Karlsruhe Institute of Technology

Probabilistic Forecasting Based on Hydrometeorological Ensembles

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Stephan Hemri
aus
Dietikon (ZH, Schweiz)

Tag der mündlichen Prüfung: 18. Mai 2016
Referent: Prof. Dr. Tilmann Gneiting
Korreferent: Dr.-Ing. Uwe Ehret

Abstract

Over the last two decades the paradigm in hydrometeorological forecasting has shifted from deterministic to probabilistic. Numerical weather prediction (NWP) models are run increasingly as ensemble forecasting systems, which provide a finite sample of forecast scenarios for atmospheric variables like near-surface temperature or precipitation. Hydrologists use such NWP ensemble forecasts as input to hydrological models in order to obtain a sample of river runoff scenarios. Predictive skill of hydrometeorological ensemble forecasts can typically be improved by statistical post processing. The objective of this thesis is twofold. The first goal is to compare raw ensemble forecasts with probabilistic forecasts that have been obtained by state-of-the-art post processing approaches. In particular, the temporal evolution of the gap in skill between meteorological raw ensemble and post processed forecasts is assessed over the period from 2004 to 2014. For some applied problems appropriate post processing methods do not exist. Accordingly, the second goal is to develop novel post processing approaches, which is summarized next.

Two methods to post process ensemble forecasts for the discrete and bounded weather variable of total cloud cover (TCC) are developed. Applying them to TCC ensemble forecasts from the European Centre for Medium-Range Weather Forecasts improves forecast skill significantly. River runoff is an inherently multivariate process with typical events lasting from hours in case of floods to weeks or even months in case of droughts. This calls for multivariate post processing techniques that yield well calibrated forecasts in univariate terms and at the same time ensure realistic temporal dependence structures. To this end, methods originally developed for meteorological variables are adapted such that their application to hydrologic ensemble forecasts leads to an improvement in forecast skill, while ensuring temporal dependences inherent to river runoff.

Zusammenfassung

Über die letzen zwei Jahrzente hat ein Paradigmenwechsel im Gebiet der hydrometeorologischen Vorhersagen stattgefunden, infolgedessen deterministische Vorhersagen mehr und mehr durch probabilistische Vorhersagen ersetzt werden. Numerische Wettervorhersagesysteme werden zunehmend verwendet, um Ensemblevorhersagen zu generieren. Solche Ensemblevorhersagen bilden eine endliche Stichprobe von Vorhersageszenarien für Variablen wie Temperatur in Bodennähe oder Niederschlag. Hydrologen verwenden solche Ensemblevorhersagen als Eingangsdaten für hydrologische Modelle, womit Abflussszenarien erzeugt werden. Die Vorhersagegüte hydrometeorologischer Ensemblevorhersagen kann mittels statistischer Nachbearbeitung verbessert werden. Die vorliegende Arbeit verfolgt zwei Ziele. Das erste Ziel ist es, Ensemblevorhersagen mit mittels "state-of-the-art" Nachbearbeitungsmethoden generierter probabilistischen Vorhersagen zu vergleichen. Insbesondere wird untersucht, wie sich die Differenz in der Vorhersagegüte zwischen meteorologischen Ensemblevorhersagen und nachbearbeiteter Vorhersagen über den Zeitraum von 2004 bis 2014 entwickelt. Für einige angewandten Probleme existieren keine adäquaten Nachbearbeitungsmethoden. Dementsprechend, beinhaltet das zweite Ziel die Entwicklung neuartiger Nachbearbeitungsmethoden, die im Folgenden erwähnt werden.

Zwei Methoden zur Nachbearbeitung von Vorhersagen des Gesamtbewölkungsgrads (TCC), der eine diskrete und beschränkte Wettervariable darstellt, werden entwickelt. Die Anwendung dieser Methoden auf TCC Ensemblevorhersagen vom Europäischen Zentrum für mittelfristige Wettervorhersage führt zu einer signifikanten Verbesserung der Vorhersagegüte. Abfluss ist ein inhärent multivariater Prozess mit stark variierenden Ereignislängen von wenigen Stunden im Falle von Hochwasserspitzen bis hin zu Wochen oder sogar Monaten im Falle von Trockenperioden. Dies erfordert Nachbearbeitungsmethoden, die zu marginal kalibrierten Vorhersagen führen und gleichzeitig die zeitliche Abhängigkeitsstruktur richtig abbilden. Hierzu werden Methoden, die für meteorologische Variablen entwickelt wurden, so angepasst, dass sie zu einer Verbesserung der Vorhersagegüte führen und gleichermaßen die abflusstypischen zeitlichen Abhängigkeitsstrukturen erhalten bleiben.

Acknowledgments

First of all, I am grateful to Tilmann Gneiting for the supervision of this thesis, his great support, including the large number of discussions that helped to overcome the many obstacles encountered. I am likewise indebted to Uwe Ehret who kindly agreed to be referee and provided important hydrological expertise.

I gratefully acknowledge the support of the Klaus Tschira Foundation and the German Federal Institute of Hydrology (BfG) for funding. The Heidelberg Institute of Theoretical Studies is thanked for providing a great research environment and administrative support. Furthermore, I am grateful for the opportunity to have worked at the European Centre for Medium-Range Weather Forecasts (ECMWF) as a visiting scientist from March to July 2014. Hydrological and meteorological datasets have been provided by the BfG and ECMWF, respectively.

Moreover, I thank Bastian Klein, Dennis Meißner, and Dmytro Lisniak for providing challenging post processing related hydrological problems, for fruitful discussions, and their great support and collaboration. Likewise I am grateful to Konrad Bogner, Thomas Haiden, Florian Pappenberger, David Richardson, and Florence Rabier for the numerous discussions, their expertise, and great support during my stay at ECMWF. Further, I like to thank Maxime Taillardat for detecting an error in one of the papers this thesis is based on. Alexander Jordan, Roman Schefzik, and Michael Scheuerer is thanked for sharing their mathematical expertise in numerous discussions. I am grateful to Thordis Thorarinsdottir for sharing her EMOS scripts as well as to Michael Scheuerer and David Möller for their contributions, upon which I could base parts of my studies at ECMWF. Furthermore, I like to thank Sándor Baran, Werner Ehm, Kira Feldmann, Fabian Krüger, Sebastian Lerch, Evgeni Ovcharov, and Peter Vogel for fruitful discussions. Additionally, I am grateful to Pierre Pinson and anonymous reviewers for their helpful comments during the peer review processes of the papers, on which large parts of this thesis are based on.

Moreover, I like to thank my friends for distraction from scientific work and the great cycling, hiking, and skiing trips in my leisure time. Last but not least, I am indebted to my parents and my sister for their unconditional support.

Statement on journal papers

This thesis is to a large extent based on the papers by Hemri et al. (2014a), Hemri et al. (2014b), Hemri et al. (2015), Richardson et al. (2015) and Hemri et al. (2016) and on the handbook article by Hemri (2016). For all the listed papers the statistical analyses were performed by myself. Likewise the writing was essentially done by myself. The hydrological datasets for Hemri et al. (2014a) and Hemri et al. (2015) were prepared by Dmytro Lisniak, the meteorological datasest used in Hemri et al. (2014b), and Hemri et al. (2016) were pre processed by Thomas Haiden. The following list relates the papers to the sections of this thesis:

- Hemri et al. (2014a): Sections 2.1.3 and 4.2
- Hemri et al. (2014b): Sections 2.1.1 and 3.1
- Hemri et al. (2015): Sections 2.1.3, 2.2, 2.3, and 4.3
- Hemri (2016): Sections 1.2, 2.1.1, 2.1.3, 2.3, and 4.1
- Hemri et al. (2016): Sections 2.1.2, 2.3 and 3.2

Contents

Chapter 1

Introduction

1.1 Motivation and outlook

Over the last two decades the paradigm in weather forecasting has shifted from deterministic to probabilistic (see e.g. Palmer (2000) and Hamill et al. (2000)). Accordingly, numerical weather prediction (NWP) models have been run increasingly as ensemble forecasting systems. The goal of such ensemble forecasts is to approximate the forecast probability distribution by a finite sample of scenarios (Leith, 1974). This sample provides an estimate of forecast uncertainty. Global ensemble forecast systems, like the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, are prone to probabilistic biases, and are therefore not reliable. They particularly tend to be underdispersive for surface weather parameters (Bougeault et al., 2010; Park et al., 2008). Probabilistic hydrologic, i.e. river runoff, forecast models, which are driven by ensemble weather forecasts, tend to inherit the inadequate representation of forecast uncertainty. In order to correct for underdispersion and bias in NWP ensembles, statistical post processing methods have been developed, of which ensemble model output statistics (EMOS, Gneiting et al., 2005) is among the most widely applied. EMOS yields a parametric forecast distribution by linking its parameters to ensemble statistics. Due to its versatility and low computational cost, we focus on EMOS in the studies presented in this thesis. In general, post processing methods like EMOS convey considerable improvements in forecast skill. The main goal of post processing is to achieve well calibrated and yet sharp probabilistic predictions (Raftery et al., 2005; Gneiting et al., 2007a). In case of well calibrated forecasts, the theoretical levels of prediction intervals are equal to the relative frequency of the observations to lie within the corresponding forecast intervals. Sharpness relates only to the forecasts and denotes how "narrow" prediction intervals are at a given nominal level.

In this work several advances in statistical post processing are presented. The main goal is to improve skill of probabilistic forecasts for different atmospheric variables and runoff. Already available, state-of-the-art statistical post processing methods are used when they are applicable. Otherwise, novel methods are

developed. In the following, we summarize these contributions.

The first study presents the work by Hemri et al. (2014b) which applies statistical post processing to ensemble forecasts of near-surface temperature, 24-hour precipitation totals, and near-surface wind speed from the global model of the ECMWF. The main objective is to evaluate the evolution of the difference in skill between the raw ensemble and the post processed forecasts. Reliability and sharpness, and hence skill, of the former is expected to improve over time. Thus, the gain by post processing is expected to decrease. Based on ECMWF forecasts from January 2002 to March 2014 and corresponding observations from globally distributed stations we generate post processed forecasts using EMOS for each station and variable. Given the higher average skill of the post processed forecasts, we analyze the evolution of the difference in skill between raw ensemble and EMOS. This is discussed in detail in Section 3.1.

The second study presents an approach by Hemri et al. (2016) to post process ensemble forecasts for the discrete and bounded weather variable of total cloud cover. Two methods for discrete statistical post processing of ensemble predictions are tested: The first approach is based on multinomial logistic regression, the second involves a proportional odds logistic regression model. Both methods are applied to TCC raw ensemble forecasts from the ECMWF. The performance of the TCC post processing methods is assessed based on a stationwise post processing scheme that covers forecasts for a global set of 3330 stations over the period from January 2007 to March 2014. This is discussed in detail in Section 3.2.

In hydrologic forecasting systems, data below or above certain threshold values are subject to increased uncertainty. This may be due to very uncertain or not defined data, when, for instance, exceeding the range of the measured pairs of gauge levels and runoff values, on which the rating curve is based on. In the third study, which is based on Hemri et al. (2014a), a post processing method is presented that is tailored to the left censored runoff values encountered in the forecasting system of the German Federal Institute of Hydrology (BfG). On the basis of EMOS, we develop a censored EMOS method that is able to cope with censored data. The censored EMOS method is applied to ensemble runoff forecasts for the gauge Friedrichsthal, river Wied, and the gauge Altenahr, river Ahr, which both are sub-catchments of river Rhine. Censored EMOS forecasts are then verified for the period from November 2008 to October 2011 over the entire forecast horizon from 1 to 114 hours using several different statistical measures. This is discussed in detail in Section 4.2.

The analyses on multivariate post processing of hydrologic ensemble forecasts by Hemri et al. (2015) are presented in the fourth study. Runoff is an inherently multivariate process with typical events lasting from hours in case of floods to weeks or even months in case of droughts. This calls for multivariate post

processing techniques that yield well calibrated forecasts in univariate terms and ensure a realistic temporal dependence structure at the same time. To this end, the univariate EMOS post processing method is combined with two different copula approaches that ensure multivariate calibration throughout the entire forecast horizon. The domain of this study covers three sub-catchments of the river Rhine that represent different sizes and hydrological regimes: the Upper Rhine up to the gauge Maxau, the river Moselle up to the gauge Trier, and the river Lahn up to the gauge Kalkofen. In this study the two approaches to model the temporal dependence structure are ensemble copula coupling (ECC: Schefzik et al. (2013)), which preserves the dependence structure of the raw ensemble, and a Gaussian copula approach (GCA: Pinson and Girard (2012)), which estimates the temporal correlations from training observations. This is discussed in detail in Section 4.3.

Additional to the main studies mentioned above, a few smaller studies focusing on hydrological forecasts, are presented as well. These studies are preliminary in that we recommend further research in order to either confirm or refute our findings. The first smaller study in Section 4.4 concerns an analysis of hydrological regime dependent post processing. The second one, which is presented in Section 4.5, summarizes a method to convert probabilistic runoff forecasts into deterministic forecasts in a sound way. Finally, Section 4.6 assesses whether the skill of seasonal hydrological forecasts can be improved by state-of-the-art statistical post processing methods.

From a historical point of view, statistical post processing methods have emerged from hydrometeorological raw ensemble forecasts that have become increasingly affordable over the last few decades. In order to place the topic of post processing into a broader perspective, an overview of raw ensemble forecasting is provided next in Section 1.2. The statistical post processing and verification methods needed for the studies mentioned above are presented in Section 2. This is followed by Chapters 3 and 4 that discuss the meteorological and hydrological studies in detail. Along with a short outlook on further research concluding remarks are provided in Chapter 5.

1.2 Raw ensemble forecasting

In this section, an introduction to (raw) ensemble forecasting is given, which follows closely Hemri (2016).

1.2.1 Introduction to ensemble forecasting

Despite of the uncertainty inherent to any forecasting problem, deterministic forecasts have been the state of the art in hydrometeorological forecasting over many decades. Even with the best physical models substantial predictive uncertainty remains. Predictive uncertainty denotes the uncertainty conditional on the fore-

caster's expertise and the information set available (Krzysztofowicz, 1999; Todini, 2008). In order to assess predictive uncertainty, the paradigm in hydrometeorological forecasting has shifted from deterministic to probabilistic forecasting over the last two decades (see e.g. Palmer (2000) and Hamill et al. (2000)). The first meteorological ensemble prediction systems (EPS) have been developed in the early 1990s. For atmospheric variables like temperature, air pressure, wind speed, or precipitation, an EPS provides an estimate of their predictive distribution. This estimate is obtained by running the same NWP model multiple times with different initial conditions and/or model variants. Hence, in ideal settings, ensemble forecast members can be interpreted as random samples from the unknown predictive distribution. Or in other words, an ensemble of parallel forecast runs may also be understood in a probabilistic manner through its empirical cumulative distribution function. The increasing availability of meteorological ensemble predictions gave rise to the development of hydrologic ensemble forecasts (Cloke and Pappenberger, 2009).

1.2.2 The poor person's approach

The first EPS methods like the ones implemented in the 1990s by the U.S. National Centers for Environmental Prediction (NCEP) and the ECMWF took only account for the uncertainty about initial states. As stated by Ziehmann (2000), these models completely neglected model uncertainty. A simple alternative to EPS forecasting is to account for NWP model uncertainty by combining the predictions from several independent weather centers without applying any modifications to the actual forecasts. This almost cost free procedure is referred to as the poor person's approach (Arribas et al., 2005; Atger, 1999; Ebert, 2001; Ziehmann, 2000). The ensemble size of global poor person's ensembles constructed from deterministic NWP forecasts is usually limited to a few members, because there are only a few weather centers that run global atmospheric models. Nevertheless, poor person's approaches proved to perform well in comparison with EPS ensemble forecasts from the NCEP and the ECMWF. Poor person's ensembles often lack a correct representation of spread, but they usually perform quite well in terms of forecast resolution. Here, resolution refers to the ability of a model to issue case dependent forecasts that differ from climatological forecasts. Refer to Hersbach (2000) for further details on forecast resolution.

1.2.3 Recent developments in ensemble forecasting

Let us now focus again on the topic of uncertainty quantification by EPS forecasting. Within a single EPS, combining several model runs that have been generated in slightly different ways, i.e. perturbations in initial states and/or modified model parameters, accounts for the corresponding sources of uncertainty. Additionally, it can be accounted for model formulation uncertainty by combining ensemble forecasts issued by different weather centers to a multi-model ensemble.

Obviously, this applies the idea of the poor person's ensemble to EPS forecasts. Hence, a multi-model raw ensemble is a physically based approach to quantify multifaceted sources of uncertainty, namely the uncertainties in initial conditions, parameterizations, model structure, and data assimilation methods of the different meteorological ensembles. The THORPEX Interactive Grand Global Ensemble (TIGGE: Bougeault et al. (2010); Park et al. (2008); Richardson et al. (2005)) project ensemble is the most prominent example of such a multi-model combination. The global TIGGE ensemble, which currently comprises the EPS forecasts from ten different operational centers, exhibits high forecast skill. According to Hagedorn et al. (2012) a reduced TIGGE ensemble consisting only of the four, often considered to be the best, EPSs provided by the Canadian Meteorological Center, the NCEP, the UK Met Office, and ECMWF showed an improved performance on the global domain for 850 hPA temperature and 2 meter temperature compared to the best single-model EPS, the ECMWF EPS. For precipitation the same reduced TIGGE ensemble performed even better compared to the reforecast calibrated ECMWF EPS (Hamill, 2012). Moreover, the results indicated that statistical post processing of the reduced TIGGE ensemble did not provide as much improvement as post processing of the ECMWF EPS did. Based on these results, Hamill (2012) concluded that "all operational centers, even ECMWF, would benefit from the open, real-time sharing of precipitation data and the use of reforecasts". Of course, multi-model approaches can also be applied on regional domains. The Grand Limited Area Model Ensemble Prediction System (GLAMEPS), for instance, combines four regional EPS forecasts over the European domain (Iversen et al., 2011). Furthermore, multi-model approaches have also been used for seasonal forecasting. Palmer et al. (2004) summarize the development of the European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER). As part of the DEMETER project a multi-model ensemble seasonal forecasting system based on seven global atmospheric models has been tested. The results of these tests indicate that such a multi-model combination approach leads to more reliable seasonal-to-interannual predictions. Detailed analyses on the performance of the DEMETER multi-model ensemble can be found in Hagedorn et al. (2005), Doblas-Reyes et al. (2005) and Weinheimer et al. (2005).

Moving on to river runoff, Cloke and Pappenberger (2009) provide a review on ensemble flood forecasting. Most uncertainty in hydrological forecasting is related to uncertainty in the meteorological inputs for forecast horizons beyond 2-3 days. Meteorological uncertainty can be quantified by using meteorological EPS forecasts as inputs to the hydrologic models. As for the meteorological variables, probabilistic hydrologic forecasts benefit from multi-model ensemble forecasting. Therefore, hydrologic ensemble forecasts are often driven by input from several atmospheric models, issued by different weather centers, of which each is either deterministic or probabilistic (Bartholmes et al., 2009; Thielen et al., 2009).

Hydrological multi-model ensemble forecasts are used operationally by dif-

ferent flood warning services. For instance, the European Flood Alert System (EFAS) uses two deterministic models, i.e. the high-resolution run of the ECMWF and the deterministic model of the German Meteorological Service (DWD), and two ensemble models, i.e. the 51 member ECMWF ensemble and the 16 member Consortium for Small-Scale Modeling (COSMO) ensemble (Bartholmes et al., 2009; Thielen et al., 2009; European Flood Awareness System, 2014). On a regional scale there are a lot of similar flood alert systems. For instance, for river Sihl, which drains a pre-alpine sub-catchment of the river Rhine catchment, an operational hydrologic ensemble prediction system based on meteorological input from the COSMO ensemble and from the COSMO-7 deterministic model by MeteoSwiss is run routinely (Addor et al., 2011).

The above mentioned hydrological ensemble forecasts account only for uncertainties in the meteorological part of the runoff generation process. Georgakakos et al. (2004) performed the first study which quantified the uncertainty in hydrologic model structure by multi-model combination. Their multi-model ensemble consisted of ensemble members stemming from both calibrated and uncalibrated deterministic hydrologic models. In this context, calibration refers to the adjustment of the hydrological model parameters like, for instance, the percolation rate to the catchment of interest, and not to calibration in a statistical sense. If not indicated otherwise, henceforth the term calibration refers to statistical calibration, which is introduced in Section 2.3.1. The multi-model ensemble by Georgakakos et al. (2004) performed quite well in terms of calibration and its mean performed better than the best single model in terms of quadratic error, which is in line with the results from the meteorological studies on the poor person's ensemble. Zappa et al. (2011) have introduced a framework that investigates the relative contributions of meteorological inputs, initial conditions, and hydrologic model parameter estimates to the total predictive uncertainty. Within this framework, a large multi-model ensemble is constructed, which consists of any permutation of the meteorological raw ensemble members, of the weather radar precipitation field ensemble members, which account for uncertainty in initial conditions, and of an ensemble of equifinal parameter sets.

Chapter 2

Methods

In this chapter, methods for both univariate and multivariate statistical post processing are presented in detail. Along with a short overview over different post processing approaches, the methods used in Chapters 3 and 4 are discussed in Sections 2.1 and 2.2. This is followed by an overview over the verification measures used in this thesis in Section 2.3. For the sake of readability methodological details specific to particular studies are omitted here and introduced along with the corresponding studies. Note that all analyses have been performed using the statistical software `R` (R Development Core Team, 2014).

2.1 Univariate post processing

2.1.1 Introduction to post processing of multi-model ensemble forecasts

One of the first methods for post processing of multi-model forecasts was the multi-model superensemble by Krishnamurti et al. (1999, 2000), which is a regression technique that is closely related to the poor person's approach introduced in Section 1.2.2. In short, the superensemble is a statistical technique that fits a multiple regression model with the members of a poor person's ensemble as predictors and the observations as dependent variable. Accordingly, parameters have to be estimated based on training data. The coefficients of the superensemble are estimated separately for each location and each variable. Here, location refers to both the geographical location and the vertical position in the atmosphere. Since the coefficients of the superensemble model reflect the performance of the different members in the poor person's ensemble, they can also be understood as weights assigned to the different members. Predictions from such a model proved to outperform any of the poor person's ensemble members in terms of root mean squared error (Krishnamurti et al., 1999, 2000). They also outperformed the ensemble mean and the mean of individually bias corrected members of the poor person's ensemble. The superensemble technique is deterministic in that it generates a deterministic forecast of increased skill by post processing the output from

an ensemble forecast, which in most cases consists of a poor person's ensemble.

State of the art techniques for univariate, i.e. each lead-time and each location is considered independently, probabilistic multi-model combination and simultaneous statistical post processing include Bayesian model averaging (BMA) developed by Raftery et al. (2005), and the ensemble model output statistics (EMOS) method introduced by Gneiting et al. (2005). An illustrative example of both methods is given in Figure 2.1. Subsequent to introductions to exchangeable ensemble members, BMA and EMOS are discussed in detail. This is followed by an introduction to discrete post processing of TCC ensemble forecasts in Section 2.1.2 and by a discussion of methods that are tailored to hydrologic forecasts in Section 2.1.3. Note that the description of univariate post processing methods closely follows Gneiting (2014) and Hemri (2016).

Exchangeable ensemble members

The concept of exchangeable ensemble members is introduced here. As already stated, meteorological EPSs give an estimate of the forecast uncertainty by providing a finite sample of forecast scenarios. Each scenario is represented by an ensemble member. In case of an exchangeable ensemble, like for instance the COSMO Limited Area Ensemble Prediction System (COSMO-LEPS: Montani et al., 2011), the ensemble members lack individually distinguishable physical features. Statistical post processing methods have to take account of exchangeable ensemble members (Fraley et al., 2010; Gneiting et al., 2005). Therefore, parameters of the BMA and EMOS models are constrained to be equal within each exchangeable group.

Bayesian model averaging

BMA is a general method that has been developed originally in order to assess, and include, model uncertainty in situations where several competing models are available to predict the same variable. Closely following Hoeting et al. (1999), a brief summary of BMA is given here. Let y be the variable to be predicted, e.g. an atmospheric variable or river runoff, and $\boldsymbol{r}$ be the available data, e.g. a raw ensemble forecast. With $\mathcal{M}_1, \ldots, \mathcal{M}_M$ being the M competing models, the posterior distribution of y given $\boldsymbol{r}$ can be written as

$$p(y|\boldsymbol{r}) = \sum_{m=1}^{M} p(y|\,\mathcal{M}_m, \boldsymbol{r}) p(\mathcal{M}_m\,|\boldsymbol{r}), \tag{2.1}$$

which corresponds to a weighted average of the of the posterior distributions $p(y|\,\mathcal{M}_m, \boldsymbol{r})$. The posterior model probabilities $p(\mathcal{M}_k\,|\boldsymbol{r})$ can be understood as model weights, which in turn are given by

$$p(\mathcal{M}_k\,|\boldsymbol{r}) = \frac{p(\boldsymbol{r}|\,\mathcal{M}_m)p(\mathcal{M}_m)}{\sum_{\ell=1}^{M} p(\boldsymbol{r}|\,\mathcal{M}_\ell)p(\mathcal{M}_\ell)}, \tag{2.2}$$

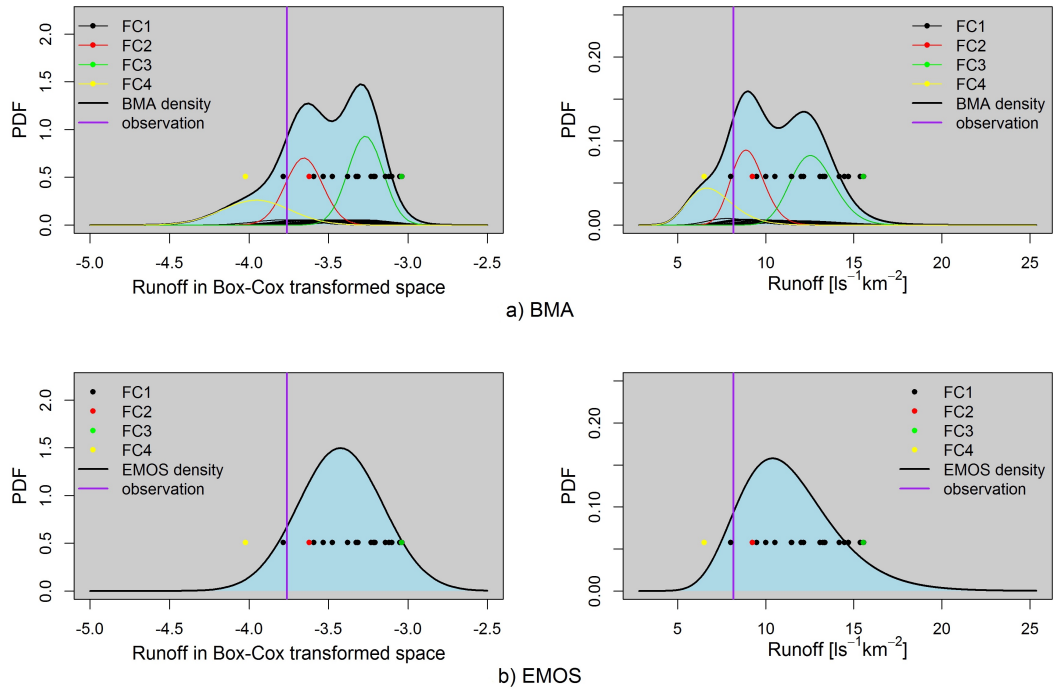


Figure 2.1: Examples of a) BMA and b) EMOS predictive probability density functions (PDFs). The panels on the left show the PDFs in the Box-Cox transformed space, those in the panels on the right are in the original space. FC1 to FC4 refer to different forecast models, of which FC1 is an ensemble of size 16 and the others are deterministic. Refer to Figure 4.7 in Section 4.3 for details on these forecast models. The horizontally aligned dots show the values of the raw ensemble members. Figure taken from Hemri (2016).

where $p(\mathcal{M}_m)$ denotes the prior probability that $\mathcal{M}_m$ is the true model. The likelihood under model $\mathcal{M}_m$ can be calculated as

$$p(\boldsymbol{r}|\,\mathcal{M}_m) = \int p(\boldsymbol{r}|\boldsymbol{\theta}_m, \mathcal{M}_m) p(\boldsymbol{\theta}_m|\,\mathcal{M}_m) \mathrm{d}\boldsymbol{\theta}_m, \tag{2.3}$$

where $\boldsymbol{\theta}_m$ denotes the parameter vector for model $\mathcal{M}_m$. The vector $\boldsymbol{\theta}_m$ may, for instance, refer to the parameters of a particular post processing model, $p(\boldsymbol{r}|\boldsymbol{\theta}_m, \mathcal{M}_m)$ denotes the likelihood under model $\mathcal{M}_m$ and parameter value $\boldsymbol{\theta}_m$, and $p(\boldsymbol{\theta}_m|\,\mathcal{M}_m)$ is the prior density of $\boldsymbol{\theta}_m$ in model $\mathcal{M}_m$. In the following, the BMA variant by Raftery et al. (2005), which is used for post processing of hydrometeorological ensemble forecasts, is presented.

For a variable of interest, y, BMA links the raw ensemble forecast $\boldsymbol{r} = r_1, \ldots, r_M$ of size M to a mixture distribution of the form

$$y \mid \boldsymbol{r} \sim \sum_{m=1}^{M} w_m g_m(y \mid r_m), \tag{2.4}$$

where $y \mid \boldsymbol{r}$ denotes the predictive distribution of y conditional on the raw ensemble forecast $\boldsymbol{r}$. In the standard BMA approach the kernel densities $g_m(y \mid r_m)$ are parametric and depend on the raw ensemble member r_m in suitable ways. The weights $w_1, \ldots, w_M \geq 0$ sum to 1 and reflect the relative performances of the corresponding raw ensemble members over the training period.

As stated in Section 1.2.3 state-of-the-art probabilistic forecasts of weather variables and river runoff are usually based on multi-model approaches. The members of a particular meteorological centers ensemble are in general exchangeable. In the case of river runoff forecasts, the corresponding hydrologic ensemble members are exchangeable as well. Fraley et al. (2010) discuss the adaptation of the basic Gaussian BMA model in Equation (2.4) to ensembles with exchangeable members. Their model is given by

$$y \mid r_{1,1}, \ldots, r_{1,N_1}, \ldots, r_{M,1}, \ldots, r_{M,N_M} \sim \sum_{m=1}^{M} \frac{w_m}{N_m} \sum_{n=1}^{N_m} g_{m,n}(y \mid r_{m,n}), \tag{2.5}$$

where the members $r_{m,1}, \ldots, r_{m,N_m}$ are the exchangeable members of model m. The ensemble size N_m of model m equals one for a deterministic model like the ECMWF high resolution (HRES) run. For the 50 member ECMWF ensemble (ENS) it would equal 50. BMA methods for normal and gamma distributed kernel densities are implemented in the `R`-package `ensembleBMA` (Fraley et al., 2015).

Ensemble model output statistics

Gneiting et al. (2005) introduced the ensemble model output statistics (EMOS) or non-homogenous regression (NR) method, which models the predictive distribution as a single parametric distribution of the general form

$$y \mid \boldsymbol{r} \sim g(y \mid \boldsymbol{r}), \tag{2.6}$$

where we use the same notation as above for BMA (cf. Equation (2.4)). EMOS variants based on many different distributions g are applied in the studies in Chapters 3 and 4. For the sake of simplicity, we assume here a Gaussian density to be appropriate for the variable to be forecast and refer to Chapters 3 and 4 for examples of other distributions. The normal EMOS predictive distribution is

$$y \mid \boldsymbol{r} \sim \mathcal{N}(\mu, \sigma^2), \tag{2.7}$$

where $\mu = a_0 + a_1 r_1 + \cdots + a_M r_M$ and $\sigma^2 = b_0 + b_1 s^2$, where s^2 denotes the ensemble variance

$$s^2 = \frac{1}{M} \sum_{m=1}^{M} (r_m - \bar{r})^2, \tag{2.8}$$

with ensemble mean $\bar{r} = 1/M \sum_{m=1}^{M} r_m$.

In case of exchangeable members the mean parameter μ is given by $\mu = a_0 + a_1\bar{r}_1 + \cdots + a_M\bar{r}_M$, where $\bar{r}_1, \ldots \bar{r}_M$ are the means of each set of exchangeable ensemble members (i.e. the members stemming from the same EPS) given by

$$\bar{r}_m = \frac{1}{N_m} \sum_{n=1}^{N_m} r_{m,n}, \tag{2.9}$$

and the ensemble variance s^2 by

$$s^2 = \frac{1}{\sum_{m=1}^{M} N_m - 1} \sum_{m=1}^{M} \sum_{n=1}^{N_m} (r_{m,n} - \bar{r})^2, \tag{2.10}$$

where $\bar{r} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} r_{m,n} \; / \sum_{m=1}^{M} N_m$ denotes the ensemble mean. The coefficients $a_0 \in \mathbb{R}$, $a_1, \ldots, a_M, b_0, b_1 > 0$ are estimated by numerical optimization over a training period. To this end, a target function, which depends on the model coefficients and the observations, is minimized. Usually, the continuous ranked probability score (CRPS: Matheson and Winkler (1976); Hersbach (2000)) is well suited for that purpose. More details on the CRPS can be found in Section 2.3.1 about verification. For EMOS based on a Gaussian distribution, functions for model fitting are available in the `R` package `ensembleMOS` (Yuen et al., 2013).

The EMOS model can alternatively be formulated as an extended logistic regression (ExtLR) model (Wilks, 2009). In order to obtain a complete predictive distribution, ExtLR uses the threshold to be forecast, y, as an additional predictor. The ExtLR forecast cumulative density function (CDF) F is given by

$$F(y) = \frac{\exp(a_0 + a_1 r_1^{\alpha} + \cdots + a_M r_M^{\alpha} + b y^{\beta})}{1 + \exp(a_0 + a_1 r_1^{\alpha} + \cdots + a_M r_M^{\alpha} + b y^{\beta})} \quad \text{for} \quad y \geq 0, \tag{2.11}$$

where $\alpha > 0$ and $\beta > 0$ are fixed coefficients and $\boldsymbol{r} \in \mathbb{R}^M$ is the vector of predictors. Though originally developed for meteorological variables, ExtLR can also be used to post process hydrologic ensemble forecasts. For instance, Fundel and Zappa (2011) apply ExtLR to hydrological reforecasts following Wilks (2011) who uses the ensemble mean $\bar{r}$ and spread s as predictors. The corresponding ExtLR model can be written as

$$F(y) = \frac{\exp(a_0 + a_1 \bar{r}^{\alpha_1} + a_2 s^{\alpha_2} + b y^{\beta})}{1 + \exp(a_0 + a_1 \bar{r}^{\alpha_1} + a_2 s^{\alpha_2} + b y^{\beta})}, \tag{2.12}$$

where $\alpha_1, \alpha_2, \beta$ have to be determined based on the forecaster's knowledge and a_0, a_1, a_2, b are estimated, for instance, by maximum-likelihood. A more recent development of ExtLR allows for interaction between predictor $\boldsymbol{r}$ and threshold y (Ben Bouallègue, 2013).

2.1.2 Total cloud cover

As stated in Section 1.1 statistical post processing methods for TCC should take account of the discrete nature of the reported TCC data. Hence, among the different post processing methods those which contain some kind of a "logistic regression" core are expected to perform quite well. In the following, two discrete statistical post processing methods are presented: a method based on multinomial (or polytomous) logistic regression (MLR: Agresti and Kateri (2011)), and a method based on proportional odds logistic regression (POLR: Walker and Duncan (1967); McCullagh (1980); Ananth and Kleinbaum (1997); Messner et al. (2014)). Despite their differences, MLR and POLR are closely related.

MLR is a direct generalization of binary logistic regression. In the case of TCC, the sample space is restricted to the discrete observations of cloudiness which take values in $\Omega = \{0, 1/8, 2/8, \ldots, 1\}$. The elements of Ω, also called octas, refer to the proportion of the sky covered by clouds. In the following, the different TCC states are denoted by $z_j \in \Omega$, $j = 1, \ldots, J$. For instance, z_1 refers to a clear sky. Hence, the MLR model has to assign probabilities to the $J = 9$ different states of cloudiness based on raw ensemble statistics as predictors. Here, again a multi-model raw ensemble containing exchangeable ensemble members is considered. Like in Section 2.1.1 such a raw ensemble is given by $\boldsymbol{r} = (r_{1,1}, \ldots, r_{1,N_1}, \ldots, r_{M,1}, \ldots, r_{M,N_M})$, where the members $r_{1,1}, \ldots, r_{m,N_m}$ are the exchangeable members of model m and N_m denotes its size. Following Wilks and Hamill (2007), and Hamill et al. (2008) we link the ensemble spread to the MLR model using the ensemble variance as an additional predictor. The ensemble variance s^2 is given by Equation (2.10). Inspired by Scheuerer (2014) we have also tested the ensemble mean difference as a more robust alternative to the ensemble variance, which – at least in the settings of the study discussed in Section 3.2 – did not improve forecast skill. Again inspired by Scheuerer (2014) we introduce the predictors f_0 and f_1, which denote the ratio of ensemble members equal to zero or one, respectively. For instance,

$$f_0 = \frac{1}{\sum_{m=1}^{M} N_m} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \mathbb{1}_{[r_{m,n}=0]}, \tag{2.13}$$

where $\mathbb{1}_{[\cdot]}$ denotes the indicator function. Accordingly, the vector of predictors is given by

$$\boldsymbol{x} = (1, \bar{r}_1, \ldots, \bar{r}_M, s^2, f_0, f_1)^T, \tag{2.14}$$

where $\bar{r}_1, \ldots \bar{r}_M$ are the means of each set of exchangeable ensemble members. Selecting now a TCC state $z_0 \in \Omega$ as a pivot, the MLR model based on a random variable Z can be written as

$$\log \frac{P(Z = z_j \neq z_0)}{P(Z = z_0)} = \boldsymbol{\beta_j} \boldsymbol{x}, \tag{2.15}$$

where $\boldsymbol{\beta_j}$ is the vector of coefficients for state z_j. Though any state z_j could be used as pivot state z_0, it is most natural to set $z_1 = z_0$. Then, model (2.15) has to be fitted $J-1$ times such that the probabilities sum up to 1. Using a suitable training period, this model can easily be estimated using the function `multinom` of the `R` package `nnet` (Ripley and Venables, 2014).

POLR (Walker and Duncan, 1967; McCullagh, 1980; Ananth and Kleinbaum, 1997) is an alternative to MLR. POLR is well suited for ordinal data like TCC. Since it assumes proportional odds, it requires fewer free parameters. This allows to add an interaction term to the set of predictors used in the MLR model. This term represents the interaction between the ensemble variance s^2 and the deviation of $\bar{r}^*$ from 0.5, where $\bar{r}^* = \frac{\sum_{m=1}^{M} \bar{r}_m}{M}$. The rationale behind this is to map s^2 to the variance of the post processed ensemble in a more natural way than in the MLR model. More specifically, the interaction term is defined as $I := s^2 \text{sign}(d) d^2$, where $d = (\bar{r}^* - 0.5)$. This formulation is expected to shift extreme TCC forecasts towards the center, if s^2 is large, and at the same time $\bar{r}^*$ is close to zero or one. Let $\pi_j = P(Z \leq z_j)$ be the cumulative predictive probability for TCC states. Then, the POLR model can be written as

$$\text{logit}(\pi_j) = \log \frac{\pi_j}{1-\pi_j} = \theta_j - \boldsymbol{\beta x}, \tag{2.16}$$

where the coefficient θ_j takes a different value for each state z_j and $\theta_1 < \theta_2 < \cdots < \theta_J$ are strictly ordered. The coefficients $\boldsymbol{\beta} = (\beta_{\bar{r}_1}, \ldots, \beta_{\bar{r}_M}, \beta_{s^2}, \beta_{f_0}, \beta_{f_1}, \beta_I)$ do not change with state. Additionally, $\beta_{\bar{r}_1}, \ldots, \beta_{\bar{r}_M}$ are constrained to be non-negative. This is ensured by estimating the model iteratively. In each iteration step negative estimates for any $\beta_{\bar{r}_m}$, $m = 1, \ldots, M$, are set to zero and the model is re-estimated without the corresponding predictor $\beta_{\bar{r}_m}$. This iterative procedure stops as soon as $\min(\hat{\beta}_{\bar{r}_1}, \ldots, \hat{\beta}_{\bar{r}_M}) \geq 0$. As stated above, the assumption of proportional odds makes POLR much sparser than MLR. For the MLR model $(p+1)(|\Omega|-1)$ coefficients have to be estimated, where $|\Omega|$ is the number of different states and p denotes the number of predictors not counting the intercept. In case of the POLR model we need only $p + |\Omega|$ coefficients. POLR is implemented in the function `polr` of the `R` package `MASS` (Venables and Ripley, 2002). Example forecasts obtained by the POLR approach are presented in the context of the study in Section 3.2.

2.1.3 River runoff

River runoff data are undoubtedly non-Gaussian. In order to be able to resort to post processing methods relying on Gaussian distributions, both the observations and the raw ensemble predictions have to be transformed such that forecast errors are approximately normally distributed. Like Duan et al. (2007) we use the Box-Cox transformation (Box and Cox, 1964) for that purpose. Details on how the Box-Cox transformation has been implemented for the studies presented in

Sections 4.2 and 4.3 can be found in Appendix A.2.1.

Truncated normal EMOS

In order to avoid positive probabilities for unrealistically high runoff forecasts, the normal EMOS method presented in Section 2.1.1 has to be replaced by a truncated normal EMOS method that is closely related to the one proposed by Thorarinsdottir and Gneiting (2010). The main difference is that the left-truncation is replaced by a right-truncation. Given a raw ensemble $\boldsymbol{r} = r_{1,1}, \ldots, r_{M,N_M}$, one first applies a Box-Cox transformation h that yields Box-Cox transformed ensemble members $f_{m,n} = h(r_{m,n})$. Accordingly, the Box-Cox transformed mean of a set of exchangeable members would be $\bar{f}_m = N_m^{-1} \sum_{n=1}^{N_m} f_{m,n}$. The upper limit b of the predictive truncated normal distribution has to be selected based on the forecaster's expertise. Depending on the actual Box-Cox transformation parameter estimate and the properties of the catchment of interest a lower limit might be needed as well. With $\mathcal{N}^b(\mu, \sigma^2)$ denoting a right truncated normal distribution with support $(-\infty, b]$ the truncated EMOS predictive density for the variable of interest y, here Box-Cox transformed runoff, can be written as

$$p(y \mid f_{1,1}, \ldots, f_{1,N_1}, f_{2,1}, \ldots, f_{M,N_M}) = \mathcal{N}^b(\mu, \sigma^2), \tag{2.17}$$

where $\mu = a_0 + a_1\bar{f}_1 + a_2\bar{f}_2 + \cdots + a_M\bar{f}_M$ and $\sigma^2 = c_1 + c_2 s^2$ depends on the ensemble variance s^2. With this model formulation, truncated EMOS also accounts for heteroscedasticity, i.e. heterogeneity in variances. Constraints on the parameters are: $a_0 \in \mathbb{R}$ and $a_1, a_2, \ldots, a_M, c_1, c_2 \in \mathbb{R}_+$. The ensemble statistics μ and s^2 are computed using Equations (2.9) and (2.10), respectively.

Censored EMOS

As already mentioned, raw ensemble forecasts for runoff gauges with censored observations and forecasts such as the rivers Wied and Ahr need a specific, i.e. censored, post processing method. In order to obtain a post processing method suitable for censored data, a flexible, yet not too complex, post processing method should be adapted, such that censored data are handled properly by the statistical model. Truncated normal EMOS proved to be a good starting point to develop a censored EMOS method. Here, we present an EMOS model that is based on a truncated normal distribution with point mass at the censoring threshold. Details on such censored distributions can be found in Gneiting et al. (2004). The censored EMOS approach moves the forecast density mass below the threshold of zero (since negative runoff has zero probability) to a point mass at zero, whereas a similar truncated normal approach would shift this density mass to the interval $[0, \infty)$. The idea of censored raw ensemble forecasts and censored post processing are illustrated in Figures 2.2 and 2.3, respectively. Both figures are based on the raw ensemble forecasts issued on 9 May 2009 for river Wied at gauge

Friedrichsthal. Details on the raw ensemble and the underlying meteorological input models can be found in the studies presented in Chapter 4. The runoff data from the catchments, considered in the just mentioned studies, need to be Box-Cox transformed with a quite extreme transformation parameter λ in order to meet the assumption of normality. As for truncated normal EMOS, back-transforming the post processed censored EMOS forecast to the original space may lead to positive probabilities for unrealistically high runoff values. This can be avoided by using an EMOS approach based on a left censored and right truncated normal distribution. For technical simplicity, the raw ensemble forecasts have to be transformed in a way such that censoring at zero makes sense. Therefore, the raw ensemble runoff means are Box-Cox transformed and shifted leading to transformed ensemble means $\bar{f}_m = h(\bar{r}_m) - h(d)$, where d denotes the lower threshold, i.e. the value at which left censoring is applied. This leads to the CDF

$$F(y) = \begin{cases} 0 & \text{if } y < 0, \\ \frac{\Phi(\frac{y-\mu}{\sigma})}{\beta} & \text{if } 0 \leq y \leq v \\ 1 & \text{if } y > v, \end{cases} \tag{2.18}$$

where Φ denotes the CDF of the standard normal distribution and $\beta = \Phi\big(\frac{v-\mu}{\sigma}\big)$ is the cumulative density at the transformed upper threshold v. The variance σ^2 depends on the raw ensemble in the same way as described above for the truncated normal EMOS model. However, the parameterization of the location parameter μ has to be adapted such that point masses $p > 0.5$ at the censoring threshold are allowed. To this end, the intercept parameter a_0 is now allowed to take values in $\mathbb{R}$. Additionally, the ratio of ensemble means $\bar{f}_m$ that are equal to the censoring threshold is used as an additional predictor. With this, the location parameter depends on

$$\mu = a_0 + a_1\bar{f}_1 + a_2\bar{f}_2 + \cdots + a_M\bar{f}_M + a_{M+1}\pi_0, \tag{2.19}$$

where π_0 is defined as the ratio of ensemble means equal to zero,

$$\pi_0 = \frac{1}{M}\sum_{m=M}^{I} \mathbb{1}_{\{\bar{f}_m=0\}}. \tag{2.20}$$

Though this model works quite well, the model may benefit from a modification of the parametrizations for parameters μ and σ, such that, despite of the application of a truncated normal distribution, the expected value of the forecast density equals to the weighted mean of the raw ensemble means $\bar{f}_m$. Differently from the parameterization in Equation (2.19), for this approach censored EMOS is applied without explicit correction of systematic errors (bias correction) in order to enhance numerical stability of the model parameter estimation algorithms. If needed, any kind of bias correction may be added prior to the estimation of the EMOS parameters. The location parameter μ is now estimated using numerical optimization and has to fulfill

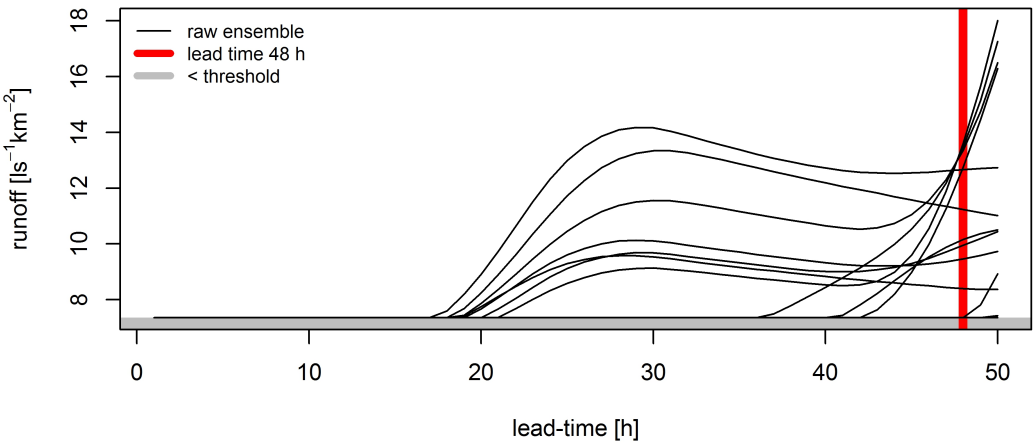


Figure 2.2: Example of a censored raw ensemble forecast covering lead times 1-50 h. Each of the 18 trajectories corresponds to a particular member of a multi-model raw ensemble forecast. Note that the trajectories may coincide at the threshold. Figure taken from Hemri et al. (2014a).

$$\mathbb{E}[Y|Y < v] = \sum_{m=1}^{M} w_m \bar{f}_m + a\pi_0 = \mu - \sigma \frac{\varphi\left(\frac{v-\mu}{\sigma}\right)}{\Phi\left(\frac{v-\mu}{\sigma}\right)}, \tag{2.21}$$

where w_m are weights with $\sum_{m=1}^{M} w_m = 1$. The parameterization of the variance parameter σ^2 depends on whether all ensemble members are equal to the lower threshold, i.e. the left-censoring threshold, or not. In the former case $\sigma^2 = c_1$ is used, since $s^2 = 0$, and in the latter case $\sigma^2 = c_1 + c_2 s^2$. The term $\sigma\left[\varphi\left(\frac{v-\mu}{\sigma}\right)\right] / \left[\Phi\left(\frac{v-\mu}{\sigma}\right)\right]$ in Equation (2.21) corrects the location parameter μ such that truncation of the forecast density does not lead to any systematic bias.

2.2 Multivariate extensions

Recently different methods to incorporate multivariate dependence structures into the post processing of ensemble forecasts have been proposed. Let us first have a look at non-parametric reordering approaches that comprise mainly the Schaake shuffle (Clark et al., 2004) and ensemble copula coupling (ECC: Schefzik et al. (2013)). Both approaches implicitly rely on empirical copulas. The notion of a copula is critical in Sklar's theorem (Sklar, 1959), which states that any L-variate CDF F with margins $F_1, \ldots, F_L$ can be represented by

$$F(y_1, \ldots, y_L) = C(F_1(y_1), \ldots, F_L(y_L)), \tag{2.22}$$

where $y_1, \ldots, y_L \in \mathbb{R}$ and C: $[0,1]^L \to [0,1]$ is a multivariate CDF with standard uniform marginal distributions. In case of ensemble forecasts, the rank order

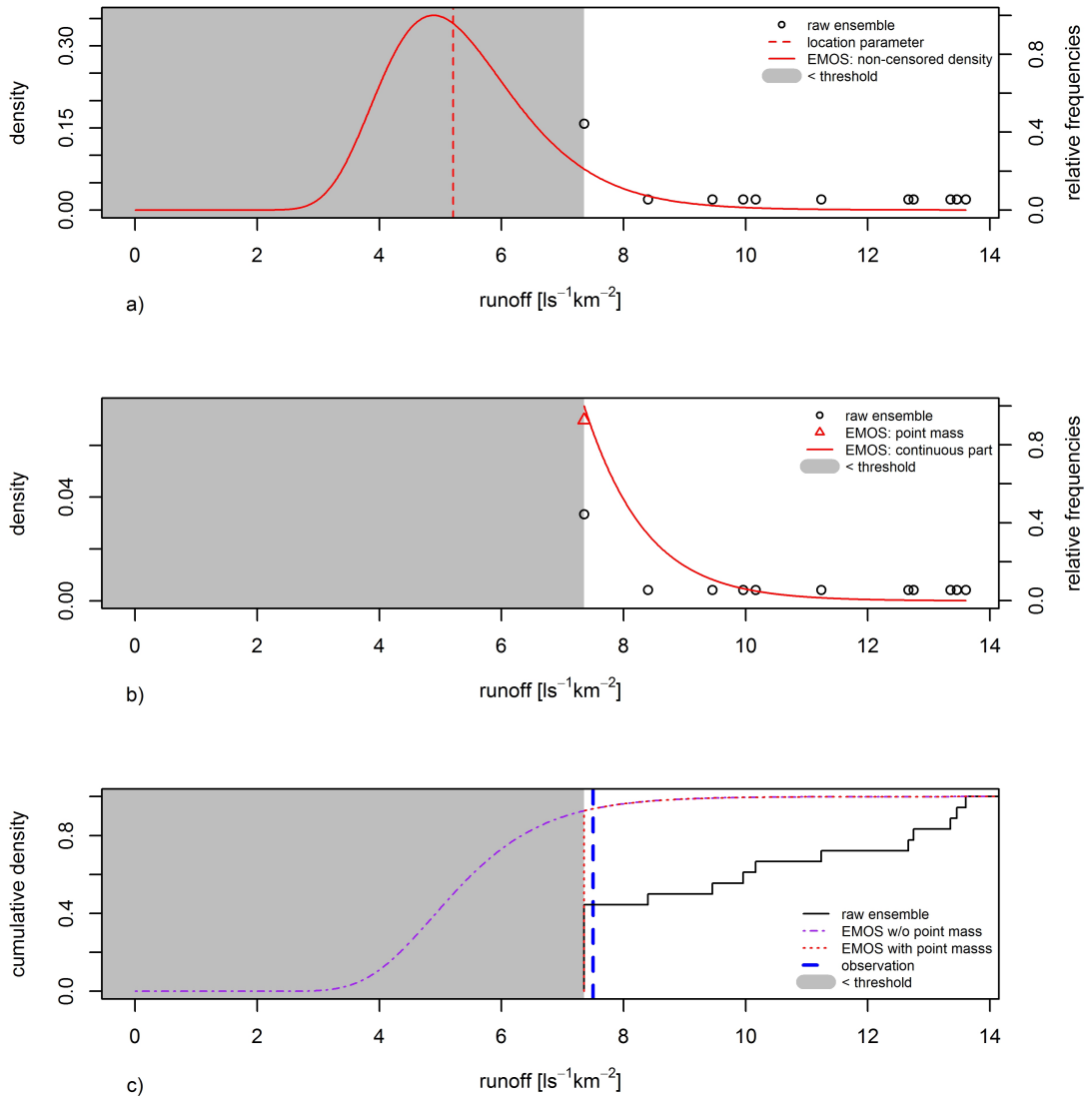


Figure 2.3: Example of a 48 h censored EMOS forecast. From top to bottom: non-censored EMOS PDF along with the relative frequencies of raw ensemble values; censored EMOS PDF with point mass at the threshold; CDF of non-censored EMOS (EMOS w/o point mass), censored EMOS (EMOS with point mass), the raw ensemble and the verifying observation. Figure taken from Hemri et al. (2014a).

structure of the ensemble members defines an empirical copula over the forecast margins, which correspond to particular lead times and locations.

2.2.1 Schaake shuffle

Before introducing the multivariate extensions used for multivariate post processing of river runoff forecasts in the study presented in Section 4.3, we discuss first the Schaake shuffle, which can somehow be understood as a precursor of ECC. The Schaake shuffle transfers historical spatio-temporal dependence structures to the forecasts of interest. For simplicity, it is assumed here that only one variable, like for instance runoff, precipitation, or temperature, is of interest. Of course, the method could easily be used to model dependence structures between different variables. Additionally, the difference between single model and multi-model ensembles is ignored here. Let $\boldsymbol{R}_{m,t,s}$ be the raw ensemble forecast array at a specific day. The index $m = 1, \ldots, M$ refers to the ensemble members, $t = 1, \ldots, T$ to the lead times, and $s = 1, \ldots, S$ to the locations. Then a corresponding observation array $\boldsymbol{Y}_{m,t,s}$ of equal size is selected. $\boldsymbol{Y}_{m,t,s}$ is constructed by selecting the same number of historical observations as there are ensemble members. Times of day, such that lead times are reflected correctly, and locations have to be equal in $\boldsymbol{R}_{m,t,s}$ and in $\boldsymbol{Y}_{m,t,s}$. In Clark et al. (2004) the dates of the historical observations are selected such that they match the calendar day of the forecast of interest by ± 7 days, regardless of the year. The multi-index $\ell = (s, t)$ defines the margins at which univariate statistically post processed ensemble forecasts are available. For a given location s and lead time t, i.e. margin ℓ, the Schaake shuffle can be summarized as follows:

1. Sort the forecast vector $\boldsymbol{R}_\ell = (r_1^\ell, \ldots, r_M^\ell)$ such that $\tilde{\boldsymbol{R}}_\ell = (\tilde{r}_1^\ell, \ldots, \tilde{r}_M^\ell) = (r_{(1)}^\ell, \ldots, r_{(M)}^\ell)$, with $r_{(1)}^\ell \leq r_{(2)}^\ell \leq \cdots \leq r_{(M)}^\ell$.

2. Sort the observation vector $\boldsymbol{Y}_\ell = (y_1^\ell, \ldots, y_M^\ell)$ such that $\tilde{\boldsymbol{Y}}_\ell = (\tilde{y}_1^\ell, \ldots, \tilde{y}_M^\ell) = (y_{(1)}^\ell, \ldots, y_{(M)}^\ell)$, $y_{(1)}^\ell \leq y_{(2)}^\ell \leq \cdots \leq y_{(M)}^\ell$, and denote the corresponding ranks by rk_m^ℓ.

3. Construct the reordered forecast vector $\boldsymbol{R}^{ss} = \left(\tilde{r}_{\mathrm{rk}_1^\ell}^\ell, \ldots, \tilde{r}_{\mathrm{rk}_M^\ell}^\ell\right)$.

The above reordering procedure is applied to all margins ℓ. With this, the Schaake shuffle preserves the Spearman rank correlation structure between the margins.

2.2.2 Ensemble copula coupling

The non-parametric ECC approach reorders samples from the predictive densities that may, for instance, have been obtained by EMOS. An illustrative example of how ECC retains the rank order structure of the raw ensemble, while still following the post processed marginal distributions, is given in Figure 4.7 in Section 4.3.3. For simplicity, only dependences between different lead times, which are now denoted by $l = 1, \ldots, L$ are considered here. Among the different ECC approaches discussed in detail in Schefzik et al. (2013) only ECC-T is applicable to the strongly auto-correlated hourly runoff predictions, i.e. yields realistic runoff

trajectories. The other approaches lead to unrealistic jumps between consecutive lead times due to the processes used for the selection of samples from the predictive distributions. For ECC-T one first fits a parametric density function to the raw ensemble for each verification day and lead time separately. Then, one checks to which quantiles of these density functions the raw ensemble members correspond. By doing this over the entire range of lead times a set of trajectories of probabilities is obtained. Each trajectory corresponds to a raw ensemble member. Hence, the trajectories inherit the rank order structure from the raw ensemble. By extracting the corresponding quantiles from the post processed predictive distributions one eventually obtains the ECC-T trajectories. Technically, ECC-T operates as follows:

1. Assign unique and ordered indices $1, \ldots, K$ to the, possibly Box-Cox transformed, raw ensemble members $f^l_{m,n}$, so that the ensemble can be rewritten as $(f^l_1, \ldots, f^l_K)$. Even though the actual order of the indices does not matter, the index assigned to a particular raw ensemble member has to remain constant over all lead times.

2. Obtain for each lead time the reordered EMOS forecasts

$$\hat{y}^l_k = \hat{F}^{-1}_l(\hat{S}_l(f^l_k)), \quad k = 1, \ldots, K, \quad l = 1, \ldots, L, \tag{2.23}$$

   where $\hat{S}_l$ is the fitted CDF of a suitable parametric distribution to the, possibly Box-Cox transformed, raw ensemble and $\hat{F}^{-1}_l$ denotes the inverse of the marginal post processed CDF.

2.2.3 Gaussian copula approach

GCA is a parametric approach for modelling the correlation structure among different lead times. One first estimates a parametric correlation function from training data and then the respective multivariate normal distribution of dimension equal to the number of lead times. By sampling several times from this distribution and then evaluating the CDF of the univariate standard normal distribution at the sampled values, trajectories of probabilities are obtained. Extracting the corresponding quantiles from the post processed forecast distributions results in the GCA trajectories. The rank order structure of the GCA trajectories is independent from the rank order structure of the raw ensemble. For an example illustrating this we refer to Figure 4.7 in Section 4.3.3. Technically, GCA can be summarized as follows:

1. Calculate the empirical correlogram among lead times 1 to L from the observations in the training period.

2. Fit a correlation function to the empirical correlogram.

3. Sample K realizations, $(x^1_k, x^2_k, \ldots, x^L_k)$ with $k = 1, \ldots, K$, from a standard L-variate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu} = \mathbf{0}, \Sigma)$ with diagonal elements $\Sigma_{l,l} = 1$ and correlation structure from 2.

4. Using the inverse, $\hat{F}_l^{-1}$, of the post processed marginal forecast CDF for each individual lead time, multivariate scenarios with marginal distributions inherited from the univariate fits are obtained as

$$\hat{y}_k^l = \hat{F}_l^{-1}\big(\Phi\big(x_k^l\big)\big). \tag{2.24}$$

2.3 Verification

In the following, methods for the verification of probabilistic forecasts are discussed. For methodological details specific to the different case studies refer to the corresponding sections in Chapters 3 and 4.

2.3.1 Univariate verification

As stated in Gneiting et al. (2007a) probabilistic forecasts should be (statistically) well calibrated and yet sharp. In practice, univariate calibration is assessed via the probability integral transform (PIT: Dawid (1984); Diebold et al. (1998); Gneiting et al. (2007a)). The PIT value z for a particular verification day and lead time is defined as the value of the predictive CDF evaluated at the observation. According to Rosenblatt (1952) well calibrated continuous forecasts imply that $z \sim \mathcal{U}(0,1)$. In the present context, this means that the observations should look like random samples from the predictive distribution. When translating into bins and calculating the relative frequencies over the entire verification period, the PIT can be visualized by a histogram as shown in Figure 2.4 (Hamill, 2001). A flat histogram indicates well calibrated forecasts, whereas underdispersion is indicated by a $\cup$-shape and overdispersion by an $\cap$-shape, respectively.

In the context of hydrometeorological forecasting the above notion of calibration, which is also called probabilistic calibration, is the most important one. In some cases, marginal calibration may be of interest in practice as well. According to Gneiting and Katzfuss (2014) a forecast with CDF F is marginally calibrated if the mean forecast CDF equals the marginal CDF of the respective observations. An example of how to assess marginal calibration in the case of total cloud cover forecasts is provided in Appendix A.1.2. Sharpness refers to how focused probabilistic forecasts are. Sharpness can, for instance, be assessed by verifying the widths of prediction intervals at a given nominal level, e.g. in many cases the centered 90 % prediction intervals. The narrower those intervals, the sharper is the forecast. An example of box plot like diagrams that assess sharpness can be found in Gneiting et al. (2007a).

Besides the appealing concepts of calibration and sharpness, a representation of forecast skill in terms of a scalar number is desirable. To this end, many different scoring rules have been proposed. For the verification of probabilistic forecasts, proper scoring rules should be applied as they "encourage the forecaster

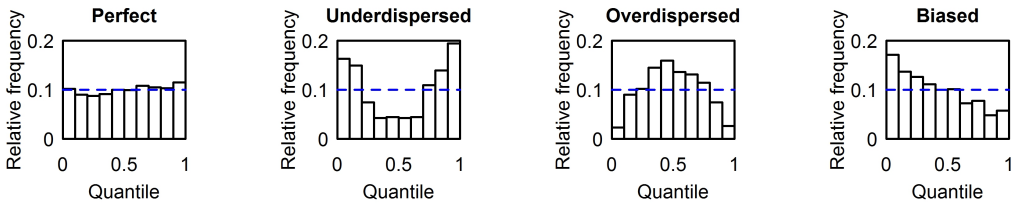


Figure 2.4: Left to right: PIT histograms for a well calibrated, an underdispersed, an overdispersed, and a biased forecast. Figure taken from Hemri et al. (2014a).

to make careful assessments and to be honest" (Gneiting and Raftery, 2007). If a forecaster issues the predictive distribution F and the event y materializes, a scoring rule $S(F, y)$ can be understood as the reward of the forecaster. Denoting the forecasters true belief by G, the negatively oriented, i.e. the lower the score the more skillful is the forecast, scoring rule $S(F, \cdot)$ is proper, if

$$\mathbb{E}_G[S(G, Y)] \leq \mathbb{E}_G[S(F, Y)], \tag{2.25}$$

for all F, G, where $\mathbb{E}_G[S(\cdot)]$ denotes the expectation of $S(\cdot)$ under G. If

$$\mathbb{E}_G[S(G, Y)] < \mathbb{E}_G[S(F, Y)], \tag{2.26}$$

holds for any $F \neq G$, $S(F, y)$ is called strictly proper. In hydrometeorological applications, the most widely used proper scoring rules are the CRPS (Matheson and Winkler, 1976; Hersbach, 2000)) and the logarithmic (log) score (Good, 1952). For a forecast with predictive CDF F, the CRPS is given by

$$\begin{aligned} \mathrm{CRPS}(F, y) &= \int_{-\infty}^{\infty} \left[F(u) - \mathbb{1}_{[u \geq y]}\right]^2 du \\ &= \mathbb{E}_F \mid Y - y \mid - \frac{1}{2} \mathbb{E}_F \mid Y - Y' \mid, \end{aligned} \tag{2.27}$$

where $Y, Y' \sim F$ are independent random variables with finite mean (Gneiting and Raftery, 2007). For a forecast with density p, the log score is given by

$$\log(p, y) = -\log p(y). \tag{2.28}$$

Skill of forecasts for dichotomous variables can be evaluated in a proper way using the Brier score (Brier, 1950), of which the CRPS is a generalization. The Brier score is given by

$$\mathrm{BS} = (p - o)^2, \tag{2.29}$$

where p denotes the predicted probability for an event x to occur, and o is an indicator function that returns 1 if the event x materializes and 0 otherwise.

In order to rank different forecasters, the scores are usually averaged over the verification period leading to the average score S_V

$$S_V = \frac{1}{V} \sum_{v=1}^{V} S(F_v, y_v). \tag{2.30}$$

The concept of skill scores has been developed in order to be able to compare forecasts for distinct sets of forecasts situations (e.g. Murphy (1973) and Gneiting et al. (2007a)). Hence, computing skill scores may, for instance, be beneficial when comparing competing hydrological forecast methods at different gauges. Following Gneiting and Raftery (2007) skill scores are calculated using

$$S_V^{\text{skill}} = \frac{S_V^{\text{f}} - S_V^{\text{c}}}{S_V^{\text{o}} - S_V^{\text{c}}}, \tag{2.31}$$

where S_V^{o} is the average score of an ideal forecast, i.e. a forecast that always assigns probability 1 to the value that materializes. The average scores S_V^{f} and S_V^{c} refer to the forecast of interest and the reference forecast, respectively. In practice, the reference forecast corresponds often to the climatology, i.e. the empirical distribution obtained from past observations. The skill score S_V^{skill} attains values in $(-\infty, 1]$ and is positively oriented. Forecasts with skill equal to the skill of the reference forecast lead to a skill score of zero.

2.3.2 Multivariate verification

As already mentioned, it is crucial to ensure a realistic correlation structure among the forecast margins in case of issuing multivariate predictions. To this end, Gneiting et al. (2008) proposed the multivariate rank histogram, which is a visual method to assess multivariate calibration, i.e. the correct representation of the dependence structure among different margins by the forecasts. For visual verification in high dimensional settings, as it is typically the case in the field of hydrometeorological forecasting, Thorarinsdottir et al. (2014) proposed the average rank and the band depth rank histogram. These rank histogram methods are further developments of the concept of the multivariate rank histogram. Given univariate calibration, they can be used to detect unrealistic correlation structures among lead times and/or locations of the forecast distribution. Following Thorarinsdottir et al. (2014) the average and the band depth rank histogram can be obtained as follows:

- Obtain M randomly sampled forecast trajectories (here over lead times 1 to 114 h, i.e. dimension $L = 114$) from the multivariate predictive distribution, where M corresponds to the size of the raw ensemble.
- Add the observed trajectory to the set of sampled forecast trajectories, leading to the set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M, \boldsymbol{x}_{M+1}\}$ of trajectories of dimension L with $\boldsymbol{x}_m = (x_{m1}, \ldots, x_{mL})$ for $m = 1, \ldots, M+1$. The observed trajectory $\boldsymbol{y}$ is now denoted by $\boldsymbol{x}_{M+1}$.

- Calculate pre-ranks using either the average pre-rank function

$$\rho_S(\boldsymbol{x}) = \frac{1}{L}\sum_{l=1}^{L} \text{rank}_S(x_l), \tag{2.32}$$

  or the band depth pre-rank function

$$\rho_S(\boldsymbol{x}) = \frac{1}{L}\sum_{l=1}^{L}\Big[\text{rank}_S(x_l)[m - \text{rank}_S(x_l)] + [\text{rank}_S(x_l) - 1]\sum_{m=1}^{M+1} \mathbb{1}_{[x_{ml}=x_l]}\Big],$$

  where $\text{rank}_S(x_l)$ denotes the rank of member $\boldsymbol{x}$ at lead time l.

- Obtain the rank of $\boldsymbol{x}_{M+1}$ by first calculating $\rho_S(\boldsymbol{x}_{M+1})$ and then determining its rank in $\{\rho_S(\boldsymbol{x}_1), \ldots, \rho_S(\boldsymbol{x}_M), \rho_S(\boldsymbol{x}_{M+1})\}$ with ties resolved at random.

Calculating the above ranks for each day in the verification period allows to plot PIT-like histograms. Though they look like univariate PIT histograms, their interpretation is somewhat different. Assuming the forecasts to be marginally, i.e. for each individual lead time, well calibrated, $\cup$-shaped average or band depth rank histograms indicate too low correlations among lead times, whereas $\cap$-shaped histograms indicate too high correlations. But note that these histograms are highly sensitive to marginal miscalibration. Refer to Thorarinsdottir et al. (2014) for further details.

For multivariate assessment of forecast skill Gneiting and Raftery (2007) proposed the energy score (ES) given by

$$\text{ES}(F, \boldsymbol{y}) = \mathbb{E}_F||\boldsymbol{X} - \boldsymbol{y}|| - \frac{1}{2}\,\mathbb{E}_F||\boldsymbol{X} - \boldsymbol{X'}||, \tag{2.33}$$

where $||\cdot||$ denotes the Euclidian norm. Here, $\boldsymbol{X}$ and $\boldsymbol{X'}$ are independent random vectors following the predictive distribution F with finite first moments and the observation vector is denoted by $\boldsymbol{y}$. The ES is negatively oriented, proper, and a generalization of the CRPS. If the forecast is available as an ensemble of size M with ensemble member trajectories $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M \in \mathbb{R}^L$, according to Gneiting et al. (2008) the ES can be calculated as

$$\text{ES}(F, \boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M}||\boldsymbol{f}_j - \boldsymbol{y}|| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}||\boldsymbol{f}_i - \boldsymbol{f}_j||. \tag{2.34}$$

In case of a deterministic forecast trajectory $\boldsymbol{f}$ the ES reduces to the Euclidian norm,

$$\text{ES}(f, \boldsymbol{y}) = ||\boldsymbol{f} - \boldsymbol{y}||. \tag{2.35}$$

Hence, the ES may be used to compare multivariate density forecasts, discrete ensemble forecasts, and deterministic forecasts (Gneiting et al., 2008). The ES discriminates well between forecasts with different mean vectors, and shows satisfying discrimination ability with regard to variance specification. But its ability to detect errors in the correlation structure is quite poor (Pinson and Girard, 2012; Pinson and Tastu, Pinson and Tastu; Scheuerer and Hamill, 2015). Scheuerer and Hamill (2015) recently developed the p-variogram score as a complement to the ES. Its main advantage is the much better discrimination ability between correct and misspecified correlation structures. The p-variogram score of order p is defined by

$$S_{\gamma_p}(F, \boldsymbol{y}) = \sum_{i,j=1}^{L} w_{ij} \big(|y_i - y_j|^p - \mathbb{E}_F |X_i - X_j|^p \big)^2, \tag{2.36}$$

where F denotes the L-variate predictive distribution, $\boldsymbol{y}$ is the observation vector of length L, X_i and X_j are the i-th and j-th component of a random vector $\boldsymbol{X} \sim F$, and $w_{ij} \geq 0$ are weights. As proposed in Scheuerer and Hamill (2015) pairs of far distant lead times are down-weighted in order to increase the signal to noise ratio. This is done by setting w_{ij} to be proportional to the inverse distance between i and j. Additionally, they have demonstrated by simulation experiments that setting $p = 0.5$ leads to a good discrimination ability of the p-variogram score.

Chapter 3

Meteorological ensemble post processing

3.1 Trends in the predictive performance of raw ensemble weather forecasts

As stated in Chapter 1 univariate statistical post processing conveys considerable improvements in skill to hydrometeorological ensemble forecasts. The study at hand assesses the evolution of skill of both raw ensemble and EMOS post processed forecasts from the ECMWF ensemble over the period from January 2004 to March 2014. The following sections follow mainly the work by Hemri et al. (2014b).

3.1.1 Introduction

NWP models are under continuous development. Hence, the forecast skill of the ECMWF ensemble, of which a detailed description can be found in Molteni et al. (1996) and Buizza et al. (2007), improves over time (Buizza et al., 1998, 2007; Richardson et al., 2013). These improvements may either stem from a reduction of probabilistic biases or from an increase in potential skill. The former directly competes with statistical post processing, whereas the latter corresponds to an increased information content of the raw ensemble. If most of the improvement is due to the reduction of probabilistic biases and the raw ensemble forecasts continue to improve in the future, the gap in skill between raw ensemble and post processed forecasts is expected to decrease over time. Eventually, the raw ensemble forecasts may become reliable and unbiased, and hence the gap in skill will be closed. However, if most of the improvement is due to an increase in potential skill, statistical post processing will keep adding skill in the future.

In this study we analyze the evolution of the global performance of the operational ECMWF raw ensemble and the corresponding post processed EMOS forecasts for 2-meter temperature (T2M), 24-hour precipitation (PPT24), and

10-meter wind speed (V10). We verify the forecasts against globally distributed surface synoptic observations (SYNOP) data over a period of about 10 years. We firstly evaluate the monthly average skill for both the raw and the EMOS forecasts. In order to assess the extent to which the results depend on the choice of the post processing method, BMA is additionally applied to the T2M raw ensemble forecasts. We will use the negatively oriented CRPS as a measure of skill. As the CRPS assesses both reliability and sharpness and is a proper scoring rule (Gneiting and Raftery, 2007), we rely on it for model fitting and verification throughout this study. Note that skill and reliability are linked in that given constant sharpness an improvement in reliability leads to an improvement in skill and vice versa. We finally analyze the evolution of the gap in CRPS between raw ensemble and post processed forecasts.

After presenting the dataset in Section 3.1.2, we summarize the methods for post processing and for the assessment of the global skill evolution in Section 3.1.3. In Section 3.1.4 the results are shown. This is followed by a discussion in Section 3.1.5 along with some concluding remarks.

3.1.2 Data

We have selected a large number of synoptical observation (SYNOP) stations for verification to perform a study which covers the entire globe as ECMWF forecasts are issued on the global domain. SYNOP stations with suspicious or too many missing data are removed from the dataset following the approach used by Pinson and Hagedorn (2012) with some modifications. The main criterion for removal of a station from the dataset for a particular variable is the percentage of data points that are equal to the previous ten data points. If this exceeds 20 % a station is considered to be unreliable. In case of PPT24 and V10 this is applied only for non-zero values. Additionally, T2M stations with values outside the range [-70°C, 60°C], PPT24 stations with values outside [0 mm, 1826 mm] and V10 stations with values outside [0 m/s, 113.2 m/s] are removed. Those ranges extend from the lowest to the highest measurements recorded on earth. With these removal criteria 4160 out of 4586, 2917 out of 2956, and 4387 out of 4509 stations are considered to be of reasonable quality for T2M, PPT24, and V10, respectively.

In this study we focus on observations for 12:00 UTC and ECMWF ensemble forecasts initialized at 12:00 UTC with lead times of 3, 6, and 10 days. This selection of forecast ranges covers the transition from higher predictability at lead time 3 d to considerably lower predictability at 10 d. The raw ensemble consists of the ECMWF high-resolution (HRES), the corresponding 50 member ensemble (ENS) and the control (CTRL) runs. During the time period considered (1st January 2002 to 20th March 2014) the forecast model, which is the same for ENS, HRES, and CTRL, has undergone several upgrades. Additionally, the ENS has been reconfigured several times over that period. The ECMWF ensemble system is described in detail in Molteni et al. (1996) and Buizza et al. (2007). Since

for the post processed forecasts some data has to be put aside for training (see Section 3.1.3), the verification periods for the following analyses are somewhat shorter and extend from January 2004 to March 2014 for T2M and V10, and from January 2007 to March 2014 for PPT24.

3.1.3 Methods

Based on the EMOS method introduced in Section 2.1.1, the EMOS variants for T2M, PPT24, and V10 are now presented in detail.

EMOS for T2M

For T2M forecasts g (cf. Equation (2.6) in Section 2.1.1) is a normal density distribution with mean m and variance σ^2. In order to account for seasonality, we use here a variant of the original EMOS approach similar to the one proposed by Scheuerer and Büermann (2014) where the departures of observed temperatures from their climatological means are related to those of the forecasts. Specifically, let $T = \{t_1, \ldots, t_n\}$ be a training period of n days preceding the forecast initialization and denote by r_{tk} the forecast of the k-th ensemble member and by y_t the observation on day $t \in T$. As a first step, we fit a regression model

$$y_{t_j} = c_0 + c_1 \sin\left(\frac{2\pi j}{365}\right) + c_2 \cos\left(\frac{2\pi j}{365}\right) + \varepsilon_{t_j}, \quad j = 1, \ldots, n \tag{3.1}$$

which captures the seasonal variation of T2M. The residual terms ε_{t_j} are likely correlated over time, but for simplicity an ordinary least squares fit is performed. We denote by $\tilde{y}_t$ the fitted value of this periodic regression model on day t and interpret it as the climatological mean temperature on this day. This model can easily be extrapolated to future days $t_{d+1}, t_{d+2}, \ldots$ The above regression includes both a sine and a cosine term which is equivalent to a cosine model with variable phase and amplitude. Since $j = 1, \ldots, n$ is just a numbering of the days in T, different training periods have different phase parameters and hence c_1 and c_2 evolve over the calendar year. We fit the same type of model also to the ensemble mean, control, and high resolution run and obtain climatological means $\tilde{r}_{\overline{\text{ENS}},t}$, $\tilde{r}_{\text{CTRL},t}$, and $\tilde{r}_{\text{HRES},t}$. The mean of the forecast distribution is then

$$m = \tilde{y} + a_1(r_{\text{HRES}} - \tilde{r}_{\text{HRES}}) + a_2(r_{\text{CTRL}} - \tilde{r}_{\text{CTRL}}) + a_3(r_{\overline{\text{ENS}}} - \tilde{r}_{\overline{\text{ENS}}}). \tag{3.2}$$

The variance of the forecast distribution is linked to the raw ensemble by

$$\sigma^2 = b_0 + b_1 s^2, \tag{3.3}$$

where $s^2 = \frac{1}{K}\sum_{k=1}^{K}(r_k - \frac{1}{K}\sum_{k=1}^{K} r_k)^2$. The parameters $\boldsymbol{\theta}_{\text{T2M}} = (a_1, a_2, a_3, b_0, b_1)^T$ are constrained to be non-negative, and hence $a_k / \sum_{k=1}^{K} a_k$ can be understood as the weight of model k.

EMOS for PPT24

For PPT24 we use the EMOS approach proposed by Scheuerer (2014), where g is a left-censored (at zero) generalized extreme value (GEV) distribution. While the shape parameter ξ of the GEV is kept constant ($\xi = 0.2$), the location and the scale parameters m and σ are linked to the raw ensemble via

$$m = a_0 + a_1 r_{\text{HRES}} + a_2 r_{\text{CTRL}} + a_3 r_{\overline{\text{ENS}}} + a_4 \pi_0, \tag{3.4}$$

$$\sigma = b_0 + b_1 \text{MD}_r, \tag{3.5}$$

where π_0 is the fraction of ensemble members predicting zero precipitation and $\text{MD}_r := K^{-2} \sum_{k,k'=1}^{K} |r_k - r_{k'}|$ is the ensemble mean difference. Again, the parameters are denoted by $\boldsymbol{\theta}_{\text{PPT24}} = (a_0, \ldots, a_4, b_0, b_1)^T$. The parameters a_1, a_2, a_3, b_0, b_1 are constrained to be non-negative, and hence the normalized parameters a_1 to a_3 can be understood as weights.

EMOS for V10

For V10 we use a modified version of the EMOS model based on a left-truncated (at zero) normal distribution by Thorarinsdottir and Gneiting (2010). A truncated normal distribution on the square root transformed space seems to be an appropriate choice for g, as it outperformed both the untransformed truncated normal model and a model with predictive gamma distributions in preliminary tests. We model the distribution of $\sqrt{y}$ by a truncated normal distribution with parameters

$$m = a_0 + a_1 \sqrt{r_{\text{HRES}}} + a_2 \sqrt{r_{\text{CTRL}}} + a_3 \sqrt{r_{\overline{\text{ENS}}}} \tag{3.6}$$

$$\sigma^2 = b_0 + b_1 \text{MD}_{\sqrt{r}}, \tag{3.7}$$

where $\text{MD}_{\sqrt{r}} := K^{-2} \sum_{k,k'=1}^{K} |\sqrt{r_k} - \sqrt{r_{k'}}|$. The parameters $\boldsymbol{\theta}_{\text{V10}} = (a_0, \ldots, a_3, b_0, b_1)^T$ are constrained to be non-negative, thus the normalized parameters a_1 to a_3 can be understood as model weights.

Model fitting and evaluation

For all three variables the parameter vector $\hat{\boldsymbol{\theta}}$ is estimated by CRPS (cf. Equation (2.27) in Section 2.3.1) minimization over the training period T. The training period for each verification day consists of the n days preceding the initialization date. Tests using a subset of European stations indicate that for T2M forecasts a training period of 720 days is appropriate, while for PPT24 and V10 training periods of 1816 and 365 days, respectively, performed best. Following Scheuerer

(2014) we try to avoid overfitting by using the parameter estimates $\hat{\boldsymbol{\theta}}_{t-1}$ as starting values for the estimation of $\hat{\boldsymbol{\theta}}_t$ for verification day t and then stopping the optimization process after a few iterations. This sliding window model fitting approach generally results in good parameter estimates, but it may be affected by sudden changes in the raw ensemble models during the training period. Nevertheless, the good performance of the post processed forecasts as shown in Section 3.1.4 indicates that this effect can be neglected for the majority of stations.

A closed-form expression for the CRPS for the normal model for T2M can be found in Gneiting et al. (2005). For the censored GEV model used for PPT24 a closed-form expression has been derived by Friederichs and Thorarinsdottir (2012) and Scheuerer (2014). For the square root transformed truncated normal model used for V10 the CRPS can be calculated using formulae by Gneiting et al. (2004). With $q = \Phi(-\mu/\sigma)$, $p = 1 - q$, and $w = (\sqrt{y} - \mu)/\sigma$ the CRPS can be written as

$$\begin{aligned}\mathrm{CRPS}(y, \mu, \sigma) \;=\; & \frac{\sigma}{p^2}\Big(\sigma - \frac{2\mu}{\sqrt{\pi}}\Big) - 2\sigma^2\left\{\frac{w^2}{2} - \frac{1}{p}\Big[(w^2-1)\Phi(w) + w\varphi(w)\Big] + \right. \\ & \left.\frac{qw^2}{p}\right\} - 2\sigma\mu\left\{w - \frac{2}{p}\Big[w\Phi(w) + \varphi(w)\Big] + \frac{2qw}{p}\right\} + \\ & \frac{q\sigma^2}{p^2}\Big[-\frac{1}{q}\varphi\Big(\frac{-\mu}{\sigma}\Big)^2 + q\Big(\frac{\mu^2}{\sigma^2} - 1\Big)\Big] + \frac{2\sigma\mu}{p^2\sqrt{\pi}}\Phi\Big(-\frac{\sqrt{2}\mu}{\sigma}\Big) - \frac{\mu^2 q^2}{p^2},\end{aligned} \tag{3.8}$$

where Φ and φ denote cumulative and probability density functions of the standard normal distribution, respectively.

BMA for T2M

As T2M predictions can be described well by a normal distribution, BMA parameters can be estimated easily using the `R` package `ensembleBMA` (Fraley et al., 2015). Hence, for T2M BMA can be used as an alternative to EMOS even on the global set of stations. Resuming the BMA model for raw ensembles with exchangeable members (cf. Equation (2.5)) and with an additional bias correction the BMA model is parameterized by

$$y|\boldsymbol{r} \sim \sum_{m=1}^{M} w_m \sum_{n=1}^{N_m} g(y \mid a_{m0} + a_{m1} r_{m,n}, \sigma_m), \tag{3.9}$$

where M is the number of subgroups of the ensemble within which all members are exchangeable and N_m is the number of members in this group (Fraley et al., 2010). In case of T2M g is a normal kernel distribution and $w_1, \ldots, w_M$ are model weights. The parameters $\hat{a}_{m0}, \hat{a}_{m1}$, $i = \mathrm{HRES}, \mathrm{CTRL}, \mathrm{ENS}$, are estimated by linear regression and $\hat{w}_m, \hat{\sigma}_m$ by the Expectation-Maximization algorithm (Dempster

et al., 1977; McLachlan and Krishnan, 1997). The BMA models for this study are fitted using a training period of 365 days prior to the verification day. The estimates for day $t-1$ are used as starting values for the estimation of the parameter values for day t.

Global CRPS analysis

As stated above, the main objective of this study is to analyze whether the gap in CRPS between the raw ensemble and the post processed forecast narrows over time. This is assessed stationwise using both a parametric and a non-parametric approach. For the former, we fit the following regression model to the monthly time series of CRPS differences ($\Delta\text{CRPS}_t = \text{CRPS}_{\text{raw},t} - \text{CRPS}_{\text{EMOS},t}$),

$$\Delta\text{CRPS}_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (3.10)$$

where ΔCRPS_t is the predictand, t is now time in months, and σ^2 denotes the error variance. For the latter, we use Kendall's correlation coefficient τ and the associated test statistic (Mann, 1945) as implemented in the `R` package `Kendall` (McLeod, 2011). In order to correct for seasonal effects, we calculate the τ statistic using the residuals of the model

$$\Delta\text{CRPS}_t = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi t}{12}\right) + \gamma_2 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (3.11)$$

Note that negative values of τ indicate a negative trend and positive values a positive one. Figures 3.1 a) and b) show the regression lines estimated by the model described by Equation (3.10) for monthly averages of ΔCRPS and the corresponding Kendall's τ test statistics for an example with decreasing and increasing gap.

3.1.4 Results

General features of ΔCRPS

Before assessing the stationwise evolution of ΔCRPS over time, we consider first the evolution of global average CRPS values of both raw ensemble and EMOS forecasts. As shown in Figures 3.1 c) to k) the average CRPS for both forecasts increases with increasing lead time regardless of the variable of interest. Note that all three variables exhibit seasonal oscillations in average CRPS. In case of T2M and V10 post processing by EMOS obviously improves the average CRPS, whereas for PPT24 the improvement is much smaller relative to its seasonal oscillations in average CRPS. In any case, further analyses on the temporal evolution of ΔCRPS should correct for seasonal effects. Note that ΔCRPS depends on the performance of the post processing method selected. If alternative post processing methods perform better, ΔCRPS will be further increased by using them.

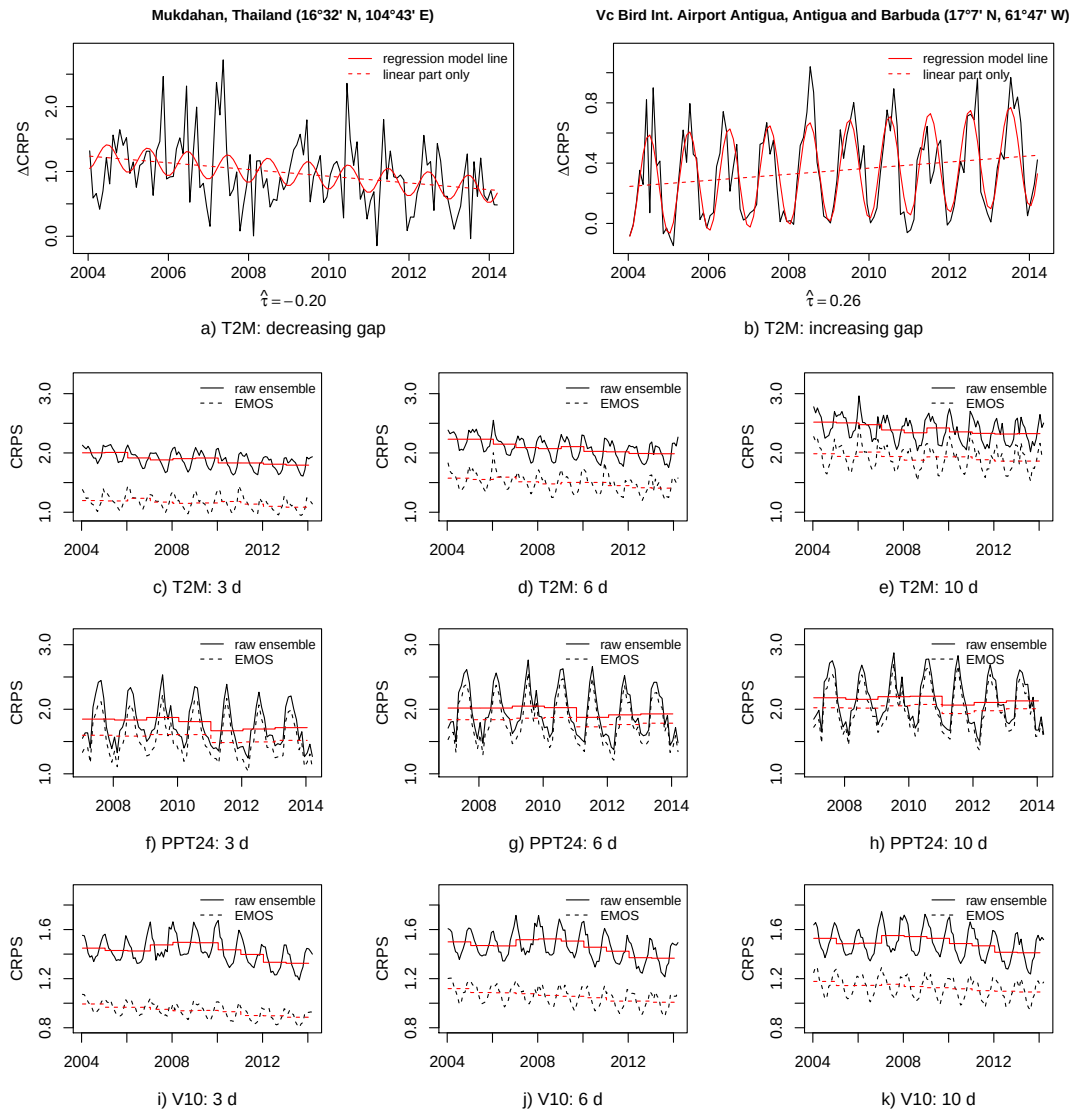


Figure 3.1: a) and b) show monthly averages of ΔCRPS between raw ensemble and EMOS forecasts with a lead time of 6 d for example stations with a decreasing and an increasing gap for T2M. The red solid lines correspond to the fits of the regression model stated in Equation (3.10); the red dashed lines to their linear parts. c) to k) depict the monthly (in black) and yearly (in red) global average CRPS of the raw ensemble and EMOS forecasts for T2M, PPT24, and V10. Figure taken from Hemri et al. (2014b).

Let us now focus on a stationwise analysis. According to the box plots on the panels on the left of Figures 3.2 a) to c) more than 95 % of the stations benefit from EMOS in terms of ΔCRPS averaged over the entire verification period

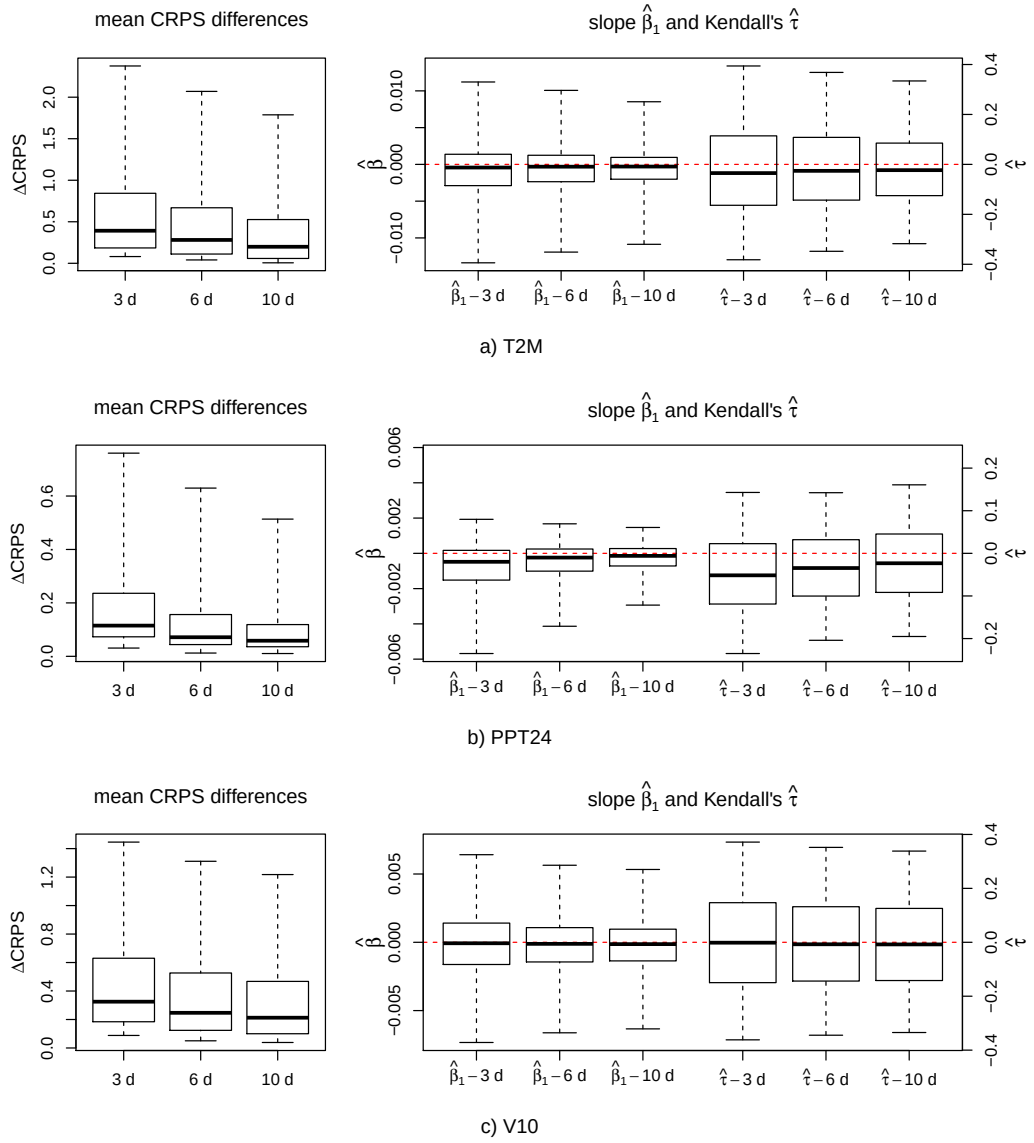


Figure 3.2: Box plots over all stations representing the 5, 25, 50, 75, and 95 % quantiles of the average CRPS differences between raw ensemble and EMOS forecasts (left panels), and (right panels) the slope coefficients of the linear model fits and the Kendall's τ statistics of monthly ΔCRPS averages. Depicted are a) T2M, b) PPT24, and c) V10; the red dashed lines on the right-hand panels indicate the zero line. Figure taken from Hemri et al. (2014b).

regardless of lead time and variable of interest. Note also the positive skewness and the decrease in ΔCRPS with increasing lead time. The box plots on the panels on the right of Figures 3.2 a) to c) describe the empirical distributions among the set of all stations considered of the slope coefficients $\hat{\beta}_1$ and the $\hat{\tau}$ test statistics of ΔCRPS against time for the parametric and the non-parametric model, respectively. For T2M and, in particular, PPT24 negative trends are more

common than positive trends, whereas the corresponding box plots for V10 are almost symmetric around the zero line. In general the medians of the $\hat{\beta}_1$ and the $\hat{\tau}$ values seem to converge to zero with increasing lead time.

Though not discussed in Hemri et al. (2014b), the spatial distribution of the relative improvement in skill by EMOS is also relevant. Neglecting the changes in CRPS over time, the relative change in average CRPS by applying EMOS compared to the raw ensemble can easily be assessed in a stationwise manner. As shown in Figure B.1 in Appendix B.1 and also discussed in Richardson et al. (2015) EMOS improves forecast skill of T2M forecasts considerably at almost all stations for lead times up to about 5 days. At a forecast horizon of 6 days there is an increasing number of stations that exhibit only a very small improvement in CRPS by EMOS, which are located mostly in Eastern Europe and North America. At a lead time of 10 days, a considerable number of stations shows no improvement or even a deterioration. Nonetheless, EMOS improves forecast skill at the majority of stations even for the very long forecast horizon of 10 days. In general, the improvement by EMOS is smaller, but still considerable, in case of PPT24 as can be seen from Figure B.2 in Appendix B.1. Note that already at a forecast lag of 3 days there are several stations that exhibit a deterioration in skill. These stations are located to a large extent in North Africa and the Arabian Peninsula. As for T2M, also for PPT24 there is an increasing number of stations with only a very small improvement in CRPS by EMOS at a forecast lag of 6 days. But compared to the results for T2M these stations are more numerous and spread more widely. Going from a forecast lag of 6 to 10 days, the picture does not change much in case of PPT24. The number of station with only a very small improvement or even a deterioration in skill by applying EMOS increases again. But the vast majority of stations still shows an improvement. For V10 the improvement in skill by EMOS is considerable at the vast majority of stations over the entire forecast horizon as can be seen from Figure B.3 in Appendix B.1. Like for T2M and PPT24, the extent of this improvement decreases with increasing forecast lag. However, even at a lead time of 10 days skill is increased by EMOS at all stations. The improvements are greatest over Brazil, India, and (South) East Asia, whereas they are rather moderate over Western Europe and North America.

Are there any significant temporal trends?

The above results indicate a tendency of a decrease in ΔCRPS over time at least for T2M and PPT24. In the following we check the percentages of stations with decreasing, an absence of, or increasing trend in ΔCRPS over time at a significance level of 0.05. In order to be more confident about the results this analysis is performed using both the parametric regression model and the non-parametric Kendall's τ correlation coefficient test. As already mentioned both approaches correct for seasonal effects. Furthermore, in case of T2M the same analysis has been performed additionally using BMA instead of EMOS in order to relax the

dependence on one particular post processing method. As shown in Table 3.1 the stations with no significant trend outnumber the stations with either negative or positive trend for all three variables and lead times considered. Note that the percentage of stations without any significant trend increases with increasing lead time. In line with the results shown in Figure 3.2, significantly negative trends are more common than positive ones for T2M and PPT24. The difference between the number of stations with negative and those with positive trend reduces with increasing lead time, but is still greater than zero for a 10 day forecast. Note that the high number of non-significant stations in case of PPT24 is likely to be due to the high variability of precipitation amounts, and hence variability of CRPS values, which leads to a large residual standard error in case of the parametric regression model and to a lot of pairs (a pair denotes here a value of ΔCRPS and its associated time stamp) opposite to the estimated direction in case of the τ test statistics. In case of V10 the stations with a negative trend and those with a positive trend are almost equally frequent regardless of the lead time. The global distributions of stations with no, significantly negative, and significantly positive trend in ΔCRPS are shown in Figures B.4 and B.5 in Appendix B.1.

Additionally to the analyses presented above that can also be found in Hemri et al. (2014b), the evolution of the weights assigned to HRES by EMOS is analyzed here. For each verification year the empirical distribution of all weights assigned to HRES pooled over all verification days and the global set of stations is obtained. Figure 3.3 shows the mean weights against lead time as a separate curve for each verification year for T2M and V10. For 2013 the $0.05, 0.25, 0.5, 0.75,$ and 0.95 quantiles of the distribution of weights are shown as well. For visualization purposes, the weights are given as equivalent number of ENS members. For instance, if EMOS assigns equal weights to the HRES run, the CTRL run and the 50 ENS runs, this weight measure would equal one. As expected the weight of HRES decreases with increasing lead time. This is in line with the gradually decreasing predictability over the forecast horizon (Richardson et al., 2015). The gain by running the model at high-resolution in comparison with the lower resolution ENS runs decreases with increasing uncertainty. But more importantly, the weights assigned to HRES generally decrease over the years. For V10 the importance of HRES mostly decreases from year to year. In case of T2M this decrease is not so obvious, in particular at the first two lead days. However, the average weight assigned to HRES during the last two years, 2012 and 2013, is considerably lower than for any of the previous years.

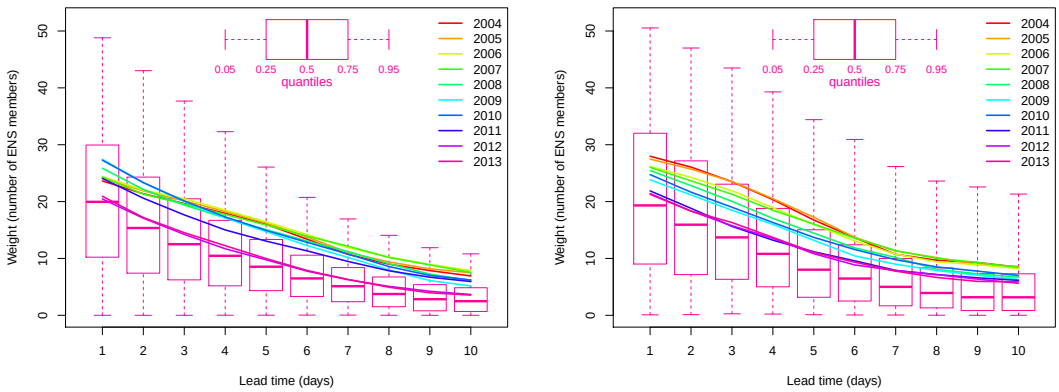
3.1.5 Discussion

According to the above analyses the gap in CRPS between the raw ensemble and the EMOS forecasts remains almost constant over time. For T2M and PPT24 ΔCRPS shows a slightly decreasing tendency. The higher the lead time the less accentuated is this tendency. For V10 such a tendency cannot be detected. The parametric regression model and the non-parametric Kendall's τ test yield similar

Table 3.1: Percentages of stations showing no, negative, or positive trend in ΔCRPS[a]

			parametric model				Kendall's τ statistics			
			T2M		PPT24	V10	T2M		PPT24	V10
			EMOS	BMA	EMOS	EMOS	EMOS	BMA	EMOS	EMOS
forecast lead time	3 d	no significant trend	42 %	42 %	76 %	41 %	44 %	43 %	77 %	42 %
		negative trend	34 %	34 %	19 %	31 %	32 %	32 %	18 %	29 %
		positive trend	24 %	25 %	5 %	28 %	24 %	25 %	5 %	29 %
	6 d	no significant trend	46 %	48 %	82 %	43 %	48 %	49 %	82 %	45 %
		negative trend	31 %	28 %	14 %	31 %	29 %	27 %	13 %	29 %
		positive trend	23 %	24 %	4 %	26 %	23 %	24 %	5 %	27 %
	10 d	no significant trend	54 %	58 %	83 %	45 %	54 %	58 %	82 %	46 %
		negative trend	27 %	23 %	11 %	31 %	26 %	23 %	11 %	28 %
		positive trend	19 %	18 %	6 %	25 %	20 %	19 %	7 %	26 %

[a] Percentages of stations (totals are 4160 (T2M), 2917 (PPT24), and 4387 (V10)) showing no, negative, or positive trend in monthly ΔCRPS values against time at a significance level of 0.05. Table taken from Hemri et al. (2014b).



a) T2M b) V10

Figure 3.3: Weights assigned to HRES by EMOS for a) T2M and b) V10 against lead time for different verification years. The solid lines correspond to the yearly averages, while the box plots show the 0.05, 0.25, 0.5, 0.75, and 0.95 quantiles of the empirical distribution of HRES weights in 2013.

results. Hence, a linear model that is overlaid by seasonal fluctuations seems to be reasonable. Note that the skill of the raw ensemble and the EMOS forecasts may sometimes be negatively affected by upgrades to the atmospheric model. Model upgrades may deteriorate raw ensemble skill at some individual stations. For instance, a resolution increase may introduce new issues with statistical downscaling of the forecasts to some specific observation sites. But more importantly, the skill of the post processed forecasts can be lowered dramatically if a model update happens between the training and the verification period. These issues may result in positive trends in ΔCRPS. Ideally, post processing would be based on a cascade of reforecasts. That is, for each atmospheric model version, training of the post processing model would be done using a corresponding time series of reforecasts made with that same model version. Furthermore, the observations may be affected by measurement errors. If these errors change over time, they may also influence the estimates of the trends in ΔCRPS. As the problems introduced by statistical downscaling may be mitigated by verifying against model analysis, a similar study that replaces observations by model analysis, as proposed by Ghelli and Lalaurette (2000) and Pappenberger et al. (2009), may give further insights.

Additionally, verification scores are affected by ensemble size (e.g. Richardson (2001)). Let us assume the hypothetical case of a perfect forecast distribution that equals the distribution of the stochastic process of interest. A raw ensemble forecast sampled from this forecast distribution would then be reliable by definition. Nevertheless, the raw ensemble would only be a step-wise approximation to the underlying forecast distribution. This would lead to an under-performance

of the raw ensemble compared to the underlying forecast distribution in terms of CRPS, because CRPS is a proper skill score (Gneiting and Raftery, 2007). This has to be kept in mind when comparing raw ensemble CRPS values with those values obtained from continuous forecast distributions. But note that this does not mean that the continuous forecast distributions obtained by post processing equal the underlying distribution mentioned above. Ferro et al. (2008) discuss the effect of ensemble size on CRPS. Hence, further analyses on the gap in skill between raw ensemble and post processed forecasts may benefit from taking this effect into account.

From the above we conclude that the probabilistic skill of both the raw ensembles and the EMOS forecasts improves over time. The fact that the gap in skill has remained almost constant, especially for V10, suggests that improvements to the atmospheric model have an effect quite different from what calibration by statistical post processing is doing. That is, they are increasing potential skill. Thus this study indicates that (a) further model development is important even if one is just interested in point forecasts, and (b) statistical post processing is important because it will keep adding skill in the foreseeable future.

3.2 Discrete post processing of total cloud cover ensemble forecasts

As stated in Chapter 1 the discrete nature of total cloud cover observations calls for discrete post processing methods. The MLR and POLR methods, which have been introduced in Section 2.1.2, are suitable for this purpose. In the following sections, which closely follow Hemri et al. (2016), we present a study that assesses the performance of MLR and POLR for ECMWF TCC forecasts on the global domain.

3.2.1 Introduction

Forecasts of total cloud cover (TCC) are an important part of numerical weather prediction (NWP) both in terms of model feedbacks and with respect to forecast users in areas such as energy demand and production, agriculture, and tourism. In NWP models cloud cover affects the evolution of the model state through feedback loops on radiative fluxes and heating rates (Köhler, 2005; Haiden and Trentmann, 2015). Predictions of energy demand and production rely in part on TCC forecasts. Photovoltaic energy forecasting in particular relies on accurate predictions of solar irradiance, which is on a day-to-day basis mainly determined by variations in TCC (Taylor and Buizza, 2003; Pelland et al., 2013). Observational astronomy depends on reliable TCC forecasts (Ye and Chen, 2013). Other applications of TCC forecasts can be found in agriculture, where they may facilitate irrigation scheduling (Diak et al., 1998), in avalanche forecasting, where

the amount of radiational cooling influences the stability of snow packs (McClung, 2002), and in leisure activities where cloudiness influences, for example, the amount of sun protection required (Dixon et al., 2008).

(Total) cloud cover is defined as "portion of the sky cover that is attributed to clouds" (American Meteorological Society, 2015). Obviously, TCC takes values in $[0, 1]$, and unlike other weather variables, such as temperature or precipitation, TCC is reported and forecast on a discrete space with only a small number of possible values. Usually, observers report TCC as values in $(0, 1, 2, \ldots, 8)$, henceforth called octas. At the European Centre for Medium-range Weather Forecasts (ECMWF) probabilistic TCC forecasts are provided as direct output from the NWP ensemble. The skill of NWP TCC forecasts in the short and medium range is low compared to the forecasts for other meteorological variables like 6 h accumulated precipitation, geopotential, 2-meter temperature, or 10-meter wind speed (Köhler, 2005). In 2004 the high resolution (HRES) ECMWF TCC forecasts showed skill compared to persistence only up to forecast day 3 over Europe. Furthermore, Haiden and Trentmann (2015) showed that the skill of 24-h HRES TCC forecasts verified against a set of European stations improved little over the last decade.

The limited skill of direct model output TCC point forecasts is partly due to a representativeness mismatch between models and observations. Areas covered by visual observations typically vary in scale from 10 to 100 km, depending on visibility and topography (Mittermaier, 2012). Automated observations as derived from ceilometers measure cloud cover directly overhead. Depending on the wind speed in the cloud layer the scanned area may or may not be representative of the model grid-scale. Temporal variability of cloudiness on hourly and subhourly scales presents an additional challenge for predicting instantaneous TCC. As shown by Haiden et al. (2015), the forecast range over which there is positive skill relative to persistence increases from 2-3 days to 5 days if daytime averages rather than instantaneous values of TCC are considered.

The potential benefits of skillful TCC forecasts together with the relatively low performance of state-of-the-art NWP TCC point forecasts, i.e. forecasts interpolated from the NWP model grid to specific sites, motivates the development of statistical methods to post process raw ensemble TCC forecasts. In this study, we focus on post processing of the global point forecasts of TCC from the ECMWF ensemble forecast system. For this purpose, we have developed variants of MLR and POLR that are suitable for TCC post processing (cf. Section 2.1.2). In order to put these models into a broader context, we will now give an overview on similar, mostly "logistic regression" based, post processing methods, of which several approaches have been proposed in the field of meteorological forecasting over the last 15 years.

Applequist et al. (2002) applied logistic regression to produce forecasts of

precipitation threshold exceedance probabilities. Hamill et al. (2004) used logistic regression to obtain probabilistic forecasts of temperature and precipitation from ensemble model output statistics. Wilks (2009) proposed extended logistic regression (ELR) as a further development of the approach by Hamill et al. (2004) that provides full predictive distributions from ensemble model output statistics. ELR has been used to post process NWP ensemble precipitation (and much less frequently also wind speed) forecasts in many studies. Schmeits and Kok (2010) compared raw ensemble forecasts from a 20 year ECMWF precipitation reforecast dataset with Bayesian model averaging (Raftery et al., 2005) and ELR. While ELR outperformed the raw ensemble only slightly in case of area-mean precipitation amounts, area-maximum forecast skill was significantly improved by ELR. Furthermore, ELR performed considerably better than BMA and equally well as a modified BMA approach by Schmeits and Kok (2010). A similar study by Roulin and Vannitsem (2012) showed that applying ELR led to substantially improved skill and mean error of ECMWF precipitation ensemble forecasts for two catchments in Belgium. Likewise, Ben Bouallègue (2013) confirmed the good performance of ELR. However, there are also studies that reveal the limitations of ELR. In a study comparing 8 different post processing methods for (ensemble) precipitation forecasts over South America, ELR ranks in the upper mid-range among the methods considered (Ruiz and Saulo, 2012). Hamill (2012) showed that ELR improved skill of ECMWF precipitation ensemble forecasts considerably over the United States, but that the multi-model ensemble consisting of the ensemble forecasts from the ECMWF, the UK Met Office, the U.S. National Centers for Environmental Prediction (NCEP), and the Canadian Meteorological Center could not be improved much by ELR. Scheuerer (2014) was able to outperform ELR by applying an ensemble model output statistics approach (Gneiting et al., 2005) based on a generalized extreme value distribution. Messner et al. (2014) applied ELR, censored logistic regression, and POLR to ECMWF ensemble wind speed and precipitation forecasts. Their study revealed the good performance of POLR for discrete, categorical sample spaces. However, we are not aware of any study that post processes TCC ensemble forecasts based on a logistic regression approach.

First, the TCC dataset used for this study is presented in Section 3.2.2. This is followed by Section 3.2.3 that discusses the different forecast models and methods. Section 3.2.4 provides an in-depth presentation of the results, which is followed by a brief discussion in Section 3.2.5.

3.2.2 Data

Like in Section 3.1 (see also Hemri et al. (2014b)) the TCC dataset used in this study consists of stationwise daily time series from January 2002 to March 2014 of forecast/observation pairs at 12:00 UTC for lead times up to ten days. As ECMWF forecasts are issued on the global domain, we have selected 3435 surface synoptic observations (SYNOP) stations that cover the entire globe (except from

Australia, which does not report at 12:00 UTC) as observational dataset. Stations with unreliable observation time series are detected and removed according to the following scheme, which is a modification of the approach by Pinson and Hagedorn (2012):

- Count the number of days with observed values that are equal to the observations from the previous ten days. If this number exceeds 20 % of the length of the time series, a station is considered to be unreliable.

- Additionally, remove stations with recorded observations outside the range [0,1].

After removing the unreliable stations, 3330 are left for the following analyses.

3.2.3 Methods

Training and verification periods

Prior to introducing the different forecast models, the training periods used for estimation of the parameters of the statistical post processing models are presented here along with the corresponding verification periods. In line with the study on trends in predictive performance of raw ensemble weather forecasts presented in Section 3.1 rather long training periods of up to five years are applied. Accordingly, the verification period extends from January 2007 to March 2014. The corresponding training periods are selected in a non-seasonal and in a seasonal way. In case of the non-seasonal approach, for any verification day x the corresponding training period covers the five calendar years prior to the day x. For instance, for a random verification day x in 2009, say 27 June 2009, the corresponding training period lasts from 1 January 2004 to 31 December 2008. The same training period would apply for any other verification day in 2009. In case of the seasonal approach, the block-wise training periods from the non-seasonal approach are additionally differentiated according to the season of the verification day. For this study, we divide the year into two seasons (April to September and October to March).

Climatological and uniform forecasts

Climatological and uniform forecasts are used as reference. The climatological forecasts are constructed stationwise in the same way as the seasonal training periods. That is, for each verification day the climatological forecast corresponds to the empirical distribution of all TCC observations in the same season (winter half-year or summer half-year) within the five calendar years prior to the verification day. The uniform forecasts simply assign a probability of 1/9 to each TCC level in $(0, 1, \ldots, 8)$ irrespective of station climatology and NWP model output.

Raw ensemble forecasts

The ECMWF TCC forecasts used in this study are issued daily from 1 January 2002 to 20 March 2014 at 12:00 UTC and cover the lead times $1, 2, \ldots, 10$ days. In the following, we will focus mostly on the lead times 3, 6, and 10 d, which reflect sequentially decreasing predictability and are representative for the other lead times. As already stated, the ECMWF EPS consists of the HRES run, of the 50 member ENS, and the CTRL run. As in the two previous case studies, this 52 member ensemble is used as raw ensemble.

MLR and POLR

In this study MLR and POLR post processing (see Section 2.1.2 for details) is applied to the ECMWF raw ensemble, which is now written as

$$\boldsymbol{r} = (r_1, \ldots, r_K) = (r_{\text{ENS},1}, \ldots, r_{\text{ENS},50}, r_{\text{HRES}}, r_{\text{CTRL}}), \tag{3.12}$$

where K denotes ensemble size. Accordingly, the first three predictors in the MLR model are the mean of the ENS runs $\bar{r}_{\text{ENS}}$, the HRES run r_{HRES}, and the CTRL run r_{CTRL} leading to a vector of predictors $\boldsymbol{x} = (1, \bar{r}_{\text{ENS}}, r_{\text{HRES}}, r_{\text{CTRL}}, s^2, f_0, f_1)^T$. As mentioned above, both non-seasonal and seasonal post processing approaches are applied. In case of MLR we test a non-seasonal approach with block-wise training periods (MLR-B) and a seasonal approach with seasonal block-wise training periods (MLR-S). As for MLR, the non-seasonal POLR model is denoted as POLR-B, and its seasonal counterpart as POLR-S h, where h indicates that it is the full model with all predictors, i.e. $\boldsymbol{x} = (1, \bar{r}_{\text{ENS}}, r_{\text{HRES}}, r_{\text{CTRL}}, s^2, f_0, f_1, I)^T$. In order to find the best set of predictors, various POLR-S models with different sets of predictors are tested. They are listed in Table 3.2.

In order to allow numerically trouble-free verification, any forecast distribution $P(Z = z_j)$, where z_j with $j = 1, \ldots, 9$ denotes the different cloud cover states, is slightly modified subsequent to model fitting. Namely, unrealistically low forecast probabilities p_j for cloud cover state z_j are avoided by setting p_j to $p_j = \max(\mathrm{P}(z_j), p')$, where $p' = 1 - (1 - \alpha)^{\frac{1}{T}}$ and T is the length of the training period. The parameter α denotes the probability that state j is observed at least once during a period of length T, i.e. $\alpha = 1 - (1 - p')^T$. For this study, we deliberately set $\alpha = 0.01$, which leads to $p' = 5.5 \cdot 10^{-6}$ for the non-seasonal models and $p' = 1.1 \cdot 10^{-5}$ for the seasonal models. In case of a forecast distribution with $p' > \mathrm{P}(z_j)$ for at least one state z_j, the probabilities $p_{i \neq j}$ have to be adjusted slightly such that $\sum_{j=1}^{9} p_j = 1$. We apply this correction to all considered predictive distributions $P(Z = z_j)$ including the raw ensemble and the climatological forecasts.

Table 3.2: Overview of the different POLR-S model variants, where $\bar{r}_{\text{ENS}}$, r_{HRES}, and r_{CTRL} are always included, s^2 denotes the ensemble variance, f_0 and f_1 represent the proportion of ensemble members equal to zero and one, respectively, and I is the interaction term between s^2 and the squared, sign adjusted difference of $\bar{r}^*$ from 0.5. Table taken from Hemri et al. (2016).

model	$\bar{r}_{\text{ENS}}$	r_{HRES}	r_{CTRL}	s^2	f_0	f_1	I
POLR-S a	✓	✓	✓				
POLR-S b	✓	✓	✓	✓			
POLR-S c	✓	✓	✓		✓	✓	
POLR-S d	✓	✓	✓				✓
POLR-S e	✓	✓	✓	✓	✓	✓	
POLR-S f	✓	✓	✓	✓			✓
POLR-S g	✓	✓	✓		✓	✓	✓
POLR-S h	✓	✓	✓	✓	✓	✓	✓

Example forecasts

Before discussing the results in Section 3.2.4, four subjectively selected example forecasts for Vienna, Austria, are presented in Figure 3.4 to highlight typical properties of post processing of TCC forecasts. Vienna was chosen as a location in Europe that is situated in the broad transition zones from maritime to continental in winter, and from mediterranean to temperate in summer. As a result, it experiences a rich and complex cloud climatology which is additionally modulated by orographic effects due to its proximity to the European Alps. For illustrative purposes, raw ensemble forecasts are compared with the corresponding seasonal POLR forecasts that use the complete set of predictors (POLR-S h). The raw ensemble and POLR-S h bear strong resemblances. However, POLR-S h seems to move some weight from the extremes (0 or 8 octas) towards the more moderate levels of cloudiness (1 to 7 octas).

3.2.4 Results

After having introduced the different forecast models, we first evaluate forecast skill of these models. This is followed by an in-depth assessment of calibration and sharpness of a selected set of models. For a fair comparison of verification scores, raw ensemble and post processed forecasts have to be mapped to the space of the observations. The function selected to map raw ensemble and post processed forecasts to the observation space influences most of the verification measures. Hence, it is important that the mapping function mimics the procedure of TCC observers, who have to give 1/8 as soon as a little cloud appears, even if the TCC is only 1 percent, and have to give 7/8 as soon as there is a little gap somewhere in the cloud layer. This is ensured by applying a non-equidistant mapping function,

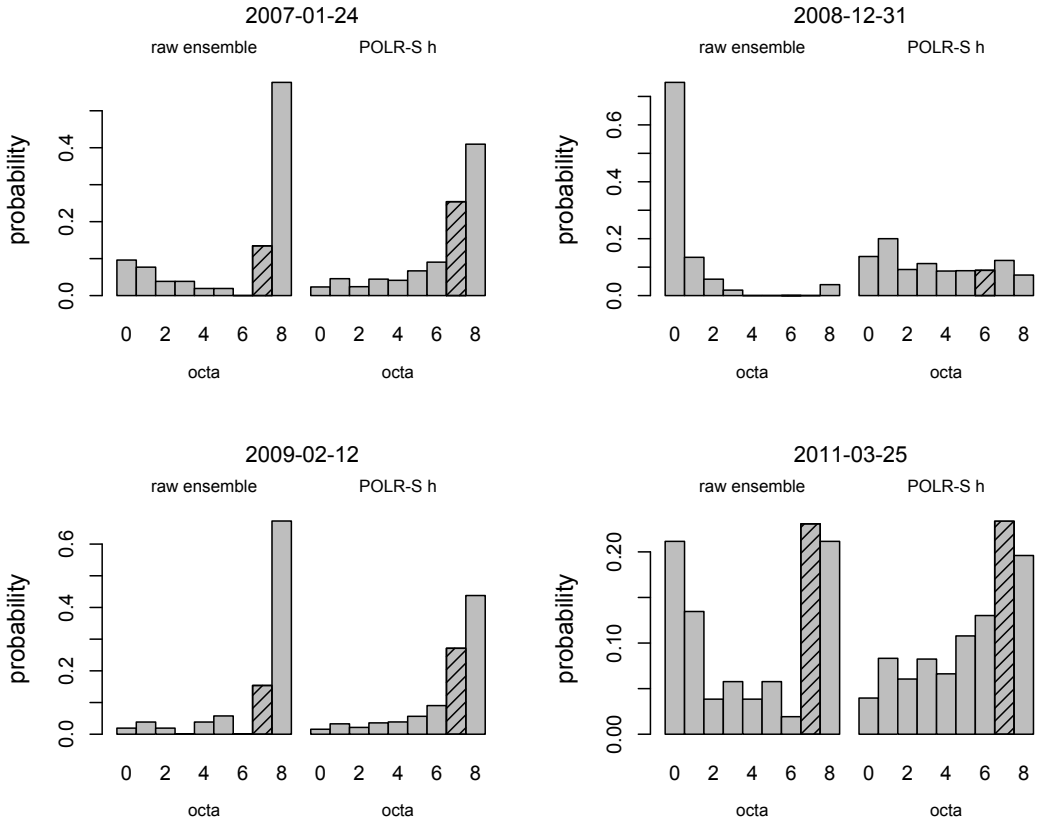


Figure 3.4: Example forecasts for Vienna comparing the raw ensemble and the POLR-S h predictions with a lead time of 6 days. The TCC class in which the observation falls is shaded in black. Figure taken from Hemri et al. (2016).

for which the details can be found in Appendix A.1.1.

Forecast skill

Average skill of the different TCC forecast models is assessed using the log score and the CRPS averaged over the entire verification period and all stations. As TCC is a discrete variable, the ranked probability score (RPS: Epstein (1969); Murphy (1969)) could be used instead of the CRPS. But since the ordered categories of TCC are not equidistant in the dataset at hand (see above and Appendix A.1.1), RPS and CRPS would differ slightly. For this study, we have decided to use the CRPS, because it allows direct skill comparison with continuous TCC forecasts, which may become available in future (see also Section 3.2.5). Both log score and CRPS are proper scoring rules that are negatively oriented, i.e. the lower the score the higher is the forecast skill. While the log score is a local scoring rule that takes only the forecast probability of the materializing observation into account, the CRPS is sensitive to distance in that forecasts with high probabilities attributed to values close to the materializing observation are considered

to be skillful (Gneiting and Raftery, 2007). Mathematical formulations of both scores are given in Section 2.3.1.

According to Table 3.3, the raw ensemble outperforms climatological and uniform forecasts in terms of CRPS for lead times of 1, 3, and 6 days, but not for 10 days. Log scores have not been calculated for the raw ensemble, because they would not be meaningful. All MLR and POLR models outperform the climatological, uniform, and raw ensemble forecasts in terms CRPS at all lead times. In terms of log score all MLR and POLR perform better than the climatological and the uniform forecasts. In case of MLR, the seasonal model slightly outperforms its non-seasonal counterpart in terms of CRPS, while the log score tends to prefer the non-seasonal model. For POLR log score and CRPS are more consistent in that both scores indicate a slightly better skill of the seasonal model. This is also reflected in Figure 3.5, which shows averaged log score and CRPS values including their associated 90 % confidence intervals for the raw ensemble (only CRPS), MLR-B, MLR-S, POLR-B, and POLR-S h, i.e. the full model (cf. Table 3.2). The 90 % confidence intervals are obtained by block bootstrapping (Künsch, 1989) with block resamples following a geometric distribution with mean $|V|^{1/3}$, where $|V|$ is the length of the verification period. The block bootstrapping method is implemented in the `R` package `boot` (Canty and Ripley, 2014). Comparing POLR-B with MLR-B and POLR-S h with MLR-S reveals a slight advantage of POLR over MLR. Additionally, POLR allows to make a statement on the relative performance of the $\bar{r}_{\text{ENS}}$, r_{HRES}, and r_{CTRL} runs. Due to the non-negativity constraint, the estimates $\hat{\beta}_{\bar{r}_{\text{ENS}}}$, $\hat{\beta}_{r_{\text{HRES}}}$, and $\hat{\beta}_{r_{\text{CTRL}}}$ can be interpreted as relative weights. As shown in Figure 3.6, $\bar{r}_{\text{ENS}}$ contributes most to the POLR forecast distribution over all lead-times, while r_{CTRL} contributes least. The high resolution run r_{HRES} shows a quite high average weight at the short lead times, but its importance decreases with increasing forecast lag. This is in line with the findings by Richardson et al. (2015) that the decreasing predictability leads to more need for the full ensemble distribution with increasing forecast lag. Note that we have also tested a POLR variant without any constraint on the coefficients. This approach did not only destroy the physical interpretability of the coefficients, but also did not lead to an improvement in forecast skill. Likewise, the coefficients of the MLR model cannot be interpreted easily. As forecast skill, physical interpretability, and model sparsity all favor POLR over MLR, the remainder of this study focuses on POLR. Knowing that the seasonal POLR models perform best, the different seasonal POLR models are now compared. Comparing the models POLR-S a to POLR-S h it becomes clear that in addition to $\bar{r}_{\text{ENS}}$, r_{HRES}, and r_{CTRL} the fraction of zero, f_0, and complete, f_1, TCC have to be included in the model. Models c, e, g, and h fulfill this requirement.

In order to assess the importance of s^2 and the interaction term I, we perform an in-depth comparison of predictive skill of the models c, e, g, and h. As the mean verification scores are almost equal, statistical testing is required in order to be able to make sound statements on relative model performances. To this end, a

Table 3.3: Means of log scores and CRPS values over the entire verification period and all stations[a]. Table taken from Hemri et al. (2016).

model	log score				CRPS			
lead time	1d	3d	6d	10d	1d	3d	6d	10d
raw ensemble	-	-	-	-	0.154	0.154	0.168	0.184
uniform	2.20	2.20	2.20	2.20	0.226	0.226	0.226	0.226
climatology	1.76	1.76	1.76	1.76	0.177	0.177	0.177	0.177
MLR-B	1.50	1.55	1.64	1.72	0.119	0.128	0.148	0.166
MLR-S	1.50	1.55	1.65	1.73	**0.117**	0.127	**0.147**	**0.165**
POLR-B	1.49	1.54	1.64	1.72	0.119	0.128	0.148	0.166
POLR-S a	1.49	1.54	1.63	**1.70**	0.121	0.129	0.148	**0.165**
POLR-S b	1.49	1.54	1.63	**1.70**	0.120	0.129	0.148	**0.165**
POLR-S c	**1.47**	**1.52**	**1.62**	**1.70**	0.118	0.127	**0.147**	**0.165**
POLR-S d	1.48	1.53	1.63	**1.70**	0.119	0.128	0.148	**0.165**
POLR-S e	**1.47**	**1.52**	**1.62**	**1.70**	**0.117**	0.127	**0.147**	**0.165**
POLR-S f	1.48	1.53	1.63	**1.70**	0.119	0.128	0.148	**0.165**
POLR-S g	**1.47**	**1.52**	**1.62**	**1.70**	**0.117**	0.127	**0.147**	**0.165**
POLR-S h	**1.47**	**1.52**	**1.62**	**1.70**	**0.117**	**0.126**	**0.147**	**0.165**

[a] In each column the best value is shown in bold. Log scores have not been calculated for the raw ensemble.

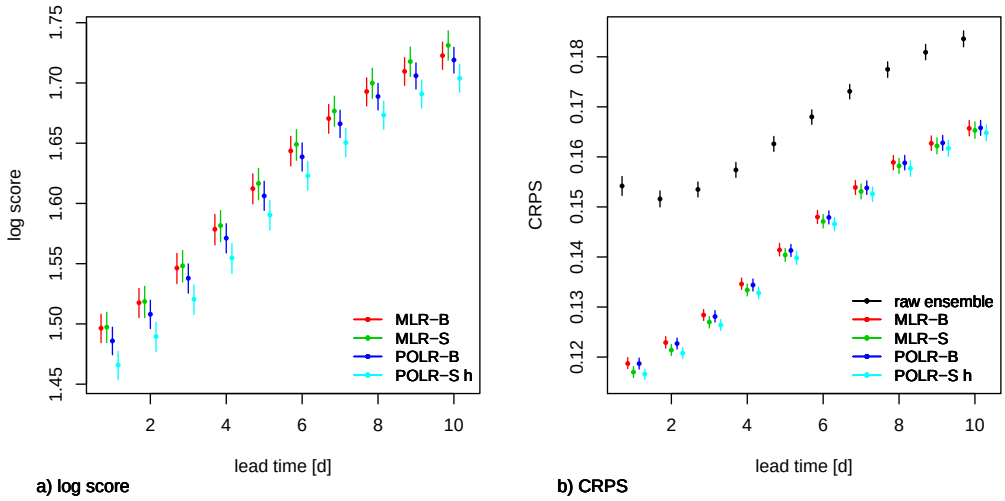


Figure 3.5: Means of log scores and CRPS values over the entire verification period and all stations for the raw ensemble (only CRPS), MLR-B, MLR-S, POLR-B, and POLR-S h. The centered 90 % confidence intervals have been obtained by block bootstrapping. Figure taken from Hemri et al. (2016).

stationwise assessment of significant changes in CRPS and/or log score has been performed using block bootstrapping. In order to combine log score and CRPS, three cases are distinguished:

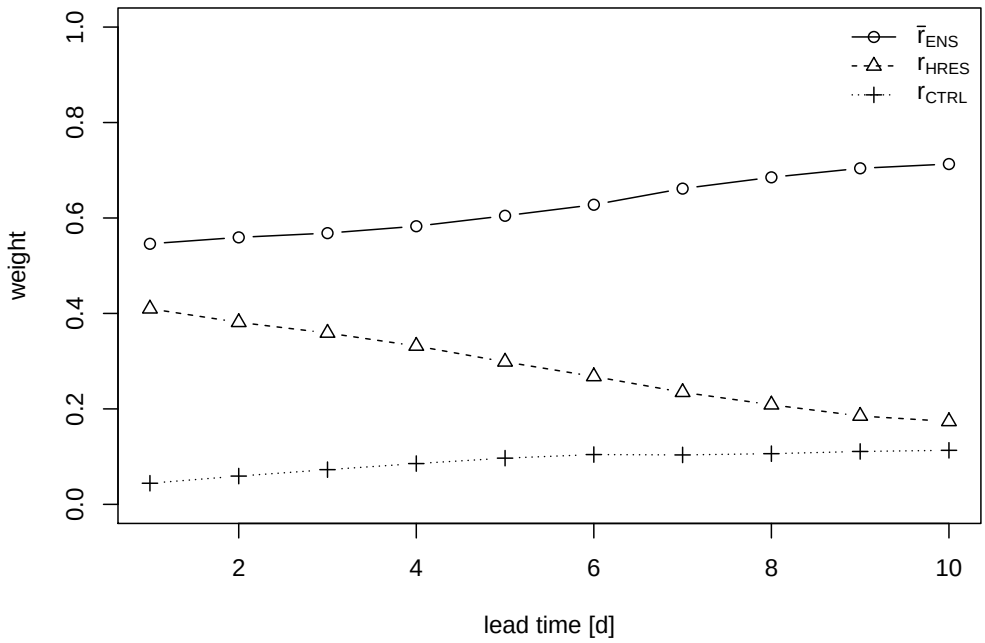


Figure 3.6: EMOS weights pooled over all stations and training periods of $\bar{r}_{\text{ENS}}$, r_{HRES}, and r_{CTRL}. Figure taken from Hemri et al. (2016).

i) Deterioration: at least one of the two scores (CRPS or log score) is deteriorated, while the other is not improved.

ii) No clear-cut difference: either both scores indicate no change in forecast skill or one of the two scores is improved, while the other is deteriorated.

iii) Improvement: at least one of the two scores is improved, while the other is not deteriorated.

As we are comparing changes in CRPS and log score simultaneously, a correction for multiple comparison has to be applied. We set the target Type I error to 0.05, i.e. $\alpha = 0.05$. In order to achieve $\alpha = 0.05$, a Bonferroni correction is applied (Bonferroni, 1936). In the present example $\alpha' = 0.0255$ is used in the individual tests for changes in CRPS and changes in log score, respectively. As for the above confidence interval calculations, the block bootstrapped tests for significant changes in CRPS and/or log score are based on block resamples following a geometric distribution with mean $|V|^{1/3}$. Models c, e, g and h are now compared using a forward selection approach. As reported in Table 3.4 adding the interaction term I leads to greater improvements in skill than adding s^2 at a forecast lag of three days. The full model h with an additional inclusion of s^2 shows a slightly increased skill relative to model g. Hence, the full model h should be preferred in case of short forecast lags. At a forecast lag of 6 days no clear

Table 3.4: From left to right: Percentage of stations with a deterioration, no clear-cut difference, or an improvement in skill when adding s^2 to POLR-S c resulting in POLR-S e, when adding I to POLR-S c resulting in POLR-S g, and when adding s^2 to POLR-S g resulting in POLR-S h. Table taken from Hemri et al. (2016).

	e to c			g to c			h to g		
	det	no diff	impr	det	no diff	impr	det	no diff	impr
lag 3 day	2.5	78.9	18.6	0.9	73.3	25.9	3.6	84.0	12.4
lag 6 day	4.7	89.9	5.4	3.5	93.3	3.2	6.1	90.3	3.6
lag 10 day	5.9	92.0	2.1	7.2	92.3	0.5	7.5	91.0	1.5

difference can be observed between the different model versions. At a very long lead time of 10 days the simplest model c seems to perform best. We subjectively select the full model h for the further analyses, because it performs best at the short lead times, which are also those with the highest predictability.

Calibration and sharpness

Keeping the improvement in skill by TCC post processing in mind, calibration and sharpness are now assessed in more detail. Calibration is the degree of statistical consistency between predictive distributions and observations, and is verified by means of PIT histograms. Figure 3.7 compares the PIT histograms of the raw ensemble, MLR-B, POLR-B, and POLR-S h predictions at forecast lead times of 3, 6, and 10 days. Flat PIT histograms indicate well calibrated forecast distributions, whereas a $\cup$-shape is a sign of underdispersion, and a $\cap$-shape is a sign of overdispersion. Pooled over all stations, all post processed models are well calibrated. The raw ensemble forecasts are clearly underdispersive at a forecast lag of 3 days and only slightly underdispersive at a forecast lag of 6 days. At a forecast lag of 10 days the PIT histogram of the pooled raw ensemble forecasts is somewhat unclear. Nevertheless, it is still less well calibrated than the corresponding post processed forecasts.

Sharpness is assessed here by an evaluation of the variances, and the widths of the centered 90 % prediction intervals, pooled over all stations and verification days. As shown in Figure 3.8, the raw ensemble provides the sharpest forecasts at a forecast horizon of 3 days. At lead times of 6 and 10 days the sharpness of the raw ensemble and the post processed forecasts is quite poor. However, all post processed models are sharper than the raw ensemble. This result is somewhat surprising in that statistical post processing improves both calibration and sharpness. Further insight into this can be obtained by assessing marginal calibration (Gneiting et al., 2007a). A forecast is marginally well calibrated if the average predictive CDF over all verification days equals the empirical CDF of the observations. A marginally well calibrated forecast leads to a horizontal marginal calibration graph. Details on the marginal calibration graph can be

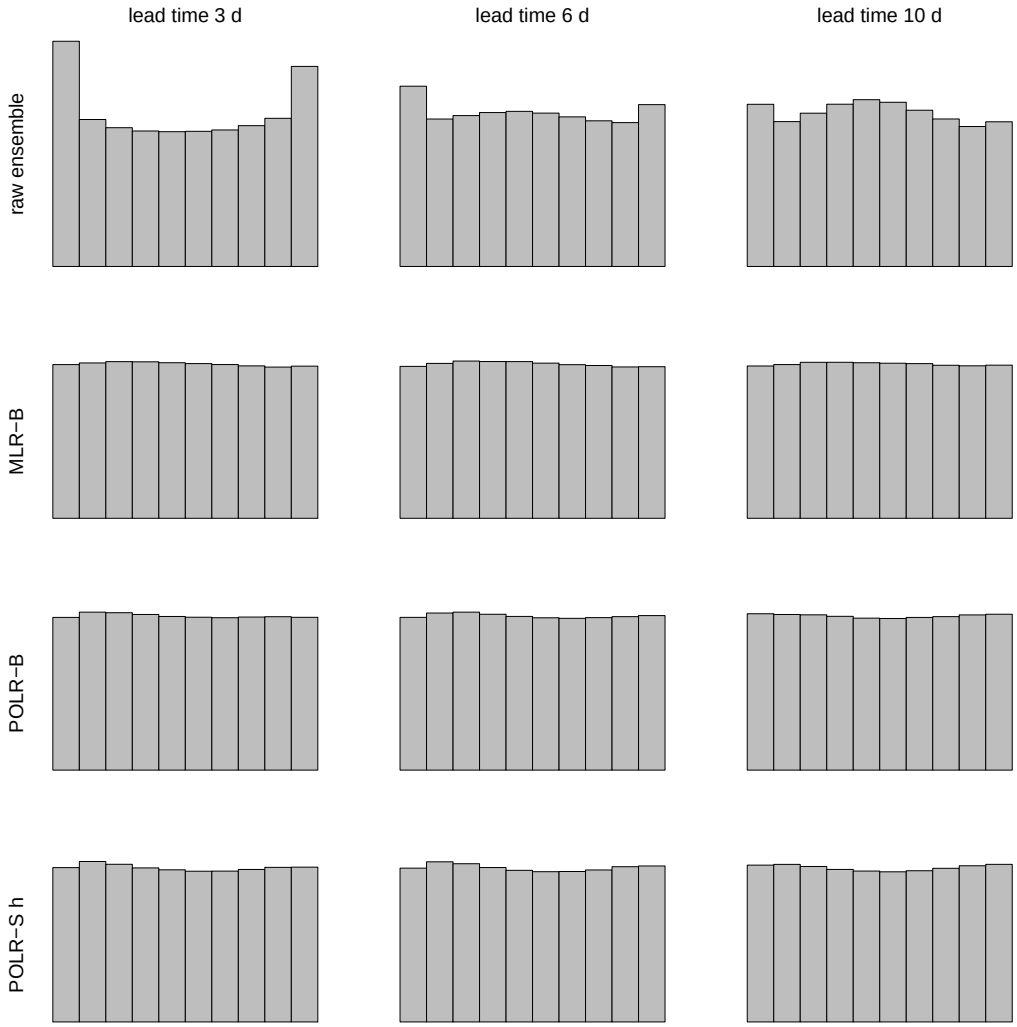


Figure 3.7: Histograms of the PIT values pooled over all stations and verification days for the raw ensemble, MLR-B, POLR-B, and POLR-S h. Figure taken from Hemri et al. (2016).

found in Appendix A.1.2. Figure 3.9 shows such graphs for the climatological, the raw ensemble, and the POLR-S h forecasts for a selection of European stations with different TCC climate. As expected, the climatological forecasts show almost perfect marginal calibration. The raw ensemble exhibits poor marginal calibration, even though it is mapped to the observation space in a sound way (see above and Appendix A.1.1). It assigns too much weight to TCC values of 0 or 8 octas irrespective of station and lead time. Brussels provides a good example of this. The most frequently observed TCC value is seven octas. However, the raw ensemble assigns forecast weight rather to 8 octas as can be seen from the accentuated negative peak in the marginal calibration graph. POLR-S h performs as well as the climatological forecasts in terms of marginal calibration. Hence, post processing conveys a significant improvement in marginal calibration.

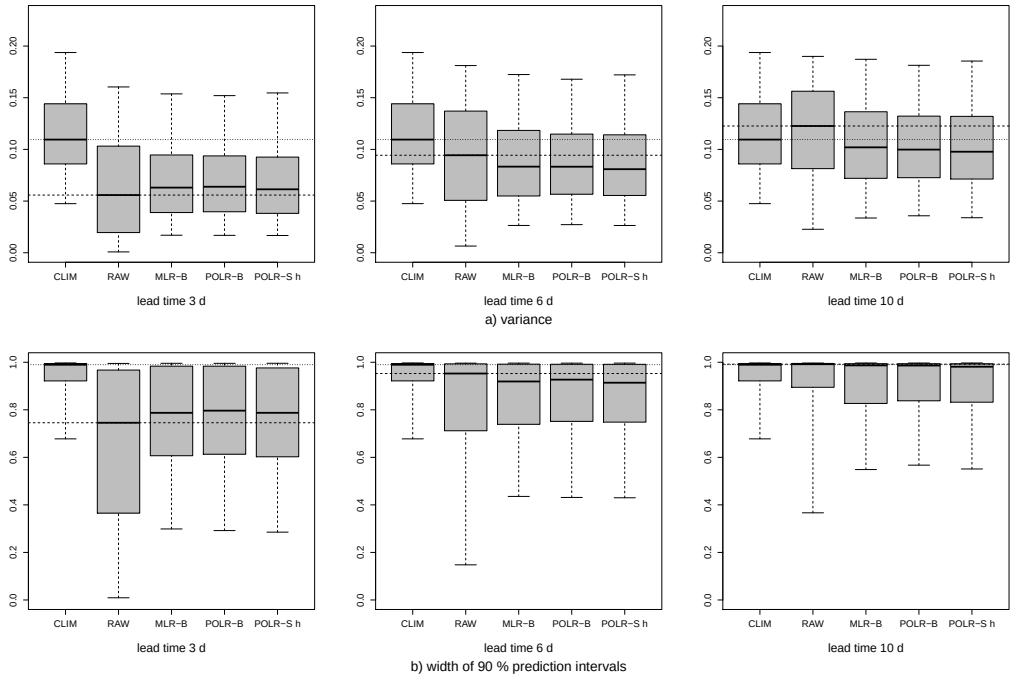


Figure 3.8: Box plots showing the 5, 25, 50, 75, and 95 % quantiles of the empirical distribution of a) the forecast variances and b) the widths of the centered 90 % prediction intervals pooled over all stations and all verification days for the climatological, the raw ensemble, MLR-B, POLR-B and POLR-S h forecasts. The horizontal dashed (dotted) line corresponds to the 50 % quantile of the empirical distribution of the corresponding statistic of the raw ensemble (climatological) forecasts. Figure taken from Hemri et al. (2016).

3.2.5 Discussion

Both MLR and POLR prove to be useful methods for post processing of raw ensemble TCC forecasts. The results indicate that on average POLR with seasonally estimated model parameters performs best. This post processing method clearly improves forecast calibration. In order to achieve well calibrated forecasts, sharpness has to be reduced at the shorter forecast horizon of 3 days. But surprisingly, sharpness can be improved by post processing for the longer forecast lags of 6 and 10 days. Keeping in mind the paradigm stated by Gneiting et al. (2005, 2007a) that the goal of statistical post processing is to maximize sharpness subject to calibration, the simultaneous improvement in calibration and sharpness is very desirable. This is mostly due to the tendency of the raw ensemble to assign too much weight to cloud cover states of zero and eight octas.

The methods presented in this study are designed to post process discrete TCC raw ensemble forecasts against SYNOP observations. Depending on the region, TCC observations are recorded automatically or manually, with different observation error characteristics. According to Mittermaier (2012) automated

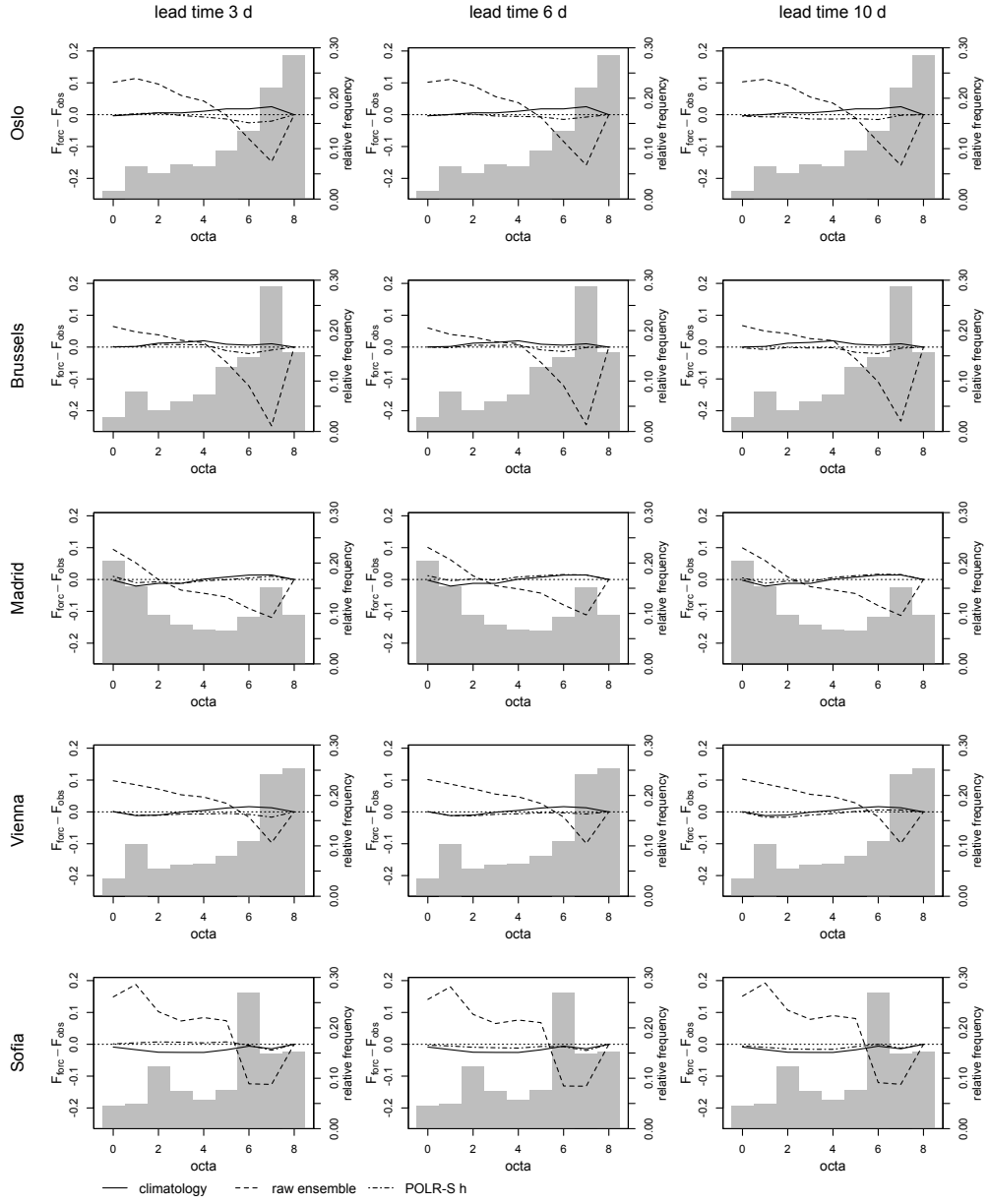


Figure 3.9: Marginal calibration plots comparing the climatological, the raw ensemble, and the POLR-S h forecasts at lead times of 3, 6, and 10 days at different stations in Europe. The observed climatology over the verification period is visualized by the barplots showing the relative frequencies of the different TCC classes. Figure taken from Hemri et al. (2016).

observations may underestimate the amount of high cloud (cirrus), while for human observers there is a tendency to underestimate cloud cover states of 0 and 8 octa. This may partly explain the poor marginal calibration of the raw ensemble when compared to the observations. However, a comparison of results at individual stations with manual observations and with automated observations

did not reveal a systematic difference in marginal calibration. As human TCC observers are increasingly replaced by automated observations (Wacker et al., 2015), one would need to know the exact date at which a particular station has been changed from manual to automated for a more detailed analysis of this effect. Currently, SYNOP observations of total cloud cover are mainly automated in western Europe, North America, Australia and New Zealand, Japan, South Africa and Antarctica. Due to the increasing number of automated stations, continuous TCC observations may become more widely available in future. As the ECMWF raw ensemble provides TCC forecasts that are continuous on the unit interval, this would allow for continuous verification and post processing of TCC raw ensemble predictions and probably further enhance forecast skill. A continuous post processing method for predictions of visibility, which is a bounded variable like TCC, has already been implemented by Chmielecki and Raftery (2011).

TCC can be differentiated into low, medium, and high level clouds. Predictive skill of NWP cloud cover forecasts can be different depending on cloud level. For instance, in the lowlands of the greater Alpine region, the ECMWF HRES model underestimates persistent low stratus (Haiden and Trentmann, 2015). It might be possible to reduce such systematic biases by cloud level specific post processing. Though a direct inclusion of low, medium, and high level cloud forecasts as predictors in the POLR model, cf. Equation (2.16), did not lead to any improvement in forecast skill (results not shown here), further analyses may be beneficial. In particular, a separate post processing of each cloud level with training observations differentiated according to cloud level may further increase forecast skill.

To summarize, considering the global set of SYNOP stations covered by this study post processing of discrete TCC raw ensemble predictions using readily available methods can improve forecast skill significantly. Hence, post processing helps to improve the generally low predictive performance of raw ensemble TCC forecasts. Additionally, this study identified the seasonal POLR model as the most skillful TCC post processing approach.

Chapter 4

Hydrological ensemble post processing

4.1 Scientific setting

4.1.1 Motivation

Reliable hydrologic forecasts are crucial for a wide range of activities like, for instance, the operation of hydropower plants, shipping, flood prevention, and leisure activities. Information about the predictive uncertainty of the predictand (i.e. runoff, water level) is required for rational decision making. As already mentioned in Section 1.2, predictive uncertainty is defined as the uncertainty of a future realization of a predictand, the quantity of interest, conditional on all available information and knowledge (Krzysztofowicz, 1999; Todini, 2008). The available knowledge about the future realization in hydrologic forecasting is generally embedded in one or more hydrological model forecasts. As already mentioned, one of the main sources of uncertainty is the meteorological uncertainty of the short- to medium-range development of weather patterns. Usually, the hydrological forecast ensembles inherit the biases and underdispersion of the meteorological input ensembles (Bougeault et al., 2010; Park et al., 2008). Additionally, hydrological uncertainties, like the level of ground water storage or uncertainties in the hydrological model formulation, are typically neglected in rainfall-runoff modelling. Accordingly, statistical post processing of the hydrologic ensemble forecasts is needed in order obtain an estimate of predictive uncertainty and improve forecast skill.

4.1.2 Univariate post processing

In order to put the hydrological case studies of this chapter into a broader context, a selection of studies that focus on statistical post processing of hydrologic ensemble forecasts is presented first. Earlier studies on statistical post processing proposed Bayesian models to quantify the uncertainties of hydrological forecasts. Krzysztofowicz (1999, 2002) introduced the Bayesian forecasting system (BFS)

to produce probabilistic forecasts from deterministic hydrological forecasts. The hydrological uncertainty processor (HUP) that aggregates the hydrological model uncertainties is a component of the BFS (Krzysztofowicz and Kelly, 2000). Reggiani et al. (2009) extended the HUP for post processing of ensemble forecasts for the river Rhine on the Dutch-German border. Madadgar et al. (2012) post processed ensemble forecasts by applying copula techniques that fit a bivariate distribution to forecasts and observations. Bayesian model averaging (BMA: Raftery et al. (2005)) has been used for the probabilistic combination of (ensemble-) runoff forecasts in many cases. For instance, Ajami et al. (2007) or Duan et al. (2007) showed that the combination of hydrologic forecasts using BMA led to both, quantitative statements on prediction uncertainty and improvements in terms of deterministic verification measures. As already stated in Section 2.1.1, Fraley et al. (2010) introduced an adapted BMA version that is able to take account of ensemble forecasts with exchangeable members as typically encountered with meteorological ensemble forecasts. Recent developments allow to use flexible predictive distributions (Parrish et al., 2012; Rings et al., 2012) and to post process forecasts over an entire range of lead times simultaneously (Hemri et al., 2013; Engeland and Steinsland, 2014). Other alternatives for statistical post processing are the model conditional processor (Todini, 2008; Coccia and Todini, 2011), which has recently been extended to handle ensembles (Todini et al., 2015), and quantile regression (Weerts et al., 2011). A non-parametric approach for the post processing of hydrological ensemble forecasts which is similar to indicator co-kriging was proposed by Brown and Seo (2010). This list gives an overview over the different post processing methods used in hydrology, but is by no means exhaustive.

4.1.3 Seamless prediction

Seamless prediction, i.e. consistent prediction over successive lead times, is of increasing importance in the field of hydrometeorological forecasting and the main topic of the study by Hemri et al. (2015) presented in Section 4.3. For instance, Palmer et al. (2008) motivate the use of seamless predictions by the verification of climate models. Based on the premise that the fundamental physical processes of seasonal forecasts and decadal climate projections are similar, probabilistic climate forecasts can be calibrated according to the validation results of the seasonal predictions of the corresponding models. In meteorology a seamless prediction system is designed to cover the time span from weather to climate predictions. However, in hydrology seamless predictions span a somewhat shorter time horizon from nowcasting flash floods to seasonal drought predictions (Yuan et al., 2014). Short range hydrologic forecasts may benefit from a blending of precipitation nowcasts and forecasts. Kober et al. (2012, 2014) and Scheufele et al. (2014) propose and apply such a blending method using a weighting function that depends on lead time and the conditional square root of the ranked probability score. Hydrologic model runs based on seamless meteorological predictions can

be expected to be seamless as well. If hydrologic ensemble forecast trajectories, which have been obtained by statistical post processing, are used as inputs to a hydrodynamic model or for river routing, it is crucial to avoid discontinuities in the marginal predictive distributions. Naive approaches smooth the parameter estimates of the univariate model fits. For instance, in case of EMOS the estimates of the parameters can be smoothed using cubic smoothing splines as implemented in the study in Section 4.3. More sophisticated approaches would be based on simultaneous parameter estimation over the entire range of lead times. To the authors knowledge there are no studies addressing this in the context of hydrological post processing, though several methods used for spatially adapted post processing of meteorological forecasts have been developed. Such methods can often be transferred to temporal problems. For instance, the locally adaptive EMOS method (Feldmann et al., 2015; Scheuerer and Büermann, 2014) could probably be modified in such a way that simultaneous parameter estimation over the entire range of lead times becomes feasible.

4.2 Ascertainment of probabilistic runoff forecasts considering censored data

As stated in Chapter 1 uncertainty of hydrologic forecasts is increased below or above certain thresholds. The BfG forecasting system for river Rhine handles uncertainty of very low runoff values by censoring runoff at a lower threshold. In the study by Hemri et al. (2014a) a method for post processing of left censored hydrologic data has been developed and applied to two test catchments. The following sections on censored post processing closely follow Hemri et al. (2014a).

4.2.1 Introduction

As stated above, the BfG forecasting system for river Rhine includes left-censored (model-) data, i.e. values below a certain threshold are replaced by this threshold. Censoring is an appropriate method to deal with very uncertain or not defined data in hydrologic real-time applications. Censoring may apply to both very low runoff values (e.g. in the case of unreliable rating curves between water level and runoff and in the case of impounded waters) and very high runoff values (e.g. when the range of the rating curve is exceeded). The goal of this study is to apply the censored EMOS model from Section 2.1.3 to censored raw ensemble forecasts. The following analyses focus on the rather small catchments of the rivers Ahr (gauge Altenahr) and Wied (gauge Friedrichsthal), because both rivers feature high proportions of censored data.

After a short description of the data used in this study in Section 4.2.2, an overview of the different censored forecasting methods will be given in Section 4.2.3. The results in Section 4.2.4 are followed by a short discussion in Section 4.2.5.

4.2.2 Runoff data

The following analyses are based on forecast and observation data from the gauges Friedrichsthal (Wied) and Altenahr (Ahr) with an hourly temporal resolution. The catchment areas amount to 680 km^2 and 746 km^2 for the rivers Wied and Ahr, respectively. Figure 4.1 depicts the location of the two catchments within the catchment of river Rhine and their topography. In case of both catchments the operational forecasting system of the BfG censors the large proportion of runoff observations that are below a global threshold of 5 m^3/s prior to any further processing (i.e. primarily statistical forecast corrections). Over the study period from 1 November 2008 to 31 October 2011 55 % of the observations are censored at gauge Friedrichsthal and at gauge Altenahr censoring amounts to 72 %. Likewise, the corresponding runoff forecasts from the hydrologic model are left-censored.

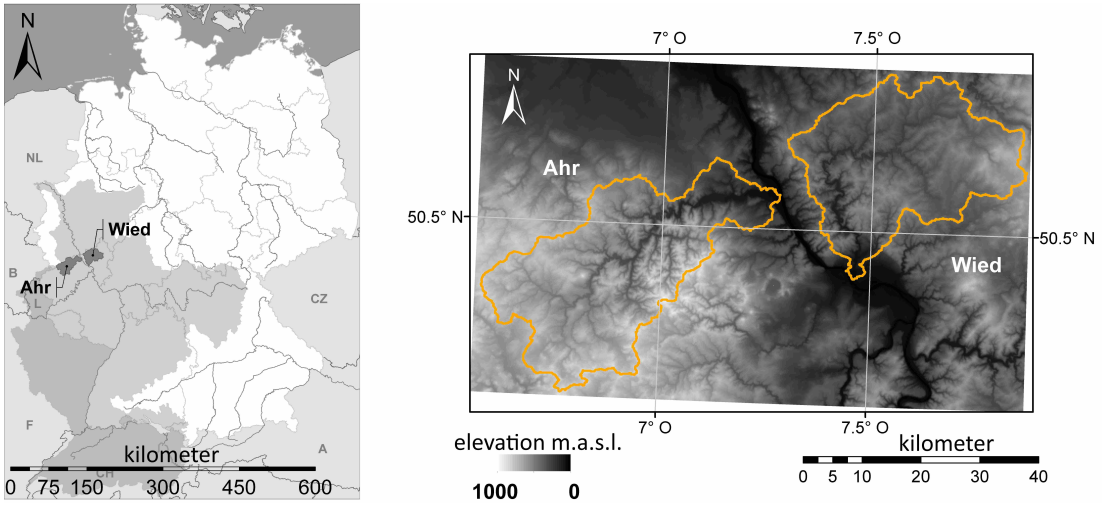


Figure 4.1: Location of the catchments of the rivers Wied and Ahr in the Rhine basin and equal-area projection of both areas' digital elevation models. Figure taken from Hemri et al. (2013).

4.2.3 Methods

Raw ensemble forecasts

At the BfG the conceptual, semi-distributed rainfall-runoff model HBV-96 (Bergström, 1995; Lindström et al., 1997) is used for operational runoff forecasting. The Rhine river basin is divided into 134 sub-basins which are further subdivided into hydrological response units (HRU) according to land use and elevation classes. The hydrologic processes for runoff formation are calculated on those HRU's (Meißner and Rademacher, 2010). The model calculates runoff with a temporal resolution of 1 h using temperature and precipitation fields that have

Table 4.1: Meteorological deterministic and ensemble forcing models[1]

name	number of members	max forecast lag	resolution
COSMO-LEPS	16	132 h	10 km
DWD-GME	1 (deterministic)	174 h	20 km
DWD-MER	1 (deterministic)	78 h (174 h)	7 km (20 km)

[1] DWD-MER stands for a model run based on COSMO-EU forcing up to lead-time 78 h and on DWD-GME thereafter. Table taken from Hemri et al. (2014a).

been interpolated over the sub-basins as meteorological input. Runoff forecasts are obtained by running the hydrological model with the meteorological forecasts from several different NWPs. NWP models can be either deterministic or probabilistic. In the first case, uncertainty is neglected and a single forecast trajectory is provided. In contrast, ensemble forecasts try to represent uncertainty by several model runs with different initial conditions, boundary conditions and physical parameter values. The term ensemble forecast comprises the collection of these distinct model runs. In current operational use at the BfG the ensemble forecasts from the hydrological model HBV-96 are used as boundary conditions and lateral inflows to a hydrodynamic model to calculate water level forecasts for stations along the river Rhine.

All hydrological forecast data used here is generated by hindcasting the hydrological model with archived operational meteorological forecasts. As summarized in Table 4.1, the meteorological forcing models vary in both the number of ensemble members and the forecast time horizon. The hydrologic model is run with these models as forcing leading to a hydrologic 18 member multi-model ensemble, referred to as the raw ensemble. It is composed of the 16 COSMO-LEPS members (Montani et al., 2011), and the two deterministic models DWD-GME (Majewski et al., 2002, 2012) and DWD-MER. The hydrologic model provides hourly forecasts up to 174 h (DWD-GME and DWD-MER) and 114 h (COSMO-LEPS). DWD-MER uses meteorological forcing from the COSMO-EU model (Steppeler et al., 2002; Schulz and Schättler, 2011) up to lead time 78 h, and data from the DWD-GME model from lead time 79 h. Hence, two members of the raw ensemble rely on the same meteorological inputs from lead time 79 h onwards. The hydrologic forecasts are initialized on a daily basis from 1 November 2008 to 25 January 2011 at 06:00 UTC. The initial conditions of the hydrologic model are generated by a continuous simulation up to the forecast issue date using observed meteorological input. Finally, the raw ensemble runoff forecasts are statistically corrected based on the observations available up to the forecast date using an autoregressive model (Boersen and Weerts, 2005). For the following analyses, only lead times up to 114 h are considered because of dropping out ensemble members. At lead time 114 h the forecast horizon of COSMO-LEPS is reached. Hence, it drops out of the raw ensemble. This problem of dropping out ensem-

ble members would require a more detailed analysis (see also Hemri et al. (2013)).

Climatological forecasts

Climatological forecasts serve as reference forecasts for forecast verification. In order to account for seasonal runoff variation daily climatological forecasts are calculated. Based on an hourly time series of observed runoff from 1 November 1998 to 31 October 2008, the climatological forecasts are obtained by calculating the empirical distribution of the observations that lie within ± 15 days of the calendar date of the verification day, but not in the same year. The dependence of runoff on the time of day is neglected. The drawback of mixing different times of day is more than compensated for by the increase in the sample size from which the climatology is constructed. The climatological forecasts are probabilistic which results directly from their construction that is based on empirical distributions of historical observations.

Censored EMOS

As mentioned in Section 4.2.1 post processing of the ensemble runoff forecasts for the rivers Wied and Ahr is based on the censored EMOS approach that has been presented in Section 2.1.3. Prior to any statistical model fitting, training and and corresponding verification periods have to be selected. Here, the verification set comprises forecast/observation pairs from 1 November 2008 to 31 October 2011. As the forecasts cover lead times from 1 to 114 h, the forecast initialization dates range from 1 November 2008 to 25 October 2011. That is, the verification set consists of 1085 initialization days and 114 lead times. As the behavior of the hydrological system varies over the year (e.g. snow accumulation in winter, snow melt in spring, low flow situations in summer), the parameters of the EMOS model have to be estimated for each meteorological season separately. That is, for the verification of a forecast issued on a particular date the forecast/observation pairs issued on days that are in the same season but not in the same year are used as training data. For instance, if the forecast to be verified is issued in March 2009, the training period comprises the forecasts issued in spring 2010 and spring 2011. Such training periods are constructed for each verification day. Examples of combinations of verification and training periods are listed in Table 4.2. These pairs of verification and training periods are used for this study and the study on multivariate post processing of runoff ensemble forecasts presented in Section 4.3.

The censored EMOS model is then fitted to Box-Cox transformed pairs of raw ensemble forecasts and observations over the training periods using minimum CRPS estimation. The estimated Box-Cox parameter $\hat{\lambda}$ are -0.31 and -0.42 for the rivers Wied and Ahr, respectively. We estimate two slightly different censored EMOS models. The first, which is called the naive approach henceforth, does not apply the correction for μ described in Equation (2.21) in Section 2.1.3. The

Table 4.2: Examples of pairs of verification and training periods [1]

verification period	training period
November 2008	SON 2009, SON 2010, SO 2011
DJF 2008/2009	DJF 2009/2010, DJF 2010/2011
MAM 2009	MAM 2010, MAM 2011
.	.
.	.
.	.
SO 2011	November 2008, SON 2009, SON 2010

[1] SON denotes September, October, November; SO September, October; DJF December, January, February; MAM March, April, May. Table taken from Hemri et al. (2015).

second, called the μ-corrected approach in the following, includes this correction term.

4.2.4 Results

Averaged over the entire verification period, raw ensemble as well as the naive and the μ-corrected EMOS methods perform well compared to the climatological forecasts. Figure 4.2 a) reveals the clear-cut skill improvement in terms of CRPSS, i.e. the skill score calculated from the CRPS based on Equation (2.31) in Section 2.3.1, by EMOS compared to the raw ensemble. In case of both catchments Wied and Ahr, CRPSS of the two EMOS methods and the raw ensemble is comparable for the first 10 to 15 lead times, whereas CRPSS benefits from EMOS at higher forecast lags. The naive and the more sophisticated μ-corrected EMOS method show hardly any differences. Because of the large number of censored observations at both gauges, an evaluation of the predicted censoring probabilities is now performed by means of the Brier skill score (BSS), i.e. the skill score associated to the Brier score. Here, the two dichotomous events considered for calculation of the BSS are censoring, i.e. runoff up to 5 m^3/s, or no censoring. As shown in Figure 4.2 b) BSS cannot be improved by any of the two EMOS methods. In contrast to the improvements in terms of CRPPS, EMOS even deteriorates BSS for forecast lags beyond 50 hours in case of the river Ahr.

Subsequent to the assessment of forecast skill, let us now take a closer look at calibration and sharpness. As depicted by the $\cup$-shaped 3D PIT histograms in Figure 4.3 a) the raw ensemble forecasts for both censored catchments are clearly underdispersed over the entire forecast horizon. Nevertheless, these forecasts are quite well calibrated compared with the forecasts for the uncensored and considerably larger catchments of the Upper Rhine, Moselle, and Lahn as will be shown in Figure 4.9 in Section 4.3. As shown in Figure 4.3 b), post processing using

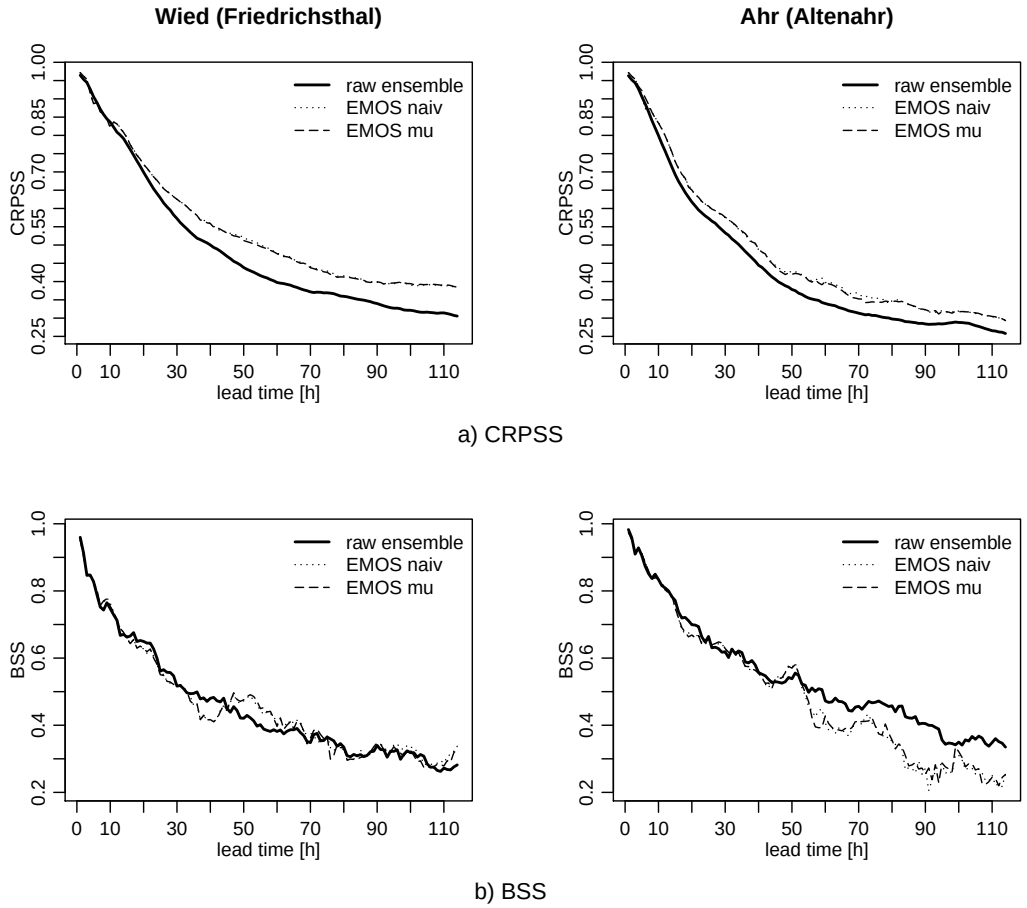


Figure 4.2: CRPSS and BSS values at the censoring threshold of 5 m^3/s for the rivers Wied at Friedrichsthal and Ahr at Altenahr. Due to a modification in the calculation of the climatological reference forecasts, the corresponding figures in Hemri et al. (2014a) show different CRPSS and BSS values.

the censored EMOS leads to well calibrated forecasts for both catchments. Note that in case of censoring probabilities greater than 0.1, censored observations are assigned proportionally to one of the possible quantile intervals. If, for instance, the forecast censoring probability is 0.15, then a censored realization is assigned at a ratio of two thirds to the first decile and the remaining one third to the second decile.

Given the well calibrated EMOS forecasts, sharpness is assessed now. To this end, the empirical distribution of the lower one-sided 90 % prediction intervals is constructed from the entire verification period. Following Gneiting et al. (2007a), important quantiles of that distribution are plotted in a box plot like manner. As shown in Figure 4.4, censored EMOS deteriorates sharpness only slightly. For both gauges, the most substantial difference can be seen from the 95 % quantiles of the just mentioned empirical distribution. This indicates that censored EMOS

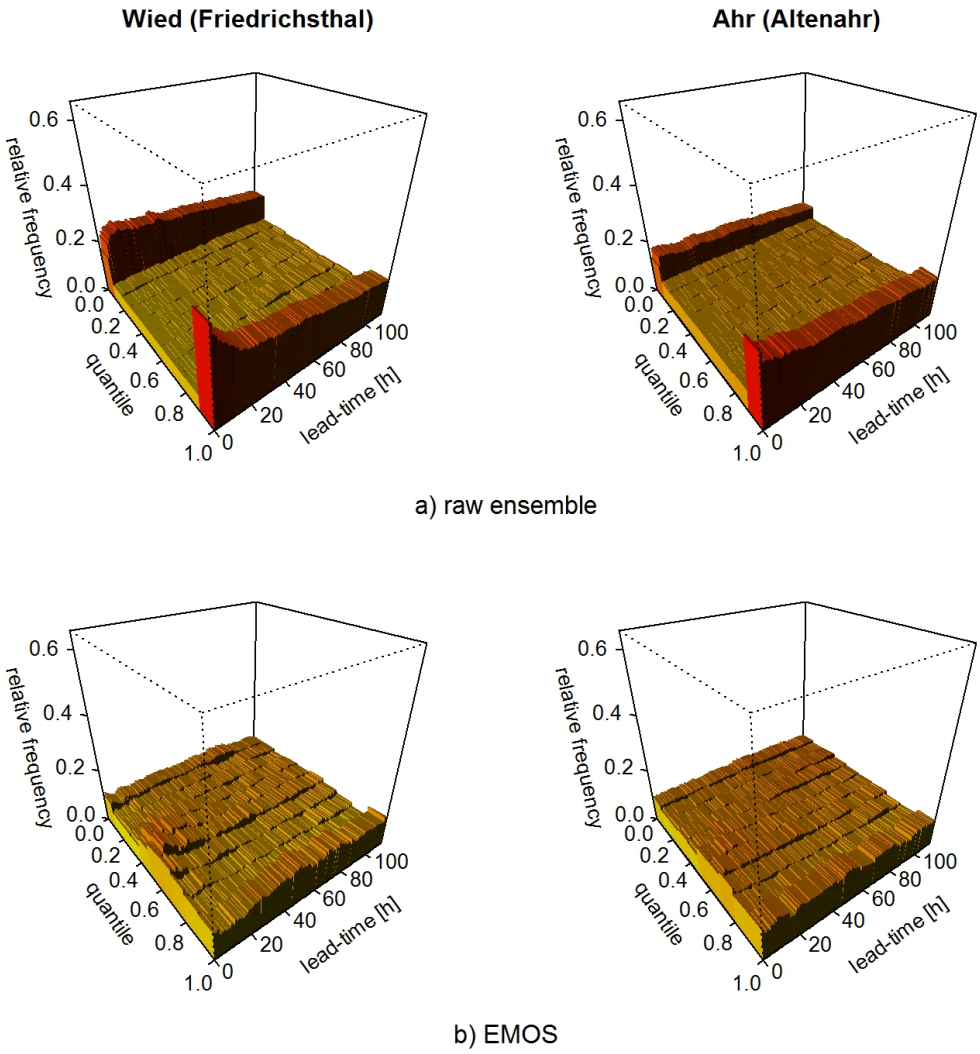


Figure 4.3: 3D PIT histograms for a) the raw ensemble and b) the censored EMOS forecasts with naively estimated variance. Figures taken from Hemri et al. (2014a).

augments forecast uncertainty particularly when raw ensemble uncertainty is already quite high. In case of the 24 h forecasts for river Wied, also the 50 % and 75 % quantiles are higher for EMOS than for the raw ensemble. Besides from that, differences in sharpness can hardly be detected. Furthermore, the 25 % quantile of the distribution of interval widths is zero for both methods at both gauges and all considered lead times. This reflects the effect of censoring on forecast sharpness. In line with the improved calibration, censored EMOS also increases forecast coverage. The rather small differences in sharpness compared to the raw ensemble, indicate that on average censored EMOS leads to quite sharp forecasts.

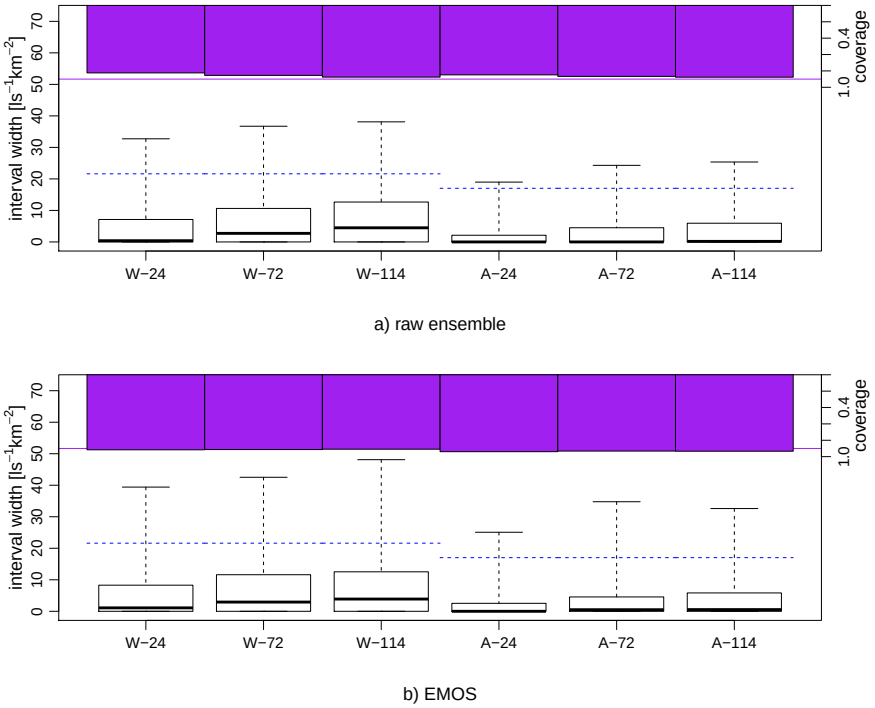


Figure 4.4: Comparison of sharpness and coverage of the raw ensemble and EMOS forecasts at lead times 24, 72, and 114 h. The blue dotted line segments show the medians of the empirical distribution of the lower one-sided 90 % prediction interval widths of the daily climatological forecasts. The purple lines correspond to the nominal coverage. The box plots show the 5, 25, 50, 75, and 95 % quantiles of the widths of lower one-sided 90 % prediction intervals from a) the raw ensemble and b) the EMOS forecasts for the rivers Wied (indicated by "W-xx", where xx denotes lead time), and Ahr ("A-xx"). Figure taken from Hemri et al. (2014a).

Example forecast

The verification results show that censored EMOS is able to post process ensemble forecasts that are affected by censoring. As an example of a forecast at the transition form censored to uncensored runoff, the raw ensemble forecasts, which are interpreted as quantiles from a probability distribution, for river Wied initialized on 7 November 2010 are shown along with the corresponding μ-corrected censored EMOS predictive distribution in Figure 4.5. In this example the marginal EMOS forecasts are well adjusted. However, it shows also the need for a multivariate EMOS method, because the forecast EMOS distribution exhibits quite sharp changes in marginal distributions from lead time to lead time that cannot be explained physically.

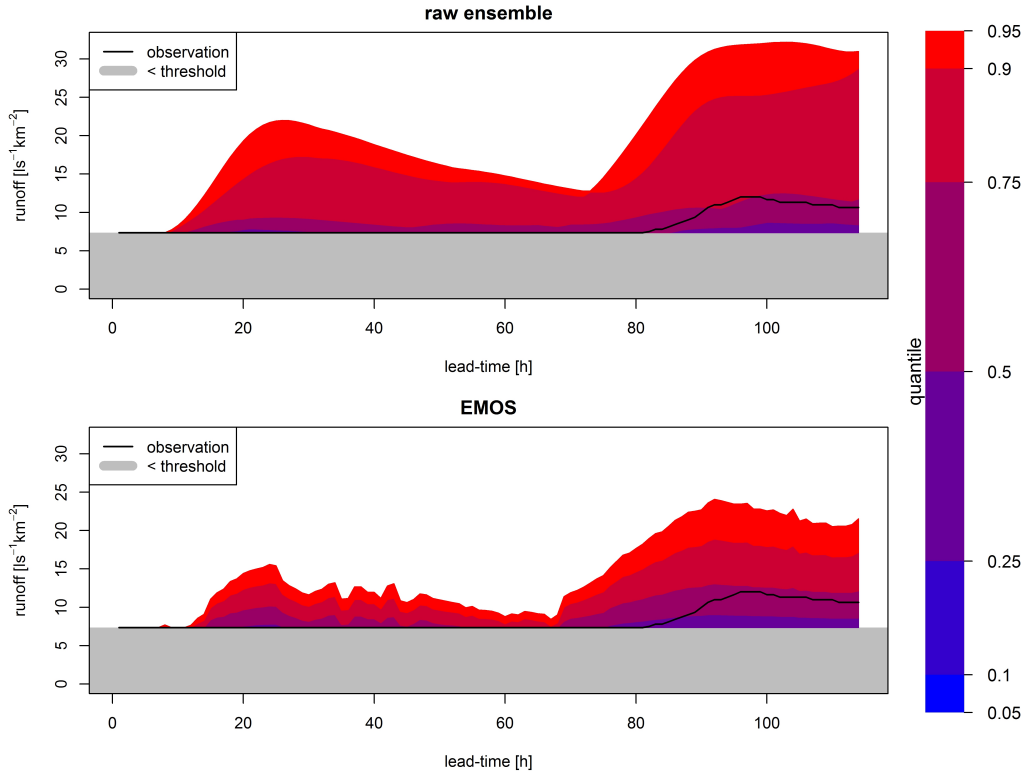


Figure 4.5: Raw ensemble and censored EMOS forecasts for river Wied at Friedrichsthal initialized on 7 November 2010 at 06:00 UTC. The quantiles of the raw forecasts have been obtained by linear interpolation between adjacent ensemble members. Figure taken from Hemri et al. (2014a).

4.2.5 Discussion

The verification analyses performed above have shown that censored EMOS improves the raw ensemble forecasts for the rivers Wied and Ahr in terms of CRPSS and calibration without deteriorating sharpness much. This is especially the case for forecast lags greater than one day. Accordingly, censored EMOS proves to be a useful approach to post process ensemble runoff forecasts at gauges with a large proportion of censored runoff values. Since censoring just shifts the density distribution below (or above) a deliberately selected censoring threshold to a point mass at the threshold value, it would be worth evaluating the general performance of censored EMOS to predict the probability of falling below (or exceeding) different thresholds. For instance, an artificially right censored EMOS method could lead to improvements in probabilistic flood forecasting. In particular in case of measurement uncertainties that are typical for flooding conditions, censored EMOS would allow to analyze questions like "What is the probability that runoff will exceed notification stage I, i.e. 823 m^3/s, at the gauge Trier in 48 hours?" adequately. Hence, following-up studies on left and right censored EMOS covering different catchments are crucial to gain further insight into censored EMOS.

The unsatisfactory results of censored EMOS in terms of BSS for river Ahr suggest that censored EMOS should be further improved. Furthermore, alternative censored methods should be taken into account. Nevertheless, censored EMOS is a promising approach, because it is a relatively simple post processing approach, of which the parameters can be estimated with low computational cost. A censored BMA approach could be an obvious alternative to censored EMOS. However, BMA would imply a much more complex modelling process. BMA based on a mixture of truncated normal kernel distributions (Baran, 2014) may be a good starting point for the development of a censored BMA method for ensemble river runoff forecasts. Furthermore, this study reveals the need for a multivariate post processing method that allows to introduce a realistic correlation structure between different forecast lags and avoids too wiggly patterns in the marginal distributions from lead time to lead time. Such multivariate post processing methods for uncensored catchments are discussed in the following section.

4.3 Multivariate post processing techniques for probabilistic hydrological forecasting

4.3.1 Introduction

As already mentioned in Chapter 1, runoff is an inherently multivariate process with typical events lasting from hours in case of floods to weeks or even months in case of droughts. This calls for multivariate post processing techniques that yield well calibrated forecasts in univariate terms and ensure a realistic temporal dependence structure at the same time. In the study presented here, which closely follows Hemri et al. (2015), multivariate post processing techniques are adapted and applied to multi-model river runoff ensemble forecasts.

The first but minor goal of this study is to achieve well calibrated and yet sharp marginal predictive densities. Here, the term marginal refers to the univariate predictive distribution for a particular lead time. To this end, we adapt the ensemble model output statistics (EMOS: Gneiting et al. (2005)) post processing method, which is frequently used for meteorological variables, so that it becomes suitable for probabilistic river discharge forecasts. More specifically, the truncated normal EMOS method from Section 2.1.3 is applied to ensemble runoff forecasts for three sub-catchments of river Rhine. Refer to Thorarinsdottir and Gneiting (2010) for details on the truncated EMOS approach.

According to Pinson and Girard (2012) knowing not only the marginal predictive distributions for each individual lead time, but also the dependence structure among different lead times, is crucial to making optimal decisions based on probabilistic forecasts. This applies in particular to runoff which is highly auto-

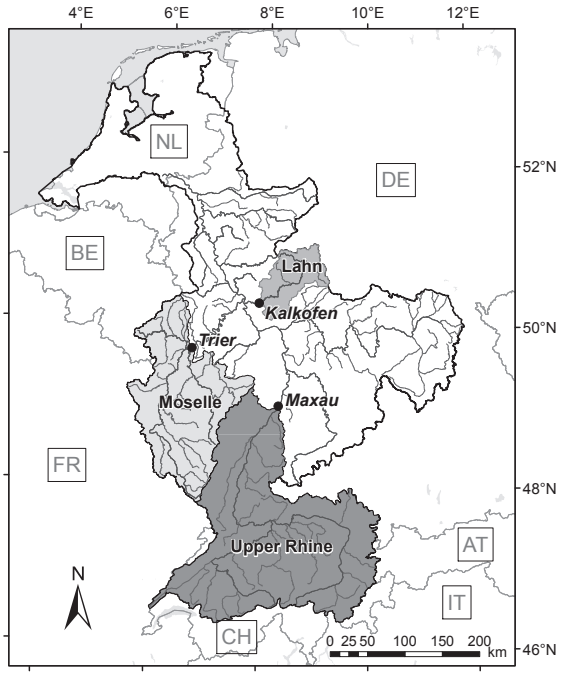


Figure 4.6: Locations of the considered sub-catchments within the Rhine river basin. Figure taken from Hemri et al. (2015).

correlated. After univariate post processing using EMOS the information about the temporal (spatio-temporal in the case of several gauges in a river basin) dependence structure of the raw ensemble is lost. If one is interested in forecast runoff trajectories, then a sound representation of the dependence structure has to be added. Forecast runoff trajectories are, for instance, required for the optimization of reservoir operation or, as in the case presented here, used as input to a hydrodynamic model to forecast water levels. Hence, the second and main goal of this study is to obtain forecasts that are not only marginally well calibrated, but from which it is possible to obtain also runoff scenarios over the entire forecast horizon. If the observed trajectory and the scenarios are likely to follow the same multivariate distribution, the forecast model is said to exhibit good multivariate calibration. This study compares two different approaches to introduce a dependence structure into the post processed forecasts: ensemble copula coupling (ECC: Schefzik et al. (2013)) and the Gaussian copula approach (GCA: Pinson and Girard (2012)). The non-parametric ECC approach is similar to the Schaake Shuffle (Clark et al., 2004) in that post processed forecast trajectories are reordered using exogenous information. The technical details of ECC, GCA, and the Schaake Shuffle have already been discussed in Section 2.2. In case of the Schaake Shuffle the ordering information stems from past observations, in case of ECC from the raw ensemble. As ECC accounts for both temporal and spatial

dependencies, it is suitable for parallel post processing of forecasts for different sub-catchments. This is required, for instance, if the post processed forecast trajectories are used as inputs to a hydrodynamic model to calculate water level forecasts further downstream. GCA is a parametric approach that estimates the correlation structure from training observations. GCA is expected to outperform ECC in cases, where a large number of forecast scenarios is required, or where it is doubtful whether the raw ensemble captures the correct correlation structure. The GCA variant described here accounts only for temporal dependencies, though it may be extended such that it is able to model spatio-temporal dependencies.

In this study EMOS, ECC, and GCA are verified based on runoff forecasts from the operational forecasting system of the German Federal Institute of Hydrology (BfG) for river Rhine (Meißner and Rademacher, 2010). Three different sub-catchments of river Rhine with different characteristics are considered: river Upper Rhine up to gauge Maxau, river Moselle up to gauge Trier, and river Lahn up to gauge Kalkofen.

In Section 4.3.2 the study areas and the observed runoff data are presented. The different types of forecasts used in this study as well as the methods used for model fitting and verification are introduced in Section 4.3.3. The results in Section 4.3.4 are followed by a discussion in Section 4.3.5.

4.3.2 Study areas and runoff data

The catchments in this study are selected such that different runoff regimes and catchment sizes are covered. Figure 4.6 shows the locations of the considered sub-catchments within the Rhine river basin. The runoff of the Upper Rhine at the gauge Maxau (referenced as Upper Rhine catchment) is dominated by the alpine part of the catchment. This explains its pronounced, single peak mountain snow (glacial-nival) regime with maximum in summer and minima in late autumn and winter. The catchments of the rivers Moselle and Lahn have a rainfall dominated runoff (pluvial) regime with maximum in winter and minimum in late summer. Catchment area as well as mean and maximum runoff are listed in the upper part of Table 4.3. Catchment area decreases in the following order: Upper Rhine $>$ Moselle $>$ Lahn. Water level measurements from 1 November 1998 to 10 January 2013, which are converted into runoff by means of rating curves, serve as observations. The forecast data are discussed in Section 4.3.3.

Table 4.3: Features of the considered catchments (top)[a] and the meteorological input models (bottom)[b]. Table taken from Hemri et al. (2015).

gauge	catchment	area [km^2]	MQ [m^3/s]	HQ [m^3/s]
Maxau (MAXA)	Upper Rhine	50196	1247	4293
Trier (TRIE)	Moselle	23857	322	2880
Kalkofen (KALK)	Lahn	5304	48	598

name	# models	lead times	spatial resolution $\sim$
COSMO-LEPS	16	1-132 h	10 km
DWD-GME	1 (deterministic)	1-174 h	20 km
DWD-MER	1 (deterministic)	1-78 h (174 h)	1-7 km (20 km)
ECMWF-HRES	1 (deterministic)	1-240h	16 km

[a] MQ (mean discharge) and HQ (maximum discharge) are calculated over the period from 01.11.1998 to 31.10.2011.

[b] DWD-MER stands for a model run based on COSMO-EU forcing up to lead time 78 h and based on DWD-GME thereafter (corresponding forecast horizon and resolution are reported within parentheses). Note also that the forecast horizon of the hydrologic forecasts based on COSMO-LEPS is only 114 h, though the meteorological model forecasts up to 132 h.

4.3.3 Methods

Raw ensemble forecasts

A detailed description of the raw ensemble forecasts has already been given in Section 4.2.3. As summarized in the lower part of Table 4.3 ECMWF-HRES, i.e. the deterministic high resolution run of the ECMWF ensemble (Molteni et al., 1996), is added to the set of meteorological input models used for the study on censored EMOS in Section 4.2. This leads to a 19 member hydrological raw ensemble forecast, of which COSMO-LEPS provides the only exchangeable group. This can also be seen from Figures 4.7 a) and b). As in Section 4.2 we consider only lead times up to 114 h for the following analysis.

Climatological forecasts

The climatological forecasts are constructed similarly to those for the study on censored EMOS, which are described in Section 4.2.3. However, now the construction of the climatological forecasts is based on an hourly time series of observed runoff from 1 November 1998 to 10 January 2013. As before, they are obtained by calculating the empirical distribution of the observations that lie within $\pm x$ days of the calendar date of the verification day, but not in the same year. But for this study we have explored different interval sizes, namely $x \in \{15, 30, 45\}$ days. This resulted in a selection of $x = 45$ for river Upper Rhine, $x = 30$ for river Moselle, and $x = 15$ for river Lahn as these values led to the best climatological forecasts in terms of CRPS (cf. Section 2.3.1 for details on the CRPS). The

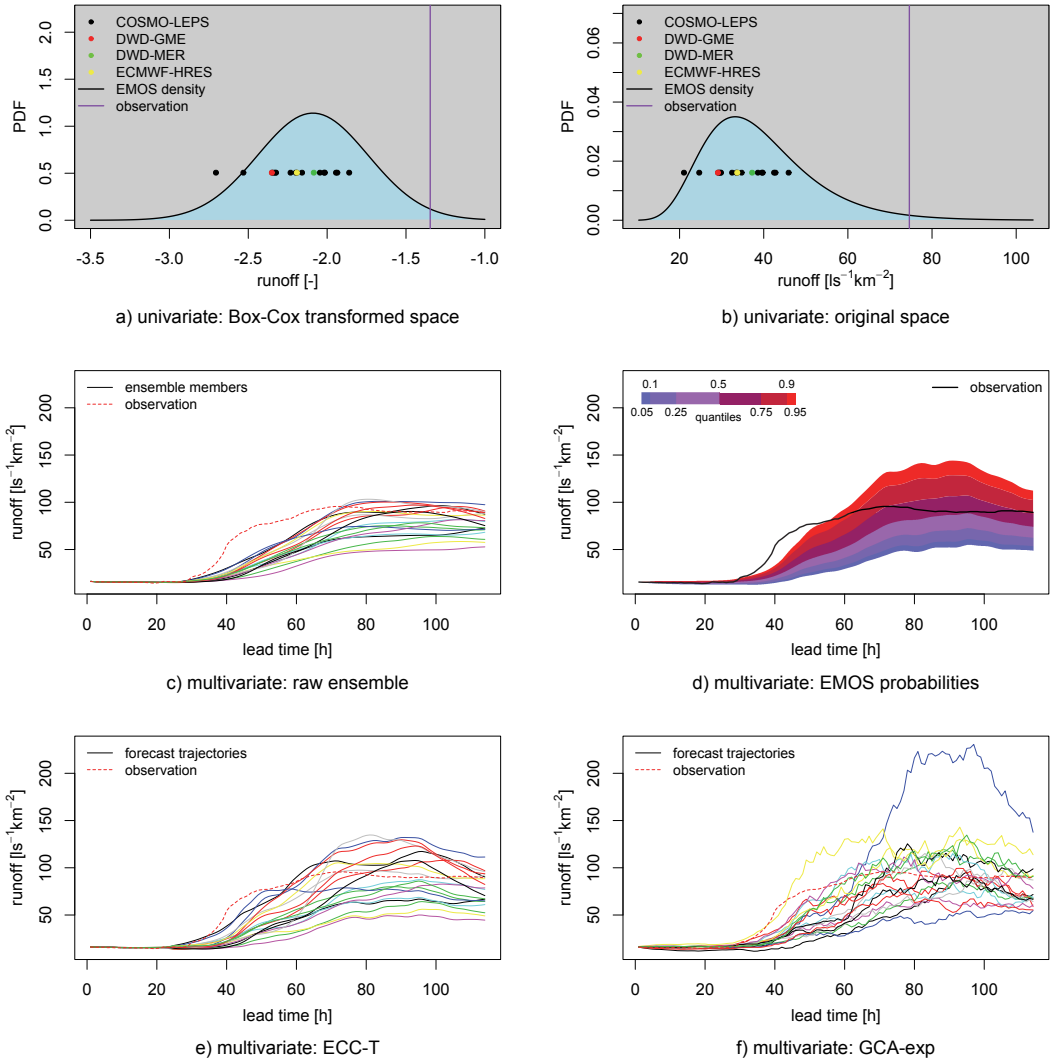


Figure 4.7: Example forecasts for river Moselle at Trier for a high flow event issued on 5 January 2011 at 06:00 UTC. Univariate raw ensemble and EMOS probability density forecasts with a lead time of 48 h are shown in subfigures a) on the Box-Cox transformed space and b) on the original space. The horizontal line of distinct dots represents the raw ensemble members, the vertical purple line shows the observed value. Subfigures c) to f) show the multivariate forecasts covering lead times 1 to 114 h. c) shows the trajectories of the raw ensemble, d) the quantiles of the EMOS forecast, and e) and f) the trajectories of the EMOS forecasts with correlation structure by ECC-T or GCA-exp, respectively. Figure taken from Hemri et al. (2015).

smaller the catchment, the narrower is the time frame to be used for calculating the climatological forecasts.

EMOS post processing

In the multi-model context of this study, EMOS converts the means of the runoff forecasts generated by the individual models (here: separate HBV-96 model runs with either COSMO-LEPS, DWD-GME, DWD-MER, or ECMWF-HRES as meteorological input) and the variance among all runoff forecast ensemble members generated by all models into a continuous predictive distribution. More specifically, EMOS predictive distributions are obtained by applying the truncated normal EMOS method (cf. Section 2.1.3) to the raw ensemble forecasts. In this approach, we apply a Box-Cox transformation prior to EMOS post processing. Details on how the Box-Cox transformation has been implemented in this study can be found in Appendix A.2. The effect of the Box-Cox transformation becomes clear from comparing the EMOS predictive densities of the same forecast on the transformed and on the original space as shown in Figures 4.7 a) and 4.7 b). As already stated, the EMOS predictive distributions are right-truncated in order to avoid - though very modest - positive probabilities for unrealistically high runoff. This limit b is set to two times the Box-Cox transform of the observations from 1 November 1998 to 31 October 2008. For the three catchments considered there is no need for a lower limit. In case of the river Moselle the estimated Box-Cox parameter $\hat{\lambda}$ is negative, which means that $-\infty$ on the Box-Cox transformed space maps to zero on the original space. $\hat{\lambda}$ is positive for the rivers Upper Rhine and Lahn, but the predictive probabilities for negative runoff are negligible. They are numerically zero for the vast majority of verification days and lead times. The highest probabilities attained are $3.7 \cdot 10^{-4}$ and $7.7 \cdot 10^{-87}$ for the rivers Upper Rhine and Lahn, respectively. Training and verification periods are selected in exactly the same way as in the study on censored EMOS (cf. Section 4.2 and Table 4.2). The parameters of the EMOS model are estimated by minimization of the CRPS over the training period.

Using formulae by Gneiting et al. (2004) the CRPS of the right truncated normal distribution verified at the observation $y \in (-\infty, b]$ can be written in closed form as

$$\mathrm{CRPS}\big[\mathcal{N}^b(\mu, \sigma^2), y\big] = \frac{\sigma}{\beta}\Big[- \frac{\Phi(\sqrt{2}v)}{\sqrt{\pi}\beta} + 2\gamma\Phi(\gamma) + 2\varphi(\gamma) - \gamma\beta\Big], \tag{4.1}$$

where $v = (b - \mu)\sigma^{-1}$, $\gamma = (y - \mu)\sigma^{-1}$ and $\beta = \Phi(v)$. Here, Φ and φ denote CDF and probability density function (PDF) of a standard normal distribution, respectively.

Multivariate extensions

In Section 2.2 the Schaake Shuffle, ECC, and GCA have been presented, which are methods to re-introduce a dependence structure into a probabilistic forecast. ECC and GCA have been selected for the study at hand. Obviously, these multivariate extensions are independent from the EMOS approach. They could, for instance, also be applied to BMA post processed probabilistic forecasts. The EMOS method presented so far, fits an independent univariate model for each forecast lead time. This approach may lead to unrealistic jumps in the marginal distributions from lead time to lead time and does not account for the correlation structure among consecutive lead times. Resolving these problems involves two steps. Firstly, jumps in the marginal distributions between individual lead times are removed by smoothing the EMOS parameters among the range of lead times. In this study, this is done by fitting a cubic smoothing spline. The smoothing parameter is estimated by leave-one-out cross-validation. This approach is implemented in the `R` function `smooth.spline`. For the rest of this study, the term EMOS predictive distribution refers to density forecasts based on the smoothed EMOS parameters. Secondly, the multivariate correlation structures are inserted by using either ECC that preserves the correlation structure of the raw ensemble forecasts or GCA that relies on the correlation structure of the training observations. In the following, we describe how ECC-T and GCA are adapted to the hydrological settings of this study.

For ECC-T, a right truncated normal distribution with mean μ_l, variance σ_l^2, and upper threshold b (cf. Equation (2.17) in Section 2.1.3) is fitted to the raw ensemble forecast $\boldsymbol{r}_l$ at each lead time l using maximum-likelihood estimation. This truncated normal distribution corresponds to the distribution $\hat{S}_l$ in Equation (2.23) in Section 2.2.2. In order to avoid unrealistically extreme quantiles in cases of very low raw ensemble spread, the variance of $\hat{S}_l$ is set to

$$\max\left\{\sigma_l^2, \left[h((1+d)\bar{r}^l) - h((1-d)\bar{r}^l)\right]^2\right\}, \tag{4.2}$$

where $\bar{r}^l$ is the mean of the raw ensemble at lead time l, h denotes the Box-Cox transformation, and d is a tuning parameter. This heuristic approach ensures that the minimal variance is linked to the mean of the raw ensemble and applicable on the Box-Cox transformed space. After having compared different example forecast trajectories and verification scores the tuning parameter d was set to $d = 0.0005$.

As a parametric alternative to ECC, we apply also the GCA method presented in Section 2.2.3. The exponential, the Matérn, and the generalized Cauchy correlation models (see Schlather (1999) for a comprehensive review of correlation functions) are candidates for the data at hand. Figure 4.8 shows the empirical correlograms and the corresponding fitted correlation functions. The correlation parameters are estimated using the `R` package `geoR` (Diggle and Ribeiro Jr, 2007; Ribeiro Jr and Diggle, 2001). In principle, GCA allows to sample infinitely many

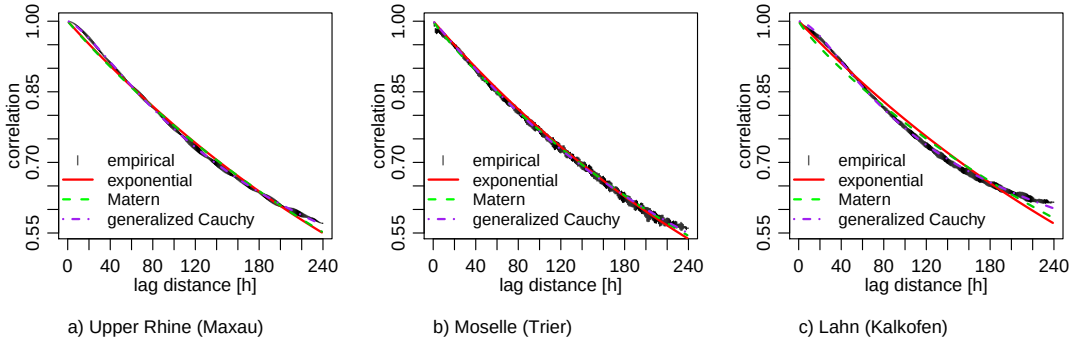


Figure 4.8: Empirical correlations and associated correlation function estimates against time lag averaged over all training periods. The black segments correspond to the range of the empirical correlation. Figure taken from the supplemental material to Hemri et al. (2015).

forecast runoff trajectories. In the study at hand the number of GCA trajectories is set to be equal to the size of the raw ensemble, i.e. $K = M$ in point iii) of the list in Section 2.2.3. Preliminary tests have shown that GCA does not depend much on the parametrization of the correlation function. For the rest of this study we consider only exponential GCA, which relies on an exponential correlation function, and hence is the simplest GCA model. Henceforth, the term GCA refers to exponential GCA.

Example forecasts

In order to illustrate EMOS, ECC, and GCA the hydrographs of an example prediction are discussed now. To this end the forecasts issued on 5 January 2011 for river Moselle have been selected, which cover a high flow event at a forecast lead time of about three days. Though the raw ensemble is able to predict the magnitude of the event, all members underestimate runoff during the rising limb of the hydrograph as shown in Figure 4.7 c). The EMOS probability forecasts shown in Figure 4.7 d) clearly improve the prediction compared to the raw ensemble. ECC-T yields quite realistic forecast trajectories with the same rank order structure as the raw ensemble. As demonstrated by the runoff trajectories in Figures 4.7 e) and 4.7 f), GCA is more flexible than ECC-T. On the one hand the 19 randomly selected quantiles, which are independent from the rank order structure of the raw ensemble, cover the observed trajectory better than the raw ensemble or ECC-T. On the other hand the forecast trajectories are a bit too wiggly. Additionally, there is a remarkably high outlier trajectory. Figures B.6 to B.22 in Appendix B.2 show similar plots for additional issue dates for all three considered catchments in low and high flow conditions.

4.3.4 Results

In the following sub-sections, truncated EMOS, ECC, and GCA are verified in detail. Refer to Section 2.3 for details on the verification methods. The considered runoff forecasts and observations are standardized by catchment size, i.e. the corresponding unit is $[\mathrm{ls}^{-1}\mathrm{km}^{-2}]$.

Univariate verification

The forecasts are now verified over the entire verification period and lead times 1 to 114 h. As a sound assessment of multivariate forecast properties relies on univariately well calibrated forecasts, we start with univariate verification of the predictive distributions for each individual lead time. Figure 4.9 a) shows the CRPSS values for the raw ensemble and EMOS forecasts with the daily climatological forecasts as reference. Skill in terms of CRPSS is much improved by post processing in case of all three catchments. The gain in skill by the raw ensemble forecasts over the climatological forecasts decreases with decreasing catchment size and increasing lead time. However, the EMOS forecasts for the river Lahn exhibit equal performance in terms of CRPSS as the EMOS forecasts for the substantially larger catchment of river Moselle.

After having discussed general prediction skill in terms of CRPSS, let us now have a closer look at calibration and sharpness. For all three catchments the raw ensemble forecasts are highly underdispersed as depicted by the 3D PIT histograms in Figure 4.9 b). With increasing lead time underdispersion slightly decreases. Note also the time-of-day-dependent oscillation of raw ensemble calibration of the forecasts for river Rhine at Maxau. This oscillation most likely arises from the intraday operation of the Swiss lakes, which are not included in the hydrological model. EMOS post processing flattens the PIT histograms regardless of catchment size and lead time. Generally, EMOS leads to well calibrated forecasts as shown in Figure 4.9 c). However, the $[0.9, 1]$ quantile interval is still overrepresented. Differences in calibration of the post processed forecasts between the different catchments can hardly be detected.

Calibration is only meaningful together with sharpness. For the assessment of forecast sharpness, the empirical distribution of the widths of the centered 90 % prediction intervals is constructed from the entire verification period. Following Gneiting et al. (2007a) important quantiles of that distribution are plotted in a box plot like manner. Figure 4.10 reveals that sharpness is clearly deteriorated by EMOS in case of river Upper Rhine. For the rivers Moselle and Lahn EMOS deteriorates sharpness at the short lead time of 24 h, whereas the effect of EMOS on sharpness for higher lead times is less pronounced. However, EMOS turns the very poor coverage of the raw ensemble forecasts into almost perfect coverage.

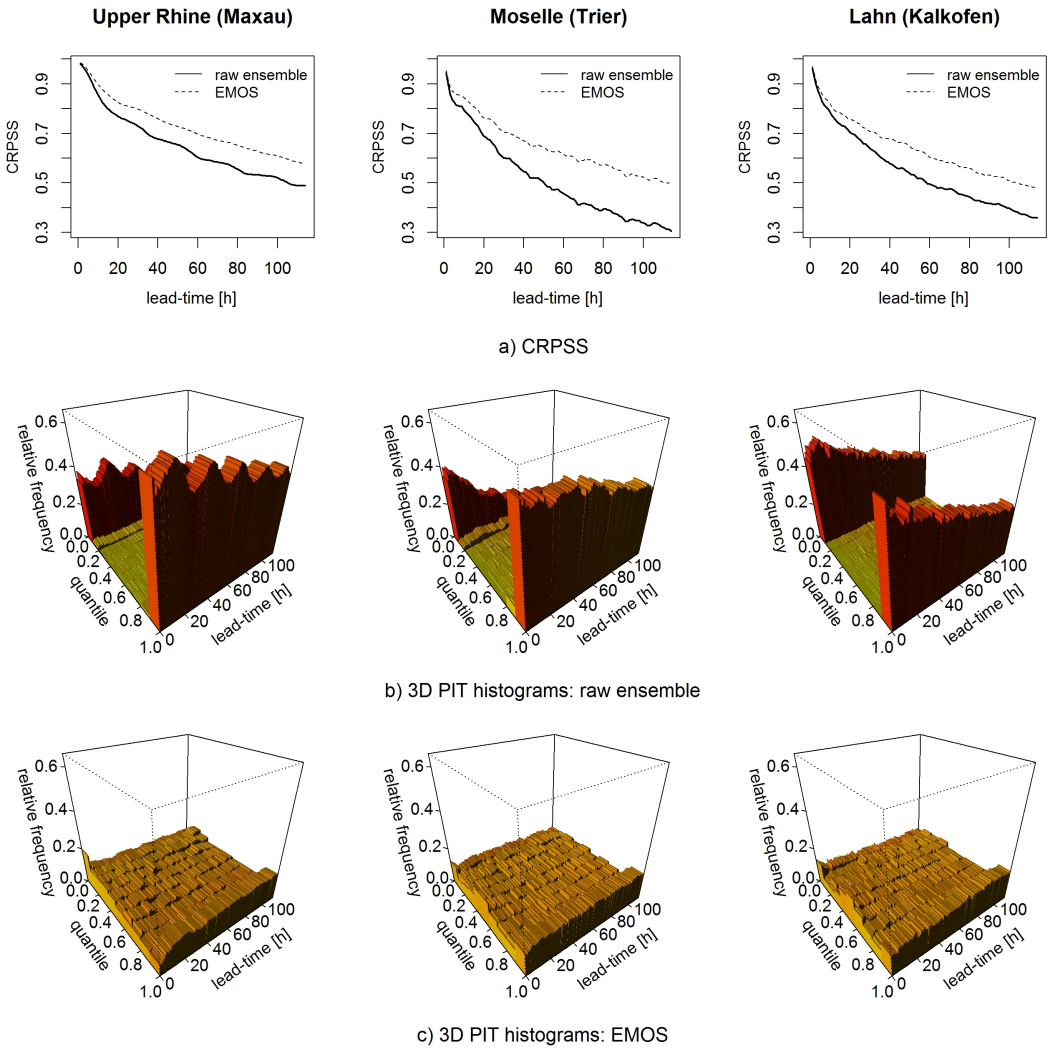


Figure 4.9: a) CRPSS against lead time of the raw ensemble and the EMOS forecasts. The daily climatological forecasts serve as reference model. b) 3D PIT histograms of the raw ensemble forecasts for the rivers Upper Rhine, Moselle, and Lahn. c) 3D PIT histograms of the truncated EMOS forecasts for the corresponding catchments. Figure taken from Hemri et al. (2015).

Multivariate verification

Even though the GCA forecasts look a bit less realistic, they perform slightly better than ECC-T in terms of multivariate statistical verification. But note that the differences in verification results between EMOS with either GCA or ECC-T are minor, compared to the differences to the raw ensemble. The average rank histograms shown in Figure 4.11 indicate that ECC-T lacks in multivariate calibration. The ∪-shaped histograms for all three catchments indicate either a too low correlation among lead times or forecast trajectories that are marginally underdispersive, in that the predictive densities for the individual lead times are too

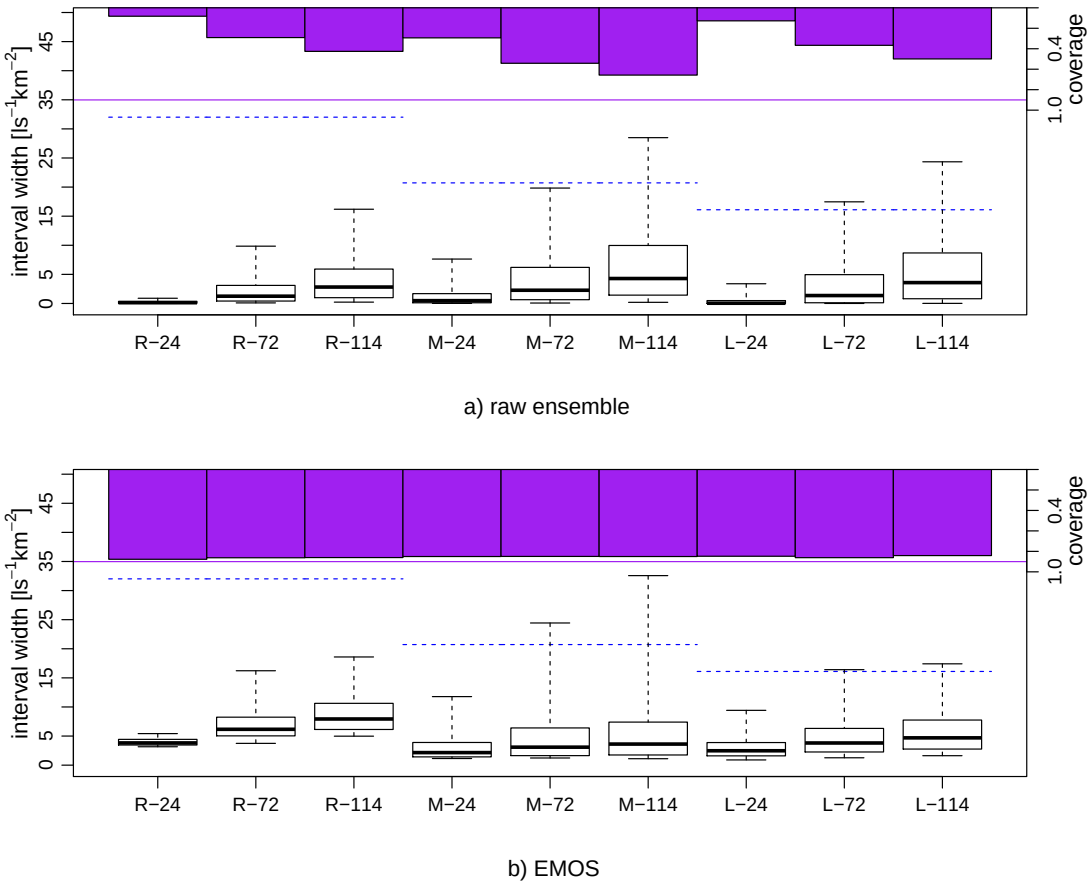


Figure 4.10: Sharpness and coverage of the raw ensemble and EMOS forecasts at lead times 24, 72, and 114 h. The blue dotted line segments show the medians of the empirical distribution of the 90 % prediction interval widths of the daily climatological forecasts. The purple lines correspond to the nominal coverage. The box plots show the 5, 25, 50, 75, and 95 % quantiles of the widths of centered 90 % prediction intervals from a) the raw ensemble and b) the EMOS forecasts for the rivers Upper Rhine (indicated by "R-xx", where xx denotes lead time), Moselle ("M-xx"), and Lahn ("L-xx"). Figure taken from Hemri et al. (2015).

narrow. The correlations of GCA are too strong as can be seen from the rather ∩-shaped histograms. In order to highlight the effects of a misspecified correlation structure, a forecast ensemble consisting of 19 runoff trajectories drawn from the marginal EMOS distributions for the individual lead times with zero correlation between lead times (INDEP) is evaluated as well in the following. According to Table 4.4, INDEP performs quite well in terms of the ES, but very poor in terms of the CRPS of the sum, minimum and maximum functionals, and the p-variogram score. The just mentioned CRPS analysis of forecast functionals refers to evaluating the CRPS of the forecasts for the sum, the minimum, and the maximum of the runoff trajectory over the entire forecast horizon. This gives insights into important multivariate properties of the forecasts. Note that for all

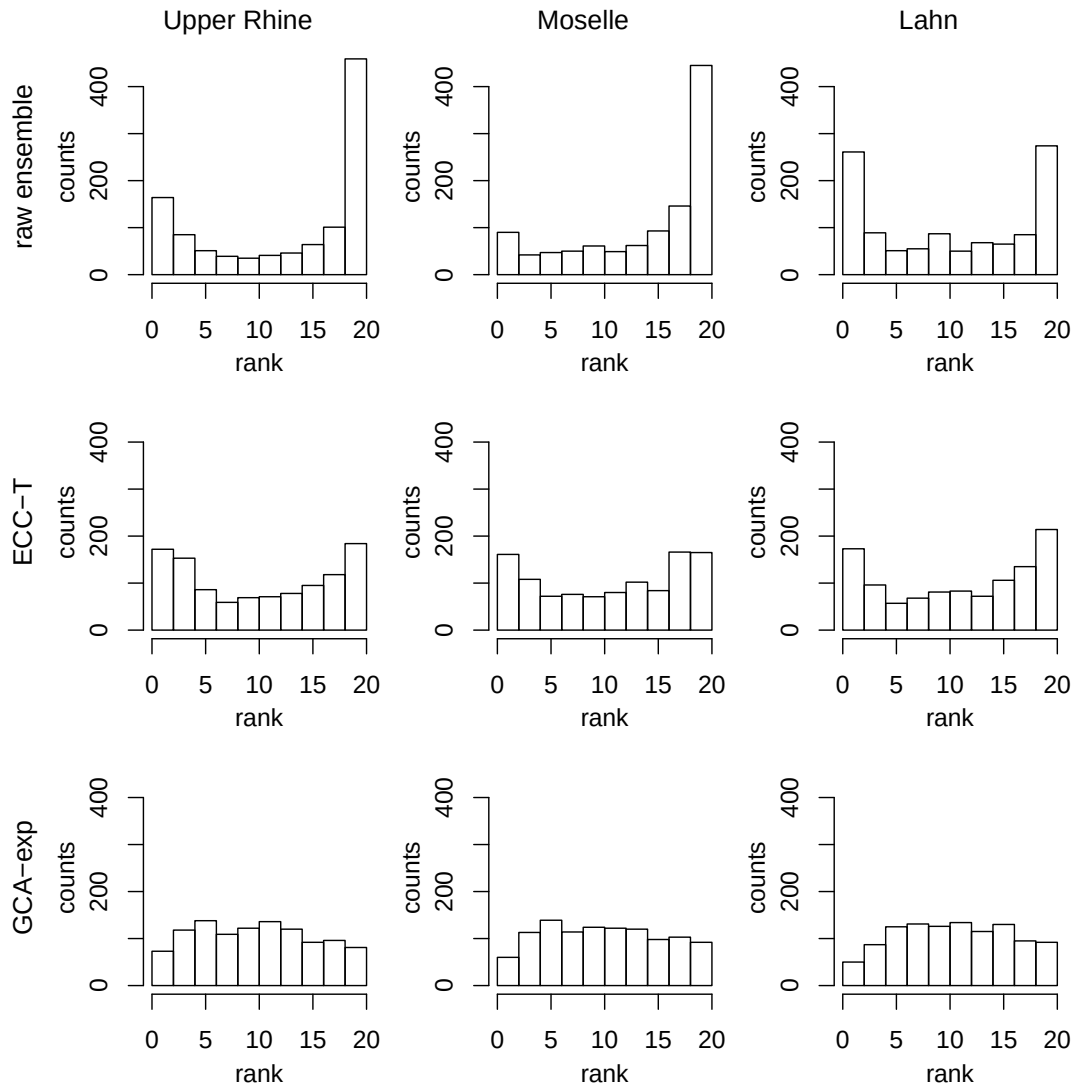


Figure 4.11: Average rank histograms comparing raw ensemble, ECC-T, and GCA with exponential covariance structure forecasts for the rivers Upper Rhine, Moselle, and Lahn. Figure taken from Hemri et al. (2015).

different forecasts the ES has been calculated based on the forecast vector of the lead times 24, 48, 72, and 96 h only in order to avoid issues of dimensionality. ECC and GCA perform better than INDEP in any combination of verification score and catchment, but the differences are very low in case of the ES. Furthermore, the p-variogram score indicates that ECC-T outperforms GCA in terms of correlation structure in case of the Upper Rhine, while GCA outperforms ECC-T for the rivers Moselle and the Lahn. The CRPS values of the minimum functional favor GCA over ECC-T for all catchments. The sum functional tends to favor GCA as well. However, ECC-T outperforms GCA in terms of the CRPS of the maximum functional in case of the large catchments Upper Rhine and Moselle.

Table 4.4: ES, CRPS and p-variogram scores[a]

		ES	CRPS_SUM	CRPS_MIN	CRPS_MAX	p-vario_0.5
Up. Rhine	raw ensemble	3.75	156	1.02	2.01	36.0
	INDEP	3.23	155	3.08	3.70	144
	ECC-T	3.18	126	0.94	**1.49**	**28.9**
	GCA-exp	**3.13**	**124**	**0.86**	1.52	29.1
Moselle	raw ensemble	3.60	147	0.94	2.24	44.1
	INDEP	3.11	144	1.48	4.44	116
	ECC-T	**3.07**	**119**	0.69	**1.97**	41.1
	GCA-exp	**3.07**	**119**	**0.62**	2.02	**40.2**
Lahn	raw ensemble	3.57	146	0.98	2.19	49.7
	INDEP	2.78	131	1.23	4.00	107
	ECC-T	2.77	109	0.70	1.57	37.5
	GCA-exp	**2.71**	**107**	**0.63**	**1.53**	**36.3**

[a] ES, CRPS of the sum, minimum, and maximum functionals as well as the 0.5-variogram scores comparing raw ensemble and EMOS forecasts with independent, ECC-T, and GCA-exp correlation structure for the rivers Upper Rhine, Moselle, and Lahn. Table taken from Hemri et al. (2015).

4.3.5 Discussion

The results confirm that univariate post processing using EMOS improves skill of the probabilistic runoff forecasts over the entire range of lead times. In particular, univariate calibration is greatly improved. However, the main focus of this study was on multivariate calibration. Our results demonstrate that temporal dependence structures can mostly be represented adequately by either ECC-T or GCA. On average, GCA performs slightly better than ECC-T in terms of statistical verification measures. However, this is expected, because GCA retains the univariate predictive distributions, while the ECC-T trajectories depend on the raw ensemble. For instance, the ECC-T spread is zero if the ensemble spread is zero even in cases where the variance of the marginal EMOS predictive distribution for the particular lead time is large. Nevertheless, in combination with the potential to model spatio-temporal dependencies between sub-catchments and lead times, EMOS with ECC-T is a suitable approach for post processing of sub-catchment ensemble forecasts. The post processed sub-catchment trajectories can then be used as boundary conditions and lateral inflows for a hydrodynamic model. In the present case, this would lead to well specified forecast scenarios of runoff, and hence also water levels, in the river Rhine. Such multivariate, probabilistic forecasts may, for instance, be useful for shipping companies. Furthermore, the results suggest that the relative performance of ECC-T compared to GCA deteriorates with decreasing catchment size. This is in line with the results by Pappenberger et al. (2010) who showed that the performance of ensemble river discharge forecasts based on similar settings of coupled atmospheric and hydrologic ensemble models decreases with decreasing catchment size. Hence, it is reasonable to assume that quality of the correlation structure of the raw ensemble

is highest for the large catchment of the Upper Rhine, moderate for the medium-sized catchment of the river Moselle, and lowest for the small catchment of the river Lahn. Further analyses are needed in order to confirm these results.

The EMOS models have been optimized on the Box-Cox transformed space. This approach has been chosen in order to be able to use Gaussian distributions. Nevertheless, one has to keep in mind that the predictive distributions have to be back-transformed. Hence, distances between equidistant quantiles on the Box-Cox transformed space are transformed to quantiles with increasing distances with increasing runoff volume on the original space. This in turn influences CRPS optimization, i.e. on the Box-Cox transformed space the lower parts of the predictive distributions have more influence on CRPS calculation than on the original space. Considering the, though slight, miscalibration of the EMOS models in the $[0.9, 1.0]$ decile, an optimization procedure that gives more weight to the higher quantiles may be desirable. A first test has shown that refining the parameter estimates on the original space may slightly increase verification scores. However, numerical CRPS optimization drastically increases computational cost. Another way to approach this problem would be to apply EMOS methods that are based on positively skewed non-Gaussian distributions. Promising approaches might rely, for instance, on generalized extreme value distributions (Scheuerer, 2014; Lerch and Thorarinsdottir, 2013).

In summary, this study confirms that EMOS along with the multivariate extensions, ECC-T and GCA, provides reasonably sharp probabilistic runoff forecasts that are well calibrated in terms of univariate calibration, and from which realistic runoff scenarios over the entire range of lead times can be extracted in a straightforward manner.

4.4 Hydrological regime dependent post processing

4.4.1 Introduction

Runoff pattern vary significantly over time due to changing hydrometeorological regimes. The EMOS post processing methods presented so far, are based on seasonal training periods. While such an approach can take account of seasonal variations, it completely neglects runoff regime. But using the same estimated EMOS parameters for low and high runoff or even flash floods may not be appropriate. Hence, forecast skill may benefit from selecting training data that are "similar" to the forecast runoff trajectory for a particular verification day. Several regime dependent approaches have already been developed in a meteorological context. Gneiting et al. (2006) developed a regime-switching forecast method designed for a wind farm in the U.S. Pacific Northwest, where it turned

out to be crucial to differentiate between easterly and westerly wind regimes. A recent approach by Junk et al. (2015) uses analog past ensemble forecasts (including the corresponding observations) as training data in EMOS post processing. The analog training data set consists of those forecasts, which are most similar to the current forecast. Lerch and Baran (2016) proposes a semi-local EMOS approach that clusters stations with similar features and then uses training data from all stations within a cluster class for model fitting. A feature could, for instance, be the station climatology or the CDF of forecast errors. In the following, we asses the potential of such a similarity based training data selection in the framework of hydrological forecasting.

4.4.2 Methods

In case of the strongly autocorrelated runoff forecast trajectories, it is crucial to select a similarity criterion that is designed to detect differences between time series. Among the different similarity criteria for hydrological trajectories listed by Ehret and Zehe (2011) dynamic time warping (DTW: Sakoe and Chiba (1978)) is chosen for this preliminary study, because it has already been implemented successfully in a hydrological context (Ouyang et al., 2010) and the corresponding algorithm is readily available in the `R` package `tsdist` (Mori et al., 2014). DTW has originally been developed for speech recognition. In order to compare word sequences spoken at different paces, it allows for stretching and compression of time series. That is, DTW considers time series to be equal if it is possible to map one into the other by stretching and compression only. More specifically, the DTW distance measure between two time series corresponds to the minimal amplitude error that is obtainable through stretching and compression. Translated to hydrology, DTW can be applied successfully in cases where the shape of the hydrograph is important, but not the actual timing (Ehret and Zehe, 2011). The following mathematical description of DTW follows closely Rabiner and Juang (1993) and Giorgino (2009). Assuming a training trajectory $\boldsymbol{x} = (x_1, \ldots, x_L)$ for lead times $1, \ldots, L$ and a verification trajectory $\boldsymbol{y} = (y_1, \ldots, y_L)$ one first computes the local dissimilarity which is given by

$$d(i,j) = f(x_i, y_j) \geq 0 \quad \text{with} \quad i,j = 1, \ldots, L, \tag{4.3}$$

where f is most commonly the Euclidian distance. DTW relies on a warping curve $\phi(k), k = 1, \ldots, K$ consisting of remapped pairs of indices from $\boldsymbol{x}$ and $\boldsymbol{y}$. Note that K denotes also the length of the remapped time series, where typically $K \neq L$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are of the same length, $\phi(k)$ is given by

$$\phi(k) = \big(\phi_x(k), \phi_y(k)\big), \tag{4.4}$$

where the functions $\phi_x(k), \phi_y(k) \in 1 \ldots L$ actually select the indices at each k. In order to avoid temporal inconsistencies, such as loops, copies of the same flood peak, and reversed peak orders, the following monotonicity constraints are introduced:

$$\phi_x(k+1) \geq \phi_x(k),$$
$$\phi_y(k+1) \geq \phi_y(k). \tag{4.5}$$

Accordingly, the warped distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be represented as

$$d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{M_\phi} \sum_{k=1}^{K} d(\phi_x(k), \phi_y(k)) m_\phi(k), \tag{4.6}$$

where m_ϕ is a weighting function and M_ϕ is a normalizing constant that ensures comparability between different paths. The DTW approach minimizes now the warped distance over all paths ϕ, that is the DTW distance is given by

$$D(\boldsymbol{x}, \boldsymbol{y}) = \min_{\phi} d_\phi(\boldsymbol{x}, \boldsymbol{y}). \tag{4.7}$$

The DTW approach is illustrated in Figure 4.12 that shows how a forecast trajectory and a corresponding training trajectory are mapped. The DTW distance would then be obtained from the difference of the mapped trajectories shown in Figure 4.12 e). Hence, hydrological regime dependent post processing can be summarized as follows:

1. Select a representative forecast trajectory from the raw ensemble to be post-processed. In the small case study presented in the following, this trajectory is obtained from the HBV-96 forecasts with ECMWF-HRES meteorological inputs covering lead times 1 to 114 h.

2. Calculate the selected distance measure, e.g. DTW distance, between the forecast trajectory and all eligible training trajectories. Typically, the set of eligible training trajectories consists of all past forecasts that stem from the same model as the forecast trajectory and do not overlap with the latter trajectory.

3. Select the T training trajectories that are most similar to the forecast trajectory, i.e. that have lowest distance measures. And use the corresponding pairs of observations and raw ensemble trajectories for estimation of the statistical post processing model.

4.4.3 Results

Here, the results from a small case study on EMOS with DTW based training periods are presented. Unless specified differently, the same data and EMOS post processing models are used as in the case study on multivariate post processing of hydrologic forecasts in Section 4.3. As mentioned above the deterministic forecasts from hydrological model runs with ECMWF-HRES meteorological inputs are used to determine the training periods. For each verification day, the $T \in \{45, 90, 180, 365\}$ dates in the dataset with the most similar ECMWF-HRES

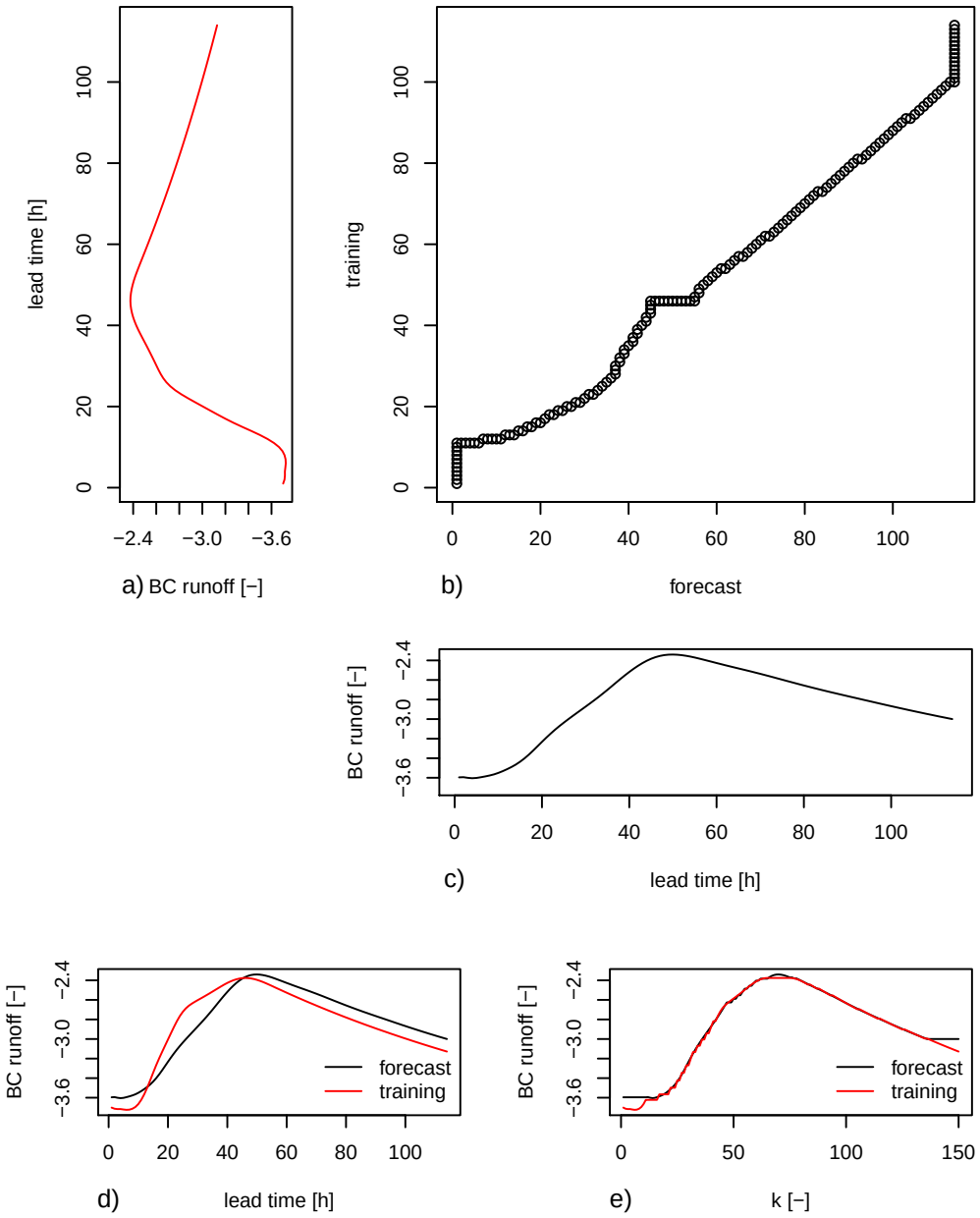


Figure 4.12: DTW illustration showing a training trajectory (a), a forecast trajectory (c), the corresponding optimal DTW path (b) as well as both training and forecast trajectory in one plot (d) and the corresponding DTW matched trajectories (e). The time series are computed from Box-Cox transformed runoff forecasts with ECMWF-HRES meteorological input for river Lahn at Kalkofen initialized on 23 January 2009 and 10 February 2009 for the training and the forecast trajectory, respectively.

forecast trajectories are used for training. In order to avoid overlapping, dates in the range of ±10 days from the verification day are not considered for train-

ing. For comparison with the reference forecast, i.e. the seasonally fitted EMOS forecasts (cf. Section 4.3), we use skill scores of the CRPS, i.e. the CRPSS and the log score for verification. The log score of the optimal forecast, i.e. S_V^o in Equation (2.31) in Section 2.3.1, would be $-\infty$. Therefore, S_V^o is arbitrarily set to -10, which corresponds to the log score of an almost optimal forecast. The length of the corresponding seasonal training periods ranges from 172 to 233 days with a median of 184 days. Hence, the seasonal approach can be compared best with the DTW approach that is based on the 180 most similar training forecasts, which is referred to as DTW EMOS 180 in the following. According to Figure 4.13 the performance of DTW EMOS 180 compared to seasonal EMOS depends on the catchment of interest. In case of the rivers Upper Rhine and Lahn DTW EMOS 180 improves forecast skill only for the first few lead times. Beyond a lead time of about 20 h the best DTW EMOS 180 variant and seasonal EMOS perform equally well in case of the Upper Rhine, while for river Lahn DTW EMOS 180 is clearly outperformed by seasonal EMOS. Only in the case of river Moselle, DTW EMOS 180 leads to an improvement in forecast skill over the entire forecast horizon. On average, DTW EMOS 45 and DTW EMOS 90, i.e. considering the 45 or 90 most similar forecasts, lead to a deterioration in terms of forecast skill compared to DTW EMOS 180. Note that this is not the case for the lead times up to 70 hours at gauge Lahn. Considering the 365 most similar training forecasts does not change forecast skill much compared to DTW EMOS 180. Small improvements can be detected for river Upper Rhine and Lahn at higher lead times.

4.4.4 Discussion

According to the above results the combination of DTW with EMOS did not lead to clear-cut improvements in forecast skill compared to the seasonally fitted EMOS predictions. Only the forecasts for river Moselle clearly benefit from DTW. In case of the river Lahn DTW even leads to a deterioration. Considering the quite different catchment features – large snowmelt dominated, large precipitation dominated, and small precipitation dominated in case of the Upper Rhine, Moselle, and Lahn, respectively – there might be a connection between catchment type and the performance of the DTW EMOS methods. There is a need for follow-up studies based on a larger set of different catchments in order to either confirm or reject this hypothesis. Furthermore, the results indicate a low performance of the DTW EMOS variants with a small number of training forecasts compared to those with a rather large number of training forecasts. This leads to the question whether it is more important to remove the very unsimilar trajectories from the training set than having only the very similar training trajectories in the training set. Additionally, it would be useful to test also alternative distance measures that may be more suitable for post processing of hydrologic ensemble forecasts.

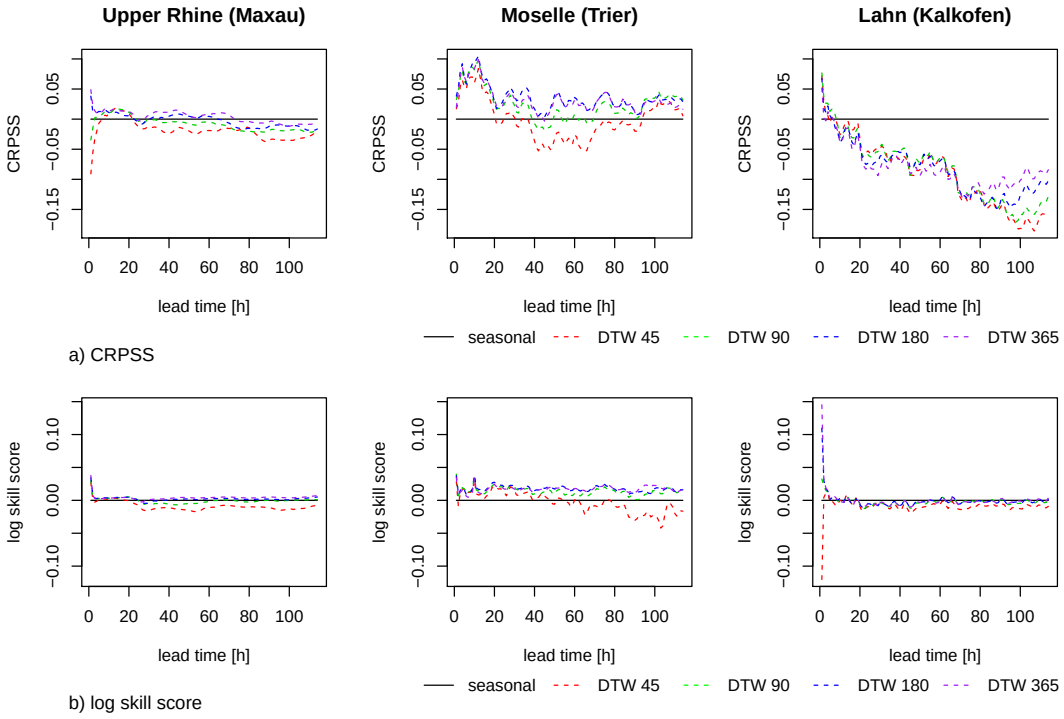


Figure 4.13: Verification scores comparing DTW EMOS with the seasonally fitted EMOS model from Section 4.3. Subfigure a) shows the CRPSS and b) the log skill score over the verification period. The seasonally fitted EMOS forecasts are the reference model. DTW EMOS is assessed depending on the length of the training period, i.e. the number of most similar training forecasts that are considered.

4.5 Deterministic evaluation of probabilistic hydrological forecasts

4.5.1 Introduction

Though probabilistic forecasts exhibit very desirable properties, there are many situations in which end-users call for a translation into deterministic forecasts (Gneiting, 2011), which are also referred to as point forecasts. For instance, the BfG seeks a deterministic water level forecast with a precision of ±10 cm in 80 % of the verification instances up to two days for river Rhine. Forecasts up to 4 days should have a precision of ±20 cm in 80 % of the cases (Meißner and Rademacher, 2010). Obviously, any deterministic hydrologic forecast model could be used to generate such a forecast if it is skillful enough. However, it is much more recommendable to make use of the information inherent to the probabilistic forecasts. In the following, we present a preliminary study on how to convert probabilistic hydrologic forecasts into deterministic forecasts.

4.5.2 Methods

In order to find a sound directive for issuing deterministic forecasts based on the EMOS predictive distributions, one needs to find a functional T for which a consistent scoring function exists that is related to the water level precision requirements of the BfG. Following Gneiting (2011) a functional T is a mapping from a class F to the real line, $F \mapsto T(F) \subseteq \mathbb{R}$. Like the proper scoring rules presented in Section 2.3.1, the scoring function $S(x, y)$ is a function of the forecast x and the event y that materializes. Furthermore, the scoring function $S(x, y)$ can as well be understood as the reward of the forecaster. Similarly to propriety (cf. Equation (2.25) in Section 2.3.1), a scoring function $S(x, Y)$ is consistent for the functional T if

$$\mathbb{E}_F[S(t, Y)] \leq \mathbb{E}_F[S(x, Y)] \tag{4.8}$$

for all F, all $t \in T(F)$ and all $x \in \mathbb{R}$ Gneiting (2011) and with $Y \sim F$. The scoring function $S(x, Y)$ is strictly consistent if

$$\mathbb{E}_F[S(t, Y)] < \mathbb{E}_F[S(x, Y)] \tag{4.9}$$

holds for all $t \in T(F)$ and all $x \notin T(F)$.

The above requirements by the BfG can be considered as an alternative, i.e. inverse, formulation of the zero-one scoring function that is given by

$$S_c(x, y) = \mathbb{1}_{[|x-y|>c]}, \tag{4.10}$$

where $c > 0$ (Gneiting, 2011). The deterministic forecast x can be understood as a functional of a predictive distribution F on the real line. According to Gneiting (2011), the respective midpoint is the optimal point forecast, namely

$$\hat{x} = \underset{x}{\operatorname{argmax}}(F(x + c) - \lim_{y \uparrow x-c} F(y)), \tag{4.11}$$

so that $\hat{x}$ is the midpoint of the interval of length $2c$ with maximal probability mass. In case of continuous predictive distributions like the truncated normal EMOS model, on which the post processing of the hydrologic ensemble forecasts at hand is based, the midpoint simplifies to

$$\hat{x} = \underset{x}{\operatorname{argmax}}(F(x + c) - F(x - c)). \tag{4.12}$$

According to Gneiting (2011), the zero-one functional S_c is consistent for the midpoint functional. Hence, point forecasts can be obtained from post processed predictive distributions obtained by methods like EMOS or BMA using the approach of Equation (4.12). The midpoint of the forecast distribution corresponds to the value at which the most mass of the density function lies within an interval of $\pm$ 10 cm or $\pm$ 20 cm, respectively. Figure 4.14 shows the forecast density of an 48 hour EMOS forecast for river Moselle at gauge Trier initialized on 11

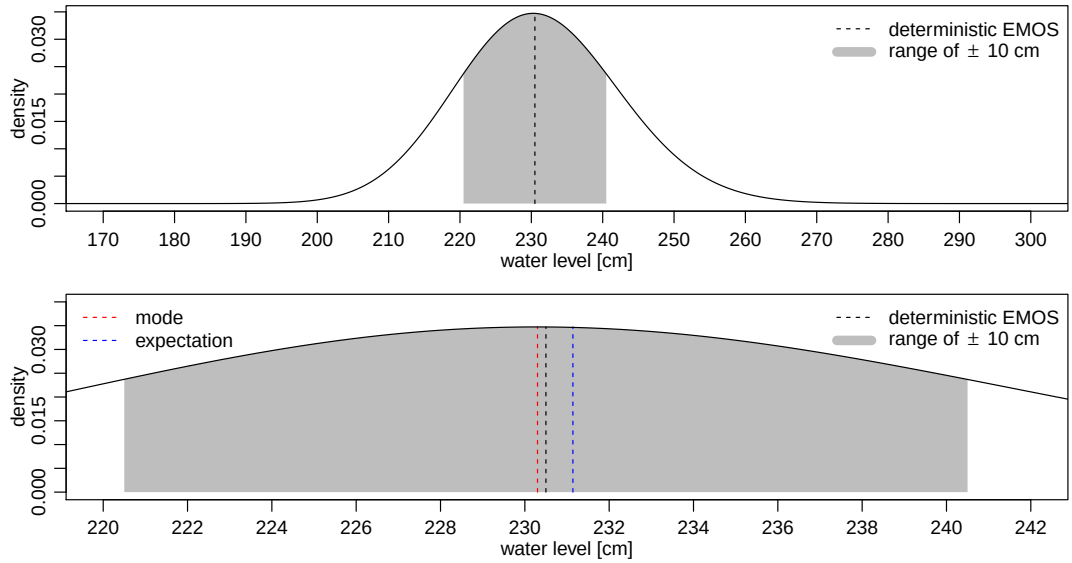


Figure 4.14: Forecast EMOS density for a 48 h forecast including the DEI forecast with its associated interval. The lower figure is a magnification of the upper one that shows the differences between DEI forecast, mode, and expectation.

September 2009 at 06:00 UTC. The midpoint of the EMOS forecast density is referred to as deterministic EMOS interval (DEI) forecast in the following. Note that it differs from both the mode and the expectation of the forecast distribution.

In the hydrological studies presented in this thesis, post processing has been applied only to runoff forecasts, but not to the corresponding water level predictions. However, as mentioned above, the deterministic forecast verification method applied by the BfG is performed on gauge levels. The water level observations that have been used to generate the hydrological ensemble forecasts used in this study are not identical with the water level observations available to us. While the former are unmodified operational measurements, the latter have undergone additional checks. For the sake of consistency, we rely on the runoff dataset and transform runoff to water level using functional rating curves. Likewise, in order to find the midpoint of the forecast distribution (cf. Equation (4.12)), water level intervals have to be back-transformed to runoff intervals. As stated in Section 2.1.3, the EMOS post processing methods used here are estimated on the Box-Cox transformed space. Hence, the runoff intervals additionally need to be Box-Cox transformed. Assuming that an EMOS forecast distribution is already available on the Box-Cox transformed space, the procedure to obtain the ±10 cm (±20 cm) DEI forecast can be summarized as follows:

1. Map the forecast CDF to the original space, i.e. runoff in m^3/s, by inverse Box-Cox transformation.

2. Map the forecast CDF from runoff to water level using fitted rating curves.

3. Find the midpoint $\hat{x}$, i.e. the center of the ± 10 cm (± 20 cm) water level interval that covers the most probability mass.

Details on the actual rating curve fitting and the effects of the rating curve and the Box-Cox transformations are discussed in Appendix A.2.2.

4.5.3 Results

The DEI forecasts are evaluated using the 10 and 20 cm criteria from above for the uncensored catchments Upper Rhine, Moselle, and Lahn. As shown in Figure 4.15, they are compared with the deterministic forecasts from DWD-GME, DWD-MER, ECMWF-HRES, and the mean of the COSMO-LEPS members. Obviously, none of the different forecasts can meet the 10 or the 20 cm criterion at forecast lags of two and four days, respectively. In order to assess the usefulness of the density mass maximization approach of Equation (4.12), the expected values of the EMOS predictive distributions are included as well. These predictions are referred to as deterministic EMOS expectation (DEE) forecasts. For the rivers Moselle and Lahn the DEI forecasts outperform all other prediction methods. Surprisingly, in case of the Upper Rhine the mean of COSMO-LEPS, DWD-GME, and DWD-MER generally outperform the DEI forecasts at higher lead times. Looking at the 10 cm criterion, this is the case for all lead times beyond 25 h. With regard to the 20 cm criterion the mean of COSMO-LEPS and DWD-MER outperform the DEI forecasts at lead times greater than 60 h, while DWD-GME outperforms the DEI forecasts only beyond 80 h. Furthermore, there is not much difference between the DEI and the DEE forecasts for the Upper Rhine. Averaged over the entire forecast horizon, the values for DEI are 0.4 % and 0.3 % better than the values for DEE with regard to the 10 cm and 20 cm criterion, respectively. In case of river Moselle both DEI and DEE meet the 10 cm and 20 cm criteria considerably more often than any of the other forecasts at lead times greater 15 h. DEI outperforms DEE only very slightly, if at all. The averaged outperformance of DEI over DEE amounts to 1.2 % for the 10 cm criterion, and to 0.8 % for the 20 cm criterion. For river Lahn the improvement of DEI and DEE over the other forecasts is much smaller. But, in case of the 10 cm criterion, the averaged relative improvement of DEI over DEE amounts to 3.6 %. Looking at the 20 cm criterion, this improvement drops to 0.7 %.

4.5.4 Discussion

On average DEI and DEE lead to a gain in deterministic forecast skill compared to the raw ensemble deterministic forecasts. The small losses in case of Upper Rhine are more than compensated by the small gains in case of river Lahn and the quite significant gains for river Moselle. Furthermore, DEI slightly outperforms DEE for the three catchments considered. Nevertheless, it is quite difficult to draw sound conclusions from the above results. There is need for a more detailed analysis of the conversion of probabilistic hydrologic forecasts into deterministic

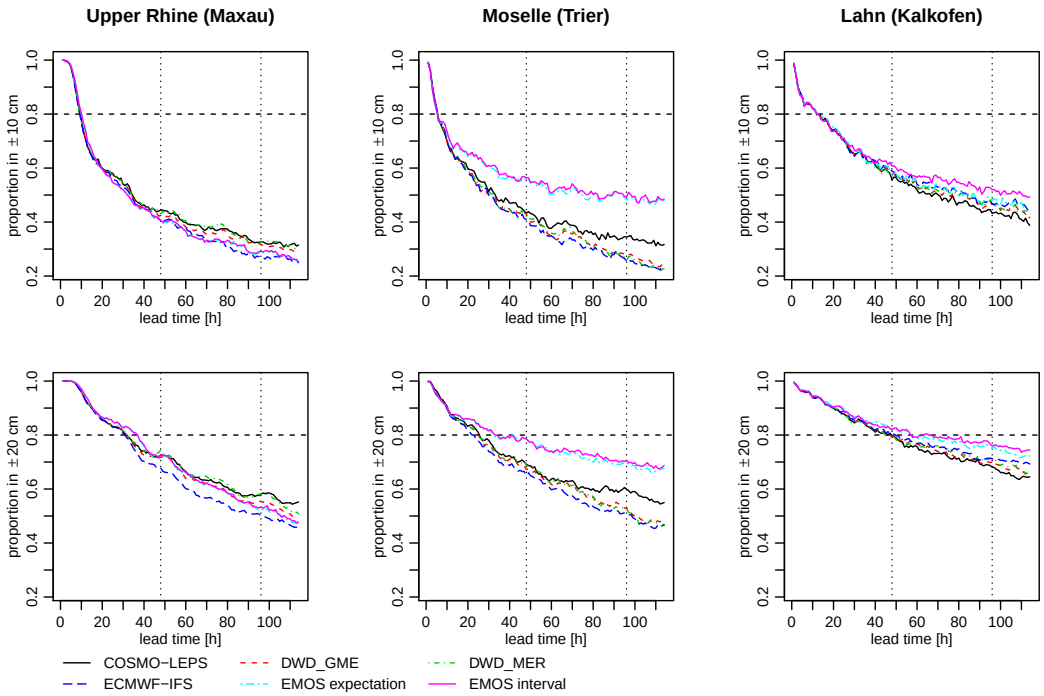


Figure 4.15: Proportion of deterministic forecasts that are not more than 10 cm (top) and 20 cm (bottom) off the verifying observation. The horizontal dashed lines indicate the 80 % target of the BfG, the vertical dotted lines are drawn at forecast lags of two and four days.

forecasts. Directions for theoretically well founded point forecasting can be found in Gneiting (2011).

4.6 Post processing of seasonal hydrological ensemble forecasts

4.6.1 Introduction

All the above case studies apply statistical post processing methods to short- to medium-range (up to 2 weeks) hydrometeorological forecasts. Though seasonal (one to several months) meteorological forecasts show only very limited skill, seasonal hydrological forecasting is somewhat more promising due to the long-term water balance memory of hydrological catchments (e.g. Hurst (1951) and Mudelsee (2007)). In this study, we rely on two ways to obtain seasonal hydrological ensemble forecasts. In the first approach, seasonal NWP forecast ensembles are used as input to the hydrological model. The second approach, is based on the past climatology of meteorological variables and summarized now. In a nutshell, the ensemble streamflow prediction (ESP: Day (1985); Wood and Lettenmaier (2008) approach can be divided into two steps. First, the hydrological model is

run using observed meteorological input variables up to the time of forecast initialization in order to obtain a sound estimate of the present model state. Then, the ESP ensemble is generated by running the hydrological model multiple times with meteorological inputs obtained by resampling from the seasonal climatologies of the meteorological input variables. The length of the resampled sequences should equal the length of the forecast horizon.

Though not many studies on post processing of seasonal hydrological forecasts have been performed up to now, Shi et al. (2008) compared the effects of statistical post processing and hydrological model calibration on forecast skill. They considered seasonal ensemble forecasts, based on a 30 member ensemble obtained by means of the EPS approach, ranging from 1 to 6 months for eight different catchments in the western U.S. Their results indicate that statistical post processing by percentile mapping (Panofsky and Brier, 1958; Wood et al., 2002) leads to seasonal forecast ensembles that perform almost equally well as those obtained by hydrological model calibration. In the following, we assess if statistical post processing adds skill to already hydrologically calibrated seasonal forecasts for the gauges Basel and Cologne (both river Rhine) as well as for the gauges Achleiten and Hofkirchen (both river Danube).

4.6.2 Data and methods

Study areas and runoff data

The sub-catchments considered cover different catchment characteristics. Their location within the catchments of the rivers Rhine and Danube is shown in Figure 4.16. As listed in Table 4.5 all catchment are rather large. In case of river Rhine, we focus on gauge Basel, which is mostly alpine dominated with a quite high average runoff relative to the catchment size and gauge Cologne that is located at the Lower Rhine and drains large parts of the river Rhine catchment. The runoff pattern at Cologne results from a mixture between the snow dominated Upper Rhine catchment and the rainfall dominated tributaries further downstream. In case of river Danube, gauge Hofkirchen drains large parts of the foothills of the Eastern Alps, which results in a mostly rainfall dominated runoff pattern with additional minor snowmelt contributions from the alpine sub-catchments. Gauge Achleiten is located just below the confluence of the Danube with river Inn. The alpine runoff pattern of river Inn with snowmelt induced peak runoff in spring/early summer strongly affects runoff of the river Danube at gauge Achleiten.

For model fitting and forecast verification two different types of monthly runoff values are used. The first approach relies on mean monthly observed runoff that is obtained from water level measurements and subsequent transformation to runoff using the corresponding rating curves. The second approach is based on mean monthly "observed" runoff that is obtained by running the hydrological models (see next paragraph for details on the hydrological models) with ECMWF

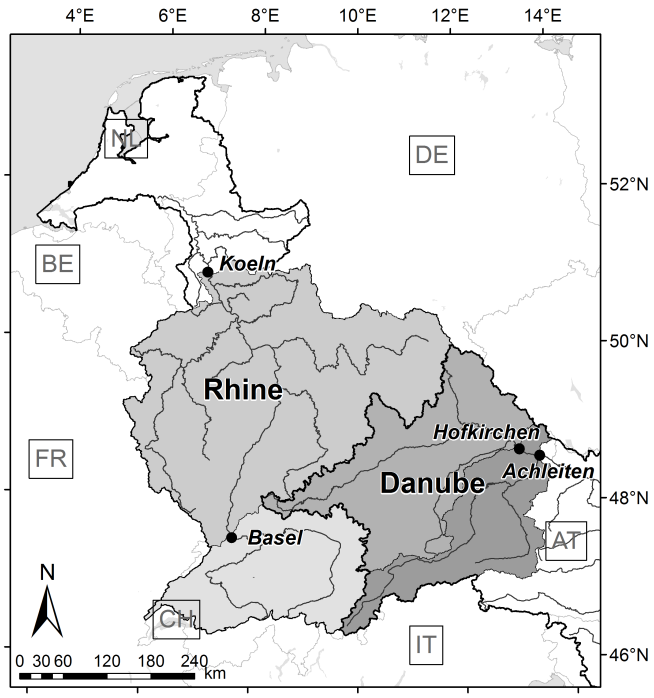


Figure 4.16: Locations of the considered sub-catchments within the basins of the rivers Rhine and Danube. Figure provided by Bastian Klein (BfG).

ERA-interim (Dee et al., 2011) analysis meteorological input. In case of the latter, uncertainties in the hydrological models and observation measurement errors can be excluded. Typically, post processing leads to considerably stronger improvements in the settings of measured observations, since many sources of systematic errors are eliminated by using simulated observations.

Seasonal hydrological raw ensemble forecasts

The seasonal hydrological raw ensemble is based on monthly hydrological hindcasts initialized at the beginning of each month from January 1981 to January 2013 covering a forecast horizon of 1 to 7 months. It consists of 48 members stemming from 15 runs, $\boldsymbol{r}_{\mathrm{EC}} = (r_{\mathrm{EC},1}, \ldots, r_{\mathrm{EC},15})$, of the hydrological model with atmospheric input from the ECMWF seasonal forecast system 4 (Molteni et al., 2011) and 33 members obtained using the ESP approach denoted by $\boldsymbol{r}_{\mathrm{ESP}} = (r_{\mathrm{ESP},1}, \ldots, r_{\mathrm{ESP},33})$. For the two gauges at river Rhine, Basel and Cologne, the seasonal hydrological forecasts are obtained by running the HBV model (cf. Section 4.2.3). For river Danube, i.e. gauges Achleiten and Hofkirchen, the spatially distributed hydrological rainfall runoff model COSERO (Continuous Semi-distributed Runoff: Nachtnebel et al. (1993)) is used, which is similar to the HBV model as, for instance, stated by Frey and Holzmann (2015).

Table 4.5: Features of the considered catchments: area, mean monthly runoff (mMQ), and maximum monthly runoff (mHQ) in the period from January 1981 to June 2013.

gauge	catchment	area [km^2]	mMQ [m^3/s]	mHQ [m^3/s]
Basel	Rhine	35897	1073	2713
Cologne	Rhine	144232	2210	5340
Hofkirchen	Danube	47609	650	1545
Achleiten	Danube	76660	1435	3407

Climatological forecasts

The climatological forecasts correspond to the empirical distribution of runoff values of the same month as the month of interest but from other years. Since we use only data from our training/verification period that covers about 30 years, the climatological forecasts are a quite rough estimate of the climatological forecast distribution and have to be interpreted as relative benchmarks that may easily be outperformed by a more elaborated climatological forecast.

EMOS post processing

Preliminary tests for the gauges Achleiten and Hofkirchen have shown that the more sophisticated BMA method is not able to outperform EMOS. Hence, we focus here on comparing different EMOS variants. The reference model is a univariate EMOS model based on either a truncated normal or a lognormal distribution. The alternative models are estimated jointly over all lead months 1 to 7. For a detailed discussion of truncated EMOS see Section 2.1.3. The estimation of the coefficients of the statistical post processing models for the forecasts initialized in month x is based on a training set that consists of all obervation/forecast pairs initialized in month x but not in the same year.

Lognormal EMOS

EMOS based on a lognormal distribution has been proposed by Baran and Lerch (2015) to post process wind speed forecasts. Here, we use it as an alternative to truncated EMOS that may be more suitable for the post processing of seasonal ensemble runoff forecasts. The EMOS predictive distribution based on a lognormal model can be written as

$$y \mid \boldsymbol{r} \sim \begin{cases} \frac{1}{y\sigma}\varphi\left(\frac{\log y - \mu}{\sigma}\right) & \text{if } y \geq 0, \\ 0 & \text{else,} \end{cases} \tag{4.13}$$

where φ denotes the probability density function of the standard normal distribution. Following Baran and Lerch (2015) the parameters μ and σ are linked to the raw ensemble by

$$\mu = \log\left(\frac{m^2}{\sqrt{v+m^2}}\right) \quad \text{and} \quad \sigma = \sqrt{\log\left(1+\frac{v}{m^2}\right)}, \tag{4.14}$$

where m and v are the mean and the variance of the predictive distribution, respectively. In this EMOS method m and v are affine functions of ensemble statistics. Here, this leads to the link

$$m = a_0 + a_1\bar{r}_{\mathrm{EC}} + a_2\bar{r}_{\mathrm{ESP}} \quad \text{and} \quad v = b_0 + b_1 s^2, \tag{4.15}$$

where $\bar{r}_{\mathrm{EC}}$ and $\bar{r}_{\mathrm{ESP}}$ denote the means of the ECMWF and the ESP seasonal ensemble forecast members, respectively and $a_0, a_1, a_2, b_0, b_1 \geq 0$.

Simultaneous parameter estimation

Since this study relies only on monthly average runoff values, the data available for training of the EMOS models consist of at most 33 forecast/observation pairs, which is rather small. In order to increase the size of the set of forecast/observation pairs available for model fitting and to smooth the EMOS parameter estimates over the different lead months, an alternative EMOS scheme has been developed for this study. The EMOS predictive distribution for a particular lead month can be written as

$$y_l \mid \boldsymbol{R} \sim g(\mu_l, \sigma_l), \tag{4.16}$$

where $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots \boldsymbol{r}_L)$ is a matrix of forecast ensembles with $\boldsymbol{r}_l$ being the forecast vector and y_l the variable of interest for lead month l with $l = 1, \ldots, L$. The function g denotes either a truncated normal or a lognormal density function. The location parameter μ_l is now parameterized as

$$\mu_l = a_0 + a_1(w_l\bar{r}_{l,\mathrm{EC}} + (1-w_l)\bar{r}_{l,\mathrm{ESP}}), \tag{4.17}$$

where the weight function w_l is given by

$$w_l = \begin{cases} \mathrm{logit}^{-1}(a_2)e^{a_3(l-1)} & \text{if } a_3 < 0 \\ 1 + [\mathrm{logit}^{-1}(a_2) - 1]e^{a_3(l-1)} & \text{else ,} \end{cases} \tag{4.18}$$

where logit^{-1} is the inverse logit transformation, i.e. $\mathrm{logit}^{-1}(\alpha) = \exp\alpha/(1+\exp\alpha)$. The parameter a_2 controls the weights of $\bar{r}_{l,\mathrm{EC}}$ and $\bar{r}_{l,\mathrm{ESP}}$ at the first lag month and a_3 the direction and rate of the change in the weights over the forecast horizon. The forecast variance σ_l^2 can be represented by either

$$\sigma_l^2 = b_0 + b_1 s_l^2 \quad \text{or} \quad \sigma_l^2 = b_0 + b_1 \log(l), \tag{4.19}$$

where s_l^2 is the raw ensemble variance at forecast lag l. The constraints on the parameters are $a_0, a_1, b_0, b_1 \geq 0$.

Model variants

Before discussing the results, we give now an overview over the different EMOS model variants that we have tested. First, the models can be divided according to distribution function and type of observations. This leads to the following four groups of models:

1. truncated normal EMOS models fitted and verified against measured observations
2. truncated normal EMOS models fitted and verified against simulated observations based on ECMWF ERA-interim meteorological input
3. lognormal EMOS models fitted and verified against measured observations
4. lognormal EMOS models fitted and verified against simulated observations based on ECMWF ERA-interim

Each of the above groups contains five different models with model configurations according to Table 4.6. Model M1 stands for separate EMOS parameter estimation for each lead month l, whereas M2 to M5 use the simultaneous parameter estimation approach described above. The model variants M4 and M5 do not apply any bias correction, i.e. $a_0 = 0$ and $a_1 = 1$ in Equation (4.17). Finally, the models can be divided according to variance specification, for M1, M2, and M4 the variance depends on the ensemble variance, whereas for M3 and M5 the variance depends on the lead month.

Table 4.6: EMOS variants for post processing of seasonal runoff forecasts

model	separate estimation	simultaneous estimation	bias correction	variance depends on
M1	✓		✓	s^2
M2		✓	✓	s^2
M3		✓	✓	$\log l$
M4		✓		s^2
M5		✓		$\log l$

4.6.3 Results

As the raw ensemble and most of the different EMOS variants exhibit positive skill compared to the reference climatology up to a forecast lag of at least three months for all considered catchments, the skill of the EMOS variants is assessed in detail in the following. In contrast with the gain in predictive skill of short

to medium-range hydrological forecasts (cf. Sections 4.2 and 4.3) by statistical post processing, skill of seasonal ensemble runoff forecasts cannot easily be improved by statistical post processing. Figures 4.17 and 4.18 show skill in terms of CRPSS of seasonal hydrologic forecasts pooled over all verification months at gauges Basel and Cologne, respectively. Obviously, skill of the raw ensemble and all EMOS variants is high compared to the climatological forecasts for the first month. Here, statistical post processing improves skill slightly compared to the raw ensemble. At higher lead times skill of the raw ensemble decreases quickly, while the post processing methods are not able to add any skill to the raw ensemble forecasts. Though they cannot outperform the raw ensemble, the lognormal EMOS models M4 and M5 perform best among the different post processing models. Figure 4.21 shows the relative performances of the raw ensemble with regard to the climatology and of the lognormal EMOS M5 forecasts compared to the raw ensemble split according to verification month and forecast lag at gauge Basel. The high CRPSS values for May and June at forecast horizons of more than one month reflect the effect of snow accumulation and snow melt on the predictability of runoff from the alpine parts of river Rhine. Though the effect is gradually attenuated further downstream, it can still be detected at gauge Cologne as shown in Figure 4.22. Independent of forecast lag and verification month, EMOS does not improve forecast skill much.

According to Figures 4.19 and 4.20 EMOS post processing seems to be more beneficial for the two sub-catchments of river Danube. At gauge Hofkirchen, upstream of the confluence with river Inn, the lognormal EMOS models M4 and M5 outperform the raw ensemble in terms of CRPSS at most of the lead times. When verified against runoff observations this holds also for gauge Achleiten, while the raw ensemble cannot be outperfomed by the EMOS models when verified against ERA-interim simulated runoff. As for the gauges Basel and Cologne, the lognormal EMOS models M4 and M5 perform best among the different post processing models at the gauges Hofkirchen and Achleiten. As shown in Figures 4.23 and 4.24, splitting again the CRPSS values according to verification month and forecast lag reveals a quite high skill of the raw ensemble forecasts at gauges Hofkirchen and Achleiten compared to the climatology in the period from roughly May to September when verified against ERA-interim simulated runoff. At gauge Hofkirchen, the raw ensemble forecasts for May underperform the climatology quite strongly when verified against measured observations. This is also reflected by the high relative skill of the lognormal EMOS model M5 compared to the raw ensemble for May. This effect may be explained partly with the extreme event in May 1999. When verified against the ERA-interim simulations, this underperformance vanishes.

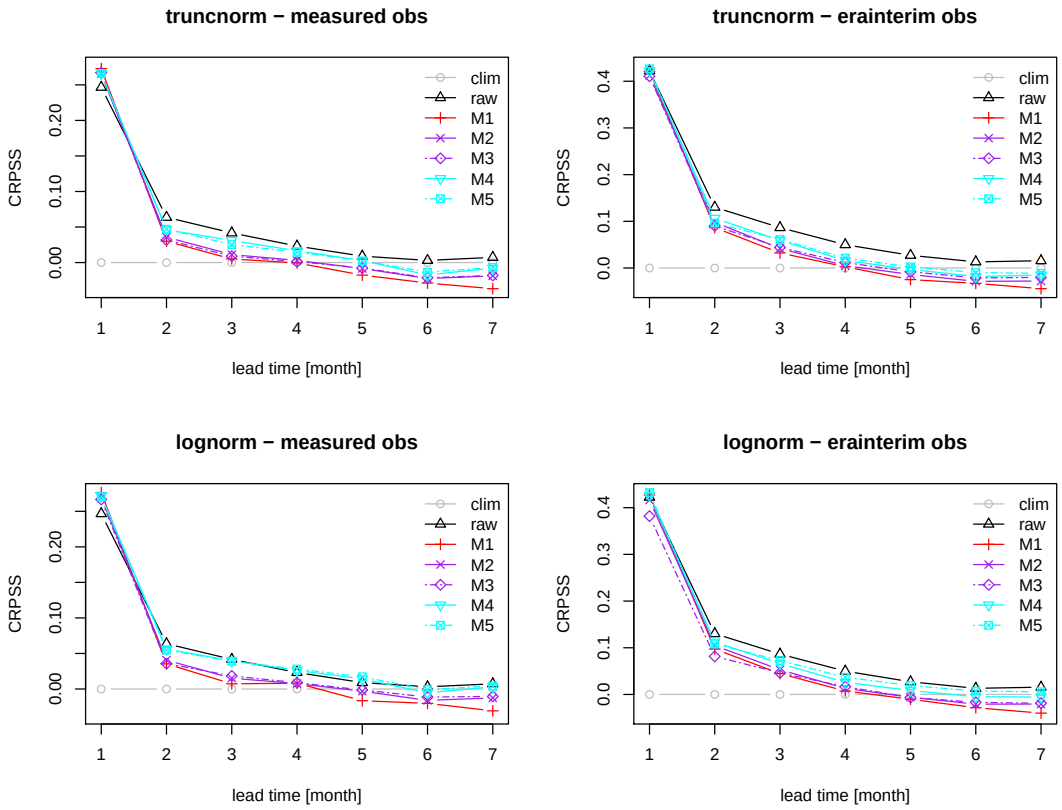


Figure 4.17: Pooled CRPSS values of the seasonal runoff forecasts at gauge Basel. The top panels show the values of the EMOS models based on a truncated normal distribution, the bottom panels the corresponding values for the models based on a lognormal distribution. The models shown in the panels on the left have been fitted and verified against measured observations, those in the panels on the right against a HBV run with ERA-interim forcing.

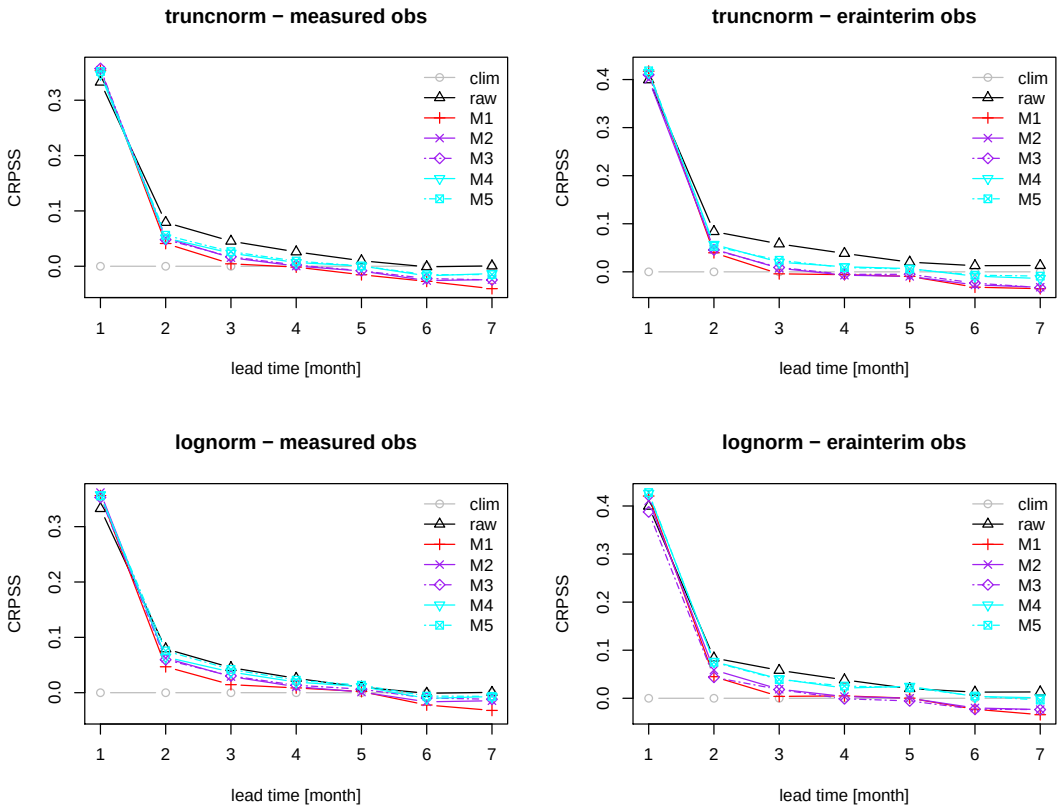


Figure 4.18: Pooled CRPSS values of the seasonal runoff forecasts at gauge Cologne. The top panels show the values of the EMOS models based on a truncated normal distribution, the bottom panels the corresponding values for the models based on a lognormal distribution. The models shown in the panels on the left have been fitted and verified against measured observations, those in the panels on the right against a HBV run with ERA-interim forcing.

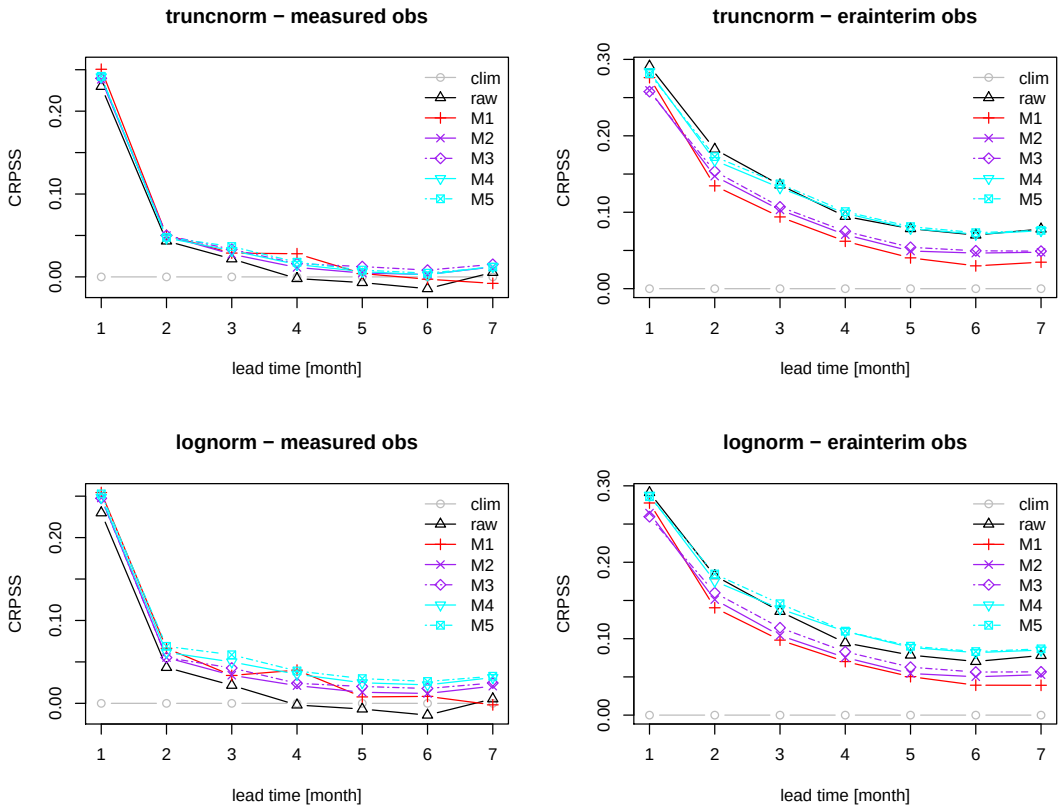


Figure 4.19: Pooled CRPSS values of the seasonal runoff forecasts at gauge Hofkirchen. The top panels show the values of the EMOS models based on a truncated normal distribution, the bottom panels the corresponding values for the models based on a lognormal distribution. The models shown in the panels on the left have been fitted and verified against measured observations, those in the panels on the right against a HBV run with ERA-interim forcing.

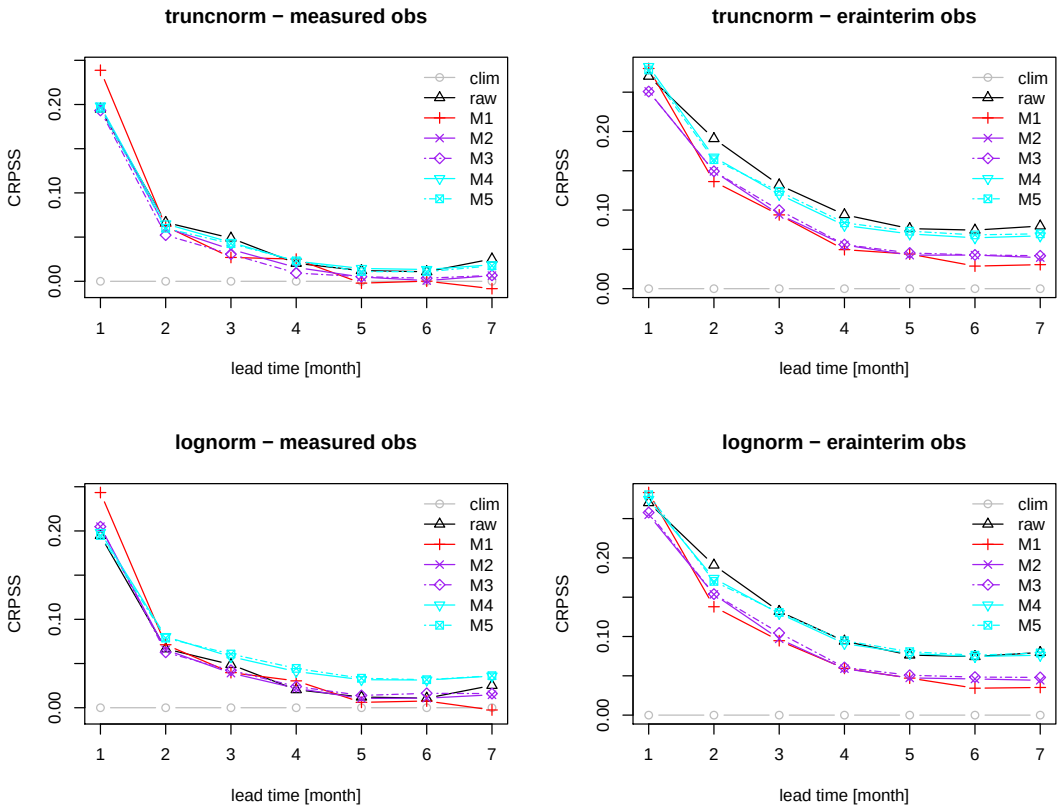


Figure 4.20: Pooled CRPSS values of the seasonal runoff forecasts at gauge Achleiten. The top panels show the values of the EMOS models based on a truncated normal distribution, the bottom panels the corresponding values for the models based on a lognormal distribution. The models shown in the panels on the left have been fitted and verified against measured observations, those in the panels on the right against a HBV run with ERA-interim forcing.

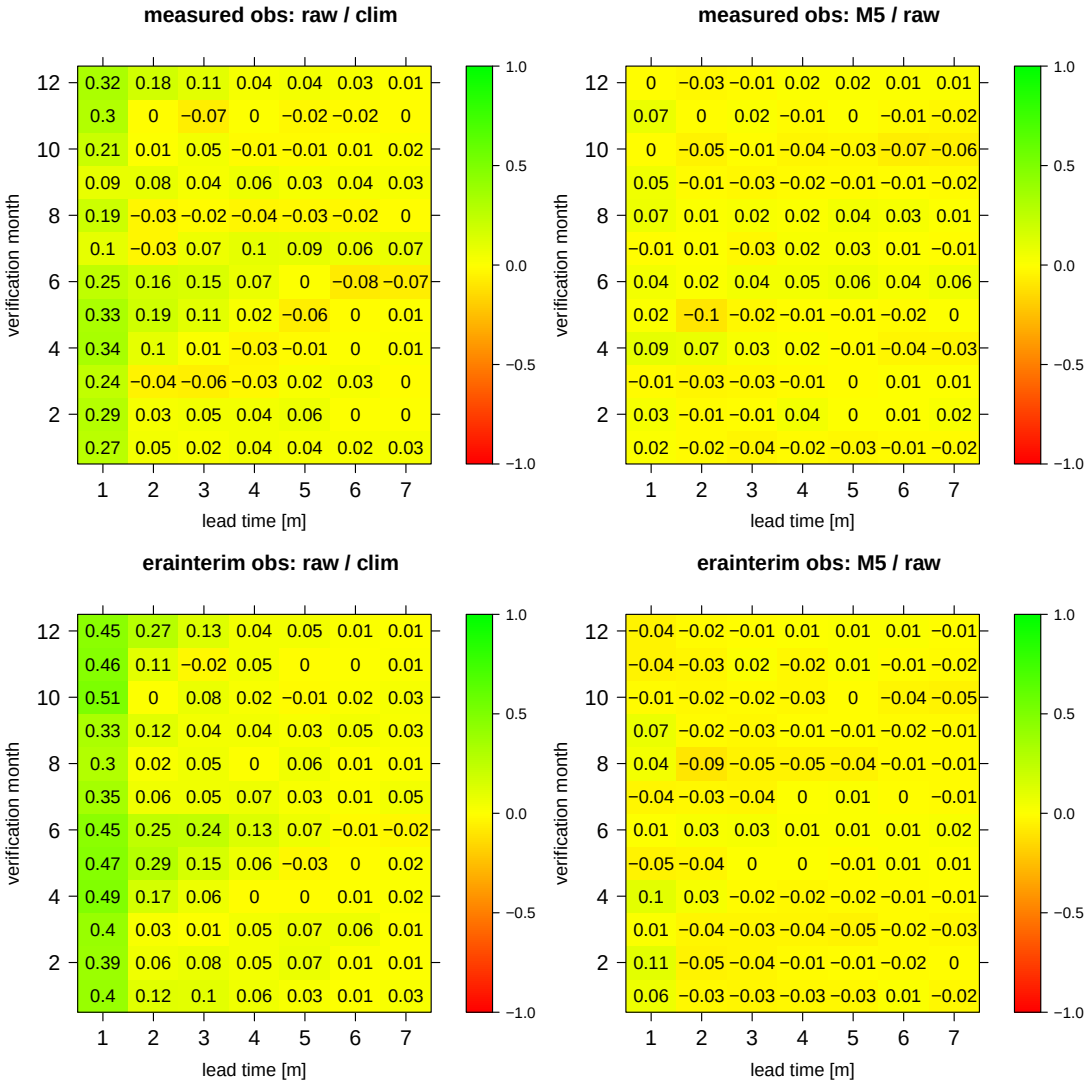


Figure 4.21: Monthly CRPSS of the seasonal runoff forecasts at gauge Basel. The results shown in the top panels are based on model fitting and verification against measured observations, those shown in the bottom panels are based on fitting and verification against a HBV run with ERA-interim forcing. The CRPSS of the raw ensemble is calculated against the monthly climatology, the CRPSS of the lognormal EMOS M5 forecasts against the raw ensemble.

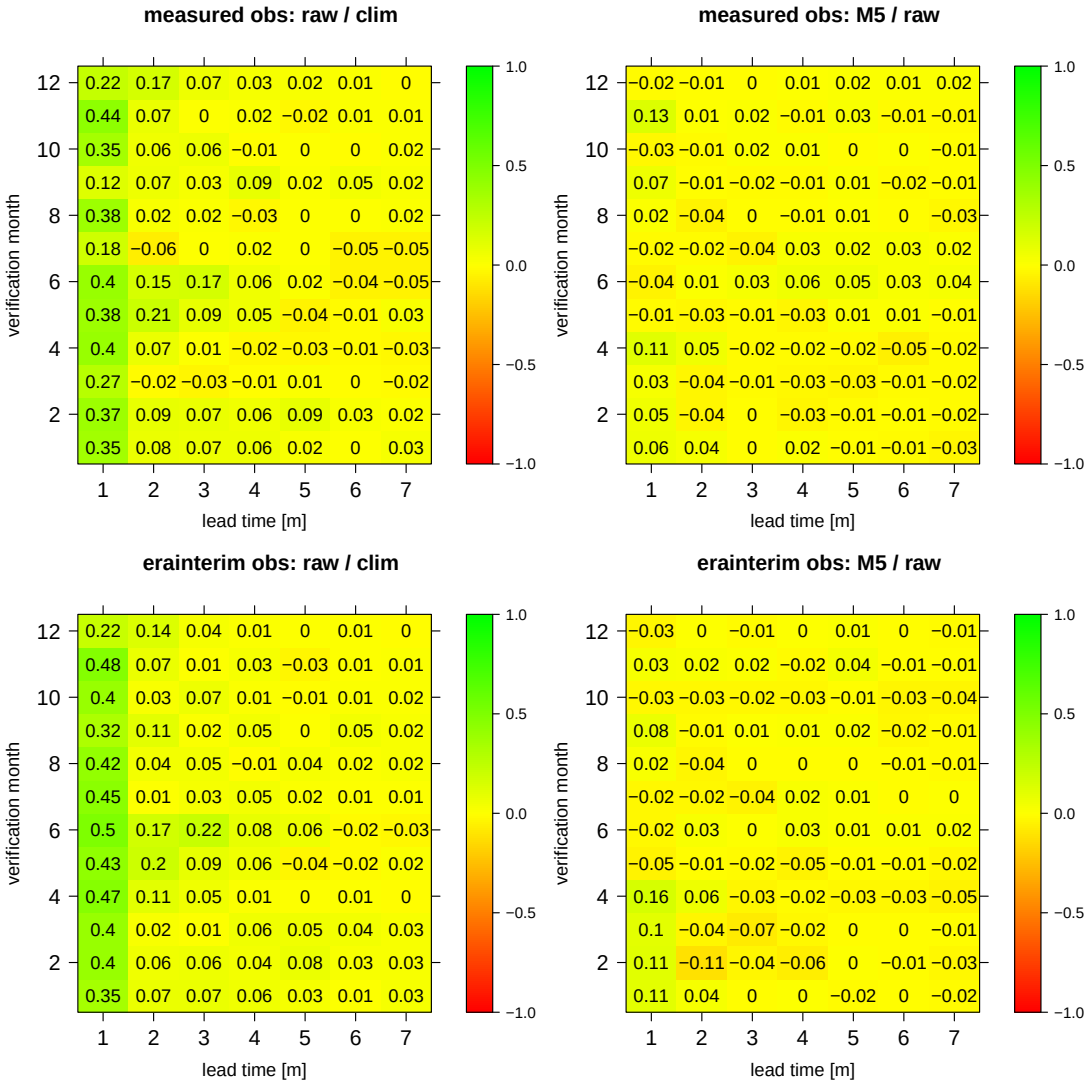


Figure 4.22: Monthly CRPSS of the seasonal runoff forecasts at gauge Cologne. The results shown in the top panels are based on model fitting and verification against measured observations, those shown in the bottom panels are based on fitting and verification against a HBV run with ERA-interim forcing. The CRPSS of the raw ensemble is calculated against the monthly climatology, the CRPSS of the lognormal EMOS M5 forecasts against the raw ensemble.

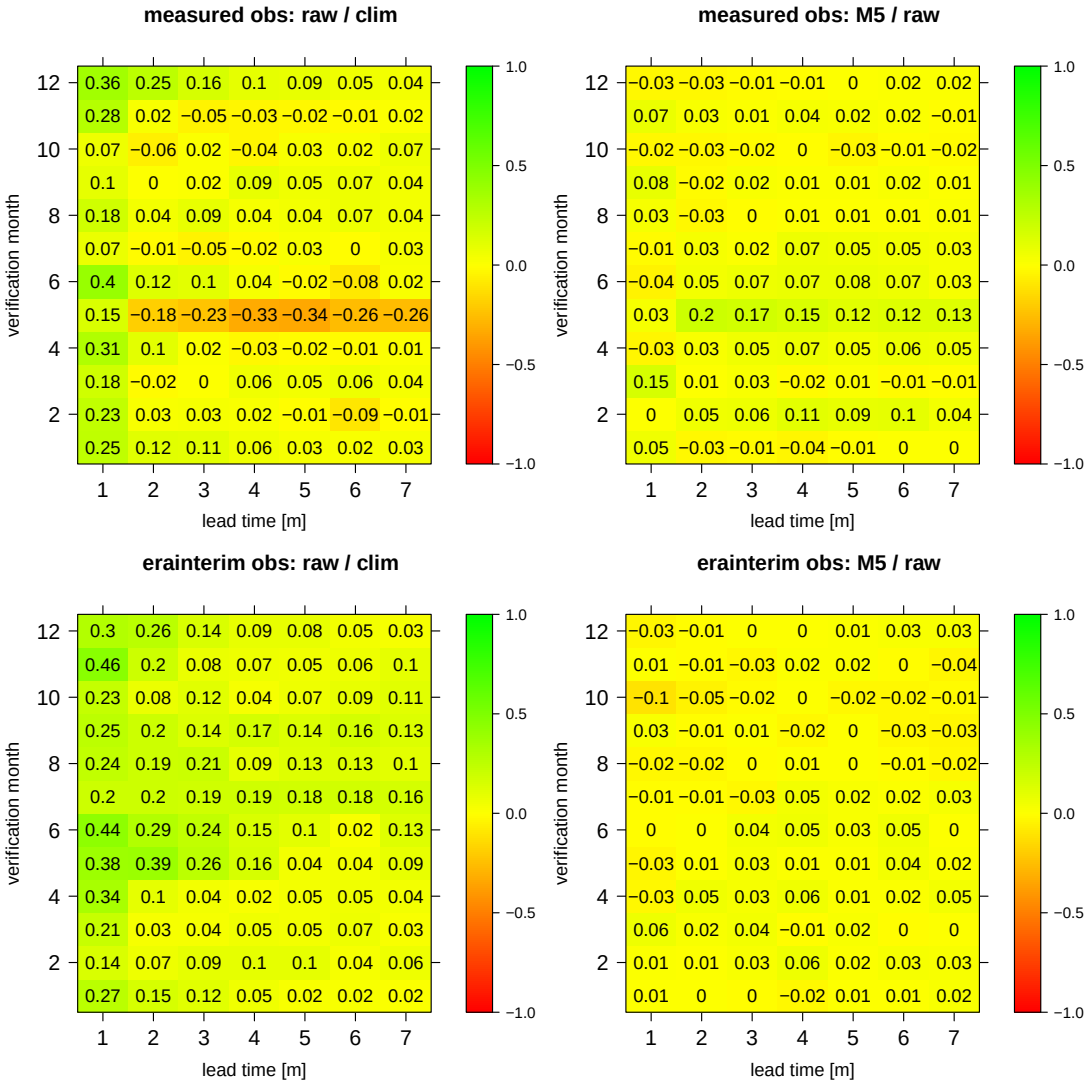


Figure 4.23: Monthly CRPSS of the seasonal runoff forecasts at gauge Hofkirchen. The results shown in the top panels are based on model fitting and verification against measured observations, those shown in the bottom panels are based on fitting and verification against a HBV run with ERA-interim forcing. The CRPSS of the raw ensemble is calculated against the monthly climatology, the CRPSS of the lognormal EMOS M5 forecasts against the raw ensemble.

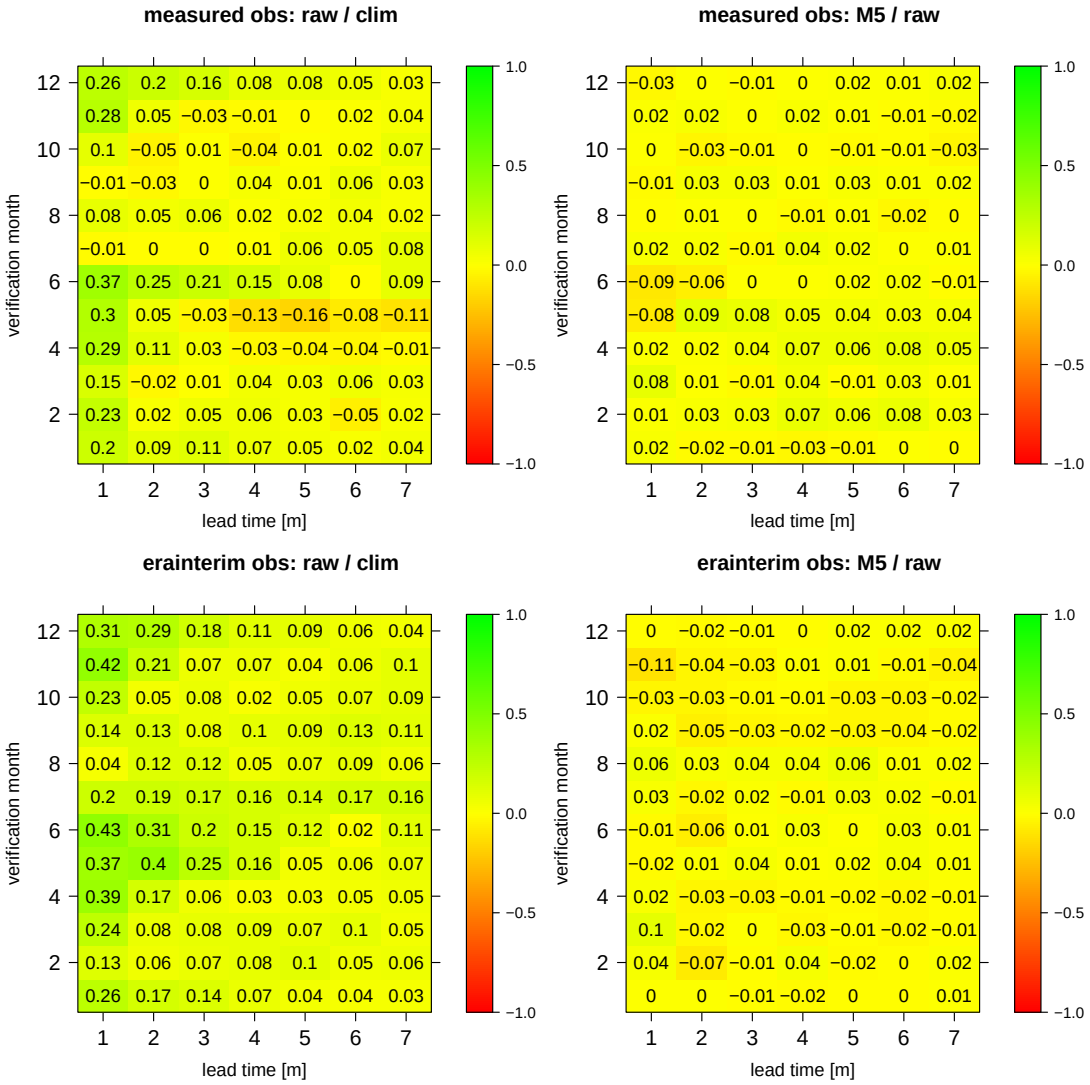


Figure 4.24: Monthly CRPSS of the seasonal runoff forecasts at gauge Achleiten. The results shown in the top panels are based on model fitting and verification against measured observations, those shown in the bottom panels are based on fitting and verification against a HBV run with ERA-interim forcing. The CRPSS of the raw ensemble is calculated against the monthly climatology, the CRPSS of the lognormal EMOS M5 forecasts against the raw ensemble.

4.6.4 Discussion

The above results indicate that statistical post processing can hardly improve forecast skill of seasonal forecasts beyond a forecast horizon of one month. At all four considered gauges, lognormal EMOS outperforms truncated normal EMOS. Furthermore, a few conclusions can be drawn from the relative performance of the different EMOS models. There is probably no need for an explicit bias correction, because M4 and M5, which only weight the means of the ESP and the EC seasonal forecast ensemble and add a variance term, perform best. Beyond a forecast lag of one month, simultaneous parameter estimation seems to be beneficial to the EMOS forecasts. The univariate approach, i.e. model M1 that fits a single EMOS model for each lead time, show a poor performance for forecast lags greater than one month. Furthermore, the raw ensemble variance does not contain much information about forecast uncertainty. This is reflected by the equal forecast skill of the EMOS models that connect forecast variance to the raw ensemble variance, i.e. M2 and M4, with the EMOS models that connect forecast variance simply to the forecast lag, i.e. M3 and M5. In summary, post processing of seasonal hydrologic forecasts cannot really add skill unless potential skill of the raw ensemble forecasts is improved in future.

Chapter 5

Conclusions and outlook

The results of the studies presented in Chapters 3 and 4 have already been discussed in the corresponding sections. The preliminary study on deterministic evaluation of probabilistic hydrologic forecasts needs further analyses before we can draw firm conclusions. Based on the results from the other case studies, we provide a brief synthesis of the covered topics and make some suggestions for further research.

The TCC post processing methods, MLR and POLR, proved to improve forecast skill significantly, when applying them to TCC raw ensemble forecasts from the ECMWF. From the post processing study on ECMWF T2M, PPT24, and V10 forecasts, we concluded that the skill gap between raw ensemble and post processed forecasts remains almost constant over time indicating that post processing will keep adding skill in the foreseeable future. Hence, we hypothesize that this would also be the case for TCC. Nevertheless, further analysis are needed in order to confirm this hypothesis. Unlike T2M and V10, PPT24 and TCC are integrated variables. In case of PPT24 precipitation is integrated over the 24 hours accumulation period and over precipitation generation processes. The ECMWF EPS differentiates between stratiform and convective precipitation. Likewise, TCC can be split into low, medium, and high level clouds. Both the ratio between stratiform and convective precipitation and the proportions of low, medium, and high level clouds depend on location, season, and, in particular, the prevailing weather regime. Hence, further analyses are needed in order to take account of the nature of such integrated variables in statistical post processing. This may lead to an additional improvement in forecast skill. A post processing method that takes account of the different types of precipitation generation processes is currently under development at the Institute for Meteorology and Climate Research of the Karlsruhe Institute of Technology. In general, weather regime dependent post processing of ensemble weather forecasts, will probably be of increasing interest in future. Of course, regime dependent post processing is not restricted to integrated variables. An example of regime dependent wind speed forecasting is provided by Gneiting et al. (2006).

From our results on regime dependent post processing of hydrological forecasts, it becomes clear that regime dependent post processing does not necessarily improve forecast skill. As precipitation is the main driver of runoff generation, hydrologic forecasts may benefit from using statistically post processed precipitation forecasts as input to the hydrological model. Additionally, discerning according to type of precipitation in the post processing process of the meteorological inputs may also be a method to achieve well calibrated hydrologic forecasts that take account of the prevailing weather regime. For instance, it is likely that the predictability of hydrologic forecasts is relatively high, when the runoff generation processes are induced by stratiform precipitation. Due to its chaotic nature predictability is expected to be lower for convective precipitation (Carbone et al., 2002). This hypothesis on predictability of stratiform and convective precipitation induced runoff should be analyzed in further studies on post processing of hydrometeorological ensemble forecasts.

The study on multivariate post processing of hydrologic forecasts reveals that both ECC and GCA are suitable for modelling the temporal dependencies of probabilistic hydrologic forecasts. Accordingly, multivariate EMOS is a good starting point for further developments. In the settings of a large river system like river Rhine with several sub-catchments spatial dependences between the gauges of different tributaries should be considered in order to obtain a better representation of the total runoff after the confluence of the tributaries. While ECC can be applied to spatio-temporal settings in a straightforward manner, a parametric model that takes account of the dependence between different gauges may be developed based on the space-time models discussed in Gneiting et al. (2007b). The further development of multivariate EMOS in time would include an extension from short- and medium-range hydrologic forecasts to seasonal forecasts. Keeping in mind that the seasonal forecasts for the catchments considered in our case study show some skill compared to the reference climatology, such an approach would lead to appealing “seamless” predictions.

Appendix A

Technical details

A.1 Discrete post processing of total cloud cover ensemble forecasts

A.1.1 TCC mapping

The SYNOP observations dataset at hand reports TCC states as values in $Z = \{0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, 1\}$. Obviously, CRPS, log score, forecast variance, and the width of the 90 % prediction interval are affected by the choice of the verification space. The ECMWF TCC raw ensemble forecasts are continuous in $[0, 1]$. The post processed MLR and POLR forecasts are given in 9 ordered categories, which can be considered octas. Raw ensemble and post processed forecasts are mapped to Z according to Table A.1.

A.1.2 Marginal calibration

Let F_v be the predictive CDF for verification day v in verification period V, then the average predictive CDF for TCC can be written as

$$\bar{F}_V(z) = \frac{1}{V} \sum_{v=1}^{V} F_v(z), \qquad z \in \{0, 1, \ldots, 8\}, \tag{A.1}$$

and the empirical CDF of the observations as

$$\hat{G}_V(z) = \frac{1}{V} \sum_{v=1}^{V} \mathbb{1}_{[z_v \leq z]}, \qquad z \in \{0, 1, \ldots, 8\}. \tag{A.2}$$

For a marginally well calibrated forecast the graph of $\bar{F}_V(z) - \hat{G}_V(z)$ describes a horizontal line at zero (Gneiting et al., 2007a).

Table A.1: Mapping of TCC raw ensemble and post processed forecasts[a]. Table taken from Hemri et al. (2016).

octa	0	1	2	3	4	5	6	7	8
z	0	0.1	0.25	0.4	0.5	0.6	0.75	0.9	1
x	0	0.01	0.1875	0.3125	0.4375	0.5625	0.6875	0.8125	0.99
y	0.01	0.1875	0.3125	0.4375	0.5625	0.6875	0.8125	0.99	1

[a] Note that x and y denote the lower and the upper limit for the mapping of forecast values to the corresponding values in the format of the SYNOP observations in Z. Unless TCC state is 8 octas, the mapping intervals are left-closed and right-open. In case of 8 octas the mapping interval is closed on both sides. The mapping intervals are not equidistant in order to mimic human observers.

A.2 Post processing of hydrologic forecasts

A.2.1 Box-Cox transformation

The Box-Cox transformation (Box and Cox, 1964) is given by

$$h(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \tag{A.3}$$

where x is on the original space and λ is the Box-Cox coefficient. For the studies in Hemri et al. (2014a) and Hemri et al. (2015) the same estimated parameter $\hat{\lambda}$ is used for both observations and forecasts. For each catchment considered the estimation has been performed using the complete time series of observations and corresponding simulations from 1 November 1998 to 31 October 2008, which corresponds to the period prior to the verification period. That is, for each catchment the estimate $\hat{\lambda}$ is constant throughout the entire study. The actual parameter estimation has been performed by minimizing the Kolmogorow-Smirnow test statistic of a normal distribution with appropriate mean and variance and the empirical distribution of the differences between the transformed observations and the corresponding hydrological model simulations using observed meteorological input by applying the `R` function `ks.test`. The estimates $\hat{\lambda}$ are -0.31, -0.42, 0.61, -0.04, and 0.03 for the rivers Wied, Ahr, Upper Rhine, Moselle, and Lahn, respectively. Another widely used alternative method to normalize the data is the normal quantile transform (NQT, see also van der Waerden (1952, 1953a,b) and Todini (2008) for an example of its application). Because of the limitations of NQT with regard to the required extrapolation beyond the maximum observed runoff, the Box-Cox transformation is preferred here.

A.2.2 Rating curve fitting

Probabilistic forecasts assign positive, though in most cases very low, probabilities to extreme outcomes. Hence, a rating curve that has been constructed from past

pairs of water level and runoff cannot cover the entire support of any predictive density that may be generated by statistical post processing. In order to resolve this problem, the rating curves for the gauges Maxau, Kalkofen, and Lahn are extended to the interval from one third of the minimum of the climatology to 3 times the maximum of the climatology, where the climatology corresponds to the empirical distribution of the hourly observations from 1 November 1998 to 31 October 2008. Additionally to the extension of the rating curve range, the fitted rating curve function needs to be monotonically increasing in order to avoid artefacts like runoff values that decrease with increasing gauge levels. A functional rating curve fulfilling these requirements can be obtained as follows:

1. Select a suitable family of rating curve functions. Usually rating curves can be approximated by

$$Q = P(G - e)^b, \tag{A.4}$$

   where Q and G denote discharge and water level, respectively, and P, e, and b are parameters that need to be estimated from pairs of water level and runoff observations (Kennedy, 1984).

2. Fit two separate rating curves: one in order to extrapolate high flows and one for low flows. The curves are estimated by minimizing the mean squared error. The curve for extrapolating at the lower end is forced to pass through the lowest measured water level/runoff pair, whereas the curve at the upper end is forced to pass through the highest measured water level/runoff pair.

3. Use the fitted rating curves to predict pairs of extreme water level and runoff values. We have predicted four pairs of water level and runoff values on both sides of the observed rating curve.

4. Fit a constrained smoothing spline function, which ensures monotonicity, to the combination of observed and extrapolated water level/runoff pairs. This can be done using the penalized splines method by Meyer (2012).

5. If derivatives of the rating curve function are needed: Replace the constrained smoothing spline function by the standard `R` function `smooth.splines` of the `stats` package. This may, for instance, be needed in order to convert probabilistic runoff forecast density distributions into probabilistic forecasts of water levels.

The rating curves for the gauges Maxau, Trier, and Kalkofen are shown in Figure A.1 a), which includes also the extrapolated pairs of water levels and runoff values as well as the penalized spline fits. The effects of the subsequent Box-Cox transformation is shown in Figure A.1 b). Note that the Box-Cox transformed values shown here, are obtained by first converting the runoff values from m^3/s to mm/h, i.e. runoff is normalized with catchment area such that it is represented as

the equivalent amount of rainfall distributed evenly over the catchment, and then applying the Box-Cox transformation. Figure A.1 c) shows the combined effect of the two steps transformation from the verification space, i.e. water level, to the forecast model space, i.e. Box-Cox transformed runoff. Considering the estimates $\hat{\lambda}$ for the Box-Cox transformation parameter that are 0.61, -0.04, and 0.03 for the gauges Maxau, Trier, and Kalkofen, respectively, it looks like the dependence of the Box-Cox transformed runoff interval width on water level is dominated by the rating curve in case of the gauge Maxau, whereas it is dominated by the Box-Cox transformation in case of the gauges Trier and Kalkofen. Given water level intervals of constant width, dominance of the rating curve corresponds to an increase in the width of the respective Box-Cox transformed runoff intervals with increasing water level. Dominance of the Box-Cox transformation corresponds to a decrease. Note that the inconsistencies at a water level of about 900 cm in the plot of the transformed runoff interval width against water level at gauge Maxau is most likely due to a small artifact produced by our actual implementation of the rating curve fitting procedure that has to be addressed in a following-up study.

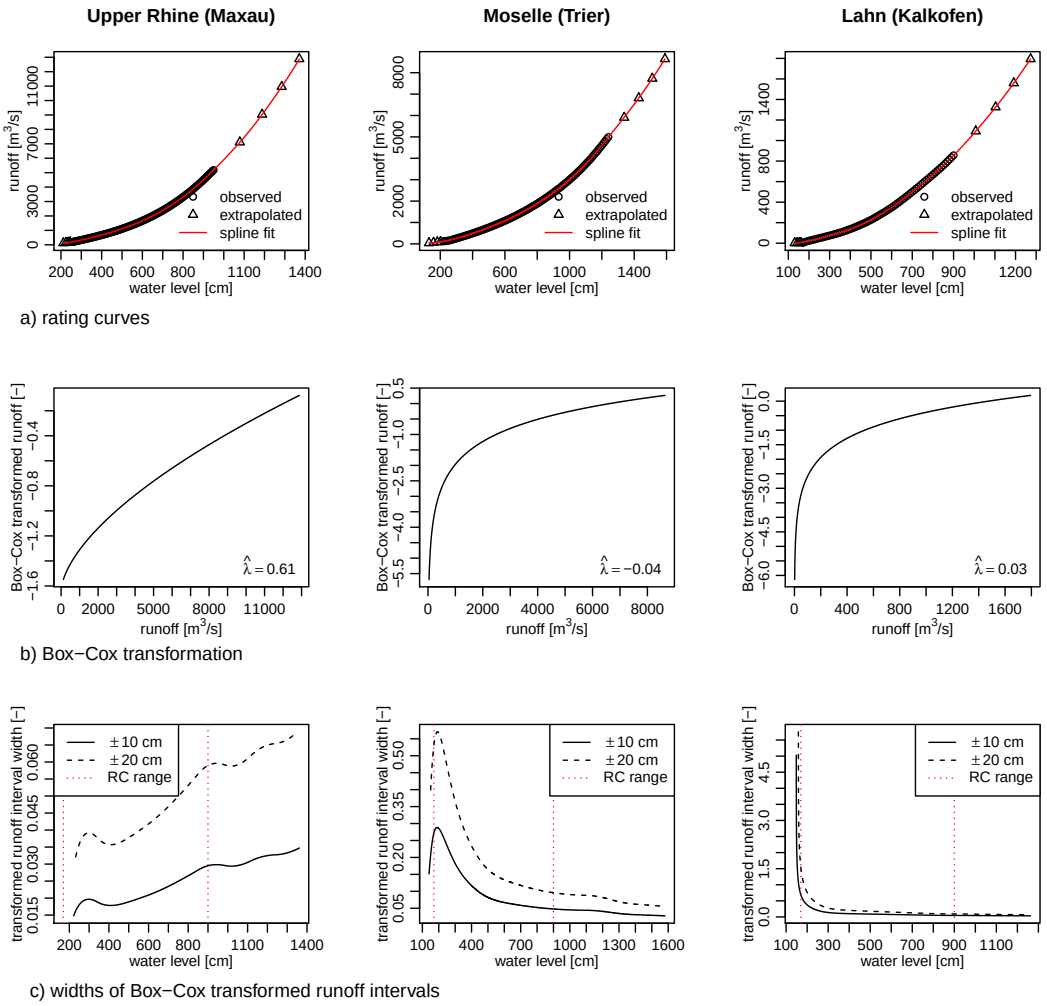


Figure A.1: Fitted rating curves (a), Box-Cox transformation curves (b), and widths of Box-Cox transformed runoff intervals (c) for the catchments Upper Rhine (Maxau), Moselle (Trier), and Lahn (Kalkofen). RC range denotes the range of observed water levels.

Appendix B

Supplemental figures

B.1 Trends in the predictive performance of raw ensemble weather forecasts

Figures B.1 to B.3 show the relative change in skill by applying EMOS for T2M, PPT24, and V10 at the global set of SYNOP stations. The global distributions of significant trends in ΔCRPS obtained by the parametric regression model and the Kendall's τ correlation coefficient are shown in Figures B.4 and B.5, respectively.

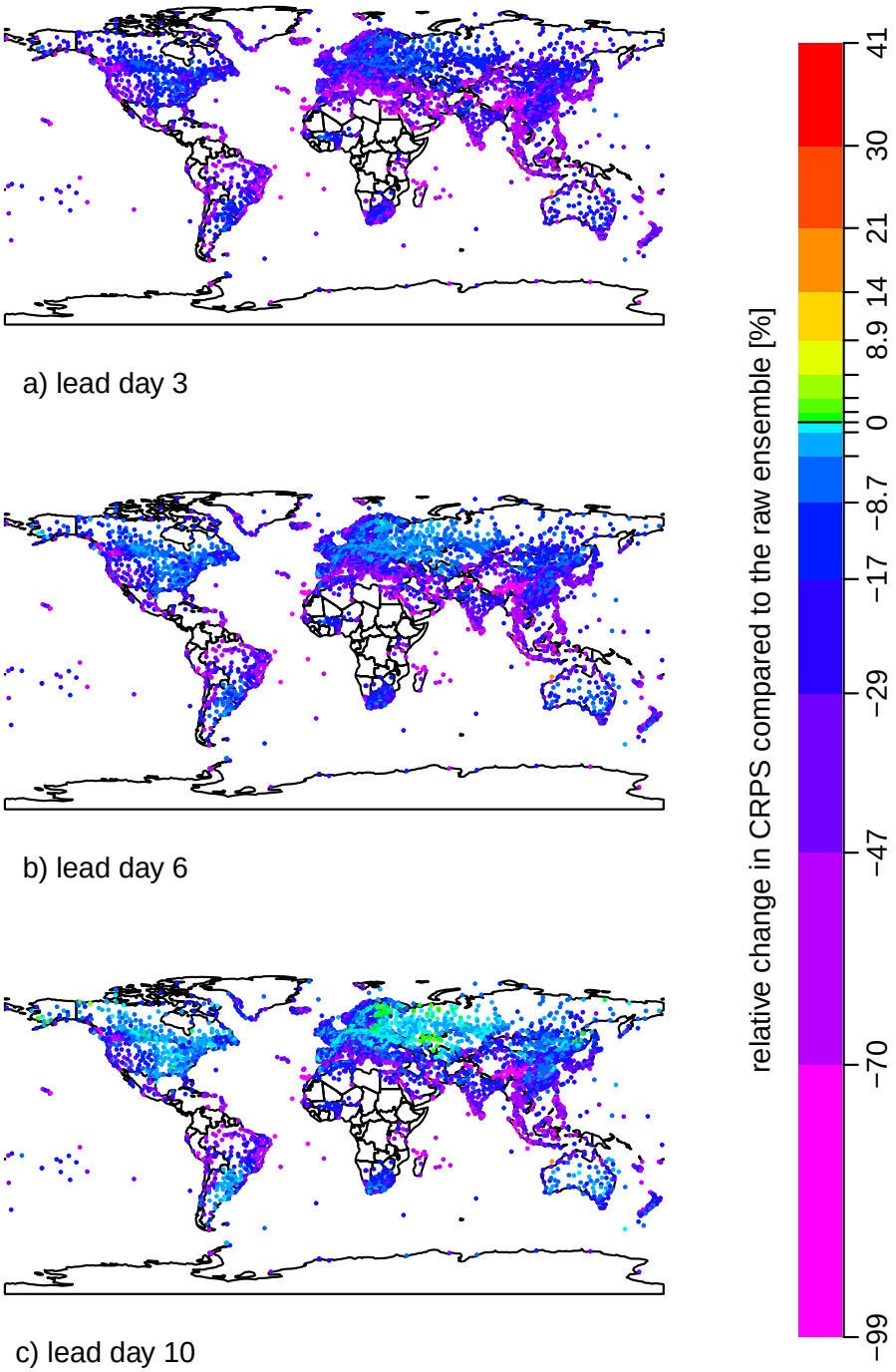


Figure B.1: Relative change (%) in CRPS by EMOS with respect to the raw ensemble at all stations for T2M for a) lead day 3, b) lead day 6, and c) lead day 10. The original figures for lead days 5 and 10 can be found in Richardson et al. (2015).

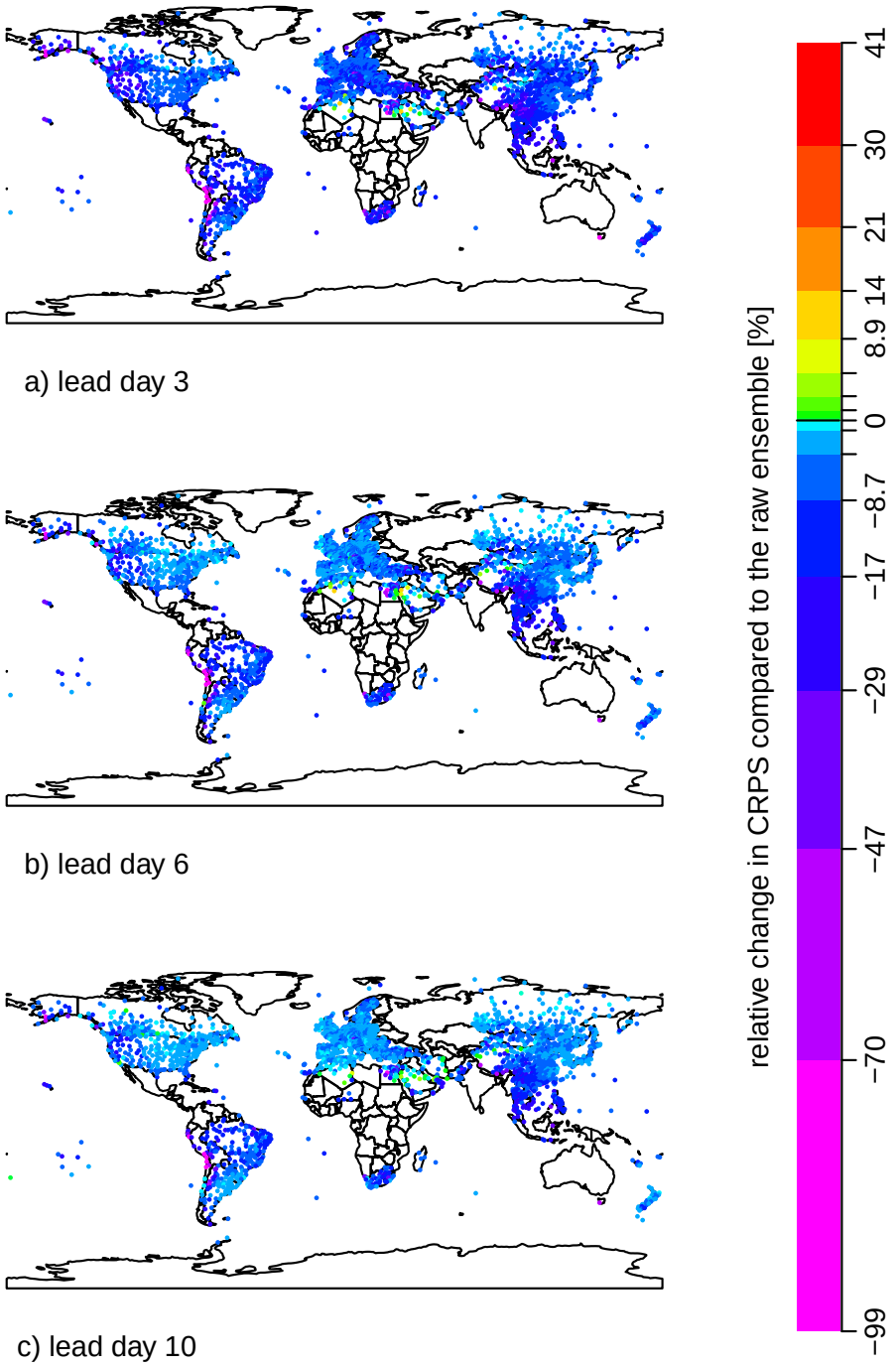


Figure B.2: Relative change (%) in CRPS by EMOS with respect to the raw ensemble at all stations for PPT24 for a) lead day 3, b) lead day 6, and c) lead day 10.

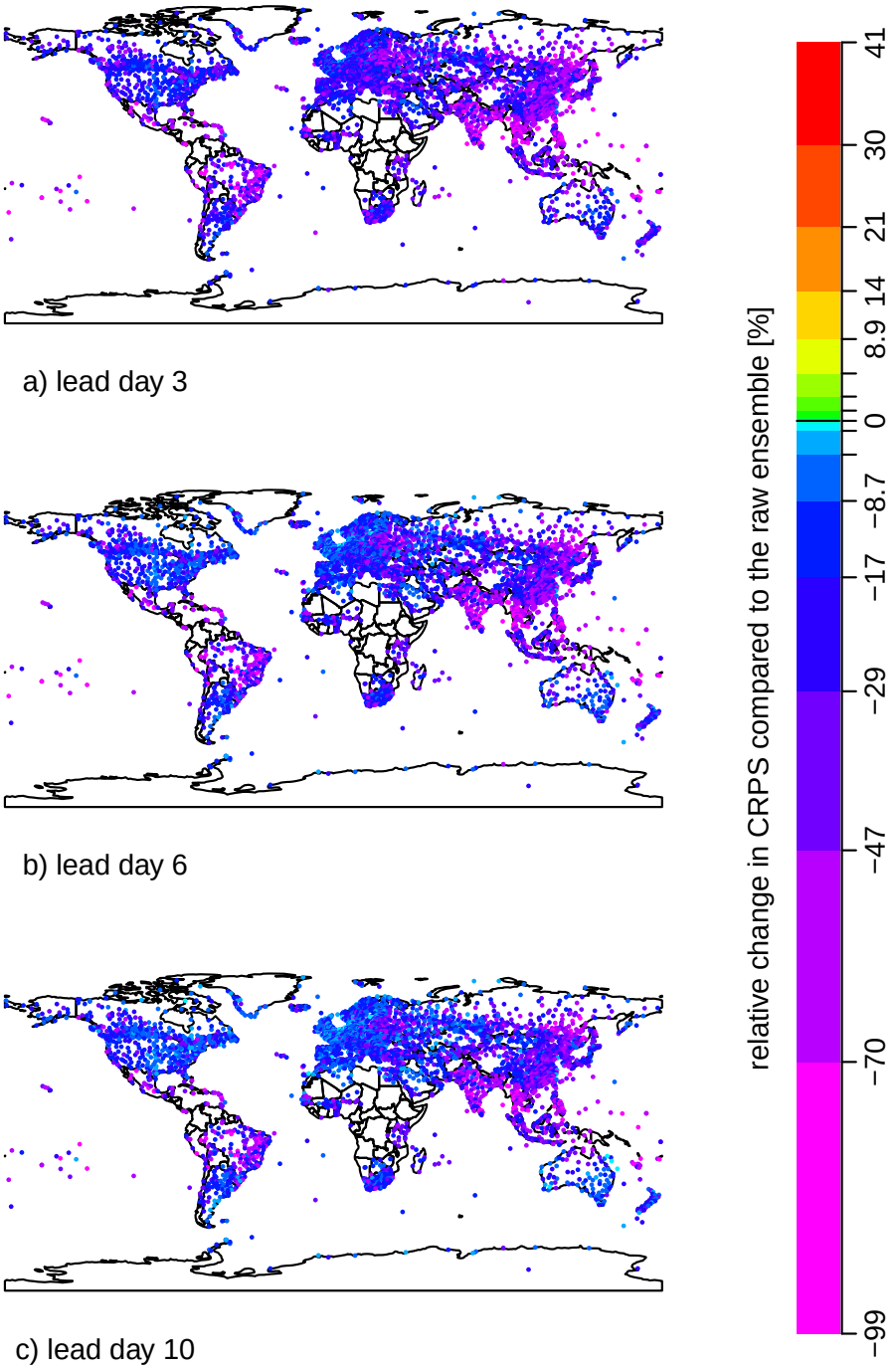
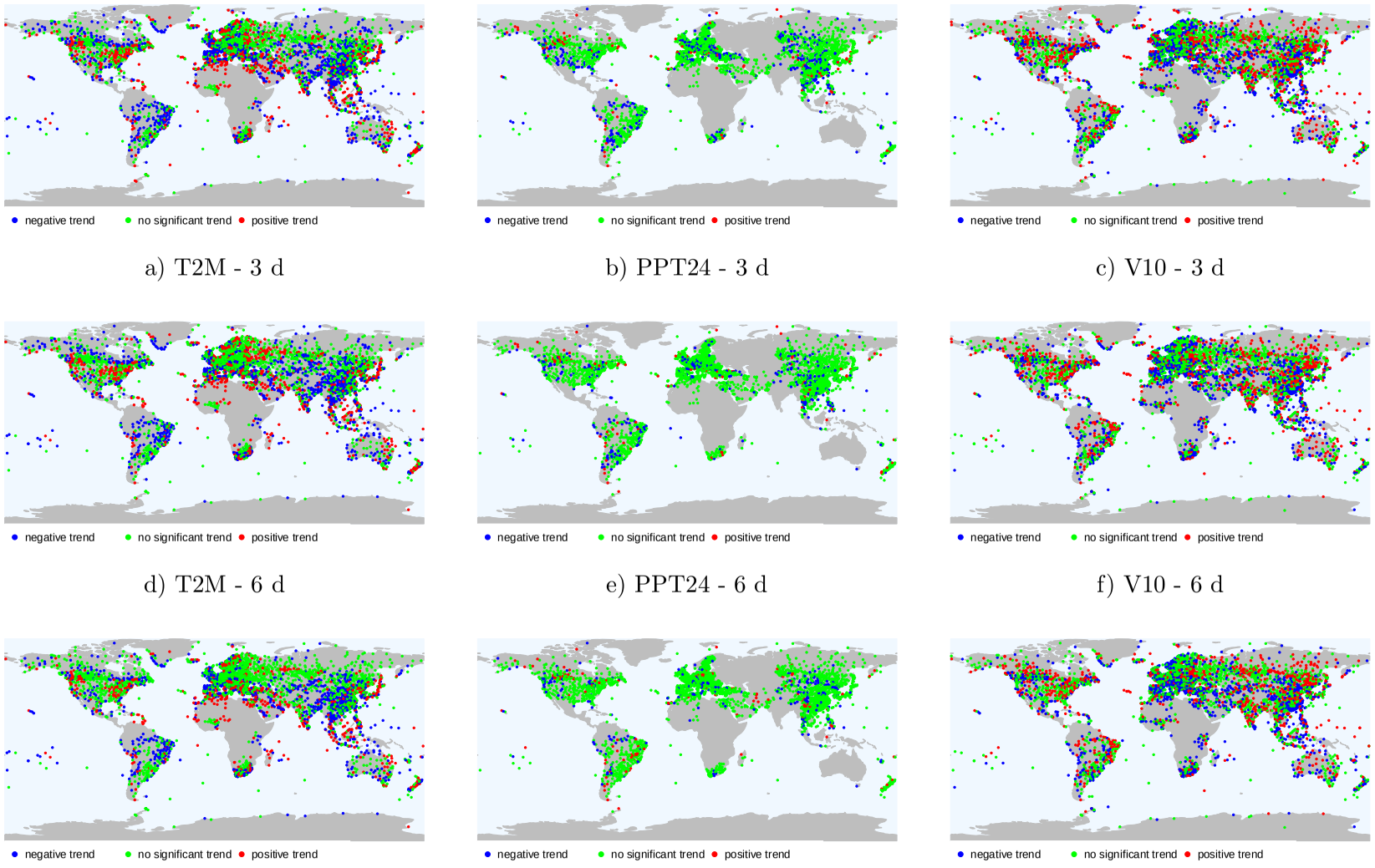


Figure B.3: Relative change (%) in CRPS by EMOS with respect to the raw ensemble at all stations for V10 for a) lead day 3, b) lead day 6, and c) lead day 10.

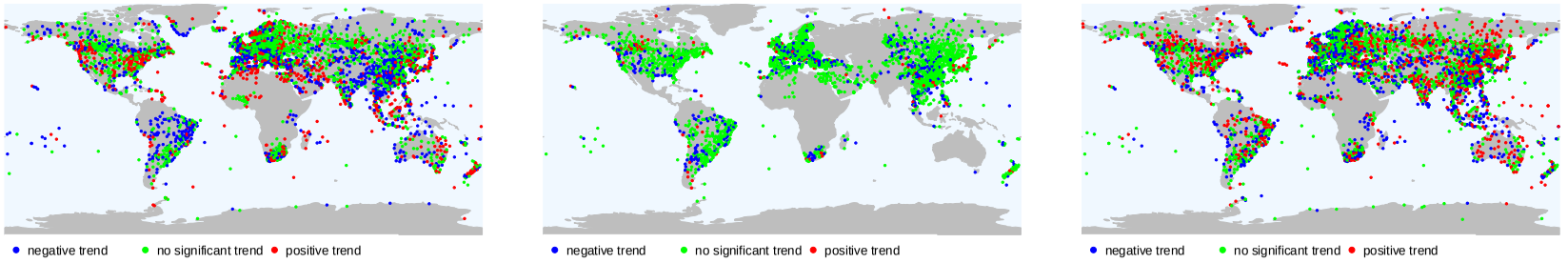


a) T2M - 3 d b) PPT24 - 3 d c) V10 - 3 d

d) T2M - 6 d e) PPT24 - 6 d f) V10 - 6 d

g) T2M - 10 d h) PPT24 - 10 d i) V10 - 10 d

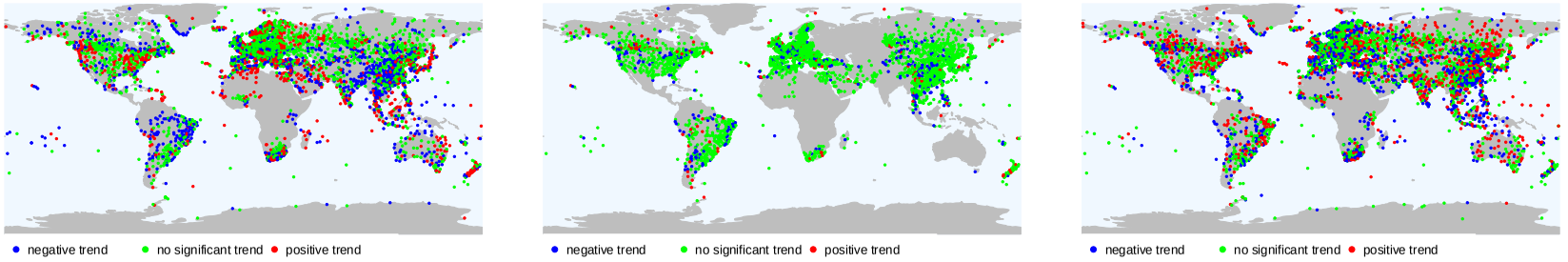
Figure B.4: Global distribution of significant trends in ΔCRPS (between the raw ensemble and the EMOS forecasts) over the verification period for T2M, PPT24, and V10. The forecast lead times considered are 3 d, 6 d, and 10 d. Significant trends are obtained using the parametric regression model with correction for seasonalities at a significance level of 0.05.



a) T2M - 3 d

b) PPT24 - 3 d

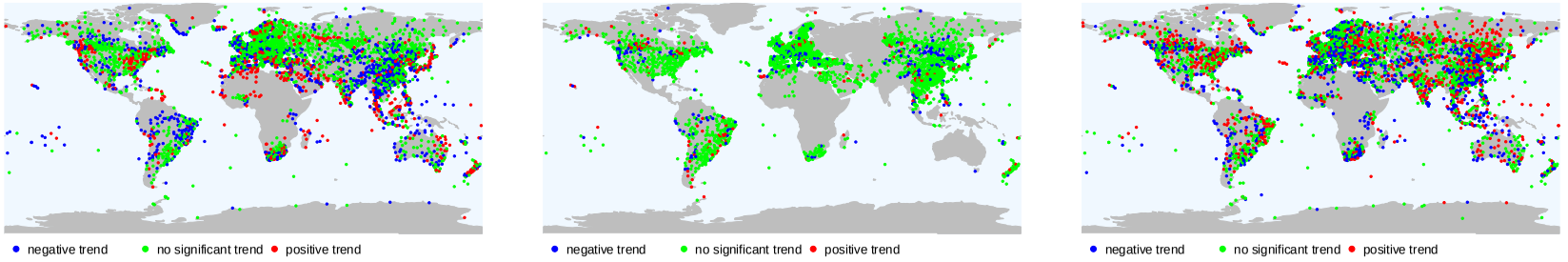
c) V10 - 3 d



d) T2M - 6 d

e) PPT24 - 6 d

f) V10 - 6 d



g) T2M - 10 d

h) PPT24 - 10 d

i) V10 - 10 d

Figure B.5: Global distribution of significant trends in ΔCRPS (between the raw ensemble and the EMOS forecasts) over the verification period for T2M, PPT24, and V10. The forecast lead times considered are 3 d, 6 d, and 10 d. Significant trends are obtained. Significant trends are obtained using the Kendall's τ correlation coefficient test with correction for seasonalities at a significance level of 0.05.

B.2 Multivariate post processing techniques for probabilistic hydrological forecasting

Figures B.6 to B.22 provide a collection of example forecasts similar to Figures 4.7 c) to f) in Section 4.3. For each catchment three examples each for low and high flow events are shown (for river Moselle one of the high flow example forecasts is actually Figure 4.7). The forecasts are issued on the dates indicated below at 06:00 UTC. The subfigures show a) the trajectories of the raw ensembles, b) the quantiles of the EMOS forecasts, and c) and d) the trajectories of the EMOS forecasts with correlation structure by ECC-T or GCA-exp, respectively.

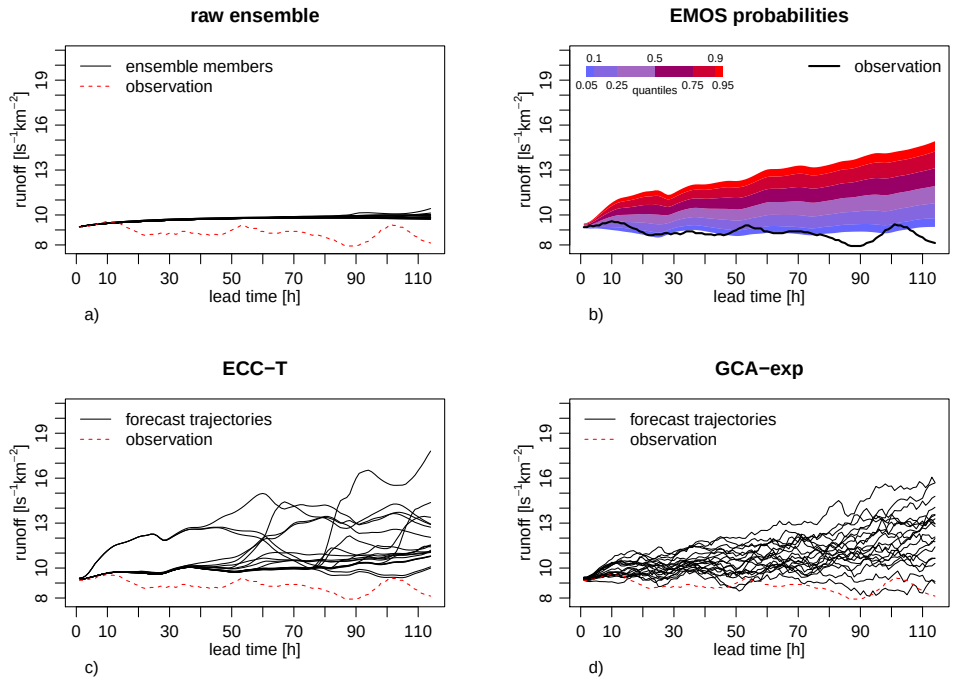


Figure B.6: Example of a low flow forecast for river Upper Rhine at Maxau issued on 1 October 2009. Figure taken from the supplemental material to Hemri et al. (2015).

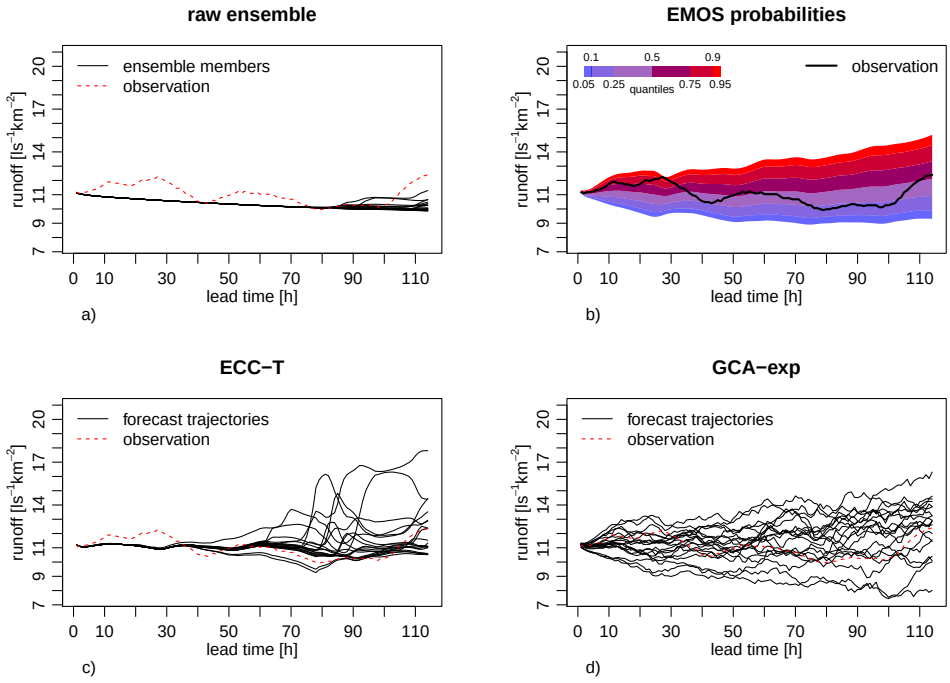


Figure B.7: Example of a low flow forecast for river Upper Rhine at Maxau issued on 29 October 2009. Figure taken from the supplemental material to Hemri et al. (2015).

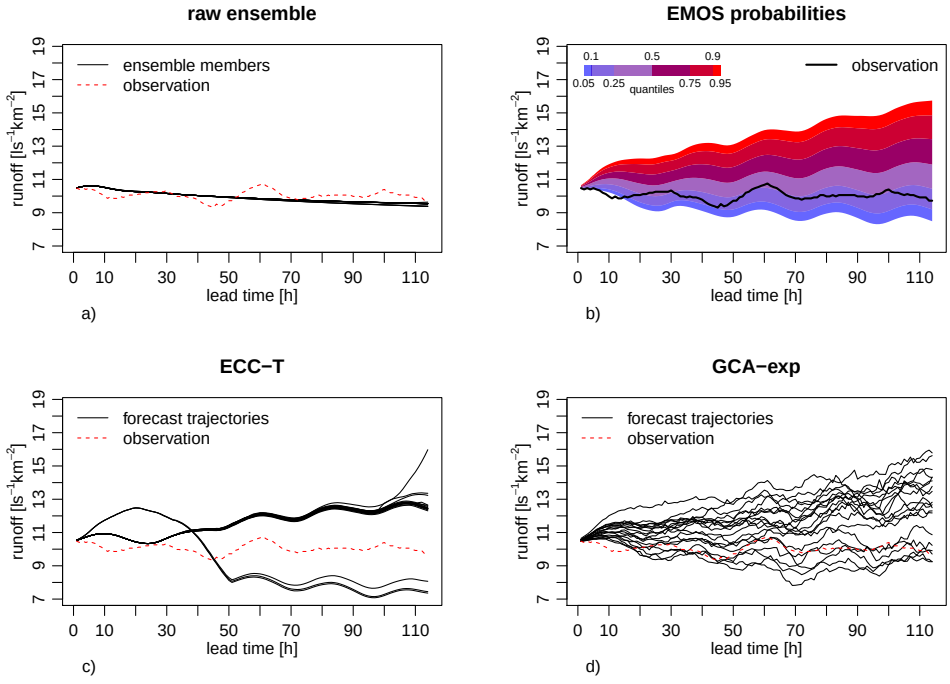


Figure B.8: Example of a low flow forecast for river Upper Rhine at Maxau issued on 6 May 2011. Figure taken from the supplemental material to Hemri et al. (2015).

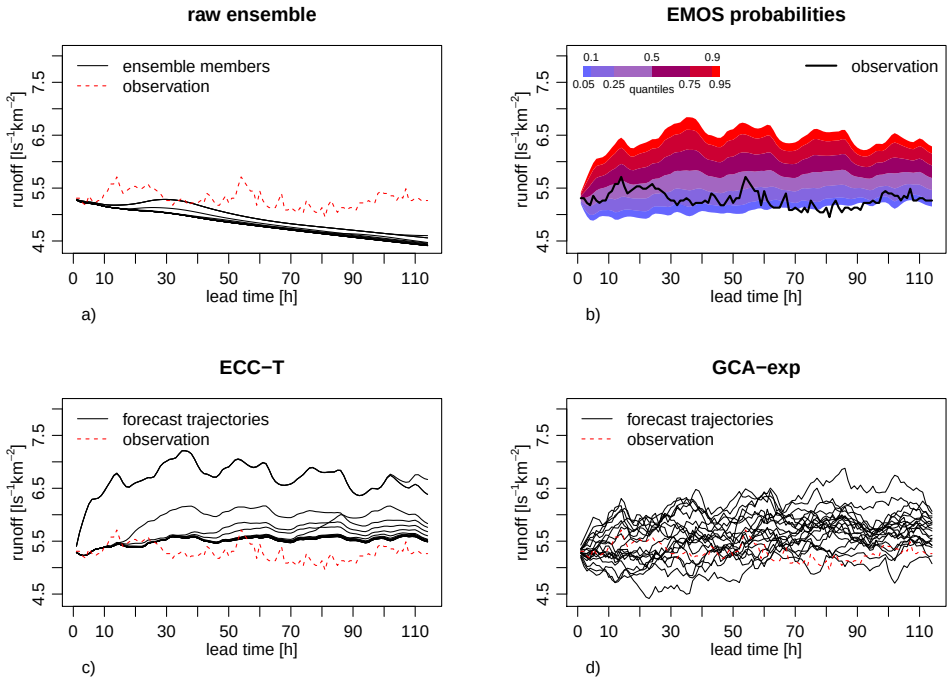


Figure B.9: Example of a low flow forecast for river Moselle at Trier issued on 26 August 2009. Figure taken from the supplemental material to Hemri et al. (2015).

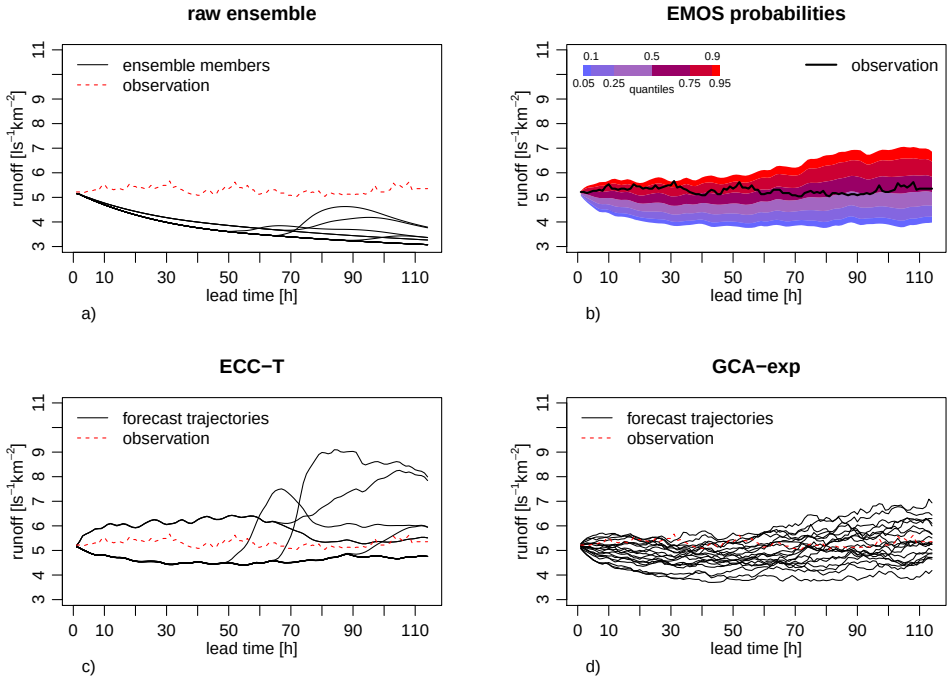


Figure B.10: Example of a low flow forecast for river Moselle at Trier issued on 7 September 2009. Figure taken from the supplemental material to Hemri et al. (2015).

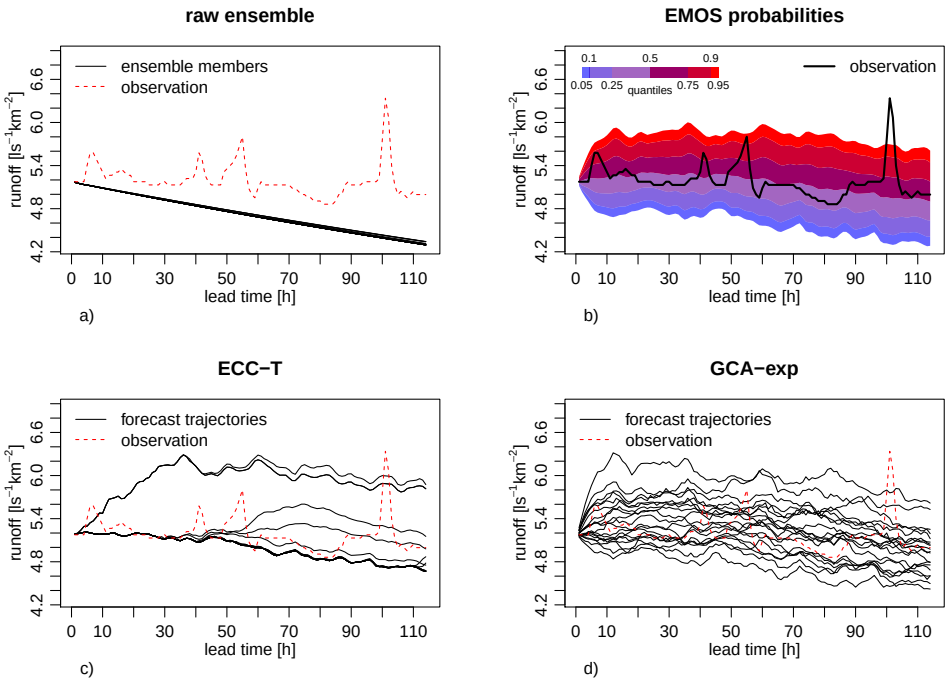


Figure B.11: Example of a low flow forecast for river Moselle at Trier issued on 26 May 2011. Figure taken from the supplemental material to Hemri et al. (2015).

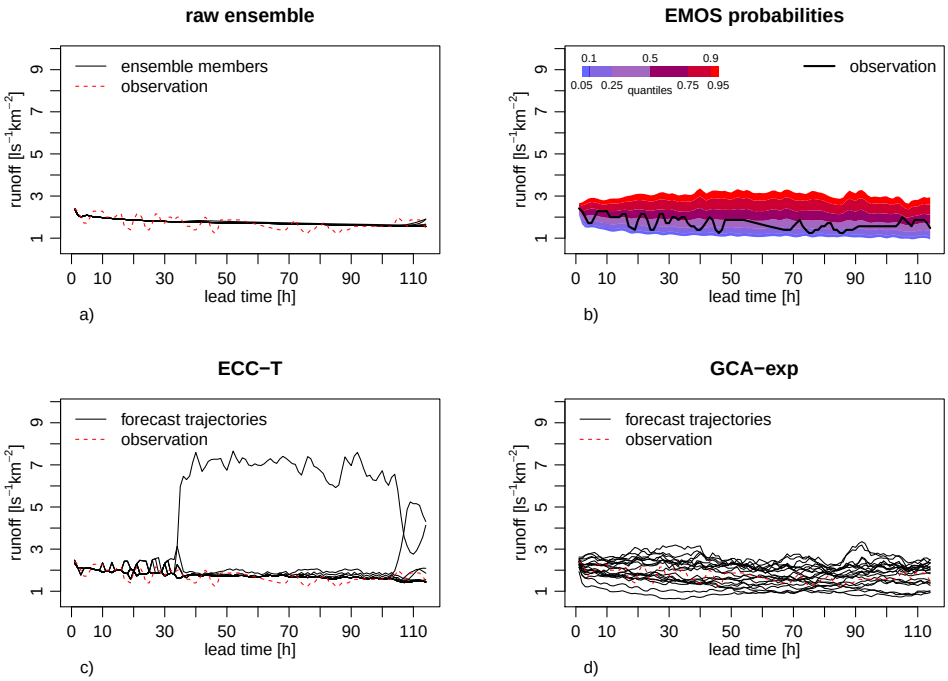


Figure B.12: Example of a low flow forecast for river Lahn at Kalkofen issued on 28 August 2009. Figure taken from the supplemental material to Hemri et al. (2015).

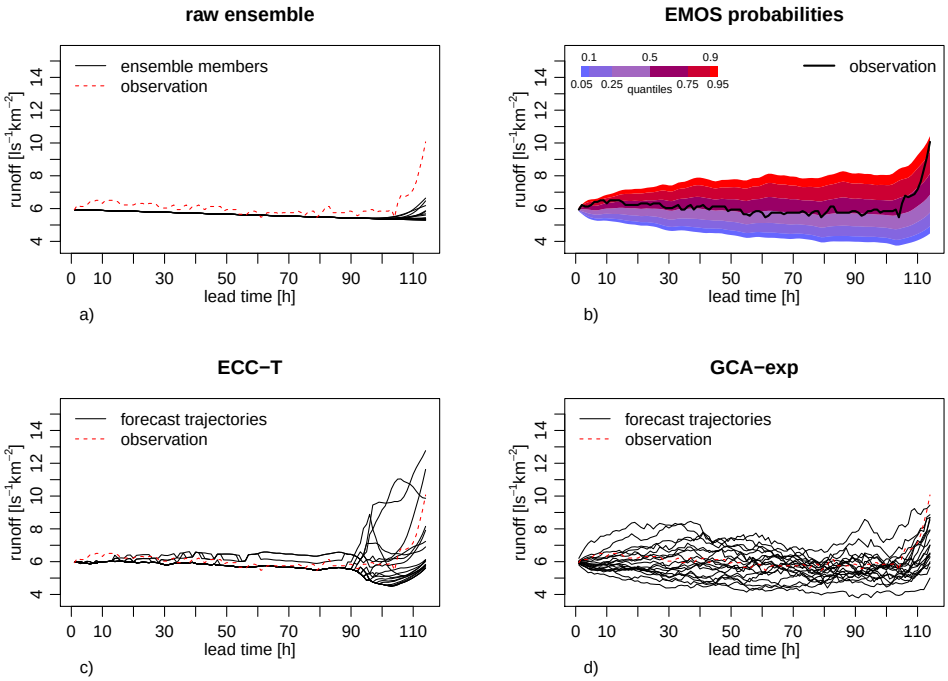


Figure B.13: Example of a low flow forecast for river Lahn at Kalkofen issued on 2 January 2011. Figure taken from the supplemental material to Hemri et al. (2015).

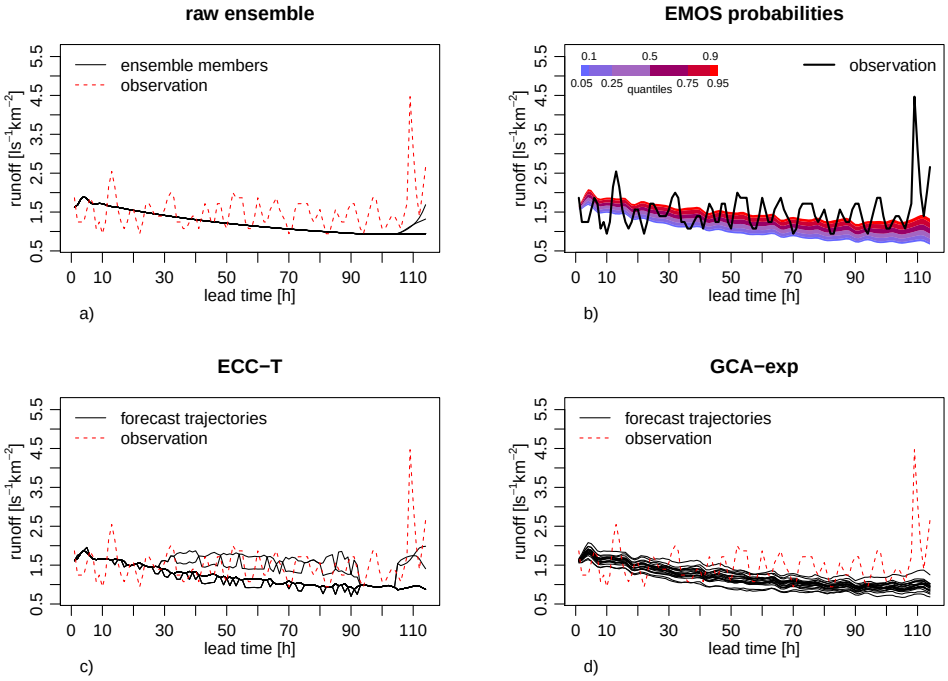


Figure B.14: Example of a low flow forecast for river Lahn at Kalkofen issued on 27 May 2011. Figure taken from the supplemental material to Hemri et al. (2015).

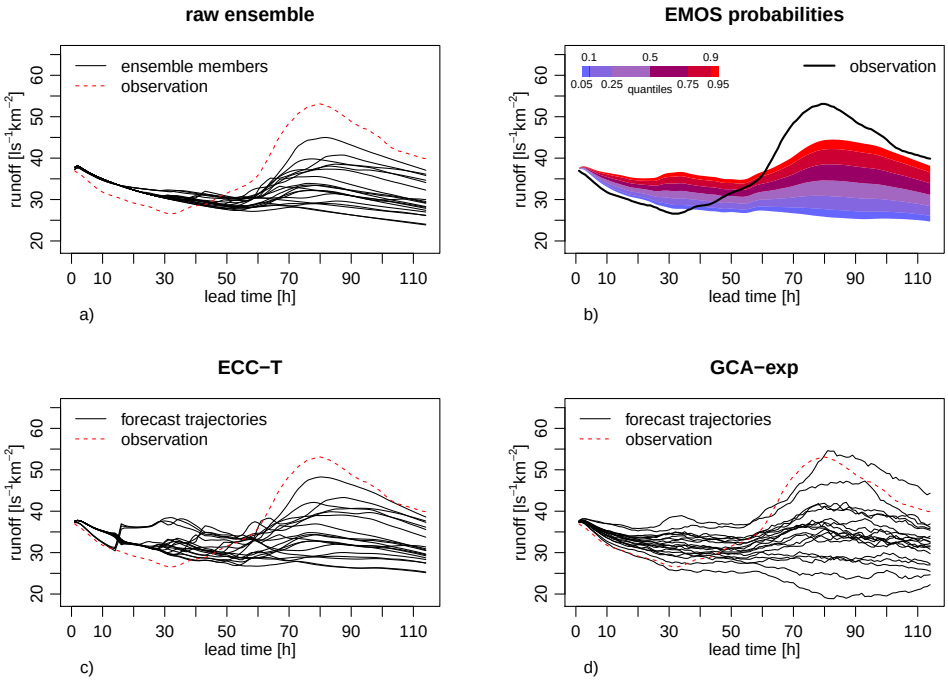


Figure B.15: Example of a high flow forecast for river Upper Rhine at Maxau issued on 16 July 2009. Figure taken from the supplemental material to Hemri et al. (2015).

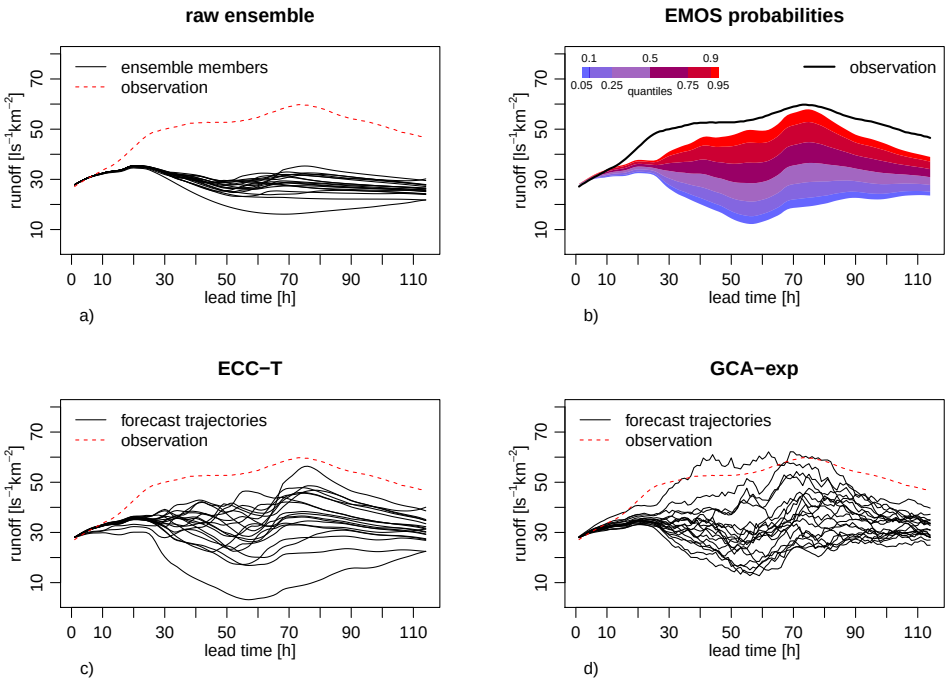


Figure B.16: Example of a high flow forecast for river Upper Rhine at Maxau issued on 7 December 2010. Figure taken from the supplemental material to Hemri et al. (2015).

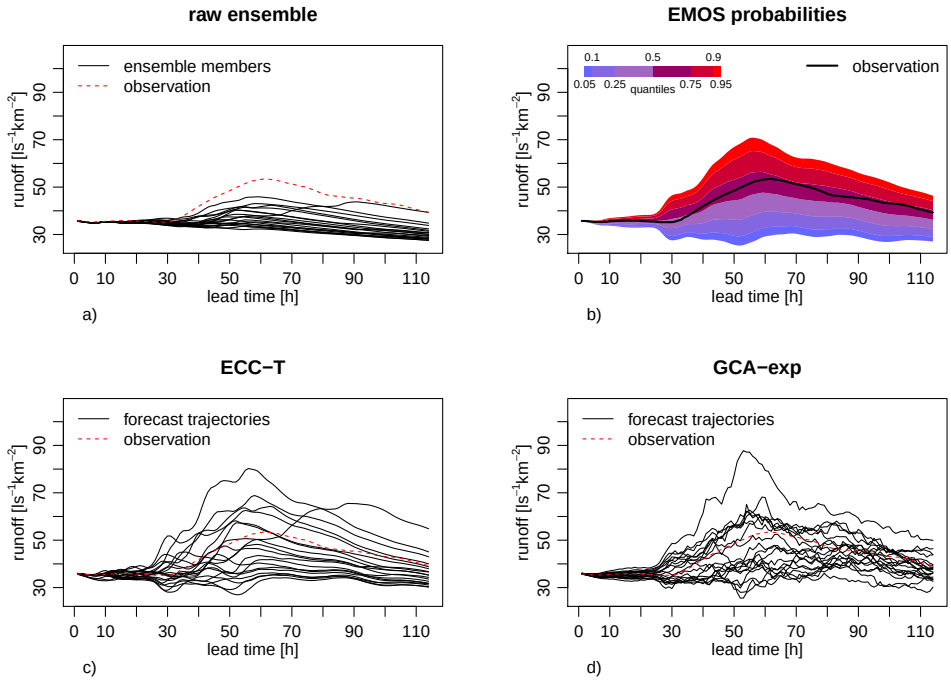


Figure B.17: Example of a high flow forecast for river Upper Rhine at Maxau issued on 12 January 2011. Figure taken from the supplemental material to Hemri et al. (2015).

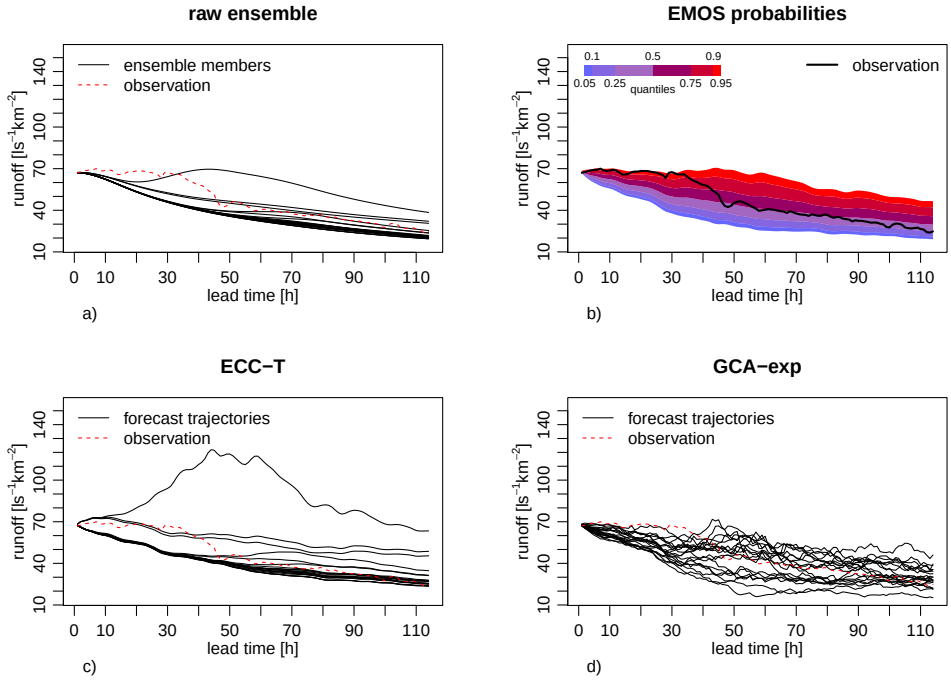


Figure B.18: Example of a high flow forecast for river Moselle at Trier issued on 10 December 2010. Figure taken from the supplemental material to Hemri et al. (2015).

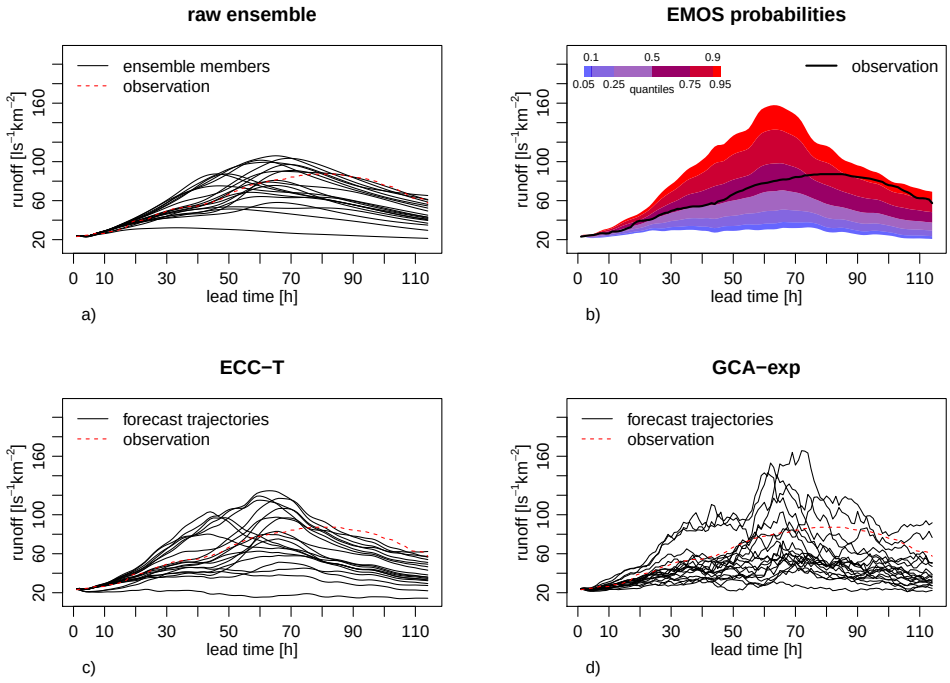


Figure B.19: Example of a high flow forecast for river Moselle at Trier issued on 21 December 2010. Figure taken from the supplemental material to Hemri et al. (2015).

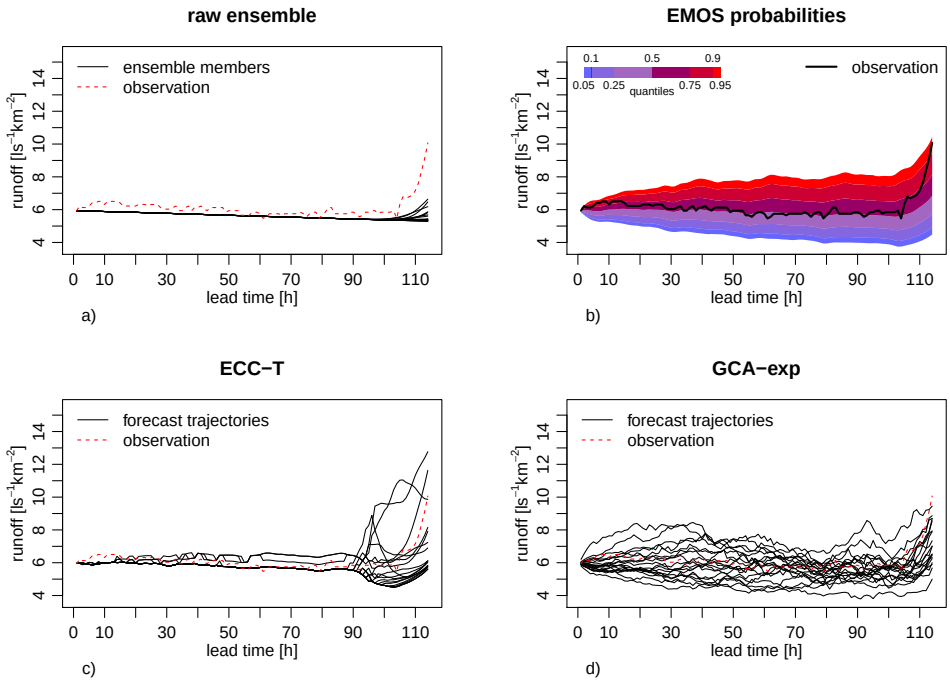


Figure B.20: Example of a high flow forecast for river Lahn at Kalkofen issued on 2 January 2010. Figure taken from the supplemental material to Hemri et al. (2015).

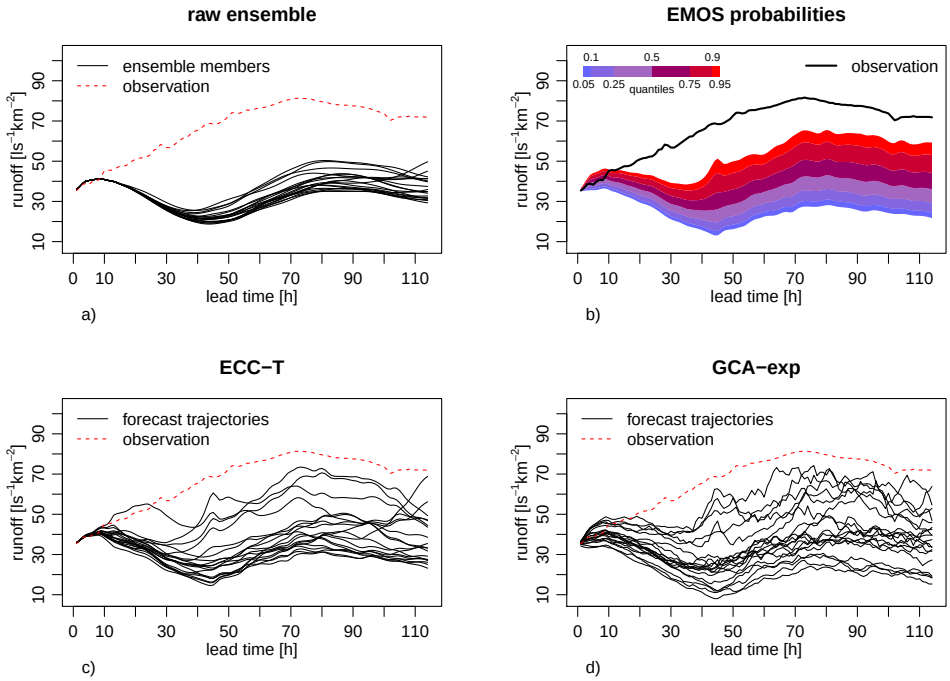


Figure B.21: Example of a high flow forecast for river Lahn at Kalkofen issued on 24 February 2010. Figure taken from the supplemental material to Hemri et al. (2015).

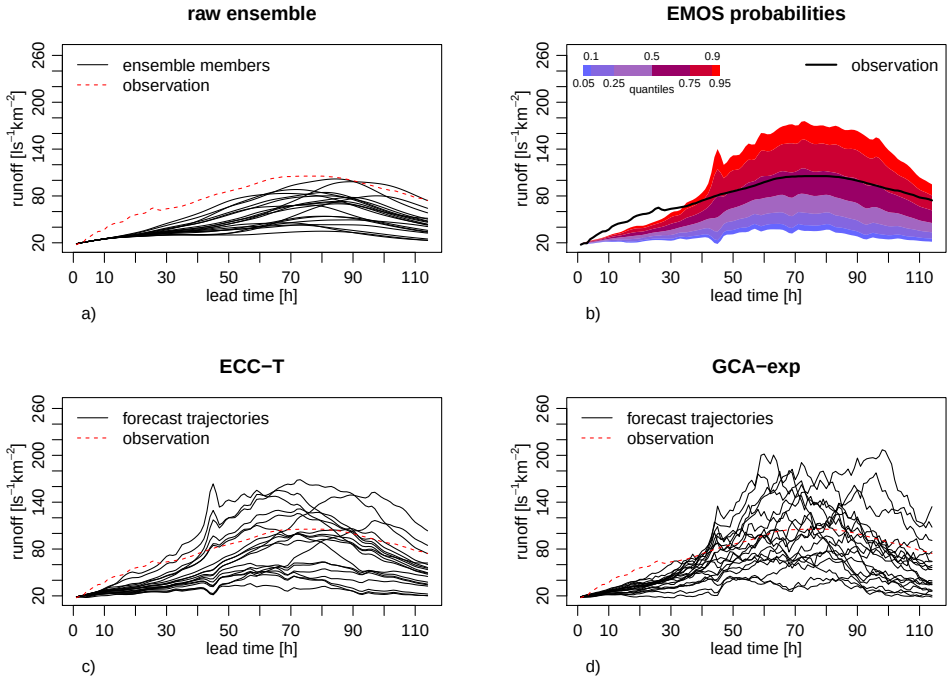


Figure B.22: Example of a high flow forecast for river Lahn at Kalkofen issued on 7 January 2011. Figure taken from the supplemental material to Hemri et al. (2015).

References

Addor, N., S. Jaun, F. Fundel, and M. Zappa (2011). An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrology and Earth System Sciences 15*, 2327–2347.

Agresti, A. and M. Kateri (2011). *Categorical Data Analysis*. Berlin, Heidelberg: Springer.

Ajami, N. K., Q. Duan, and S. Sorooshian (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research 43*(1), 1–19. DOI: 10.1029/2005WR004745.

American Meteorological Society (2015). Cloud cover. Glossary of Meteorology. http://glossary.ametsoc.org/wiki/Cloud_cover, last checked: 7 March 2016.

Ananth, C. V. and D. G. Kleinbaum (1997). Regression models for ordinal responses: A review of methods and applications. *Water Resources Research 26*(6), 1323–1333. DOI: 10.1093/ije/26.6.1323.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu (2002). Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather and Forecasting 17*(4), 783–799.

Arribas, A., K. B. Robertson, and K. R. Mylne (2005). Test of a poor man's ensemble prediction system for short-range probability forecasting. *Monthly Weather Review 133*(7), 1825–1839.

Atger, F. (1999). The skill of ensemble prediction systems. *Monthly Weather Review 127*(9), 1941–1953.

Baran, S. (2014). Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis 75*, 227–238. DOI: 10.1016/j.csda.2014.02.013.

Baran, S. and S. Lerch (2015). Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society 141*, 2289–2299.

Bartholmes, J. C., J. Thielen, M.-H. Ramos, and S. Gentilini (2009). The European flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences 13*, 141–153.

Ben Bouallègue, Z. (2013). Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting 28*(2), 515–524. DOI: 10.1175/WAF-D-12-00062.1.

Bergström, S. (1995). The HBV model. In V. Singh (Ed.), *Computer Models of Watershed Hydrology*. Water Resources Publications, Highlands Ranch, USA, pp. 443–476.

Boersen, P. and A. Weerts (2005). Automatic error correction of rainfall-runoff models in flood forecasting systems. In *Proceedings of Instrumentation and Measurement Conference 2005*, Volume 2, Ottawa, Canada, pp. 963–968.

Bonferroni, C. E. (1936). *Teoria Statistica delle Classi e Calcolo delle Probabilita*. Libreria internazionale Seeber.

Bougeault, P., Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, and S. Worley (2010). The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society 91*(8), 1059–1072. DOI: 10.1175/2010BAMS2853.1.

Box, G. and D. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society (Series B) 26*, 211–252.

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

Brown, J. D. and D.-J. Seo (2010). A non-parametric post-processor for bias-correction of hydrometeorological and hydrologic ensemble forecasts. *Journal of Hydrometeorology 11*(3), 642–665. DOI: 10.1175/2009JHM1188.1.

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart (2007). The new ECMWF VAREPS (variable resolution ensemble prediction system). *Quarterly Journal of the Royal Meteorological Society 133*(624), 681–695. DOI: 10.1002/qj.75.

Buizza, R., T. Petroliagis, T. 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.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208. DOI: 10.1137/0916069.

Canty, A. and B. Ripley (2014). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-13, http://CRAN.R-project.org/package=boot, last checked: 7 March 2016.

Carbone, R. E., J. D. Tuttle, D. A. Ahijevych, and S. B. Trier (2002). Inferences of predictability associated with warm season precipitation episodes. *Journal of the Atmospheric Sciences 59*(13), 2033–2056. DOI: 10.1175/1520-0469(2002)059<2033:IOPAWW>2.0.CO;2.

Chmielecki, R. M. and A. E. Raftery (2011). Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review 139*(5), 1626–1636. DOI: 10.1175/2010MWR3516.1.

Clark, M., S. Gangopadhyay, L. Rajagalopalan, and R. Wilby (2004). The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology 5*(1), 243–262. DOI: 10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Cloke, H. L. and F. Pappenberger (2009). Ensemble flood forecasting: a review. *Journal of Hydrology 375*, 613–626.

Coccia, G. and E. Todini (2011). Recent developments in predictive uncertainty assessment based on the model conditional processor. *Hydrology and Earth System Sciences 15*(10), 3253–3274. DOI: 10.5194/hess-15-3253-2011.

Coles, S., J. Bawa, J. Trenner, and P. Dorazio (2001). *An Introduction to Statistical Modeling of Extreme Values*, Volume 208. London: Springer.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society (Series A) 147*, 278–292. DOI: 10.2307/2981683.

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management 111*(2), 157–170. DOI: 10.1061/(ASCE)0733-9496(1985)111:2(157).

Dee, D. P., S. Uppala, and A. J. S. et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society 137*(656), 553–597. DOI: 10.1002/qj.828.

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

Diak, G. R., M. C. Anderson, W. L. Bland, J. M. Norman, J. M. Mecikalski, and R. M. Aune (1998). Agricultural management decision aids driven by real-time satellite data. *Bulletin of the American Meteorological Society 79*(7), 1345–1355. DOI: 10.1175/1520-0477(1998)079<1345:AMDADB>2.0.CO;2.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*(4), 863–883. DOI: 10.2307/2527342.

Diggle, P. J. and P. J. Ribeiro Jr (2007). *Model Based Geostatistics*. New York: Springer.

Dixon, H. G., M. Lagerlund, M. J. Spittal, D. J. Hill, S. J. Dobbinson, and M. A. Wakefield (2008). Use of sun-protective clothing at outdoor leisure settings from 1992 to 2002: serial cross-sectional observation survey. *Cancer Epidemiology Biomarkers & Prevention 17*, 428–434. DOI: 10.1158/1055-9965.EPI-07-0369.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A 57*, 234–252.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources 30*(5), 1371–1386. DOI: 10.1016/j.advwatres.2006.11.014.

Ebert, E. E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review 129*(10), 2461–2480.

Ehret, U. and E. Zehe (2011). Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences 15*(3), 877–896. DOI:10.5194/hess-15-877-2011.

Engeland, K. and I. Steinsland (2014). Probabilistic post processing models for flow forecasts for a system of catchments and several lead times. *Water Resources Research 50*(1), 182–197. DOI:10.1002/2012WR012757.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology 8*(6), 985–987. DOI: 10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

European Flood Awareness System (2014). EFAS concepts and tools. https://www.efas.eu/about-efas.html, last checked: 7 March 2016.

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir (2015). Spatial post processing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Monthly Weather Review 143*, 955–971.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications 15*, 19–24. DOI: 10.1002/met.45.

Fraley, C., A. E. Raftery, and T. Gneiting (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review 138*(1), 190–202. DOI: 10.1175/2009MWR3046.1.

Fraley, C., A. E. Raftery, J. M. Sloughter, T. Gneiting, and U. of Washington. (2015). *ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging*. R package version 5.1.1, http://CRAN.R-project.org/package=ensembleBMA, last checked: 7 March 2016.

Frey, S. and H. Holzmann (2015). A conceptual, distributed snow redistribution model. *Hydrology and Earth System Sciences 19*, 4517–4530. DOI: 10.5194/hess-19-4517-2015.

Friederichs, P. and T. L. Thorarinsdottir (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics 23*(7), 579–594.

Fundel, F. and M. Zappa (2011). Hydrological ensemble forecasting in mesoscale catchments: Sensitivity to initial conditions and value of reforecasts. *Water Resources Research 47*(W09520).

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology 298*, 222–241.

Ghelli, A. and F. Lalaurette (2000). Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter 87*, 9–17.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software 31*(7), 1–24. DOI: 10.18637/jss.v031.i07.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762. DOI: 10.1198/jasa.2011.r10138.

Gneiting, T. (2014). Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum, No. 719*, 30p.

Gneiting, T., F. Balabdoui, and A. E. Raftery (2007a). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society (Series B) 69*, 243–268. DOI: 10.1111/j.1467-9868.2007.00587.x.

Gneiting, T., M. G. Genton, and P. Guttorp (2007b). Geostatistical space-time models, stationarity, separability, and full symmetry. In B. Finkenstadt, L. Held, and V. Isham (Eds.), *Statistical Methods for Spatio-Temporal Systems*. Boca Raton: Chapman and Hall/CRC, 151–175.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*, 125–151. DOI: 10.1146/annurev-statistics-062713-085831.

Gneiting, T., K. Larson, and K. Westrick (2004). Development of next-generation wind energy forecast and optimization technologies. *Technical report*, University of Washington, Department of Statistics, Seattle, USA.

Gneiting, T., K. Larson, K. Westrick, G. M. Genton, and E. Aldrich (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method. *Journal of the American Statistical Association 101*(475), 968–979.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378. DOI: 10.1198/016214506000001437.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*(5), 1098–1118. DOI: 10.1175/MWR2904.1.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test 17*(2), 211–235. DOI: 10.1007/s11749-008-0114-x.

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

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society 132.621C*, 2925–2942. DOI: 10.1256/qj.05.235.

Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer (2012). Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society 138*(668), 1814–1827.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A 57*, 219–233.

Haiden, T., R. Forbes, M. Ahlgrimm, and A. Bozzo (2015). The skill of ECMWF cloudiness forecasts. *ECMWF Newsletter 143*, 14–19.

Haiden, T. and J. Trentmann (2015). Verification of cloudiness and radiation forecasts in the greater Alpine region. *Meteorologische Zeitschrift 25*, 3–15. DOI: 10.1127/metz/2015/0630.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review 129*, 550–560.

Hamill, T. M. (2012). Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Monthly Weather Review 140*(7), 2232–2252. DOI: 10.1175/MWR-D-11-00220.1.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker (2006). Reforecasts: an important dataset for improving weather predictions. *Bulletin of the American Meteorological Society 87*(1), 33.

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*(7), 2620–2632. DOI: http://dx.doi.org/10.1175/2007MWR2411.1.

Hamill, T. M., C. Snyder, and R. E. Morss (2000). A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Monthly Weather Review 128*(6), 1835–1851.

Hamill, T. M., J. S. Whitaker, and X. Wei (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review 132*(6), 1434–1447. DOI: http://dx.doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hemri, S. (2016). Multi-model combination and seamless prediction. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, and J. C. Schaake (Eds.), *Handbook of Hydrometeorological Ensemble Forecasting / in print*. Berlin: Springer-Verlag.

Hemri, S., F. Fundel, and M. Zappa (2013). Simultaneous calibration of ensemble river flow predictions over an entire range of lead-times. *Water Resources Research 49*(10), 6744–6755.

Hemri, S., T. Haiden, and F. Pappenberger (2016). Discrete post processing of total cloud cover ensemble forecasts. Accepted for publication in *Monthly Weather Review*.

Hemri, S., D. Lisniak, and B. Klein (2014a). Ermittlung probabilistischer Abflussvorhersagen unter Berücksichtigung zensierter Daten. *Hydrologie und Wasserbewirtschaftung 58*(2), 84–94. DOI: 10.5675/HyWa_2014,2_4.

Hemri, S., D. Lisniak, and B. Klein (2015). Multivariate post processing techniques for probabilistic hydrological forecasting. *Water Resources Research 51*(9), 7436–7451. DOI: 10.1002/2014WR016473.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden (2014b). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters 41*(24), 9197–9205. DOI: 10.1002/2014GL062472.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(2), 382–417.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers 116*, 770–808.

Iversen, T., A. Deckmyn, C. Santos, K. Sattler, J. B. Bremnes, H. Feddersen, and I.-L. Frogner (2011). Evaluation of 'GLAMEPS' – a proposed multimodel EPS for short range forecasting. *Tellus A 63*, 513–530.

Junk, C., L. D. Monache, and S. Alessandrini (2015). Analog-based ensemble model output statistics. *Monthly Weather Review 143*, 2909–2917.

Kennedy, E. J. (1984). Discharge ratings at gaging stations. In *Techniques of Water-Resources Investigations*, U.S. Geological Survey, book 3, chap. A10, 59 p. http://pubs.usgs.gov/twri/twri3-a10/, last checked: 7 March 2016.

Kober, K., G. C. Craig, and C. Keil (2014). Aspects of short-term probabilistic blending in different weather regimes. *Quarterly Journal of the Royal Meteorological Service 140*, 1179–1188.

Kober, K., G. C. Craig, C. Keil, and A. Dörnbrack (2012). Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Quarterly Journal of the Royal Meteorological Service 138*, 755–768.

Köhler, M. (2005). Improved prediction of boundary layer clouds. *ECMWF Newsletter 104*, 18–22.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science 285*(5433), 1548–1550.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate 13*(23), 4196–4216.

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research 35*(9), 2739–2750. DOI: 10.1029/1999WR900099.

Krzysztofowicz, R. (2002). Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology 268*(1–4), 16–40. DOI: 10.1016/S0022-1694(02)00106-3.

Krzysztofowicz, R. and K. S. Kelly (2000). Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resources Research 36*(11), 3265–3277. DOI: 10.1029/2000WR900108.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics 17*(3), 1217–1241.

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

Lerch, S. and S. Baran (2016). Similarity-based semi-local estimation of EMOS models. Accepted for publication in the *Journal of the Royal Statistical Society (Series C)*. Preprint available at http://arxiv.org/abs/1509.03521, last checked: 7 March 2016.

Lerch, S. and T. L. Thorarinsdottir (2013). Comparing nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus A 65*. 21206, DOI: 10.3402/tellusa.v65i0.21206.

Lindström, G., B. Johansson, M. Persson, M. Gardelin, and S. Bergström (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology 201*(1-4), 272–288. DOI: 10.1016/S0022-1694(97)00041-3.

Madadgar, S., H. Moradkhani, and D. Garen (2012). Towards improved post processing of hydrologic forecast ensembles. *Hydrological Processes 28*(1), 104–122. DOI: 10.1002/hyp.9562.

Majewski, D., D. Liermann, P. Prohl, B. Ritter, M. Buchhold, T. Hanisch, G. Paul, W. Wergen, and J. Baumgardner (2002). The operational global icosahedral-hexagonal gridpoint model GME: description and high-resolution tests. *Monthly Weather Review 130*(2), 319–338. DOI: 10.1175/1520-0493(2002)130<0319:togihg>2.0.co;2.

Majewski, D., D. Liermann, and B. Ritter (2012). Kurze Beschreibung des Globalmodells GME (20 km / L60) und seiner Datenbanken auf dem Datenserver des DWD. *Technical report*, Deutscher Wetterdienst (DWD), Offenbach, Germany.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica 13*, 245–259.

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

McClung, D. M. (2002). The elements of applied avalanche forecasting part II: The physical issues and the rules of applied avalanche forecasting. *Natural Hazards 26*(2), 131–146. DOI: 10.1023/A:1015604600361.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society (Series B) 42*, 109–142.

McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.

McLeod, A. (2011). *Kendall: Kendall rank correlation and Mann-Kendall trend test*. R package version 2.2, http://CRAN.R-project.org/package=Kendall, last checked: 31.10.2014.

Meißner, D. and S. Rademacher (2010). Die verkehrsbezogene Wasserstandsvorhersage für die Bundeswasserstraße Rhein. *KW Korrespondenz Wasserwirtschaft 3*(9), 485–491.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review 142*(8), 3003–3014. DOI: 10.1175/MWR-D-13-00355.1.

Meyer, M. C. (2012). Constrained penalized splines. *The Canadian Journal of Statistics 40*(1), 190–206. `R` code available at `http://www.stat.colostate.edu/~meyer/penspl.htm`, last checked: 7 March 2016.

Mittermaier, M. (2012). A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Quarterly Journal of the Royal Meteorological Society 138*(668), 794–1807. DOI: 10.1002/qj.1918.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society 122*(529), 73–119. DOI: 10.1002/qj.49712252905.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart (2011). The new ECMWF seasonal forecast system (System 4). *ECMWF Technical Memorandum, No. 656*, 51p.

Montani, A., D. Cesari, C. Marsigli, and T. Paccagnella (2011). Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A 63*(3), 605–624. DOI: 10.3402/tellusa.v63i3.15816.

Mori, U., A. Mendiburu, and J. A. Lozano (2014). *TSdist: Distance Measures for Time Series data.* R package version 1.2.

Mudelsee, M. (2007). Long memory of rivers from spatial aggregation. *Water Resources Research 43*(1). DOI: 10.1029/2006WR005721.

Murphy, A. H. (1969). On the "ranked probability score". *Journal of Applied Meteorology 8*(6), 988–989.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology 12*(1), 215–223. 10.1175/1520-0450(1973)012<0215:HASSFP>2.0.CO;2.

Nachtnebel, H. P., S. Baumung, and W. Lettl (1993). Abflussprognosemodell für das Einzugsgebiet der Enns und Steyr. *Technical report*, Institute of Water Management, Hydrology and Hydraulic Engineering, University of Natural Resources and Applied Life Sciences Vienna, Austria.

Ouyang, R., L. Ren, W. Cheng, and C. Zhou (2010). Similarity search and pattern discovery in hydrological time series data mining. *Hydrological Processes 24*(9), 1198–1210. DOI: 10.1002/hyp.7583.

Palmer, T. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics 63*, 71–116.

Palmer, T. N., F. J. Doblas-Reyes, R. Hagedorn, A. Alessandri, S. Gualdi, U. Andersen, H. Feddersen, P. Cantelaube, J.-M. Terres, M. Davey, R. Graham, P. Délécluse, A. Lazar, M. Déqué, J.-F. Guérémy, E. Díez, B. Orfila, M. Hoshen, A. P. Morse, N. Keenlyside, M. Latif, E. Maisonnave, P. Rogel, V. Marletto, and M. C. Thomson (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society 85*, 853–872.

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell (2008). Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bulletin of the American Meteorological Society 89*, 459–470.

Panofsky, H. A. and G. W. Brier (1958). *Some Applications of Statistics to Meteorology.* The Pennsylvania State University. 224 pp.

Pappenberger, F., A. Ghelli, R. Buizza, and K. Bódis (2009). The skill of probabilistic precipitation forecasts under observational uncertainties within the Generalized Likelihood Uncertainty Estimation framework for hydrological applications. *Journal of Hydrometeorology 33*, 807–819.

Pappenberger, F., J. Thielen, and M. D. Medico (2010). The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrological Processes 25*, 1091–116. DOI: 10.1002/hyp.7772.

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*(637), 2029–2050. DOI: 10.1002/qj.334.

Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012). Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. *Water Resources Research 48*(3), 1–18. DOI: 10.1029/2011WR011116.

Pelland, S., G. Galanis, and G. Kallos (2013). Solar and photovoltaic forecasting through post processing of the Global Environmental Multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications 21*(3), 284–296. DOI: 10.1002/pip.1180.

Pinson, P. and R. Girard (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy 96*, 12–20. DOI: 10.1016/j.apenergy.2011.11.004.

Pinson, P. and R. Hagedorn (2012). Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteorological Applications 19*(4), 484–500. DOI: 10.1002/met.283.

Pinson, P. and J. Tastu. Discrimination ability of the energy score. *Technical report*, Technical University of Denmark.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/, last checked: 7 March 2016.

Rabiner, L. and B. H. Juang (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(2), 1155–1174. DOI: 10.1175/MWR2906.1.

Reggiani, P., M. Renner, A. H. Weerts, and P. A. H. J. M. V. Gelder (2009). Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system. *Water Resources Research 45*. DOI: 10.1029/2007WR006758.

Ribeiro Jr, P. J. and P. J. Diggle (2001). geoR: a package for geostatistical analysis. *R-News 1*(2), 15–18.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society 127*, 2473–2489.

Richardson, D. S., J.-R. Bidlot, L. Ferranti, T. Haiden, T. Hewson, M. Janousek, F. Praters, and F. Vitart (2013). Evaluation of ECMWF forecasts, including 2012-2013 upgrades. *ECMWF Technical Memorandum, No. 710*, 55p.

Richardson, D. S., R. Buizza, and R. Hagedorn (2005). TIGGE – First Final Report. *WMO/TD-No.1273*. http://www.wmo.int/pages/prog/arep/wwrp/new/thorpex_publications.html, last checked: 7 March 2016.

Richardson, D. S., S. Hemri, K. Bogner, T. Gneiting, T. Haiden, F. Pappenberger, and M. Scheuerer (2015). Calibration of ECMWF forecasts. *ECMWF Newsletter 142*, 12–16.

Rings, J., J. A. Vrugt, G. Schoups, J. A. Husman, and H. Vereecken (2012). Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiment. *Water Resources Research 48*. DOI: 10.1029/2011WR011607.

Ripley, B. and W. Venables (2014). *Feed-forward Neural Networks and Multinomial Log-Linear Models*. R-package version 7.3-8, http://www.stats.ox.ac.uk/pub/MASS4/, last checked: 7 March 2016.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics 23*, 470–472.

Roulin, E. and S. Vannitsem (2012). Post processing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly Weather Review 140*(3), 874–888. DOI: 10.1175/MWR-D-11-00062.1.

Ruiz, J. J. and C. Saulo (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: sensitivity to calibration methods. *Meteorological Applications 19*(3), 302–313. DOI: 10.1002/met.286.

Sakoe, H. and S. Chiba (1978). Dynamic-programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*(1), 43–49. DOI: 10.1109/TASSP.1978.1163055.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science 28*, 616–640. DOI: 10.1214/13-STS443.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society 140*(680), 1086–1096. DOI: 10.1002/qj.2183.

Scheuerer, M. and L. Büermann (2014). Spatially adaptive post processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society (Series C) 63*(3), 405–422.

Scheuerer, M. and T. M. Hamill (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review 143*(4), 1321–1334. DOI: 10.1175/MWR-D-14-00269.1.

Scheufele, K., K. Kober, G. C. Craig, and C. Keil (2014). Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble. *Meteorological Applications 21*, 230–240.

Schlather, M. (1999). An introduction to positive definite functions and to unconditional simulation of random fields. *Technical Report ST 99-10*. Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Schmeits, M. J. and K. J. Kok (2010). A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review 138*(11), 4199–4211. DOI: 10.1175/2010MWR3285.1.

Schulz, J.-P. and U. Schättler (2011). Kurze Beschreibung des Lokal-Modells Europa COSMO-EU (LME) und seiner Datenbanken auf dem Datenserver des DWD. *Technical Report*, Deutscher Wetterdienst (DWD), Offenbach, Germany.

Shi, X., A. W. Wood, and D. P. Lettenmaier (2008). How essential is hydrologic model calibration to seasonal streamflow forecasting? *Journal of Hydrometeorology 9*(6), 1350.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris 8*, 229–231.

Steppeler, J., G. Doms, and G. Adrian (2002). Das Lokal-Modell LM. *Promet 27*(3/4), 123–128.

Taylor, J. W. and R. Buizza (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting 19*(1), 57–70. DOI: 10.1016/S0169-2070(01)00123-6.

Thielen, J., J. C. Bartholmes, M.-H. Ramos, and A. de Roo (2009). The European Flood Alert System – Part 1: Concept and development. *Hydrology and Earth System Sciences 13*, 125–140.

Thorarinsdottir, T. L. and T. Gneiting (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society (Series A) 173*, 371–388. DOI: 10.1111/j.1467-985X.2009.00616.x.

Thorarinsdottir, T. L., M. Scheuerer, and C. Heinz (2014). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*. DOI: 10.1080/10618600.2014.977447.

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management 6*(2), 123–137. DOI: 10.1080/15715124.2008.9635342.

Todini, E., G. Coccia, and E. Ortiz (2015). On the proper use of ensembles for predictive uncertainty assessment. *Geophysical Research Abstracts 17*. EGU2015-10365.

van der Waerden, B. L. (1952). Order tests for two-sample problem and their power I. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 55*, 453–458.

van der Waerden, B. L. (1953a). Order tests for two-sample problem and their power II. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 56*, 303–310.

van der Waerden, B. L. (1953b). Order tests for two-sample problem and their power III. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 56*, 311–316.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S. Fourth Edition*. New York: Springer.

Wacker, S., J. Gröbner, C. Zysset, L. Diener, P. Tzoumanikas, A. Kazantzidis, L. Vuilleumier, R. Stöckli, S. Nyeki, and N. Kämpfer (2015). Cloud observations in Switzerland using hemispherical sky cameras. *Journal of Geophysical Research: Atmospheres 120*(2), 695–707. DOI: 10.1002/2014JD022643.

Walker, S. H. and D. B. Duncan (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54*(1-2), 167–179. DOI: 10.1093/biomet/54.1-2.167.

Weerts, A., H. C. Winsemius, and J. S. Verkade (2011). Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrology and Earth System Sciences 15*(1), 255–265. DOI: 10.5194/hess-15-255-2011.

Weinheimer, A., L. A. Smith, and K. Judd (2005). A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. *Tellus A 57*, 265–279.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences (3rd ed.)*. Academic press.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications 16*(3), 361–368. DOI: 10.1002/met.134.

Wilks, D. S. and T. M. Hamill (2007). Comparisons of ensemble-MOS methods using GFS forecasts. *Monthly Weather Review 135*, 2379–2390. DOI: 10.1175/MWR3402.1.

Wood, A. W. and D. P. Lettenmaier (2008). An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical Research Letters 35*. DOI: 10.1029/2008GL034648.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier (2002). Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres 107*(D20), 4429. DOI: 10.1029/2001JD000659.

Ye, Q. Z. and S. S. Chen (2013). The ultimate meteorological question from observational astronomers: how good is the cloud cover forecast? *Monthly Notices of the Royal Astronomical Society 128*(4), 3288–3294. DOI: 10.1093/mnras/sts278.

Yuan, X., E. Wood, and M. Liang (2014). Developing a seamless hydrologic forecast system: Integrating weather and climate prediction. *Geophysical Research Abstracts 16*(EGU2014-2268).

Zappa, M., S. Jaun, U. Germann, A. Walser, and F. Fundel (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research 100*, 246–262.

Ziehmann, C. (2000). Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus A 52*, 280–299.

List of Figures

List of Tables