Gamma distributions are used as prior distributions.

Posterior The posterior distribution is given by the likelihood and the product of the prior distributions

$$p(\mathbf{B} \mid \mathbf{Y}) = L_\tau(\mathbf{Y} \mid \mathbf{B}) \prod_{j=1}^{P} \left\{ p(\boldsymbol{\beta}_\tau^{(j)}(\mathbf{r}) \mid \mu_j, \sigma_j, \Phi_j, \eta_j) p(\mu_j) p(\sigma_j) p(\Phi_j) p(\eta_j) \right\} . \tag{9.4}$$

Realizations of the posterior (9.4) are obtained by a MCMC procedure. Following the schematic overview in Fig. 9.2, each regression coefficient $j = 1, ..., P$ is updated separately. First, the parameters of the spatial process $\mu_j, \sigma_j, \Phi_j, \eta_j$ are updated one after another using a single-component Metropolis algorithm. Given the accepted process parameters, a new realization $\hat{\boldsymbol{\beta}}_\tau^{(j)}(\mathbf{r})$ of the Gaussian process is drawn conditional on the previous (initial) values of the re-

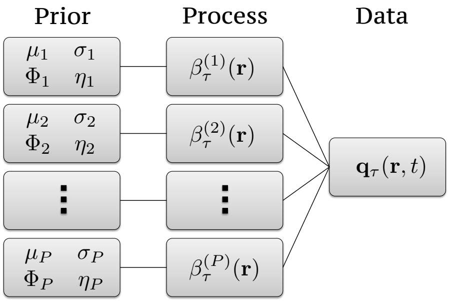

Prior
Process
Data
μ_1 σ_1 Φ_1 η_1
μ_2 σ_2 Φ_2 η_2
μ_P σ_P Φ_P η_P
$\beta_\tau^{(1)}(\mathbf{r})$
$\beta_\tau^{(2)}(\mathbf{r})$
$\beta_\tau^{(P)}(\mathbf{r})$
$\mathbf{q}_\tau(\mathbf{r},t)$

Figure 9.2.: Illustration of the Bayesian hierarchical model for spatial quantile regression. The quantile forecasts $\mathbf{q}_\tau(\mathbf{r},t)$ depend on the regression coefficients $\boldsymbol\beta_\tau^{(j)}$, with $j = 1,...,P$ the number of covariates. The regression coefficients are modeled spatially by a multivariate Gaussian process. The Gaussian process is described by a spatially constant expectation μ_j, and a parametric covariance function which is determined by the three parameters σ_j, Φ_j, η_j.

gression coefficients by a conditional simulation

$$\begin{aligned}\hat{\boldsymbol\beta}_\tau^{(j)}(\mathbf{r}) \mid \boldsymbol\beta_\tau^{(j)}(\mathbf{r}) \sim \mathcal{MVN}\Big(&\mu_j\mathbf{I}_S + \boldsymbol\Sigma^{(j)}\left(\mathcal{V}+\boldsymbol\Sigma^{(j)}\right)^{-1}\left(\boldsymbol\beta_\tau^{(j)}(\mathbf{r}) - \mu_k\mathbf{I}_S\right),\\ &\boldsymbol\Sigma^{(j)} - \boldsymbol\Sigma^{(j)}\left(\mathcal{V}+\boldsymbol\Sigma^{(j)}\right)^{-1}\boldsymbol\Sigma^{(j)}\Big).\end{aligned}$$

Here, $\mathcal{V}$ is a diagonal matrix with the proposal variance as entries on the diagonal and zeros off the diagonal. The new values $\hat{\boldsymbol\beta}_\tau^{(j)}(\mathbf{r})$ are accepted with probability α, which is estimated from the ratio of new to old posterior.

9.3.1. Spatial prediction

For a spatial prediction, the regression coefficients $\boldsymbol\beta_\tau^{(j)}(\mathbf{r})$ have to be interpolated to new spatial locations. Let s $= (s_1, ..., s_{S^\star})'$ be the locations for which we want to make predictions. A kriging approach (e.g. Banerjee et al., 2004, Sec. 2.4) is used to obtain the regression coefficients $\boldsymbol\beta_\tau^{(j)}(\mathbf{s})$. The joint distribution of regression coefficients for $\mathbf{s}$ and $\mathbf{r}$ is given by

$$\begin{pmatrix}\boldsymbol\beta_\tau^{(j)}(\mathbf{s})\\ \boldsymbol\beta_\tau^{(j)}(\mathbf{r})\end{pmatrix} \sim \mathcal{MVN}\left(\mu_j\mathbf{I}^{S+S^\star}, \begin{pmatrix}\boldsymbol\Sigma_{S^\star}^{(j)} & \boldsymbol\Sigma_{S^\star S}^{(j)}\\ \boldsymbol\Sigma_{SS^\star}^{(j)} & \boldsymbol\Sigma_{S}^{(j)}\end{pmatrix}\right).$$

The conditional expectation of $\boldsymbol\beta_\tau^{(j)}(\mathbf{s}) \mid \boldsymbol\beta_\tau^{(j)}(\mathbf{r})$, also denoted as kriging estimate, is given by

$$\mathbf{E}\left[\boldsymbol\beta_\tau^{(j)}(\mathbf{s}) \mid \boldsymbol\beta_\tau^{(j)}(\mathbf{r})\right] = \mu_j\mathbf{I}_{S^\star} + \boldsymbol\Sigma_{S^\star S}^{(j)}\left(\boldsymbol\Sigma_S^{(j)}\right)^{-1}\left(\boldsymbol\beta_\tau^{(j)}(\mathbf{r}) - \mu_j\mathbf{I}_S\right).$$

Realizations of $\boldsymbol\beta_\tau^{(j)}(\mathbf{s})$ are obtained by kriging estimates conditional on $\boldsymbol\beta_\tau^{(j)}(\mathbf{r})$ taken from the MCMC iterations.

10. Results for B(QP)^2

The Bayesian quantile regression model (BQR) and spatial quantile regression model (SQR) are explored for further enhancement of quantile forecasts from COSMO-DE-EPS. To keep the computational time at a reasonable limit, the evaluation is restricted to the summer half of the year 2011 (180 days between April and September). In preparation of a spatial forecast verification, the dataset is separated into training and verification data by stations. 457 stations are selected for statistical model training, while the remaining 622 stations are used for verification (spatial cross-validation).

10.1. Bayesian quantile regression

We start with the spatial homogeneous BQR as described in Sec. 9.2. First, the influence of more covariates on the predictive performance of quantile forecasts is investigated. While QR can largely improve the calibration, the resolution and hence the information content depends on the covariates. A broad range of variables is considered:

- covariates based on total precipitation (first-guess): mean, standard-deviation, PoP, and quantiles;
- other meteorological variables (mean and standard-deviation): CAPE, humidity divergence (tdiv_hum), total water content (twater), wind gusts 10m, as well as divergence, vorticity and omega at 850hPa;
- the station elevation.

The BQR model uses zero-mean Laplacian priors for the regression coefficients. The complexity of the model (i.e. the number of covariates) depends on the scale parameter λ. Small values of λ force a sparse selection of covariates. Most of the coefficients are centered around zero and only a few covariates are selected (e.g. have distributions which significantly differ from zero). Figure 10.1 shows the distribution of regression coefficients under a strong LASSO condition ($\lambda = 0.01$) for the estimation of the 0.9-quantile. A MCMC with $50\,000$ iterations produces realizations from the posterior of the regression coefficients. The chain converges after less than $5\,000$ iterations. The distribution for each regression coefficient is shown in Fig. 10.1 by the boxplots which are obtained from the last $10\,000$ iterations of the Markov chain. The LASSO selects the fgM, fgPoP, and the fgQ$_{0.99}$ from the total precipitation variables. Moreover, additional variables are identified as the standard-deviations of CAPE and wind gusts, as well as the mean of the total water content.

An overview about the selected variables for other quantiles is given in Tab. 10.1. Mainly three covariates are selected from the total precipitation variables for each τ. The fgPoP is

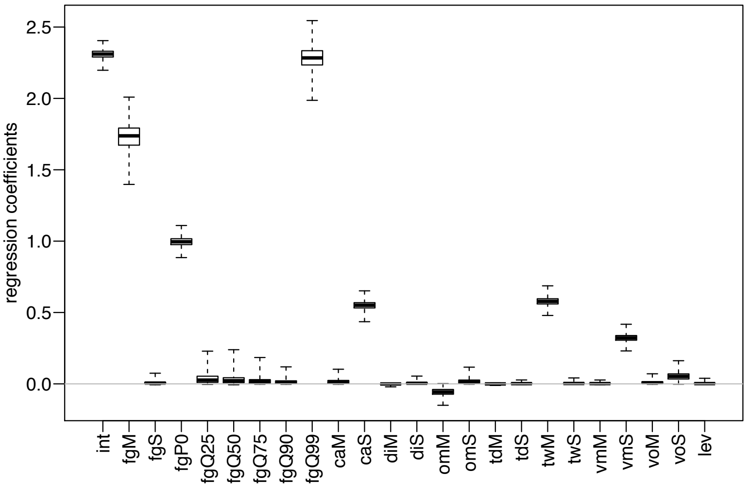


Figure 10.1.: Distribution of regression coefficients from BQR for the 0.9-quantile and a strong LASSO condition.

selected as predictor for all quantiles. The fgM is selected for quantiles with τ between 0.25 to 0.9, while the fgSd is selected for the higher quantiles with $\tau > 0.9$. The 0.25- and 0.5-quantile chose the fgQ$_{0.25}$ as covariate, whereas higher quantiles select the fgQ$_{0.99}$. Considering other meteorological variables, the lower quantiles profit from the wind gusts, while higher quantiles chose a combination of wind gusts, CAPE, and total water content.

Quantile forecasts are obtained from a non-penalized BQR model with λ set to 10. This corresponds to a flat prior for the regression coefficients. Forecasts are obtained from two sets of variables: The first set is based on the selected variables from total precipitation alone. The second set of covariates considers all selected variables from Tab. 10.1. Again, a MCMC with $50\,000$ iterations is used to obtain realizations from the posterior of the regression coefficients. The coefficients of 1000 iterations are used to perform predictions at the verifying stations for the 180 days. For verification, the QS and its decomposition is calculated for each of the 1000 forecasts schemes and is displayed in Fig. 10.2 (upper panel). The plot shows the resolution vs. the reliability, while the gray contours are lines of constant QS. A better predictive performance is indicated by a smaller reliability and a higher resolution component, and hence given by points which lie in the upper left corner of the plot region. The error bars denote the QS of the BQR and show the uncertainty due to the varying regression coefficients. For comparison, the blue squares show the QS for the classic QR model with the same set of covariates. Regression coefficients are estimated by the three-step procedure described in Sec. 5.2.2. The values for the 0.25-quantile are omitted from the plot, since they lie far beyond the values of the BQR.

We compare forecasts from BQR for the two sets of covariates in the upper panel of 10.2. Forecasts based on all selected variables show a significant better predictive performance than forecasts based solely on total precipitation. The gain in QS is mostly due to an increase in resolution. The information content is largely increased by the additional meteorological parameters. Moreover, the reliability of higher quantiles ($\tau > 0.5$) is improved. The classic QR also benefits from the additional covariates selected by the Bayesian LASSO. Forecasts from QR show a gain in resolution if more meteorological variables are included. The BQR model per-

Table 10.1.: Selected variables from the Bayesian LASSO for various quantiles with τ between 0.25 and 0.999.

$\tau =$	0.25	0.5	0.75	0.9	0.99	0.999
fg-mean	X	X	X	X		
fg-sd					X	X
fgPoP	X	X	X	X	X	X
fgQ25	X	X				
fgQ99			X	X	X	X
wind gusts (mean)	X	X				
wind gusts (sd)		X	X	X	X	
CAPE (sd)			X	X	X	X
twater (mean)			X	X	X	X
tdiv_hum (mean)					X	

forms similar to the classic QR model for forecasts based solely on total precipitation covariates. A benefit from BQR is obtained for the 0.25-quantile. Here, the restriction of the training data by the three step procedure leads to large biases in the regression coefficients for the classic QR model. However, for forecasts based on all selected covariates, the BQR yields a better predictive performance for nearly all quantiles (except the median which shows a similar performance to classic QR).

The lower panel of Fig. 10.2 shows quantile reliability plots for first-guess quantiles, as well as BQR and classic QR with all selected variables. First-guess quantiles generally overestimate low quantiles ($\tau < 0.5$) and underestimate higher quantiles ($\tau > 0.5$). Classic QR lead to a better calibration of quantiles with respect to the first-guesses. However, some miscalibration of small quantile values are obtained for all quantile levels. Small values are generally overestimated, as was already observed by the evaluation of the COSMO-DE-EPS in Section 8.3. In contrast, all BQR forecasts are well calibrated with reliability curves which are close to the diagonal. However, an overestimation of the highest forecast category can still be observed.

Table 10.2 shows the QSS for the BQR model with respect to a climatological forecasts (climatology at each station during the time period of investigation) and with respect to first-guess quantiles. BQR forecasts generally have positive skill. Compared to climatology, the improvement in skill ranges between 10% for $\tau = 0.25$ up to 60% for $\tau \geq 0.99$. The benefit of postprocessing can be seen from the QSS with first-guess as reference. The benefit is small for lower quantiles and ranges between 2 to 7% for τ between 0.25 and 0.75. For higher quantiles, the gain in predictive skill largely increases from 16% for the 0.9-quantile up to 90% for the extremal quantile. Here, postprocessing becomes indispensable to obtain skillful quantile forecasts from the ensemble.

Table 10.2.: Quantile skill score (in %) for BQR with all selected variables and with climatology and first-guess quantiles as reference forecasts.

$\tau =$	0.25	0.5	0.75	0.9	0.99	0.999
climatology	11.70	25.63	44.64	55.66	60.99	58.16
first-guess	6.74	2.16	4.61	16.37	63.59	91.21

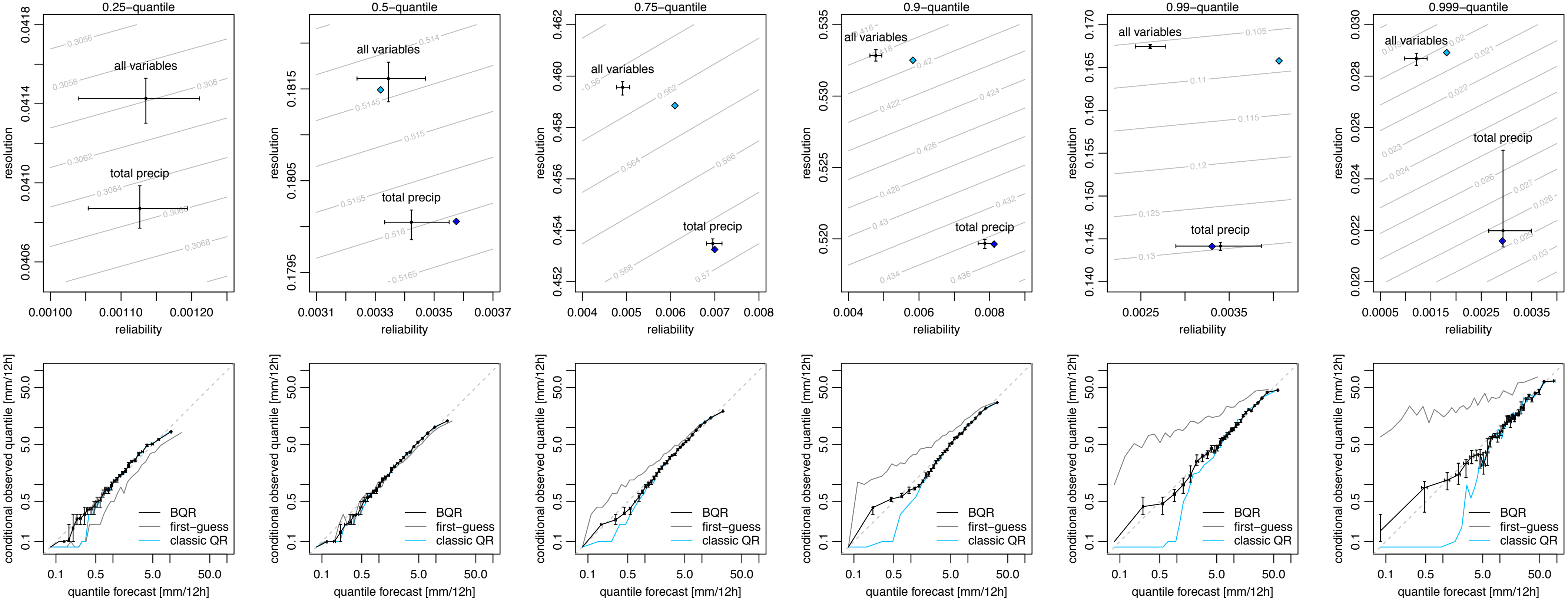


Figure 10.2.: Upper panel: QS decomposition into reliability and resolution for various quantile forecasts derived from BQR (black lines) and classic QR (blue squares) for two sets of covariates. The error bars show the 95% confidence interval for BQR. The gray contours show lines of constant QS. Lower panel: Reliability diagram for BQR and classic QR with all selected covariates and for the first-guess quantile forecasts.

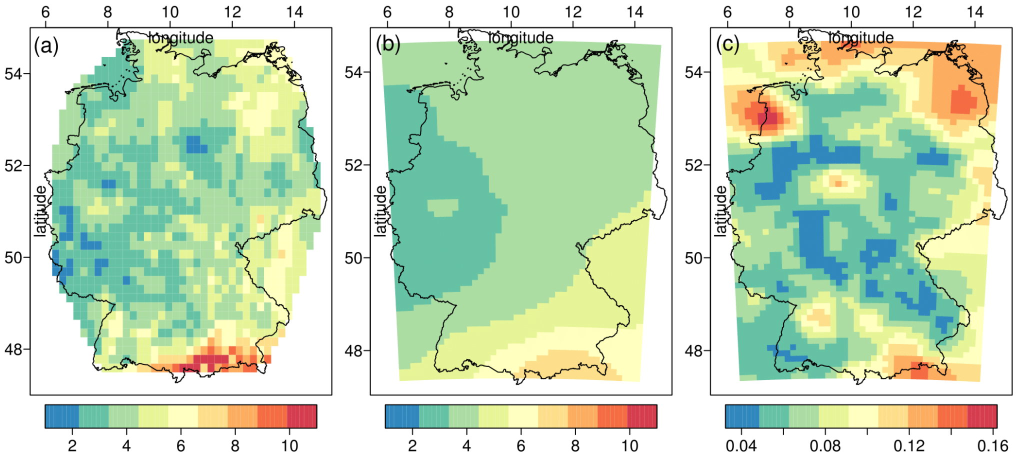


Figure 10.3.: (a) Spatial distribution of sample quantiles in mm/12h for $\tau = 0.9$, estimated at each of the 1079 stations from the 180 days between April and September 2011. The quantiles are bilinearly interpolated for enhanced visibility. (b) The mean spatial field of the intercept of SQR-0, interpolated onto the grid via kriging. The arithmetic mean is estimated from 100 kriging estimates derived from the MCMC realizations. (c) The standard-deviation of the mean spatial field.

10.2. Spatial quantile regression

We will now turn to the spatial modeling of regression coefficients and the influence on the predictive performance of quantile forecasts. The analysis is restricted to the 0.9-quantile. Three different SQR models are investigated:

- **SQR-0** without covariates - only the intercept β_0 is modeled spatially,
- **SQR-1** using 1 covariate - $\beta_0 + \beta_1 \cdot$ fgQ$_{0.99}$,
- **SQR-2** using 2 covariates - $\beta_0 + \beta_1 \cdot$ fgPoP + $\beta_2 \cdot$ fgQ$_{0.99}$.

The selection of predictors is based on the results from Section 10.1. For a better assessment of the effect of the spatial modeling, we have limited the number of covariates. A MCMC scheme with $30\,000$ iterations produces realizations of the regression coefficients β_0, β_1, β_2 at the 457 locations in the training data set. Conditional on these realizations, a kriging approach is used to obtain the regression coefficients for new locations. For SQR-0, the intercept should represent the sample quantile for the 180 days at each location. Fig. 10.3(a) shows the sample quantiles for each of the 1079 stations, interpolated onto a regular grid for enhanced visibility. For the spatial distribution of the intercept, 100 kriging estimates are calculated conditional on the MCMC realizations. The mean and standard-deviation of the spatial fields is shown in Fig. 10.3(b) and (c). The intercept shows a significant spatial structure. Lower values of 3 are obtained in western Germany, while higher values of 5 to 7 are obtained in southern Germany. The middle and north-eastern parts show values of 4, which are close to the overall climatological quantile of 4.2 mm/12h. The spatial structure is similar to those observed from the sample quantiles in (a). However, the spatial model significantly underestimates the spatial

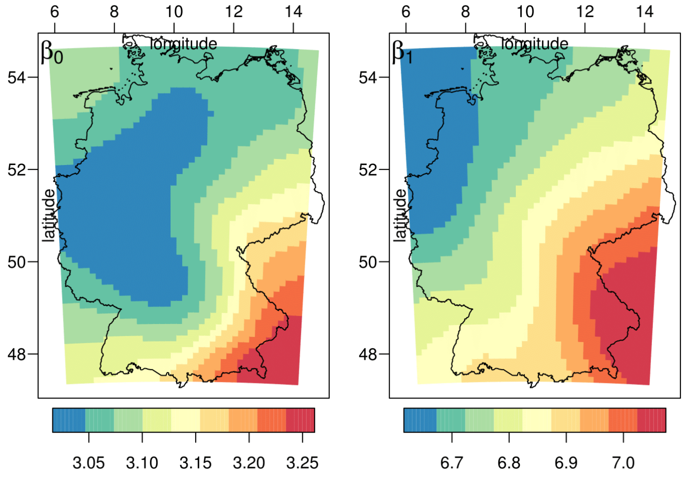


Figure 10.4.: Mean spatial field of the regression coefficients for SQR-1, interpolated onto the grid via kriging (β_0: intercept, β_1: fgQ$_{0.99}$). The arithmetic mean is estimated from 100 kriging estimates derived from the MCMC realizations.

variability. The sample quantiles have lower values of 1 mm/12h in the western part, and values up to 11 mm/12h in the southern part. Fig. 10.3(c) shows the uncertainty of the kriging estimates. Lower values are obtained for regions with higher station density. The northern part has less stations, and thus shows higher uncertainties, as well as the regions outside of Germany. However, the uncertainty of the kriging estimates is much smaller as the spatial variability[1], and the structure in Fig. 10.3(b) is highly significant.

A single covariate is used for SQR-1, and the spatial distribution of the regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ is shown in Fig. 10.4. The spatial distribution is again obtained as the arithmetic mean over 100 kriging estimates. The spatial structures are significant in the sense, that the spatial variance is significantly larger than the variance of the kriging estimates. The regression coefficients show a clear spatial structure with lower values in the north-west and higher values in the south-east of Germany. The values range between 3 and 3.3 for the intercept and between 6.5 and 7 for β_1. As a reference model, we use classic QR with spatial homogeneous coefficients, further denoted by QR-1. The regression coefficients for QR-1 are estimated as $\beta_0 = 3.48 \pm 0.03$ and $\beta_1 = 6.33 \pm 0.11$, and thus show a slight bias compared to the range of values obtained from SQR-1.

For the verification of the SQR-1 model, regression coefficients are estimated for the 622 locations omitted from the training data. Quantile forecasts are made from 1000 kriging estimates for the 180 days and 622 locations. Fig. 10.5(a) shows quantile reliability plots for SQR-01, QR-1 (with spatial homogeneous coefficients), and the first-guess quantiles as reference. The error bars denote the 95% confidence interval estimated from the 1000 SQR forecasts. The classic QR completely fails to predict lower quantile values if only the fgQ$_{0.99}$ is used as covariate. The smallest quantile forecast amounts to 0.6 mm and is largely overestimated. Higher values above 2 mm are better calibrated. The spatial modeling of regression coefficients improves the

[1] For all gridboxes in Fig. 10.3(c), the values of $1.96\sigma/\sqrt{n}$, where n is the number of kriging estimates, is much smaller than the standard deviation of the mean field.

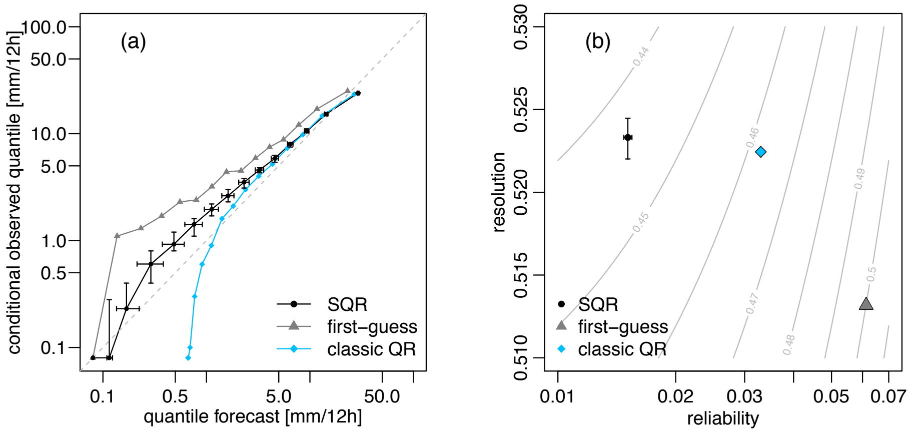


Figure 10.5.: (a) Reliability diagram for the 0.9-quantile, estimated from first-guess forecasts, classic QR with fgQ_{99} as single covariate and SQR-1. (b) The QS and its decomposition, evaluated at the verifying stations for 180 days between April and September 2011. The error bars show the 95% confidence interval for SQR. The reliability is presented on a logarithmic scale.

reliability of the quantile forecasts for lower values. The zero quantile forecasts are in good agreement with the observations. Smaller values still show some miscalibrations and an underestimation of quantile values between 0.1 and 5 mm. The first-guesses largely underestimate the 0.9-quantile over the complete range of forecast values.

The QS, reliability and resolution are displayed in Fig. 10.5(b). Compared to the first-guess, the postprocessing increases the resolution of quantile forecasts. Both models SQR-1 and QR-1 show a similar gain in resolution obtained from the covariate $fgQ_{0.99}$. The reliability of the first-guess is worse compared to SQR/QR, which results from the general underestimation of quantiles. The quantile forecasts have a better reliability after postprocessing, and the spatial modeling outperforms the classic QR approach. The QSS is calculated with respect to the first-guess forecasts, and amounts to 7.5% for the QR-1, and 11.3% for the SQR-1 model.

The SQR-2 model uses two covariates, the fgPoP and $fgQ_{0.99}$. The spatial distribution of the regression coefficients is shown in Fig. 10.6 again as the mean over 100 kriging estimates. The spatial structures are again significant with respect to the variance of the kriging estimates. The structures of the intercept (β_0) and the $fgQ_{0.99}$ (β_2) are similar to SQR-1, but the values for β_2 are a little bit lower and range between 5.5 and 6. The values for β_1 (fgPoP) range between 0.6 and 1 with lower values in the north and higher values in the south. A classic QR model with the same covariates is again used for comparison (QR-2). The regression coefficients for QR-2 are estimated from the training data as $\beta_0 = 3.12 \pm 0.05$, $\beta_1 = 0.79 \pm 0.06$, and $\beta_2 = 5.8 \pm 0.13$. In contrast to QR-1, the regression coefficients of QR2 lie in the middle of the range of values obtained from SQR-2 and show no bias.

The predictive performance of QR-2 is similar to the predictive performance of SQR-2. Fig. 10.7(a) shows the quantile reliability curves. The additional covariate fgPoP largely improves the calibration of small quantile values, and the reliability curves for both models are close to the diagonal. The QS, reliability and resolution are displayed in Fig. 10.7(b). Compared to SQR-1, the additional covariate does not increase the resolution of the quantile forecasts. But the reliability part of the QS is largely improved. The reliability decreases from about 0.015

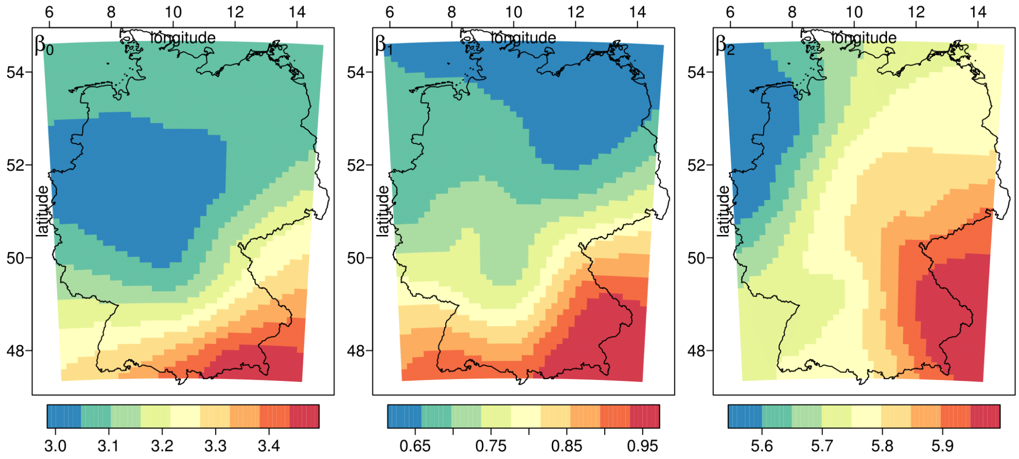


Figure 10.6.: Mean spatial field of the regression coefficients for SQR-2, interpolated onto the grid via kriging (β_0: intercept, β_1: fgPoP, β_2: fgQ$_{0.99}$). The arithmetic mean is estimated from 100 kriging estimates derived from the MCMC realizations.

for SQR-1 to 0.0044 for SQR-2. The predictive skill of forecasts from SQR-2 and QR-2 is about 13.3% compared to first-guess quantiles.

In contrast to SQR-1, there is no significant improvement of the spatial model in the case of the two covariates fgPoP and fgQ$_{0.99}$. The covariates are taken from a numerical weather prediction ensemble and already contain information about the spatial structure of precipitation. A spatial homogeneous postprocessing might be sufficient. However, if the covariates show some deficiencies in representing the spatial variability, as it is the case for the fgQ$_{0.99}$ as single covariate, the spatial model can improve the calibration of the spatial forecasts, and hence lead to a better predictive performance.

10.3. Conclusion

The Bayesian formulation of the quantile regression model is used for an enhanced postprocessing of precipitation quantiles derived from the mesoscale NWP ensemble COSMO-DE-EPS. The Bayesian model and the MCMC procedures requires a high amount of computational costs compared to the frequentistic QR explored in Chapter 8.3. Nevertheless, BQR can add significantly skill to the quantile predictions. The Bayesian LASSO allows to identify predictive covariates from a wide range of meteorological variables. Besides total precipitation, the wind-gusts, total water content, and CAPE are good predictors for precipitation quantiles. The additional covariates largely increase the information content and hence the resolution of quantile forecasts. Moreover, they improve the calibration for higher quantiles. The selected variables from the Bayesian LASSO also improve the predictive performance of the classic QR with a significant gain in resolution.

The BQR yields a better estimate of regression coefficients. The censored QR is affected by biases in the parameter estimates, which results from the three step procedure described in Section 5.2.2. The effect is stronger for the lower quantiles. Especially the 0.25-quantile benefits from BQR, but a gain in predictive performance from BQR can also be observed for the

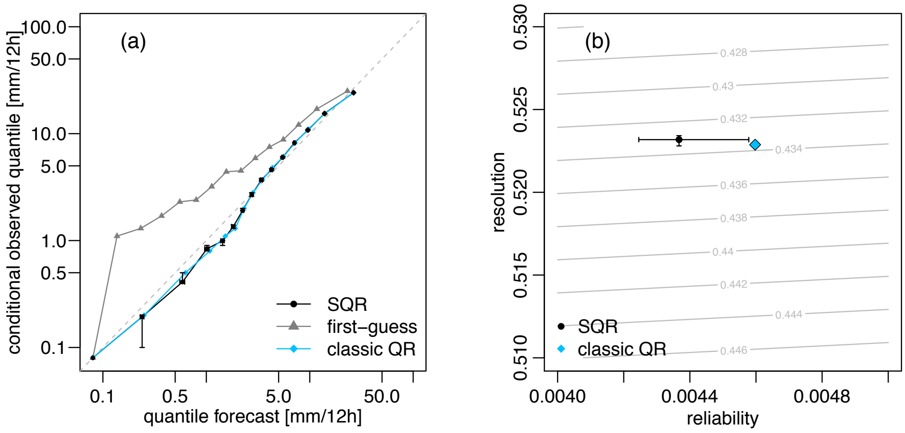


Figure 10.7.: (a) Reliability diagram for the 0.9-quantile, estimated from first-guess forecasts, classic QR with two covariates fgPoP and fg$Q_{0.99}$ and SQR-2. (b) The QS and its decomposition, evaluated at the verifying stations for 180 days between April and September 2011. The error bars show the 95% confidence interval for SQR.

higher quantiles if a larger number of covariates is used in the regression model. BQR yields a better calibration for almost all quantiles compared to classic QR. Moreover, the BQR gives a better representation of the uncertainty of the regression coefficients.

The spatial model yields significant structures of the regression coefficients over Germany. In the case of only one covariate (fg$Q_{0.99}$), the SQR leads to a large improvement in calibration compared to the classic QR model with spatially constant regression coefficients. However, the benefit of the SQR depends on the covariates. If the spatial distribution of precipitation is captured well by the covariates, e.g. as in the case of the two covariates fgPoP and fg$Q_{0.99}$, a spatial homogeneous postprocessing might be sufficient. The selection of informative covariates is thus of great importance for the predictive performance of the regression model.

Part IV.

Conclusion

11. Summary and Conclusion

Numerical weather prediction has seen great advances during the last decade. Due to a steady increase in computational power, NWP models run on even finer horizontal scales, include more complex physical and microphysical processes, and provide the user with realistic weather pattern on the km-scale. However, the validation of such small scale model output remains still a challenge in the verification and quantification of forecast uncertainty. Ensemble forecasting is the main tool to assess the uncertainty in NWP on all scales. Mainly two major sources of uncertainty are addressed: incomplete knowledge about initial conditions, as well as the model error due to discretization, parameterization and incomplete physics. However, ensemble forecasting requires a vast amount of computational time, and is still limited with respect to the ensemble size. Most raw ensembles provide probabilistic guidance for a range of meteorological variables, but still suffer from an underrepresentation of ensemble spread.

This study investigates the predictive performance of a convective-scale NWP ensemble to predict quantitative precipitation. Precipitation is still one of the most difficult weather variable to predict and is often used to measure the performance of mesoscale NWP models (Ebert et al., 2003). In a first step, ensemble forecasts have to be translated into probabilistic predictions. A predictive distribution function describes the most probable state of a variable together with its uncertainty. The formulation of a predictive distribution often relies on the specification of a parametric distribution function. The predictive performance strongly depends on how suitable the parametric distribution fits the data. Especially for precipitation, a suitable parametric distribution can hardly be obtained (Bentzien and Friederichs, 2012). A representation of quantitative precipitation in terms of quantile forecasts is thus an attractive alternative. Quantile forecasts can be graphically displayed by boxplots and are intuitive for users. Together with the probability of precipitation, they give a complete picture of forecast uncertainty. In order to investigate the predictive performance of quantile forecasts, a new framework for the verification of quantile forecasts has been developed (Bentzien and Friederichs, 2014). The quantile score decomposition gives more detailed insights with respect to calibration and information content (resolution) of quantile forecasts. The graphical presentation of quantile reliability curves can be used to explore ensemble calibration and the added value of statistical postprocessing.

11.1. Evaluation of ensemble forecasts

The convective-scale ensemble system COSMO-DE-EPS is a 20 member multi-analysis and multy-physics ensemble centered over Germany. An evaluation of precipitation forecasts based on the PIT histogram, which is a generalization of the analysis rank histogram, reveals a strong underestimation of ensemble spread. Moreover, COSMO-DE-EPS reveals a small positive bias due to an overestimation of precipitation. The method of lagged average forecasts is explored as inexpensive technique to increase the ensemble size of COSMO-DE-EPS. The so called time-lagged ensemble largely improves the representation of ensemble spread. It is remarkable that most of the benefit is already obtained from only 5 members of COSMO-DE-EPS complemented by their time-lagged model runs.

First-guess quantile forecasts are estimated from the ensemble under consideration of a spatial neighborhood of 5×5 gridboxes. Quantile reliability curves are explored and reveal further deficiencies in ensemble calibration. Quantiles for the probability level $\tau < 0.75$ are generally overestimated. Higher quantiles with $\tau > 0.75$ are in contrast largely underestimated. Merely the 0.75-quantile shows a reliability curve which is close to the diagonal and indicates good calibration. The time-lagged ensemble lead to better estimates of probabilistic forecasts with quantile reliability curves which are closer to the diagonal. However, statistical postprocessing largely improves the calibration of quantile forecasts, and becomes indispensable to obtain calibrated and skillful forecasts for higher quantiles with $\tau > 0.9$.

11.2. Ensemble postprocessing

Bentzien and Friederichs (2012) explore various techniques for statistical postprocessing of ensembles forecasts. We focus on probability and quantile forecasts, which are either derived from logistic and quantile regression for single thresholds and probability levels, or from a parametric mixture model based on LR for the probability of precipitation, a Gamma-GLM for the distribution of precipitation amounts, and a GPD for the extremal tail of the distribution. Although the mixture model gives promising results, calibrated forecasts from logistic and quantile regression are still superior in probabilistic quantitative precipitation forecasting.

Statistical postprocessing of COSMO-DE-EPS uses LR for the probability of precipitation and QR for quantiles between $\tau = 0.25$ and 0.999. Lower quantiles are frequently censored and therefore omitted from the analysis. The PoP can largely be improved by LR. The improvement is due to both, a better calibration and an increase in the resolution. Probability forecasts from higher thresholds merely profit from a better calibration, while the resolution is not further improved. Moreover, higher thresholds require a sufficiently large data set. A threshold of 10 mm/12h is exceeded in just 3% of the observations. Statistical calibration for higher thresholds suffers from sampling errors and is hence limited due to the sample size.

The calibration of quantile forecasts from COSMO-DE-EPS is largely improved by QR. The benefit is larger for lower and higher probability levels, which show larger deviations from a perfect calibrated forecast, than the 0.75-quantile, which is already good calibrated. However, in case of COSMO-DE-EPS, predictive quantile forecasts for $\tau > 0.9$ cannot be obtained from the raw ensemble. Time-lagging enables predictive quantile forecasts up to the 0.99-quantile. However, statistical postprocessing becomes indispensable for extremal quantiles.

A Bayesian formulation of the quantile regression based on Yu and Stander (2007) is used for an advanced inference of quantile forecasts. The classic QR model mainly improves the calibration of quantile forecasts, while the resolution depends on the set of predictive covariate. In the Bayesian framework, inference about variable selection can be done using independent zero-mean Laplacian priors. This is equivalent to the least absolute shrinkage and selection operator proposed by Tibshirani (1996). The Bayesian LASSO allows the selection of predictive covariates from a wide range of variables. The variance parameter of the Laplacian prior determines the sparseness of the regression coefficients. Smaller variances will lead to regression coefficients which are close to zero, and identifies only a small set of informative predictors. Most of the work concerning Bayesian quantile regression, including censored data or penalized regression for variable selection, has been done using survival data in medical or biological applications. To the best of my knowledge, this is the first attempt to use a Bayesian quantile regression model for quantitative precipitation forecasts from short-range NWP.

Predictive covariates are selected for different quantiles separately by the Bayesian LASSO. So far, only covariates based on total precipitation forecasts (mean and standard-deviation, probabilities, quantiles) are used for the postprocessing. The set of covariates now also contains other meteorological variables. In addition to the total precipitation variables, the Bayesian LASSO reveals that lower quantiles profit from wind gusts forecasts, while for higher quantiles a combination of wind gusts, CAPE, and total water content are informative predictors. The additional meteorological variables largely increase the information content of the quantile forecasts and hence lead to a much better resolution. Moreover, an improvement in reliability is obtained for higher quantiles with $\tau > 0.5$. The selected variables also improve the performance of the classic QR model. However, parameters of the censored QR might be biased through the three step procedure which is used for the estimation of regression coefficients. The Bayesian QR yields better estimates of the regression coefficients, especially for the 0.25-quantile. A benefit can also be observed for higher quantiles, if a larger number of coefficients has to be estimated. Moreover, the Bayesian QR gives a better representation of the uncertainty of the regression coefficients.

A spatial QR model is developed for a better assessment of the predictive performance of spatial quantile forecasts. So far, a spatially constant relationship is assumed between the regression coefficients and the covariates. Predictions for new locations, which are not included into the training data of the statistical model, can easily be made by the constant regression coefficients and the covariates for the new locations. Note that all covariates used in this study are taken from the NWP model output and already contain information about a spatial structure. However, the spatial modeling of regression coefficients reveals a better calibration of quantile forecasts if the covariates do not represent sufficient information about the spatial characteristics of precipitation. If the spatial distribution of precipitation is captured well by the covariates, a spatial homogeneous postprocessing might be sufficient. The selection of informative covariates is thus of great importance for the predictive performance of the regression model.

11.3. Probabilistic forecast verification

This study contributes to an enhanced framework of statistical postprocessing and probabilistic forecast verification for quantitative precipitation forecasts in terms of quantiles. Proper scores are the main tool in probabilistic forecast verification. They assign a single score value to a forecasting scheme. The propriety of the score function guarantees that the score cannot be hedged, and thus allows for a ranking of competing forecast systems, or a quantitative assessment of the added value of postprocessing. Moreover, proper scores are used in estimation problems, which is known as optimum score estimation (Gneiting and Raftery, 2007). The unknown parameters of a statistical model are estimated such that the postprocessed forecasts optimize their respective score function. In this sense, forecast verification is closely related to statistical postprocessing.

The proposal of the quantile score decomposition provides an extended framework of quantile verification, analog to the Brier score decomposition for probability forecasts. The QS decomposition enables the exploration of quantile forecasts with regard to the two forecast attributes reliability and resolution. The graphical representation of quantile reliability curves can be used for the exploration of ensemble calibration. If the ensemble suffers from an underestimation of ensemble spread, the quantile reliability curves will show deviations from the diagonal. The deviations will depend on the probability level of the quantile. Greater deviations are obtained for the outer quantiles, which suffer most from an insufficient ensemble spread. Moreover, ensembles can be compared by their information content, expressed as resolution. While reliable or calibrated quantile forecasts can be obtained from statistical postprocessing, an increase of the resolution depends on the selection of predictive covariates. The QS decomposition allows to attribute the impact of different covariates either on the resolution or the calibration of quantile forecasts.

The QS decomposition together with a Bayesian formulation of quantile regression, including the Bayesian LASSO for variable selection, provides an enhanced framework for the verification and statistical postprocessing of quantitative precipitation quantile predictions derived from mesoscale NWP ensembles.

List of Figures

List of Tables

Bibliography

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

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Monthly Weather Review*, 131, 1509–1523.

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 130, 627–646.

Baldauf, M., J. Förstner, S. Klink, T. Reinhardt, C. Schraff, A. Seifert, and K. Stephan, 2011a: *Kurze Beschreibung des Lokal-Modells Kürzestfrist COSMO-DE (LMK) und seiner Datenbanken auf dem Datenserver des DWD*. Deutscher Wetterdienst, Version 1.6.

Baldauf, M., A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer, and T. Reinhardt, 2011b: Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. *Monthly Weather Review*, 139, 3887–3905.

Banerjee, S., B. P. Carlin, and A. E. Gelfand, 2004: *Hierarchical modeling and analysis for spatial data*. Monographs on Statistics and Applied Probability, Chapman&Hall/CRC.

Barkmeijer, J., R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2333–2351.

Ben Bouallègue, Z., 2011: Upscaled and fuzzy probabilistic forecasts: verification results. *COSMO Newsletter*, 11, 124–132.

Ben Bouallègue, Z., S. E. Theis, and C. Gebhardt, 2013: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, 22, 49–59.

Bentzien, S. and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting*, 27, 988–1002.

Bentzien, S. and P. Friederichs, 2014: Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140, 1924–1934.

Bierdel, L., P. Friederichs, and S. Bentzien, 2012: Spatial kinetic energy spectra in the convection-permitting limited-area NWP model COSMO-DE. *Meteorologische Zeitschrift*, 21, 245–258.

Bollmeyer, C., J. D. Keller, C. Ohlwein, S. Wahl, S. Crewell, P. Friederichs, A. Hense, J. Keune, S. Kneifel, I. Pscheidt, S. Redl, and S. Steinke, 2015: Towards a high-resolution regional reanalysis for the European CORDEX domain. *Quarterly Journal of the Royal Meteorological Society*, 141, 1–15.

Bott, A., 1989: A positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes. *Monthly Weather Review*, 117, 1006–1016.

Bousquet, O., C. A. Lin, and I. Zawadzki, 2006: Analysis of scale dependence of quantitative precipitation forecast verification: a case-study over the Mackenzie river basin. *Quarterly Journal of the Royal Meteorological Society*, 132, 2107–2125.

Bowler, N. E., A. Arribas, K. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134, 703–722.

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Bröcker, J., 2008: Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 136, 4488–4502.

Bröcker, J., 2009: Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135, 1512–1519.

Bröcker, J., 2012: Probability forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, I. T. Jolliffe and D. B. Stephenson, Editors, Wiley, Chapter 7, 119–139.

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

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

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.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125, 2887–2908.

Buizza, R. and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *Journal of the Atmospheric Science*, 52, 1434–1456.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 124, 1935–1960.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131, 2131–2150.

Chernozhukov, V. and H. Hong, 2002: Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association,* 97, 872–882.

Clark, J. S. and A. E. Gelfand (Editors), 2006: *Hierarchical modelling for the environmental sciences*. Oxford University Press.

Coles, S., 2001: *An introduction to statistical modeling of extreme values*. Springer Series in Statistics, Springer.

Cooley, D., D. Nychka, and P. Naveau, 2007: Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102, 824–840.

Craig, G., E. Richard, D. Richardson, D. Burridge, S. Jones, F. Atger, M. Ehrendorfer, M. Heikinheimo, B. Hoskins, A. Lorenc, J. Methven, T. Paccagnella, J. Pailleux, F. Rabier, M. Roulston, R. Saunders, R. Swinbank, S. Tibaldi, and H. Wernli, 2010: Weather Research in Europe. A THORPEX European Plan. *WMO/TD-No. 1531, WWRP/THORPEX No. 14*.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmospheric Science Letters*, 5, 110–117.

Eady, E., 1949: Long waves and cyclone waves. *Tellus*, 1, 33–52.

Eady, E., 1951: The quantitative theory of cyclone development. *Compendium of Meteorology*, T. F. Malone, Editor, American Meteorological Society, Boston, 464–469.

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts. *Bulletin American Meteorological Society*, 84, 481–492.

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

Efron, B. and R. J. Tibshirani, 1993: *An introduction to the bootstrap*. Chapman&Hall/CRC.

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

Fahrmeir, L. and G. Tutz, 1994: *Multivariate statistical modelling based on generalized linear models*. Springer.

Friederichs, P., 2010: Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13, 109–132.

Friederichs, P. and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135, 2365–2378.

Fritsch, J. M. and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: a USWRP research and development strategy. *Bulletin American Meteorological Society*, 85, 955–965.

Fundel, F., A. Walser, M. A. Liniger, C. Frei, and C. Appenzeller, 2010: Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. *Monthly Weather Review*, 138, 176–189.

Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variations of lateral boundaries. *Atmospheric Research*, 100, 168–177.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004: *Bayesian data analysis*. 2nd edition, Text in Statistical Science, Chapman&Hall/CRC.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Editors), 1996: *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics, Chapman&Hall/CRC.

Gilleland, E., 2014: *R-package ”verification”: Weather forecast verification utilities*. NCAR - Research Applications Laboratory, URL `http://CRAN.R-project.org/package=verification`, version 1.41.

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

Gneiting, T., 2011b: Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27, 197–207.

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

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

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

Gneiting, T. and R. Ranjan, 2011: Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29, 411–422.

Golding, B. W., S. P. Ballard, K. Mylne, N. M. Roberts, A. Saulter, C. Wilson, P. Agnew, L. S. Davis, J. Trice, C. Jones, D. Simonin, Z. Li, C. Pierce, A. Bennett, M. Weeks, and S. Moseley, 2014: Forecasting capabilities for the London 2012 Olympics. *Bulletin American Meteorological Society*, 95, 883–896.

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., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560.

Hamill, T. M. and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review*, 125, 1312–1327.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Monthly Weather Review*, 136, 2620–2632.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, 132, 1434–1447.

Harper, K., L. W. Uccellini, E. Kalnay, K. Carey, and L. Morone, 2007: 50th anniversary of operational numerical weather prediction. *Bulletin American Meteorological Society*, 88, 639–650.

Hastings, W. K., 1970: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

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

Hoffman, R. N. and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, 35A, 100–118.

Hou, D., Z. Toth, Y. Zhu, W. Yang, and R. Wobus, 2010: A stochastic total tendency perturbation scheme representing model-related uncertainties in the NCEP global ensemble forecast system. Technical report, Environmental Modeling Center/NCEP/NOAA, Camp Springs, MD, USA.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124, 1225–1242.

Houtekamer, P. L., H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Monthly Weather Review*, 137, 2126–2143.

Hsu, W. and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2, 285–293.

Hyndman, R. J. and Y. Fan, 1996: Sample quantiles in statistical packages. *The American Statistician*, 50, 361–365.

Ji, Y., N. Lin, and B. Zhang, 2012: Model selection in binary and tobit quantile regression using the Gibbs sampler. *Computational Statistics & Data Analysis*, 56, 827–839.

Jolliffe, I. T. and D. B. Stephenson (Editors), 2012: *Forecast verification: A practitioner's guide in atmospheric science*. 2nd edition, Wiley.

Kain, J. S. and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: the Kain-Fritsch scheme. *The representation of cumulus convection in numerical models*, K. A. Emanuel and D. J. Raymond, Editors, Meteorological Monographs, Chapter 16, 165–170.

Kalnay, E., B. Hunt, E. Ott, and I. Szunyogh, 2006: Ensemble forecasting and data assimilation: two problems with the same solution? *Predictability of weather and climate*, T. N. Palmer and R. Hagedorn, Editors, Camebridge University Press, Chapter 7, 157–180.

Keller, J. D. and A. Hense, 2011: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms. *Meteorologische Zeitschrift*, 20, 107–117.

Kneib, T., 2013: Beyond mean regression. *Statistical Modelling*, 13, 275–303.

Koenker, R., 2005: *Quantile regression*. Econometric Society Monographs, Cambridge University Press.

Koenker, R., 2013: *R-package "quantreg": Quantile regression*. URL `http://CRAN.R-project.org/package=quantreg`, version 4.97.

Koenker, R. and J. A. F. Machado, 1999: Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94, 1296–1310.

Kottas, A. and M. Krnjajic, 2009: Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36, 297–319.

Kyung, M., J. Gill, M. Ghosh, and G. Casella, 2010: Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5, 369–412.

Laio, F. and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, 11, 1267–1277.

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

Lewis, J. M., 2005: Roots of ensemble forecasting. *Monthly Weather Review*, 133, 1865–1885.

Lewis, J. M., 2014: Edward Epstein's stochastic–dynamic approach to ensemble weather prediction. *Bulletin American Meteorological Society*, 95, 99–116.

Li, Q., R. Xi, and N. Lin, 2010: Bayesian regularized quantile regression. *Bayesian Analysis*, 5, 533–556.

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

Lorenz, E. N., 1963b: The predictability of hydrodynamic flow. *Transactions of the New York Academy of Sciences*, 25, 409–432.

Lorenz, E. N., 1965: On the possible reasons for long-period fluctuations of the general circulation. *WMO-IUGG Symposium on Research and Development Aspects of Long-range Forecasting*, Technical Note No. 66, WMO-No. 162.TP.79, 203–211.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307.

Lu, C., H. Yuan, B. E. Schwartz, and S. G. Benjamin, 2007: Short-range numerical weather prediction using time-lagged ensembles. *Weather and Forecasting*, 22, 580–595.

Lum, K. and A. E. Gelfand, 2012: Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, 7, 235–258.

Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlinear Processes in Geophysics*, 12, 527–536.

Masbou, M., 2008: LM-PAFOG - a new three-dimensional fog forecast model with parametrised microphysics. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin American Meteorological Society*, 83, 407–430.

Matheson, J. E. and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

McCullagh, P. and J. A. Nelder, 1989: *Generalized linear models*. 2nd edition, Monographs on Statistics and Applied Probability, Chapman&Hall/CRC.

Mellor, G. L. and T. Yamada, 1974: A hierarchy of turbulence closure models for planetary boundary layers. *Journal of the Atmospheric Sciences*, 31, 1791–1806.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953: Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1087–1092.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society*, 133, 1487–1500.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Murphy, A. H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*, 12, 595–600.

Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Weather and Forecasting*, 6, 302–307.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8, 281–293.

Murphy, A. H. and R. L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, 115, 1330–1338.

Palmer, T. N., G. J. Shutts, R. Hagedorn, F. J. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annual Review of Earth and Planetary Sciences*, 33, 163–193.

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134, 2029–2050.

Pellerin, G., L. Lefaivre, P. L. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes in Geophysics*, 10, 463–468.

Peralta, C., Z. Ben Bouallègue, S. E. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research*, 117, D07 108.

R Core Team, 2014: *R: A language and environment for statistical computing*. Vienna, Austria, R Foundation for Statistical Computing, URL `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.

Reich, B. J., H. D. Bondell, and H. J. Wang, 2010: Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11, 337–352.

Reich, B. J., M. Fuentes, and D. B. Dunson, 2011: Bayesian spatial quantile regression. *Journal of the American Statistical Association*, 106, 6–20.

Reinhardt, T. and A. Seifert, 2006: A three-category ice scheme for LMK. *COSMO Newsletter*, 6, 115–120.

Ritter, B. and J.-F. Geleyn, 1992: A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Monthly Weather Review*, 120, 303–325.

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97.

Rockel, B., A. Will, and A. Hense, 2008: The regional climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift*, 17, 347–348.

Saito, K., T. Fujita, Y. Yamada, J.-i. Ishida, Y. Kumagai, K. Aranami, S. Ohmori, R. Nagasawa, S. Kumagai, C. Muroi, T. Kato, H. Eito, and Y. Yamazaki, 2006: The operational JMA nonhydrostatic mesoscale model. *Monthly Weather Review*, 134, 1266–1298.

Schaffer, C. J., W. A. Gallus Jr., and M. Segal, 2011: Improving probabilistic ensemble forecasts of convection through the application of QPF–POP relationships. *Weather and Forecasting*, 26, 319–336.

Schättler, U., G. Doms, and C. Schraff, 2013: *A description of the nonhydrostatic regional COSMO-model. Part VII: User's Guide*. URL `http://www.cosmo-model.org`, COSMO v4.29.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Schwartz, C. S., J. S. Kain, M. C. Coniglio, S. J. Weiss, D. R. Bright, J. J. Levit, M. Xue, F. Kong, K. W. Thomas, and M. S. Wandishin, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, 25, 263–280.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Monthly Weather Review*, 139, 976–991.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review*, 132, 3019–3032.

Skamarock, W. C. and J. B. Klemp, 2008: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227, 3465–3485.

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.

Staniforth, A. and N. Wood, 2008: Aspects of the dynamical core of a nonhydrostatic, deep-atmosphere, unified weather and climate-prediction model. *Journal of Computational Physics*, 227, 3445–3464.

Stephan, K., S. Klink, and C. Schraff, 2008: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Quarterly Journal of the Royal Meteorological Society*, 134, 1315–1326.

Taddy, M. A. and A. Kottas, 2010: A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business and Economic Statistics*, 28, 357–369.

Theis, S. E., C. Gebhardt, and Z. Ben Bouallègue, 2012: *Beschreibung des COSMO-DE-EPS und seiner Ausgabe in die Datenbanken des DWD*. Deutscher Wetterdienst, Version 1.0.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, 12, 257–268.

Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, 9, 275–295.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1, 522–534.

Tibshirani, R. J., 1996: Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, 117, 1779–1800.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bulletin American Meteorological Society*, 74, 2317–2330.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Weather and Forecasting*, 8, 379–398.

Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Monthly Weather Review*, 139, 403–423.

Vogel, B., H. Vogel, D. Bäumer, M. Bangert, K. Lundgren, R. Rinke, and T. Stanelle, 2009: The comprehensive model system COSMO-ART – Radiative impact of aerosol on the state of the atmosphere on the regional scale. *Atmospheric Chemistry and Physics*, 9, 8661–8680.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, 60A, 62–79.

Weigel, A. P., 2012: Ensemble forecasts. *Forecast verification: A practitioner's guide in atmospheric science*, I. T. Jolliffe and D. B. Stephenson, Editors, Wiley, Chapter 8, 141–166.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, 13, 243–256.

Wilks, D. S., 2006b: *Statistical methods in the atmospheric sciences*. 2nd edition, International Geophysics Series, Academic Press.

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

Yu, K. and R. A. Moyeed, 2001: Bayesian quantile regression. *Statistics & Probability Letters*, 54, 437–447.

Yu, K. and J. Stander, 2007: Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics*, 137, 260–276.

Yu, K. and J. Zhang, 2005: A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, 34, 1867–1879.

Yuan, H., C. Lu, J. A. McGinley, P. J. Schultz, B. D. Jamison, L. Wharton, and C. J. Anderson, 2009: Evaluation of short-range quantitative precipitation forecasts from a time-lagged multimodel ensemble. *Weather and Forecasting*, 24, 18–38.

BONNER METEOROLOGISCHE ABHANDLUNGEN

Herausgegeben vom Meteorologischen Institut der Universität Bonn durch Prof. Dr. H. FLOHN (Hefte 1-25), Prof. Dr. M. HANTEL (Hefte 26-35), Prof. Dr. H.-D. SCHILLING (Hefte 36-39), Prof. Dr. H. KRAUS (Hefte 40-49), ab Heft 50 durch Prof. Dr. A. HENSE.

Heft 1-49: siehe `http://www.meteo.uni-bonn.de/bibliothek/bma`

Heft 50: ***Petra Friederichs***: Interannuelle und dekadische Variabilität der atmosphärischen Zirkulation in gekoppelten und SST-getriebenen GCM-Experimenten. 2000, 133 S. + VIII. € 25

Heft 51: ***Heiko Paeth***: Anthropogene Klimaänderungen auf der Nordhemisphäre und die Rolle der Nordatlantik-Oszillation. 2000, 168 S.+ XVIII. € 28

Heft 52: ***Hildegard Steinhorst***: Statistisch-dynamische Verbundsanalyse von zeitlich und räumlich hoch aufgelösten Niederschlagsmustern: eine Untersuchung am Beispiel der Gebiete von Köln und Bonn. 2000, 146 S. + XIV. € 25

Heft 53: ***Thomas Klein***: Katabatic winds over Greenland and Antartica and their interaction with mesoscale and synoptic-scale weather systems: three-dimensional numerical models. 2000, 146 S. + XIV. € 25

Heft 54: ***Clemens Drüe***: Experimentelle Untersuchung arktischer Grenzschichtfronten an der Meereisgrenze in der Davis-Straße. 2001, 165 S. + VIII. € 28

Heft 55: ***Gisela Seuffert***: Two approaches to improve the simulation of near surface processes in numerical weather prediction models. 2001, 128 S. + VI. € 25

Heft 56: ***Jochen Stuck***: Die simulierte axiale atmosphärische Drehimpulsbilanz des ECHAM3-T21 GCM. 2002, 202 S. + VII. € 30

Heft 57: ***Günther Haase***: A physical initialization algorithm for non-hydrostatic weather prediction models using radar derived rain rates. 2002, 106S. + IV. € 25

Heft 58: ***Judith Berner***: Detection and Stochastic Modeling of Nonlinear Signatures in the Geopotential Height Field of an Atmospheric General Circulation Model. 2003, 157 S. + VIII. € 28

Heft 59: ***Bernd Maurer***: Messungen in der atmosphärischen Grenzschicht und Validation eines mesoskaligen Atmosphärenmodells über heterogenen Landoberflächen. 2003, 182 S. + IX. € 30

Heft 60: ***Christoph Gebhardt***: Variational reconstruction of Quaternary temperature fields using mixture models as botanical – climatological transfer functions. 2003, 204 S. + VIII. € 30

Heft 61: ***Heiko Paeth***: The climate of tropical and northern Africa – A statistical-dynamical analysis of the key factors in climate variability and the role of human activity in future climate change. 2005, 316 S. + XVI. € 15

Heft 62: ***Christian Schölzel***: Palaeoenvironmental transfer functions in a Bayesian framework with application to Holocene climate variability in the Near East. 2006, 104 S. + VI. € 15

Heft 63: ***Susanne Bachner***: Daily precipitation characteristics simulated by a regional climate model, including their sensitivity to model physics, 2008, 161 S. € 15

Heft 64: ***Michael Weniger***: Stochastic parameterization: a rigorous approach to stochastic three-dimensional primitive equations, 2014, 148 S. + XV. open access[1]

Heft 65: ***Andreas Röpnack***: Bayesian model verification: predictability of convective conditions based on EPS forecasts and observations, 2014, 152 S. + VI. open access[1]

Heft 66: ***Thorsten Simon***: Statistical and Dynamical Downscaling of Numerical Climate Simulations: Enhancement and Evaluation for East Asia, 2014, 48 S. + VII. + Anhänge open access[1]

Heft 67: ***Elham Rahmani***: The Effect of Climate Change on Wheat in Iran, 2014, [erschienen] 2015, 96 S. + XIII. open access[1]

Heft 68: ***Pablo A. Saavedra Garfias***: Retrieval of Cloud and Rainwater from Ground-Based Passive Microwave Observations with the Multi-frequency Dual-polarized Radiometer ADMIRARI, 2014, [erschienen] 2015, 168 S. + XIII. open access[1]

Heft 69: ***Christoph Bollmeyer***: A high-resolution regional reanalysis for Europe and Germany - Creation and Verification with a special focus on the moisture budget, 2015, 103 S. + IX. open access[1]

Heft 70: ***A S M Mostaquimur Rahman***: Influence of subsurface hydrodynamics on the lower atmosphere at the catchment scale, 2015, 98 S. + XVI. open access[1]

Heft 71: ***Sabrina Wahl***: Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, 108 S. open access[1]

[1] Available at `http://hss.ulb.uni-bonn.de/fakultaet/math-nat/`

METEOROLOGISCHES INSTITUT
MATHEMATISCH NATURWISSENSCHAFTLICHE FAKULTÄT
UNIVERSITÄT BONN