Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales[☆]

J.S. Verkade[a,b,c,*], J.D. Brown[d], P. Reggiani[a,e], A.H. Weerts[a,f]

[a] *Deltares, PO Box 177, 2600 MH Delft, The Netherlands, +31.88.335.8348 (tel), +31.88.335.8582 (fax)*
[b] *Ministry of Infrastructure and the Environment, Water Management Centre of The Netherlands, River Forecasting Service, Lelystad, The Netherlands*
[c] *Delft University of Technology, Delft, The Netherlands*
[d] *Hydrologic Solutions Limited, Southampton, United Kingdom*
[e] *RWTH Aachen University, Aachen, Germany*
[f] *Wageningen University and Research Centre, Hydrology and Quantitative Water Management Group, Wageningen, The Netherlands*

Abstract

The ECMWF temperature and precipitation ensemble reforecasts are evaluated for biases in the mean, spread and forecast probabilities, and how these biases propagate to streamflow ensemble forecasts. The forcing ensembles are subsequently post-processed to reduce bias and increase skill, and to investigate whether this leads to improved streamflow ensemble forecasts. Multiple post-processing techniques are used: quantile-to-quantile transform, linear regression with an assumption of bivariate normality and logistic regression. Both the raw and post-processed ensembles are run through a hydrologic model of the river Rhine to create streamflow ensembles. The results are compared using multiple verification metrics and skill scores: relative mean error, Brier skill score and its decompositions, mean continuous ranked probability skill score and its decomposition, and the ROC score. Verification of the streamflow ensembles is performed at multiple spatial scales: relatively small headwater basins, large tributaries and the Rhine outlet at Lobith. The streamflow ensembles are verified against simulated streamflow, in order to isolate the effects of biases in the forcing ensembles and any improvements therein. The results indicate that the forcing ensembles contain significant biases, and that these cascade to the streamflow ensembles. Some of the bias in the forcing ensembles is unconditional in nature; this was resolved by a simple quantile-to-quantile transform. Improvements in conditional bias and skill of the forcing ensembles vary with forecast lead time, amount, and spatial scale, but are generally moderate. The translation to streamflow forecast skill is further muted, and several explanations are considered, including limitations in the modelling of the space-time covariability of the forcing ensembles and the presence of storages.

Keywords: bias-correction, post-processing, ensemble forecasting, uncertainty estimation, verification, Rhine

1. Introduction

Hydrologic forecasts are inherently uncertain. Uncertainties originate from the forcing data and from the initial conditions, the model structure and its parameters. Estimating the uncertainties in hydrologic forecasts yields probabilistic forecasts that form one input to risk-based decision making. While “best practice” for using these probabilistic forecasts attracts ongoing debate, there is good evidence to suggest that probabilistic forecasts could improve decision-making if used appropriately (e.g. Krzysztofowicz, 2001; Raiffa and Schlaifer, 1961; Ramos et al., 2012; Todini, 2004; Verkade and Werner, 2011).

Hydrologic models are often forced with the output from numerical weather prediction (NWP) models. As hydrologic models are sensitive to the forcing inputs, and meteorological forecasts often contain significant biases and uncertainties, the forcing data is typically an important source of bias and uncertainty in streamflow forecasting. Meteorological ensemble prediction systems (EPS) are increasingly used in hydrologic prediction (see, for example, Cloke and Pappenberger 2009 for an overview of ensemble use in flood forecasting). Examples of meteorological EPS include the National Centers for Environmental Prediction’s Global Ensemble Forecast System (GEFS; Hamill and Whitaker 2006), the UK Met Office’s Global and Regional Ensemble Prediction System (MOGREPS; Bowler et al. 2008; Schellekens et al. 2011) and the European Centre for Medium-Range Weather Forecasts’ Ensemble Prediction System (ECMWF-EPS; Buizza et al. 2007).

Due to limitations of the models and associated data, forecasts from meteorological EPS generally contain biases in the mean, spread and higher moments of their forecast distributions. These biases are manifest at temporal and spatial scales that are relevant to hydrologic predic-

[☆] Supplemental information related to this paper is available at an online data repository via `http://dx.doi.org/10.4121/uuid:56637037-8197-472b-b143-2f87adf49abc`

[*] Corresponding author

Email address: `jan.verkade@deltares.nl` (J.S. Verkade) tion. The information content in the raw forcing may contain valuable information for post-processing. A variety of techniques may be used for this, including techniques that use single-valued predictors, such as the ensemble mean of the forcing forecast (e.g. Kelly and Krzysztofowicz, 2000; Reggiani and Weerts, 2008b; Zhao et al., 2011), and techniques that use additional moments or all ensemble members, as well as auxiliary variables.

Biases in forcing ensembles propagate through the hydro-meteorological system and may, therefore, introduce biases into the streamflow predictions. Biases in streamflow forecasts are often removed through statistical post-processing[1] where, based on the historical performance of the forecasting system, operational streamflow forecasts are statistically corrected in real-time (e.g. Bogner and Pappenberger, 2011; Brown and Seo, 2013; Krzysztofowicz, 1999; Reggiani and Weerts, 2008a; Todini, 2008; Weerts et al., 2011). This correction may lump together the hydrologic and meteorological uncertainties or factor them separately (Brown and Seo, 2013). The two sources of uncertainty are lumped together by calibrating the streamflow post-processor on observed streamflow. The hydrologic uncertainties are factored out by calibrating the streamflow post-processor on simulated streamflow, i.e. on streamflow predictions with observed forcing (Seo et al., 2006; Zhao et al., 2011). In both cases, the streamflow forecasts may benefit from post-processing of the forcing forecasts. However, in separately accounting for the hydrologic uncertainties (the first case), it is assumed that the meteorological uncertainties and biases have been adequately addressed. In contrast, corrections to the streamflow should indirectly account for the meteorological biases and uncertainties if the forcing and hydrologic uncertainties are lumped together into a streamflow postprocessor.

Important questions remain about the combined benefits of forcing and streamflow post-processing in this context. For example, lumping together the forcing and streamflow uncertainties may lead to strongly heterogeneous behaviours that are difficult to model statistically. However, post-processing of forcing forecasts is generally complex and resource intensive, requiring statistical models of temporal, spatial and cross-variable relationships to which streamflow is often sensitive and for which sample sizes may be limited; in short, forcing bias correction may leave substantial residual biases and invoke imperfect models of space-time covariability.

Indeed, initial attempts to address this issue have been reported in the scientific literature. Kang et al. (2010) focused on the reduction of uncertainties by applying post-processing to predicted forcings, to predicted streamflow and both. In their study, post-processed ensemble members were re-ordered using the Schaake Shuffle prior to being used in the hydrologic and hydrodynamic models. The Schaake Shuffle aims to capture spatio-temporal patterns in the observed meteorological forcings that are lost following post-processing of the marginal distributions. The authors found that the forecasts were most skillful when combining post-processing of the forcings with post-processing of the streamflow forecasts. However, they also note that post-processing of the streamflow forecasts more effectively reduced the total uncertainty than post-processing the forcings alone. Clearly, this will depend on the relative importance of the forcing and hydrologic uncertainties in any given basin.

Zalachori et al. (2012) compared the skill of, and biases, in ensemble streamflow forecasts that were produced using different combinations of forcing and streamflow post-processing. Post-processing of meteorological forcings was performed by dressing the ensemble members with 50 analog scenarios that naturally included appropriate space-time relationships. They found that, while post-processing the forcings increased the skill of the forcing ensembles, there was little improvement in the skill of the streamflow ensembles. Also, those improvements were obscured by the effect of streamflow post-processing.

Similarly,Yuan and Wood (2012) explored the benefits of post-processing of forcing ensembles versus post-processing of streamflow ensembles, but in a different context, namely that of seasonal forecasting. They found that both post-processing of forcings and post-processing of streamflow adds skill, and when techniques are combined, skill is highest.

Several techniques have been proposed for reducing bias in forcing forecasts (Hamill, 2012). These techniques use past forecasts and observations (and possibly auxiliary variables) to estimate the parameters of a statistical model that is subsequently applied in real-time to estimate the "true" (unbiased) probability distribution of the forecast variable, conditionally upon the raw forecast (and any other predictors). Techniques include linear regression with an assumption of joint normality (e.g. Gneiting et al., 2005; Hagedorn et al., 2008; Wilks, 2006), logistic regression (Hamill et al., 2008; Wilks, 2006), quantile regression (Bremnes, 2004) and indicator co-Kriging (Brown and Seo, 2010, 2013), among others. Unsurprisingly, Wilks and Hamill (2007) conclude that no single post-processing technique is optimal for all applications.

Statistical correction of numerical weather forecasts requires a long historical record of forecasts and observations, from which the joint distribution can be estimated with reasonably small sampling uncertainty and bias. Unless explicitly accounting for non-stationarity with additional model parameters, the joint distribution should be relatively homogeneous in time. Forecasting systems, however, generally improve over time, rendering archived operational forecasts inhomogeneous. In contrast, weather

[1] In this paper, we use the term post-processing to indicate reduction of biases and/or estimation of uncertainties using statistical techniques that are applied subsequently to a model run. As such, post-processing is synonymous with bias-correction, forecast calibration, statistically correcting, and preprocessing. In hydrology, the term preprocessing is sometimes used to indicate the post-processing of meteorological forcings prior to being used in a hydrologic model.

forecasts that are retrospectively generated with a fixed numerical model ("reforecasts" or "hindcasts"), provide a reasonable platform for statistically correcting weather forecasts (Hamill et al., 2006). Available reforecast datasets include the ECMWF-EPS (Hagedorn, 2008), GFS (Hagedorn et al., 2008; Hamill and Whitaker, 2006; Hamill et al., 2008), and the more recent GEFS, for which hindcasts were recently completed (Hamill et al., 2013) and TIGGE (Hamill, 2012).

The extent to which the skill of, and biases in, streamflow forecasts can be improved through post-processing of the forcing ensembles, separately or together with streamflow post-processing, is an ongoing question and the focus of this paper. For example, these issues must be explored in basins with different hydrologic characteristics and for which the total uncertainties comprise different contributions from the meteorologic and hydrologic uncertainties, including a mixture of headwater and downstream basins. First, we evaluate the biases in the forcing ensembles at the scales used to force the hydrologic models, and how these biases translate into the streamflow ensemble forecasts. Secondly, a number of bias-correction techniques are applied to the temperature and precipitation ensembles. The post-processed forcing ensembles are used to drive the hydrologic models, which are then evaluated for any reduction in bias and increase in skill associated with the forcing post-processing. These post-processing techniques include the unconditional quantile-to-quantile transform (a correction to the forecast climatology) as well as conditional techniques such as linear regression in the bivariate normal framework and logistic regression. The streamflow ensembles are evaluated at multiple spatial scales and, crucially, by verifying against simulated streamflows (predictions made with observed forcings), in order to isolate the contribution of the forcing biases and uncertainties to the streamflow forecasts.

The structure of the manuscript is as follows. The Materials and Methods section describes (i) the techniques that have been used for post-processing of forcing ensembles, (ii) the study basin, (iii) the models and data that are used and (iv) a detailed setup of the different experiments. The results are presented in (Section 3) and subsequently discussed in (Section 4). Finally, some conclusions are drawn together with suggestions for future studies (Section 5).

2. Materials and Methods

2.1. Post-processing techniques

Several techniques were used to post-process the temperature and precipitation ensembles. Temperature ensemble forecasts were post-processed using the quantile-to-quantile transform and, separately, using linear regression. For precipitation, the quantile-to-quantile transform was used, as well as logistic regression. A brief description of each technique is provided below; more details can be found in AppendixA.

The quantile-to-quantile transform (QQT, sometimes also called Quantile Mapping or cdf-matching, e.g. Brown and Seo 2013; Hashino et al. 2007; Madadgar et al. 2012; Wood et al. 2002) is an unconditional technique insofar as the unconditional climatology of the forecasts is re-mapped to the unconditional climatology of the observations. QQT is not expected to provide post-processed ensembles that are equally skilful as those resulting from a conditional correction. However, the skill of a conditional correction may largely stem from an improvement in forecast climatology and an unconditional correction provides a valuable baseline for a more complex, conditional correction.

The conditional post-processing techniques are often applied in similar ways. For each of the forcing variables, the post-processor is configured for each lead time and each location (basin-averaged quantity) separately. A distribution of the predictand Y (observed temperature or precipitation) is sought, conditional upon a vector of predictors $\mathbf{X} = X_1, \ldots, X_m$. In this case, the predictors comprise the five (possibly biased) ensemble members of the raw forecast.

For post-processing temperature ensemble predictions, the observed and forecast temperatures are frequently assumed joint normally distributed. Linear regression is then used to estimate the mean and spread (and hence full probability distribution) of the observed variable conditionally upon the predictors (Gneiting et al., 2005; Hagedorn et al., 2008; Wilks, 2006).

Predictive distributions of precipitation are non-Gaussian (e.g. Hamill et al., 2008), and threshold-based or "non-parametric" techniques are often applied, although a meta-Gaussian approach is also possible (Wu et al., 2011). Precipitation forecasts are often biased conditionally upon observed precipitation amount (a so-called Type-II conditional bias), with overestimation of smaller observed precipitation and underestimation of larger observed precipitation. These amounts are typically important for practical applications of hydrologic forecasts (e.g. for drought and flood forecasting; see Brown and Seo 2013). Logistic regression is a common approach for post-processing of precipitation forecasts and is known to perform reasonably well in a variety of contexts (e.g. Hamill et al., 2008; Schmeits and Kok, 2010; Wilks, 2006). The technique involves estimating the probability of not-exceeding several discrete thresholds, for which the parameters of the logistic regression may be estimated separately at each threshold (standard logistic regression) or fixed across all thresholds (Wilks, 2009). In estimating the parameters separately at each threshold, the cumulative probabilities are not guaranteed to be valid in combination, and some post-correction smoothing is typically required.

A potential problem with statistically post-processing temperature and precipitation forecasts separately at each of multiple forecast lead times and locations is that space-time covariability is not adequately captured. For hydrologic applications, the space-time covariability of the forcing is important as the hydrologic model integrates the forcing both in time and in space (Clark et al., 2004).

In order to introduce appropriate space-time covariability into the post-processed forcing ensembles, the so-called "Schaake shuffle" was used here (Clark et al., 2004). For each ensemble trace, a corresponding observed time-series was obtained from the same start date in a randomly chosen historical year. The ensemble members at each forecast lead time were then assigned the same rank positions as the observations from the corresponding (relative) times in their associated historical years. The Schaake shuffle introduces (observed) rank correlations to the forecast ensemble members on the basis that spatial and temporal covariability will lead to ensemble members at nearby locations and proximate times having similar ranks within their own probability distributions. The Schaake shuffle does not, however, capture this space-time covariability conditionally upon the state of the atmosphere at the forecast issue time. Rather, it introduces space-time covariability conditionally upon forecast issue date alone (as formulated in Clark et al. 2004). Clearly, other implementations are possible, such as preservation of the rank order-relations in the raw forecasts.

2.2. Study basin: Rhine

The river Rhine runs from the Swiss Alps along the French-German border, through Germany and enters the Netherlands near Lobith, which is often considered the outflow. At Lobith, the basin area equals approx. 160,000 km^2. Snow and snowmelt have a large effect on the river Rhine's temporal streamflow patterns. During spring and early summer, more than half of the river's flow at the outlet originates from snowmelt in the Swiss Alps. Figure 1 shows the basin location, elevations and the gauged outlets of tributaries that were used in this study; the three different symbols used for the gauging stations coincide with the three spatial scales used in the analysis.

Clearly, the quality of the streamflow predictions at downstream locations is affected by the quality of the streamflow predictions at upstream locations. Ensemble streamflow predictions are therefore analysed at three spatial scales: (i) 43 outlets of basins that each have a contributing area of less than 2500 km^2; in the remainder of this manuscript, these are referred to as headwater basins (ii) four outlets of relatively large Rhine tributaries: the Main, Moselle, Neckar and Swiss Rhine, and (iii) the outlet of the river Rhine, at Lobith. Some summary statistics of the magnitudes of the contributing areas of these outlets are shown in Table 1.

Figure 2 shows the non-exceedence climatological probabilities of observed daily mean temperature, daily total precipitation and daily averaged streamflow for the three spatial scales used in the analysis. Both the "tributaries" and the "headwater" scales comprise of multiple outlets (four and 43 respectively). For these scales, the thick line designates the median location, and the thin lines designate the 10th and 90th percentiles. In the case of the four main tributaries, determining the quantiles required linear interpolation between four available data points.

Determination of temperature and precipitation at larger spatial scales has a modulating effect on extreme values of temperature and precipitation. The relatively fat tail of precipitation over the four tributaries originates from relatively high precipitation levels over the Swiss Rhine. As none of the headwater basins considered are located in that tributary basin, this fat tail is not observed in the curve for the smaller, headwater basins.

2.3. Models and data

For the temporal and areal aggregation of ensemble forcing forecasts and corresponding observations, and for retrospective generation of streamflow predictions, a Delft-FEWS forecast production system (Werner et al., 2012) was used. The system is an adapted version of the forecast production system FEWS Rivers, which is used by the Water Management Centre of the Netherlands for real-time forecasting of streamflow and water levels in the Rhine and Meuse rivers.

The system contains an implementation of the HBV rainfall-runoff model (Bergström and Singh, 1995). This is a semi-lumped, conceptual hydrologic model, which includes a routing procedure of the Muskingum type. The model schematisation consists of 134 sub-basins jointly covering the entire Rhine basin. The model runs at a daily time step. Inputs to the model consist of temperature and precipitation forcings; actual evaporation is estimated from a fixed annual profile that is corrected using temperature forecasts.

The forecasting system runs in two operating modes: historical and forecast mode. In historical mode, the hydrologic models are forced with meteorological observations for a period leading up to the forecast issue time. This ensures that the internal model states reflect the actual initial conditions of the basin as closely as possible. In forecast mode, these model states are the starting point for the model run, where the models are now forced by numerical weather predictions.

For observations of precipitation, the CHR08 dataset was used. This dataset covers the period 1961 through 2007. The CHR08 dataset was prepared specifically for the HBV model used here (Photiadou et al., 2011). The spatial scale of these CHR08 observations coincides with the 134 sub-basins used in the HBV model schematisation for the Rhine basin. Temperature observations originate from version 5.0 of the E-OBS data set; these were available from 1951 through mid 2011 (Haylock et al., 2008). Both precipitation and temperature data were available at a daily time step.

The ECMWF reforecast dataset, comprising medium-range EPS forecasts with 5 ensemble members (Hagedorn, 2008), was used for retrospective predictions of temperature and precipitation. At ECMWF, a retrospective forecast is produced every week for the same date in the 18 years preceding the current year, using the current operational model. To illustrate, on March 13, 2009, reforecasts were produced with initial conditions of March 13, 1991, March 13, 1992, and so forth until March 13, 2008. The reforecasts are produced using the operational model (currently Cy38r1 with a T639 horizontal resolution, i.e. 0.25 degrees in either direction). The set of reforecasts was thus produced using an operational model which, since the inception of the reforecasting scheme, has changed only slightly. This has little or no effect on the hydrologic model outcomes though, as was shown by Pappenberger et al. (2011). By July 2011, over 3,100 retrospective forecasts were available for use in the present study. While the forecast horizon extends to 30 days at a six hour time steps, for the present study only the first 10 days were available. Forecasts were temporally aggregated to a daily time step to match the time step used by the hydrologic model. The gridded forecasts were spatially averaged to the HBV sub basin scale.

Hourly streamflow observations for hydrologic stations within the Rhine basin were obtained from the Water Management Centre of the Netherlands. These observations were temporally aggregated to daily averages.

2.4. Experiment

Streamflow forecasts were produced with raw and post-processed forcings and verified against simulated streamflow, in order to establish the contribution of the forcing post-processing to the streamflow forecasts independently of any biases in the hydrologic model.

The baseline scenario comprised no post-processing of the forcing ensembles. Raw ensemble predictions of precipitation and temperature were used to generate streamflow ensemble predictions. In subsequent cases, temperature and precipitation ensemble predictions were statistically corrected using the techniques described in Section 2.1. These post-processed forcing ensemble predictions were then used to generate streamflow ensemble predictions. Thus, three cases were considered (Table 2): a baseline case, a case where an unconditional quantile-to-quantile transform (QQT) was applied to each variable (Case 1), and a case in which the forcing ensemble predictions were corrected using conditional techniques (Case 2). In terms of the latter, temperature ensemble predictions were statistically corrected using linear regression in the bivariate normal framework (LIN) and precipitation ensemble predictions were corrected using logistic regression (LOG). Variants of these two techniques were also considered, but not adopted. Specifically, for temperature, the assumption of homogeneous spread of the post-processed ensembles was relaxed to allow for a linear dependence on the raw ensemble spread (Gneiting et al., 2005), but without discernible benefits. For precipitation, a variant of LOG involving homogeneous parameters across all thresholds (Wilks, 2009) was evaluated, but this incurred an appreciable loss of skill.

2.5. Post-processing strategy

The parameters of any post-processor must be estimated with sample data. Both ensemble predictions and verifying observations were available for the period 1991–2007. This amounted to roughly 2,920 pairs of forecasts and observations at each forecast lead time. These pairs were not evenly distributed over the period of record due to the reforecasting procedure adopted by ECMWF.

The forcing ensembles were post-processed using the approaches described in Section 2.1 and AppendixA. Post-processing was performed separately for each of the 10 forecast lead times and 134 subbasins. Spatio-temporal covariability was then introduced via the Schaake Shuffle (Section 2.1). The post-processing was conducted within a cross-validation framework whereby separate periods of record were used to estimate the model parameters and independently verify the post-processed forecasts. Specifically, a leave-one-year-out cross-validation approach was adopted. This led to 17 separate calibrations of each post-processor, each comprising 16 years of calibration data and one year of independent prediction. The 17 years of independent predictions were then collated, verified, and used to force the streamflow models.

2.6. Verification strategy

The verification strategy focused on identifying the skill and biases in the forcing ensembles, as well as in the streamflow ensembles generated using these forcings. Skill and bias were identified with five well-known verification metrics. The correlation coefficient and the Relative Mean Error (RME) are measures of, respectively, the linear association of the forecast ensemble mean and observations and the relative bias of the ensemble mean. The (half) Brier Score (BS), the mean Continuous Ranked Probability Score (CRPS) and the area under the Relative Operating Characteristic (ROC) curve measure different attributes of the probabilistic quality of the forecasts. A short description of the latter scores is provided below, with accompanying equations given in AppendixB. Verification was performed with the Ensemble Verification System (Brown et al., 2010). The data that constituted input for verification, is posted to an online data repository (Verkade et al., 2013).

The Brier Score (Brier, 1950; Murphy, 1973; Wilks, 2001) measures the average square error of a probabilistic forecast of a discrete event. The mean CRPS (Hersbach, 2000; Stanski et al., 1989) is an integral measure of (square) probabilistic error in the forecasts across all possible discrete events. Both the BS and CRPS may be decomposed into further attributes of forecast quality by conditioning on the forecast variable (the calibration-refinement factorization). In addition, the BS may be decomposed by conditioning on the verifying observation (the likelihood-base-rate factorization). The area under the ROC curve (AUC) is a measure of event discrimination; that is, the ability of the forecasts to adequately discriminate between the exceedence and non-exceedence of a discrete threshold, such as the flood threshold.

Skill scores provide a convenient method for summarizing an improvement (or reduction) in forecast quality over a wide range of basins and conditions, as they are normalized measures. Here, both the BS and CRPS are formulated as skill scores with sample climatology as the baseline. These scores are denoted by the Brier Skill Score (BSS) and the Continuous Ranked Probability Skill Score (CRPSS), respectively. Rather than using the raw ensembles as the reference forecast, the scores are shown for the raw and post-processed ensembles with a consistent baseline, namely sample climatology. Likewise, the ROC Score (ROCS) comprises the AUC of the main forecasting system normalized by the AUC of the climatological forecast, i.e. 0.5 (AppendixB). This allows for the relative improvement of the forcing and streamflow forecasts to be identified in the context of background skill. However, some care is needed with interpretation, as sample climatology is unconditional and, therefore, increasingly (conditionally) biased towards the tails.

Conditional quality and skill was determined by calculating verification metrics for increasing levels of the non-exceedence climatological probability P, ranging from 0 to 1 (except for the BSS and ROCS, which are event skills and are, therefore, unknown for thresholds corresponding to the extremes of the observed data sample, nominally denoted by $P = 0$ and $P = 1$). Essentially, $P = 0$ constitutes an unconditional verification for continuous measures, such as the CRPSS, as all available data pairs are considered (Bradley and Schwartz, 2011), and is undefined for discrete measures, such as the BS. Conversely, at $P = 0.99$, only the data pairs with observations falling in the top 1% of sample climatology are considered; this amounts to approx. 30 pairs here. While the sampling uncertainties of the verification metrics were not explicitly evaluated here (see Brown and Seo, 2013), the results were not interpreted for thresholds larger than the 0.99 climatological probability or $\backsim$ 30 pairs.

3. Results

The results are presented in three subsections, each coinciding with one of the variables considered: temperature, precipitation and streamflow. Within those subsections, a discussion of the baseline case is followed by a discussion of the post-processed cases 1 (QQT) and 2 (conditional corrections LIN and LOG).

Correlation coefficients are very similar across cases and are mentioned in the text but not shown in tables or plots. Verification results are plotted in a series of multi-panel figures, showing RME, BSS, CRPSS and ROCS for the forecasts with lead times of 24-hours, 120-hours and 240-hours. The metrics are plotted as a function of the value of the verifying observation, expressed as a climatological probability of non-exceedence P, to allow for comparison across different locations. Most figures show results for multiple locations: thick lines indicate median values and thin lines denote the 10% and 90% quantiles of metrics over those multiple locations. Metrics pertaining to streamflow ensemble forecasts are shown across several plots, each corresponding to a spatial scale defined in Section 2.2. Note that, for ease of interpretation, all skill scores and associated decompositions are oriented to show the "best" scores at the top of the range axis and the "worst" at the bottom (Figure 3).

3.1. Ensemble temperature forecasts

Verification metrics for the ensemble temperature forecasts are shown in Figure 4. The metrics indicate that forecast quality decreases with increasing lead time, that it is conditional on the magnitude of the verifying observation and that this conditionality is more pronounced at longer lead times. This is true for both raw and post-processed temperature ensembles.

3.1.1. Raw temperature ensembles

Raw temperature ensembles show reasonably good correlation with observations. Values for the unconditional sample (at $P = 0$) range from 0.99 at the 24-hour lead time to 0.90 at the 240-hour lead time. At $P = 0.95$, these values are 0.87 and 0.18 respectively. Relative Mean Error plots indicate that for most basins, the ensemble mean underestimates the observation; this under-forecasting increases with higher values of the verifying observation. The CRPSS is largely constant, with a small dip in CRPSS values near to the median observed value. The BSS and ROCS show similar patterns, different from the CRPSS; both scores are consistently lowest at the extreme ends of the distribution.

These patterns reflect the different formulations of the verification scores and the choice of reference forecast. The BSS and ROCS measure the quality of discrete predictions, with contributions to the score being dominated by the corollary (i.e. non-occurrence) at extreme (low and high) thresholds. At longer lead times, the residual skill of the temperature forecasts is concentrated towards the median temperature, where the forecasts have least conditional bias and greatest correlation (and the occurrences and non-occurrences, by definition, contribute equally). In contrast, the CRPS is a smooth, continuous measure that factors skill across all possible thresholds for each paired sample. Since the sample climatology is unconditional by construction, the baseline forecasts will be least reliable in the tails of the climatological distribution, with large conditional biases contributing to poorer quality of the reference forecast in the tails (and hence greater relative quality of the ECMWF forecasts, whether post-processed or not).

3.1.2. Post-processed temperature ensembles

After post-processing, the correlation of the temperature ensembles with the verifying observations was virtually unchanged from the raw case. In terms of RME, BSS, CRPSS and ROCS, LIN almost always outperformed QQT (noting that QQT is a non-linear transform and may not preserve correlation), which in turn outperformed the raw ensembles. For the latter three metrics, the differences in quality are most pronounced at large values of the verifying observation.

3.2. Ensemble precipitation forecasts

Verification metrics for the ensemble precipitation forecasts are shown in Figure 5. Subsequent figures show the calibration-refinement decomposition of the CRPSS (Figure 7) and the BSS (Figure 8) as well as the likelihood-base rate decomposition of the BSS (Figure 9). Similar to the temperature figures, verification metrics are plotted as a function of observed amount, expressed as a climatological probability of non-exceedence, P. In the case of precipitation, however, the domain axis range is $[0.4, 1.0]$. As the probability of precipitation (PoP) is approx. 60% for all basins, smaller probabilities all correspond to the PoP threshold of zero preciptiation and produce identical scores. As in the case of temperature ensembles, forecast quality is seen to decrease with increasing lead time, and to be strongly conditional on the amount of precipitation.

3.2.1. Raw ensemble precipitation forecasts

Correlation between the mean of the raw precipitation ensembles and observations is largely positive, but distinctly lower than that of the temperature ensembles. Correlation deteriorates with forecast lead time and with increasing value of the observation. At $P = 0$, correlation ranges from 0.71 to 0.13 for lead times of 24-hour and 240-hours respectively. At $P = 0.95$, these values are 0.36 and 0.04 respectively.

The RME shows that the ensemble mean overestimates zero and small precipitation amounts. For increasing values of the observation, the ensemble mean increasingly underestimates precipitation. For example, at a lead time of 120 hours, the RME equals 0.07, -0.18 and -0.59 at $P = 0$, $P = 0.5$ and $P = 0.9$ respectively.

These conditional biases stem from the inability of the raw predictors used in the post-processor to correctly predict when large events occur (large relative to other events in the climatological distribution). This leads to a real-time adjustment that reflects the assumed, but wrong, conditions. Also, statistical post-processors are calibrated for good performance under a range of conditions (i.e. for unconditional skill and unbiasedness), which inevitably leads to some conditional biases. In short, some conditional bias is a "natural" consequence of post-processing with imperfect predictors and with a focus on global optimality. However, it is also a practically significant feature of these and other post-processed ensemble forecasts. While the precise description of these conditional biases will depend on the choice of measure (e.g. the RME is sensitive to skewness), the conditional biases are present, regardless of the choice of measure. Figure 6 shows the 120-hour lead time forecast error as a function of the verifying observation for a single basin. Clearly, at higher values of the observation, the ensembles consistently, and increasingly, underestimate the observed value, with insufficient spread to offset this conditional bias.

The CRPSS declines with both lead time and increasing amount of observed precipitation. The BSS and ROCS plots show similar patterns; both metrics are lowest at the tails, indicating that it is relatively difficult to distinguish between zero and non-zero precipitation and to correctly predict the occurrence of large precipitation amounts.

3.2.2. Post-processed precipitation ensembles

When moving from raw to post-processed precipitation ensembles, the correlation between the ensemble forecast and the observation is largely conserved. Only in the case of LOG does correlation drop slightly, and only at higher precipitation amounts.

Both the QQT and LOG techniques produce ensemble forecasts that are unconditionally unbiased. However, in all cases, there is an increasingly large conditional negative bias at higher precipitation amounts. At longer lead times, the RME across all cases is very similar. The raw ensembles initially show a small positive RME, which at some value of P becomes negative and then continues to drop. For non-zero precipitation, LOG shows the highest negative RME at all lead times. From Figure 6, it is clear that the post-processing methods were unable to correct for the Type-II conditional biases at high observed precipitation amounts.

For both techniques, the gain in CRPSS following post-processing is only modest or marginal at all lead times and precipitation amounts. In terms of unconditional CRPSS ($P = 0$), LOG shows the highest increase in skill at all lead times. At higher observed precipitation amounts, LOG does markedly worse than the raw and QQT ensembles due to a large, negative, conditional bias in the ensemble mean. The CRPSS of the QQT corrected ensembles are largely similar to that of the raw ensembles. The CRPSS decomposition (Figure 7) and the BSS decomposition (Figure 8) show that none of the post-processing techniques was able to consistently improve both the reliability and resolution of the precipitation ensembles. Rather, there is a trade-off whereby the post-processing generally results in improved reliability at the expense of some loss in resolution. This is different from the post-processed temperature ensembles, which showed improved reliability while consistently maintaining or improving resolution (results not shown). For precipitation, the combination of lower quality of the raw forecasts and a larger number of parameters to estimate for LOG leads to greater sampling uncertainty and weaker performance overall.

In terms of BSS, LOG consistently outperforms the raw and QQT-post-processed precipitation ensembles. As indicated in Figure 9, this is largely explained by an increase in the reliability (or reduction in Type-I conditional bias) of the precipitation ensembles following LOG. However, the RME and the likelihood-base-rate decomposition of the BSS (Figure 10) show a greater tendency of the LOG ensembles to under-forecast high observed precipitation amounts, i.e. they display a larger Type-II conditional bias.

3.3. Streamflow ensemble forecasts

Verification results for the streamflow ensembles are presented for multiple spatial scales. For the 43 headwater basins, Figure 10 shows RME, CRPSS, BSS and ROCS values. Figures 11 and 12 show calibration-refinement decompositions of the CRPSS and BSS respectively; Figure 13 shows the likelihood-base-rate decomposition of the BSS. The RME, CRPSS, BSS and ROCS values for the Main, Neckar, Moselle and Swiss Rhine tributaries and for the Rhine outlet at Lobith are shown in Figures 14 and 15 respectively.

3.3.1. Streamflow ensemble forecasts based on raw forcings

The ensemble mean of the streamflow forecasts is highly correlated with the simulated streamflow at short lead times. For example, the correlation exceeds 0.98 at $P = 0$ and, at $P = 0.95$, ranges from 0.90 to 0.98 to 0.99 for the smallest to largest spatial scales, respectively. Generally, correlation reduces with decreasing spatial scale: it is lowest for the collection of headwater basins and highest at the outlet, where the aggregate response has a modulating effect on the errors from individual basins. Correlation declines with increasing lead time and with increasing value of the streamflow simulation.

At all spatial scales, the unconditional RME is negligible at the earliest lead times, but increases with increasing lead time. For streamflows larger than the median climatological flow, the forecast ensemble mean increasingly underestimates the simulated streamflow. For example, the RME for the headwater basins (Figure 10) at a lead time of 120 hours shows a median RME of -0.07, -0.20 and -0.27 at $P = 0.5$, $P = 0.9$ and $P = 0.95$ respectively. At the outlet at Lobith (Figure 15) the corresponding values are -0.001, -0.02 and -0.03 respectively.

The patterns in BSS, CRPSS and ROCS are similar to one another and across all spatial scales and lead times. The skill is greatest for the unconditional flows at the shortest lead times and declines with increasing value of the verifying simulation, particularly above the median climatological streamflow where the conditional bias increases. The skill also increases with increasing spatial scale. For example, the median CRPSS values at $P = 0.90$ at a lead time of 120-hours are 0.54, 0.73 and 0.90 for headwaters, tributaries and outlet respectively.

3.3.2. Streamflow ensembles based on post-processed forcings

Correlations between the simulated streamflow and the forecast ensemble means are hardly affected by post-processing of the forcings. A slight reduction is observed at longer lead times and at higher quantiles of the simulation for the LIN-LOG case. At $P = 0.90$ and a 240-hour lead time, correlation drops from 0.34 for the raw case to 0.31 for the LIN-LOG case.

Unconditionally, the combinations of QQT-QQT and LIN-LOG result in RME values that are closer to 0 than those of the RAW-RAW case. However, there is a tendency for all techniques to under-forecast the higher simulated streamflows, with the greatest conditional bias for the LIN-LOG case. For example, in the LIN-LOG case at a lead time of 120 hours, the median RME for the four main tributaries increases (negatively) from -0.04 at $P = 0.50$ to -0.15 at $P = 0.90$ and -0.19 at $P = 0.95$ (Figure 14).

In terms of the BSS, CRPSS and ROCS, the streamflow ensembles derived from quantile-to-quantile transformed forcings generally show higher skill than those derived from raw forcings. However, the differences are small. The QQT-QQT ensembles show similar skill at low and moderate values of the verifying simulation only, since QQT is unable to correct for conditional biases, whether Type-I or Type-II in nature.

Generally, post-processing of the forcing variables using conditional techniques (LIN-LOG) does not result in increased skill in terms of CRPSS and BSS. Below the median climatological streamflow, skills are largely similar to those of streamflow ensembles derived from raw forcings. At higher quantiles, there is actually a small decrease of skill. An increase of skill is observed in terms of the reliability component of the BSS and in terms of the ROCS; that is, in the ability of the forecasts to discriminate between the occurrence and non-occurrence of discrete events.

4. Discussion

Several questions were posed in this case study: are the raw ECMWF-EPS temperature and precipitation ensembles biased and if so, how? Do these biases translate into streamflow biases and reduced skill? Does post-processing of the temperature and precipitation ensembles improve the quality of the forcing ensembles, and is this improvement noticeable in the streamflow ensembles?

The raw temperature and precipitation ensemble forecasts are biased in both the mean and spread. However, they are more skilful than sample climatology for shorter lead times and moderate thresholds, with reduced skill at longer lead times and for larger amounts (and for zero precipitation). The temperature ensemble forecasts are less biased and more skilful than the precipitation ensemble forecasts. The effects of these biases on the streamflow ensemble forecasts depend on the concentration time of the basins considered with a more rapid deterioration with leadtime in skill for headwaters than for downstream basins. This is largely due to the absence of hydrologic biases and uncertainties in the verification results, of which those in the initial conditions are an important part (i.e. verification was conducted against simulated streamflow).

Thus, the skill of streamflow predictions is strongly affected by the initial conditions; this effect lasts longer in larger basins.

Overall, the improvements in the ensemble forcing predictions were modest; this was especially the case for the precipitation ensembles. However, this does not imply that the forcing ensembles are nearly perfect. Rather, it suggests that no additional signal can be found in the forecast-observed residuals to improve forecast quality with the statistical techniques considered. In some cases, post-processing reduces skill; it attempts to use a signal that, in hindsight, turns out to be noise with no predictive information for future forecasts.

Post-processing of the temperature ensembles resulted in greater improvements than post-processing of the precipitation ensembles. This is not surprising, because temperatures are relatively more predictable than precipitation and the gain in skill from post-processing (with a conditional technique) partly depends on the strength of association between the forecasts and observations. Much of the improvement in the precipitation ensemble forecasts is unconditional in nature. Possibly, the improvements from conditional post-processing would be greater when calibrating on a larger dataset, as there were only ∽ 2,900 pairs available in this study, when using a more parsimonious statistical technique (e.g. Wu et al., 2011) or when supplementing the training data set at a particular location with data from other locations with similar climatologies (Hamill et al., 2008).

Application of the forcing post-processors generally results in a reduction in bias and an improvement in skill of the forcing ensembles, although the precise effects depend on forecast lead time, threshold, spatial scale and the types of bias considered. For example, while LOG generally improves the reliability of the precipitation ensembles, the ensemble mean is negatively biased with increasing observed precipitation amount, i.e. a Type-II conditional bias. Post-processing does not improve on all qualities at all lead times and at all levels of the verifying observation. Generally, but not always, post-processing improves the reliability of the forecasts, but this is sometimes accompanied by a loss of resolution or an increase in the Type-II conditional biases.

Changes in the biases and skill of the forcing ensembles cascade to the ensemble streamflow forecasts. No combination of techniques improves all forecast qualities considered at all lead times and all levels of the verifying simulation. A reduction in the unconditional bias and in the reliability of the ensemble precipitation forecasts is followed by improvements in the reliability of the streamflow ensemble forecasts. However, the trade-off between reliability and resolution is also observed in the streamflow predictions.

The improvements in precipitation and temperature do not translate proportionally into the streamflow forecasts. This may be partly explained by the strong non-linearity of the Rhine basin (due to substantial storage of water in the subsurface, in extensive snowpacks and, to a lesser degree, in reservoirs) and, accordingly, the hydrologic model. Possibly, the effects of post-processing would be stronger in basins where streamflow has a more linear response to forcing variables, e.g. in basins with less storage ,or when leadtimes are sufficiently long to allow for the stored water to reach the streamflow network. This may explain why Yuan and Wood (2012) found that in their seasonal forecasting case, post-processing of forcings leads to a more noticeable improvement of streamflow forecast skill than was found in the case described in the present manuscript.

Another potential cause of muted signal resulting from the forcing bias-correction may also be explained by inadequate modelling of the space-time covariability of the forcing forecasts. Forcing verification (as presented here, but more generally) is sensitive to the joint distribution of the forecasts and observations at specific times, locations and for specific variables. In contrast, hydrologic models are sensitive to the space-time covariability of the forcing forecasts. In this context, the use of the Schaake shuffle to recover some of this space-time covariability may be limiting. The Schaake shuffle introduces rank association only, and it introduces this only insofar as these patterns appear historically on the same or nearby dates. For example, it cannot account for more complex statistical dependencies, novel structures, or structures that are conditional upon the state of the atmosphere at the forecast issue time. These weaknesses are likely exaggerated when the forecasts have greater spread because the Schaake shuffle has greater scope to affect the space-time patterns of the ensemble traces. In order to account for more complex structures, post-processors with explicit models of space-time covariability are needed, such as geostatistical models (Kleiber et al., 2011), together with parsimonious verification techniques that are sensitive to these space-time and cross-variable relationships.

Verification against simulated streamflows allows for the hydrologic biases to be factored out of the streamflow skill associated with forcing post-processing. However, it also magnifies the resulting streamflow ensemble skill. When verifying against observations, the overall biases and uncertainty will be larger due to inclusion of the hydrologic biases and uncertainties, including those in the streamflow observations. Relatively speaking, the change in skill due to the post-processed forcings will be more difficult to detect.

The research questions posed in the introduction were addressed by looking at a selection of verification metrics. While reasonably broad, the results may be sensitive to the choice of metric. In addition, the parameters of each post-processing techniques are estimated with a particular objective function. If these objective functions are similar to the verification metrics used, it should not be surprising that a particular technique scores well in terms of that metric.

The available reforecast dataset allowed for testing our hypothesis using a reasonable number of retrospective forecasts (just over 3,100). Conditional verification however, especially of extreme events, quickly reduces the size of the subsample. In this study, the cut-off of the climatological nonexceedence probability was chosen at $P = 0.99$, which is where 1% of the available data is used for verification. This coincides with approx. 30 data pairs. If the present study would be repeated and extended by stratifications, for example on a two season basis, then the $P = 0.99$ quantile would equate to approx. 15 verification pairs, which is deemed too small for verification purposes. Conversely, if, in case of stratification, the minimum number of pairs would be kept fixed at 30, this would mean that less extreme events can be analysed only. Ideally, longer sets of reforecasts (hindcasts) would be available. Note that by the time the present manuscript was submitted for publication, the ECMWF reforecast set had been extended considerably. Even so, the authors support the call for reforecast datasets, eloquently voiced by Hamill et al. (2006).

In the current study, the improvements to streamflow accrued by post-processing of the forcing predictions were modest. Moreover, these effects may be negligible when verifying against streamflow observations. Since forcing post-processing is both labour intensive and inherently difficult for precipitation, particularly in accounting for appropriate space-time covariability, it is worth considering other methods to improve the skill of the forcing and streamflow ensembles, such as multi-model combinations, data assimilation (to improve the hydrologic initial conditions), and streamflow post-processing. For example, under conditions where forcing post-processing contributes significant skill to streamflow, it needs to be established whether that skill remains after streamflow post-processing or whether statistical post-processing can adequately remove the forcing biases via the streamflow, despite the aggregation of multiple sources of bias and uncertainty.

5. Summary and conclusions

Ensemble forecasts of temperature and precipitation were tested for biases and an attempt was made to reduce these biases through statistical post-processing. This resulted in modest improvements in the quality of the forcing ensembles. The effects on streamflow were explored by factoring out the effects of bias in the hydrologic model; that is, by verifying against simulated streamflow. In general, the improvements in streamflow quality were muted at all spatial scales considered, with explanations including a limited model of the space-time covariability of the forcing ensembles.

6. Acknowledgements

The authors would like to thank Florian Pappenberger at ECMWF for providing reforecast data. We are also grateful to Tom Hamill, Tom Pagano and an anonymous reviewer for taking the time to review an earlier version of this manuscript and for providing constructive comments. Any errors remain ours. Publication of this manuscript was partially funded by Deltares research funds. Several open source and freely (as in: without charge) available software and data sources were used for the analyses described in this manuscript: Quantum GIS (Quantum GIS Development Team, 2012), R (R. Core Team, 2012), the Ensemble Verification System (Brown et al., 2010), the Delft-FEWS forecast production system (Werner et al., 2012), the SRTM digital elevation model (Jarvis et al., 2008), CHR08 precipitation data (Photiadou et al., 2011) and E-OBS temperature data (Haylock et al., 2008).

7. Author contributions

J.S.V. and J.D.B. conceived of the study, performed the numerical modeling and wrote the paper. A.H.W. and P.R. contributed to the latter.

AppendixA. Post-processing techniques

AppendixA.1. Quantile-to-quantile transform

The quantile-to-quantile transform, also known as quantile mapping or cdf matching, is given by

$$x_{\mathrm{qqt},c} = \bar{F}_Y^{-1}\left(\bar{F}_{X_c}\left(x_c\right)\right), \tag{A.1}$$

where $\bar{F}_Y$ denotes the sample climatology of the predictand Y, or the empirical distribution of observations, $\bar{F}_{X_c}$ denotes the sample climatology of the predictor X_c and $x_{\mathrm{qqt},c}$ represents the quantile-to-quantile transformed prediction for the c^{th} member of the C–member forcing ensemble. Thus, the transform is applied to each of the C members and their C separate, but practically identical, climatologies. In general, $x_{\mathrm{qqt},c}$ will not map linearly to x_c, because the curvatures of $\bar{F}_Y$ and $\bar{F}_{X_c}$ are different.

AppendixA.2. Linear regression

Given a training data set, a simple linear regression relation is assumed to exist between observed temperature and the mean of the ensemble prediction (Wilks, 2006),

$$Y = \beta_0 + \beta_1 \bar{X} + \varepsilon, \tag{A.2}$$

where β_0 and β_1 are regression parameters to estimate and ε is a stochastic residual. This relation is sought for each location and lead time separately but subscripts denoting these are omitted from Equation A.2. The regression coefficients are found by minimising the expected square difference between the temperatures predicted by the model and observed. The regression constants are determined for each lead time and location separately.

The residuals are assumed to be Normally distributed with zero mean, μ,

$$\varepsilon = \mathrm{N}\left(\mu = 0, \sigma\right), \tag{A.3}$$

and σ given by the sample standard deviation of errors.

From this regression equation, probabilistic temperature forecasts are produced for a given value of the raw ensemble mean, $\bar{x}$, by sampling from $\mathrm{N}\left(\beta_0 + \beta_1 \bar{x}, \sigma\right)$.

AppendixA.3. Logistic regression

The conditional probability that the future amount of precipitation, Y, does not exceed a discrete threshold, y, given the raw ensemble mean, $\bar{x}$, is

$$P\left(Y \leq y | \bar{X} = \bar{x}\right) = \frac{1.0}{1.0 + \exp^{-\left(\beta_0 + \beta_1 \bar{x}^{1/3}\right)}}, \qquad \text{(A.4)}$$

where β_0 and β_1 are the parameters of the linear model to estimate through maximum likelihood. The power transformation has the effect of allowing the precipitation forecast data to be more normally distributed (Hamill et al., 2008). Similar to the experiment described in Sloughter et al. (2007), a one-third power transformation is used. Here, 200 thresholds are considered. The thresholds are then interpolated using a spline constrained to be a valid cumulative distribution function using the method described by He and Ng (1999).

AppendixB. Verification metrics

For ease of reference, the probabilistic verification metrics used in this study are briefly explained; this description is based on Brown and Seo (2013). Further details can be found in the documentation of the Ensemble Verification System (Brown et al., 2010) as well as in reference works on forecast verification by Jolliffe and Stephenson (2012) and Wilks (2006).

AppendixB.1. Relative Mean Error

The Relative Mean Error (RME, sometimes called *relative bias*) measures the average difference between a set of forecasts and corresponding observations, relative to the mean of the latter,

$$\mathrm{RME} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{Y}_i\right)}{\sum_{i=1}^{n} x_i}, \qquad \text{(B.1)}$$

where x is the observation and $\bar{Y}$ is the mean of the ensemble forecast. The RME thus provides a measure of relative, first-order bias in the forecasts. RME may be positive, zero, or negative. Insofar as the mean of the ensemble forecast should match the observed value, a positive RME denotes overforecasting and a negative RME denotes underforecasting. A RME of zero denotes the absence of relative bias in the mean of the ensemble forecast.

AppendixB.2. Brier score and Brier skill score

The (half) Brier score (BS) measures the mean square error of n predicted probabilities that Q exceeds q,

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}\left\{F_{X_i}(q) - F_{Y_i}(q)\right\}^2, \qquad \text{(B.2)}$$

where $F_{X_i}(q) = \Pr\left[X_i > q\right]$ and $F_{Y_i}(q) = \begin{cases} 1 & \text{if } Y_i > q; \\ 0 & \text{otherwise} \end{cases}$.
By conditioning on the predicted probability, and partitioning over J discrete categories, the BS is decomposed into the calibration-refinement (CR) measures of Type-I conditional bias or reliability (REL), resolution (RES) and uncertainty (UNC),

$$\begin{aligned} \mathrm{BS} \quad = \quad & \underbrace{\frac{1}{n}\sum_{j=1}^{J} N_j\left\{F_{X_j}(q) - \bar{F}_{Y_j}(q)\right\}^2}_{\mathrm{REL}} \\ & \underbrace{-\frac{1}{n}\sum_{j=1}^{J} N_j\left\{F_{Y_j}(q) - \bar{F}_{Y}(q)\right\}^2}_{\mathrm{RES}} \\ & + \underbrace{\sigma_Y^2(q)}_{\mathrm{UNC}}. \end{aligned} \qquad \text{(B.3)}$$

Here, $\bar{F}_Y(q)$ represents the average relative frequency (ARF) with which the observation exceeds the threshold, q. The term $F_{Y_j}(q)$ represents the conditional observed ARF, given that the predicted probability falls within the j^{th} of J probability categories, which happens N_j times. Normalizing by the climatological variance UNC, $\sigma_Y^2(q)$, leads to the Brier Skill Score (BSS),

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{UNC}} = \frac{\mathrm{RES}}{\mathrm{UNC}} - \frac{\mathrm{REL}}{\mathrm{UNC}}. \qquad \text{(B.4)}$$

By conditioning on the $K = 2$ two possible observed outcomes, $\{0, 1\}$, the BS is decomposed into the likelihood-base-rate (LBR) measures of Type-II conditional bias (TP2), discrimination (DIS), and sharpness (SHA),

$$\begin{aligned} \mathrm{BS} \quad = \quad & \underbrace{\frac{1}{n}\sum_{k=1}^{K} N_k\left\{\bar{F}_{X_k}(q) - \bar{F}_{Y_k}(q)\right\}^2}_{\mathrm{TP2}} \\ & \underbrace{-\frac{1}{n}\sum_{k=1}^{K} N_k\left\{F_{X_k}(q) - \bar{F}_{X}(q)\right\}^2}_{\mathrm{DIS}} \\ & + \underbrace{\sigma_X^2(q)}_{\mathrm{SHA}}. \end{aligned} \qquad \text{(B.5)}$$

Here, $\bar{F}_{X_k}(q)$ represents the average probability with which X is predicted to exceed q, given that Y exceeds q ($k = 1$) or does not exceed q ($k = 2$), where N_k is the conditional sample size for each case. The BSS is then given by

$$\begin{aligned}\mathrm{BSS} &= 1-\frac{\mathrm{BS}}{\mathrm{UNC}} \\ &= 1-\frac{\mathrm{TP2}}{\mathrm{UNC}}+\frac{\mathrm{DIS}}{\mathrm{UNC}}-\frac{\mathrm{SHA}}{\mathrm{UNC}}. \end{aligned} \tag{B.6}$$

AppendixB.3. Mean Continuous Ranked Probability Score and Skill Score

The Continuous Ranked Probability Score (CRPS) measures the integral square difference between the cumulative distribution function (cdf) of the forecast $\mathrm{F}_X(q)$, and the corresponding cdf of the observed variable $\mathrm{F}_Y(q)$,

$$\mathrm{CRPS}=\int_{-\infty}^{\infty}\{\mathrm{F}_X(q)-\mathrm{F}_Y(q)\}\,\mathrm{d}q. \tag{B.7}$$

The mean CRPS comprises the CRPS averaged across n pairs of forecasts and observations. The Continuous Ranked Probability Skill Score (CRPSS) is a ratio of the mean CRPS of the main prediction system, $\overline{\mathrm{CRPS}}$, and a reference system, $\overline{\mathrm{CRPS}_{\mathrm{ref}}}$,

$$\mathrm{CRPSS}=\frac{\overline{\mathrm{CRPS}_{\mathrm{ref}}}-\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}_{\mathrm{ref}}}}. \tag{B.8}$$

AppendixB.4. Relative Operating Characteristic score

The Relative Operating Characteristic (ROC; Green and Swets 1966) measures the trade-off between correctly forecasting that a discrete event will occur (Probability of Detection, PoD) and incorrectly forecasting that it will occur (Probability of False Detection, PoFD). This trade-off is expressed as a decision threshold, d, at which the forecast probability triggers some action. The ROC plots the PoD versus the PoFD for all possible values of d in $[0,1]$. For a particular threshold, the empirical PoD is

$$\mathrm{PoD}=\frac{\sum_{i=1}^{n} I_{X_i}\left(F_{X_i}(q)>d|Y_i>q\right)}{\sum_{i=1}^{n} I_{Y_i}\left(Y_i>q\right)}, \tag{B.9}$$

where I denotes the indicator function. The empirical PoFD is

$$\mathrm{PoFD}=\frac{\sum_{i=1}^{n} I_{X_i}\left(F_{X_i}(q)>d|Y_i>q\right)}{\sum_{i=1}^{n} I_{Y_i}\left(Y_i\leq q\right)}. \tag{B.10}$$

The ROC score measures the area under the ROC curve (AUC) after adjusting for the climatological base rate, i.e.

$$\mathrm{ROCS}=2\times(\mathrm{AUC}-0.5). \tag{B.11}$$

Bergström, S., Singh, V. P., 1995. The HBV model. In: Singh, V. P. (Ed.), Computer models of watershed hydrology. Water Resources Publications, Highlands Ranch, Colorado, United States, pp. 443–476.

Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resources Research 47 (7), W07524.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., Beare, S. E., 2008. The MOGREPS short-range ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 134 (632), 703–722.

Bradley, A. A., Schwartz, S. S., 2011. Summary verification measures and their interpretation for ensemble forecasts. Monthly Weather Review 139 (9), 3075–3089.

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

Brier, G., 1950. Verification of forecasts expressed in terms of probability. Monthly weather review 78, 1–3.

Brown, J. D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The ensemble verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environmental Modelling & Software 25 (7), 854 – 872.

Brown, J. D., Seo, D.-J., 2010. A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. Journal of Hydrometeorology 11 (3), 642–665.

Brown, J. D., Seo, D.-J., 2013. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. Hydrological Processes 27 (1), 83–105.

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

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., Wilby, R., 2004. The schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. Journal of Hydrometeorology 5 (1), 243–262.

Cloke, H., Pappenberger, F., 2009. Ensemble flood forecasting: a review. Journal of Hydrology 375 (3-4), 613–626.

Gneiting, T., Raftery, A., Westveld, A., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review 133 (5), 1098–1118.

Green, D. M., Swets, J. A., 1966. Signal detection theory and psychophysics. John Wiley & Sons, Inc., New York.

Hagedorn, R., 2008. Using the ECMWF reforecast dataset to calibrate EPS forecasts. ECMWF Newsletter 117, 8–13.

Hagedorn, R., Hamill, T., Whitaker, J., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part i: Two-meter temperatures. Monthly Weather Review 136 (7), 2608–2619.

Hamill, T., Hagedorn, R., Whitaker, J., 2008. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part II: precipitation. Monthly Weather Review 136 (7), 2620–2632.

Hamill, T., Whitaker, J., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Monthly weather review 134 (11), 3209–3229.

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

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.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y., Lapenta, W., Feb. 2013. NOAA's second-generation global medium-range ensemble reforecast data set. Bulletin of the American Meteorological Society, early online release.
URL http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-12-00014.1

Hashino, T., Bradley, A., Schwartz, S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrology and Earth System Sciences 11, 939–950.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., New, M., 2008. A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res 113, D20119.

He, X., Ng, P., 1999. COBS: Qualitatively constrained smoothing via linear programming. Computational Statistics 14, 315–338.

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

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-filled seamless SRTM data v4. http://srtm.csi.cgiar.org.

Jolliffe, I. T., Stephenson, D. B. (Eds.), 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition.

Kang, T.-H., Kim, Y.-O., Hong, I.-P., 2010. Comparison of pre- and post-processors for ensemble streamflow prediction. Atmospheric Science Letters 11 (2), 153–159.

Kelly, K., Krzysztofowicz, R., 2000. Precipitation uncertainty processor for probabilistic river stage forecasting. Water resources research 36 (9).

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F., Grimit, E., 2011. Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local bayesian model averaging. Monthly Weather Review 139 (8), 2630–2649.

Krzysztofowicz, R., 1999. Bayesian forecasting via deterministic model. Risk Analysis 19 (4), 739–749.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. Journal of Hydrology 249 (1-4), 2–9.

Madadgar, S., Moradkhani, H., Garen, D., 2012. Towards improved post-processing of hydrologic forecast ensembles. Hydrological Processes.

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

Pappenberger, F., Thielen, J., Del Medico, M., Mar. 2011. The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the european flood alert system. Hydrological Processes 25 (7), 1091–1113.
URL http://doi.wiley.com/10.1002/hyp.7772

Photiadou, C. S., Weerts, A. H., van den Hurk, B. J. J. M., 2011. Evaluation of two precipitation data sets for the rhine river using streamflow simulations. Hydrology and Earth System Sciences 15 (11), 3355–3366.

Quantum GIS Development Team, 2012. Quantum GIS Geographic Information System.

R. Core Team, 2012. R: A Language and Environment for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0.

Raiffa, H., Schlaifer, R., 1961. Applied Statistical Decision Theory. Harvard University Press.

Ramos, M. H., van Andel, S. J., Pappenberger, F., 2012. Do probabilistic forecasts lead to better decisions? Hydrology and Earth System Sciences Discussions 9 (12), 13569–13607.

Reggiani, P., Weerts, A., 2008a. A bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river rhine. Journal of Hydrology 356 (1-2), 56–69.

Reggiani, P., Weerts, A. H., 2008b. Probabilistic quantitative precipitation forecast for flood prediction: An application. Journal of Hydrometeorology 9 (1), 76–95.

Schellekens, J., Weerts, A. H., Moore, R. J., Pierce, C. E., Hildon, S., 2011. The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across england and wales. Advances in Geosciences 29, 77–84.

Schmeits, M. J., Kok, K. J., 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.

Seo, D., Herr, H., Schaake, J., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrology and Earth System Sciences Discussions 3 (4), 1987–2035.

Sloughter, J., Raftery, A., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using bayesian model averaging. Monthly Weather Review 135 (9), 3209–3220.

Stanski, H., Wilson, L., Burrows, W., 1989. Survey of common verification methods in meteorology. World Meteorological Organization Geneva.

Todini, E., 2004. Role and treatment of uncertainty in real-time flood forecasting. Hydrological Processes 18 (14), 2743–2746.

Todini, E., 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. International Journal of River Basin Management 6 (2), 123–138.

Verkade, J. S., Werner, M. G. F., 2011. Estimating the benefits of single value and probability forecasting for flood warning. Hydrology and Earth System Sciences 15 (12), 3751–3765.

Verkade, J. S., Brown, J. D., Reggiani, P., Weerts, A., 2013. Dataset for "post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales" by Verkade et al., 2013. Available from http://dx.doi.org/10.4121/uuid:56637037-8197-472b-b143-2f87adf49abc

Weerts, A., Winsemius, H., Verkade, J., 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.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2012. The delft-FEWS flow forecasting system. Environmental Modelling & Software.

Wilks, D., 2006. Statistical methods in the atmospheric sciences. Academic Press.

Wilks, D. S., 2001. A skill score based on economic value for probability forecasts. Meteorological Applications 8 (2), 209–219.

Wilks, D. S., 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. Meteorological Applications 16 (3), 361–368.

Wilks, D. S., Hamill, T. M., 2007. Comparison of ensemble-MOS methods using GFS reforecasts. Monthly Weather Review 135 (6), 2379–2390.

Wood, A. W., Maurer, E. P., Kumar, A., Lettenmaier, D. P., 2002. Long-range experimental hydrologic forecasting for the eastern united states. J. Geophys. Res 107 (D20), 4429.

Wu, L., Seo, D.-J., Demargne, J., Brown, J. D., Cong, S., Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. Journal of Hydrology 399 (3-4), 281–298.

Yuan, X., Wood, E. F., 2012. Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. Water Resources Research 48 (12).
URL http://dx.doi.org/10.1029/2012WR012256

Zalachori, I., Ramos, M.-H., Garon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Advances in Science and Research 8, 135–141.

Zhao, L., Duan, Q., Schaake, J., Ye, A., Xia, J., 2011. A hydrologic post-processor for ensemble streamflow predictions. Adv. Geosci. 29, 51–59.

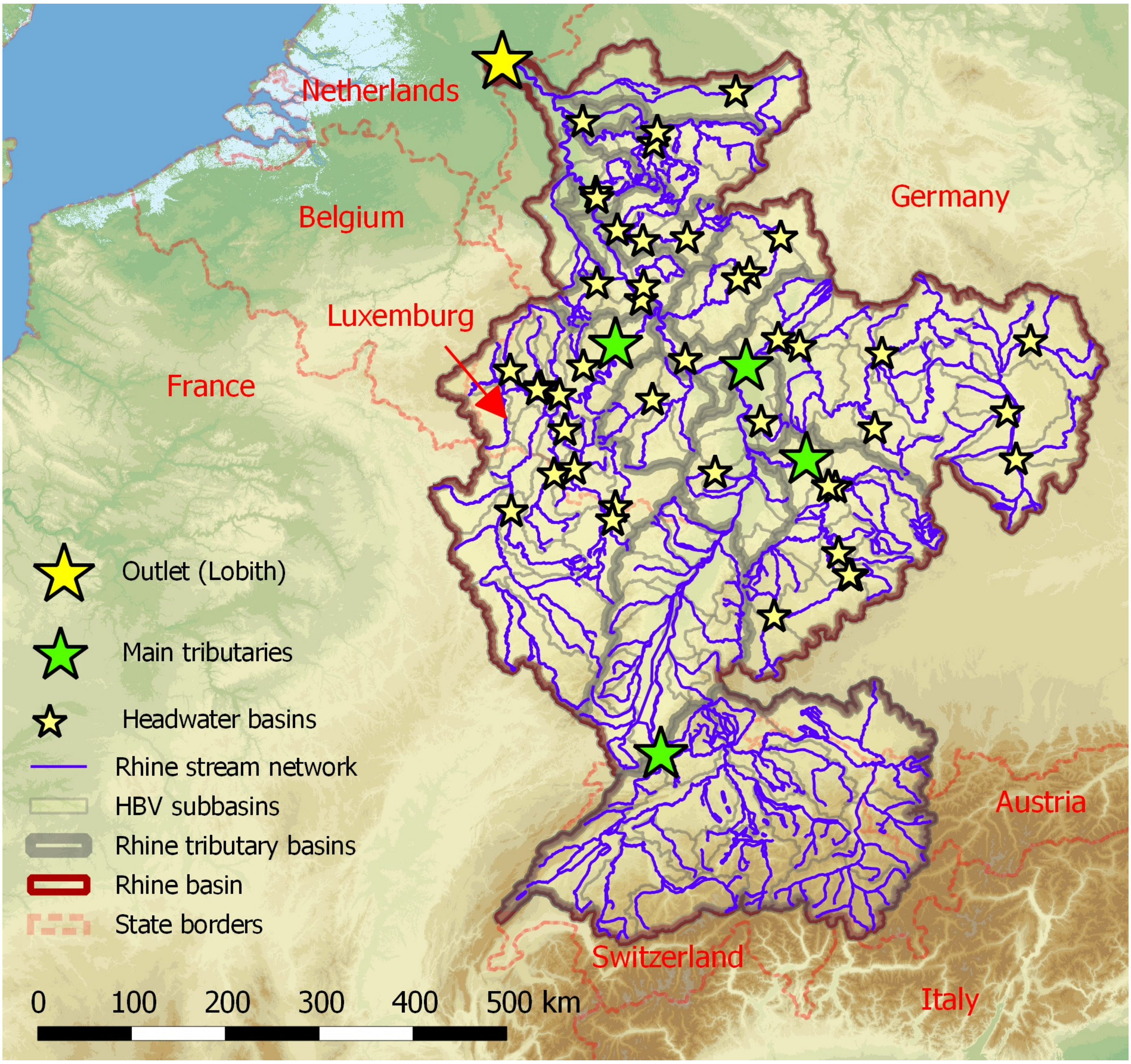

Figure 1: Location of the Rhine basin in continental Europe.

		Contributing area $[km^2]$				
Spatial Scale	#constituents	10^{th} perc	median	90^{th} perc	mean	sum
Headwater basins	43	370	929	2,008	1,142	49,125
Tributaries	4	17,972	27,767	34,035	26,507	106,029
Rhine basin	1					159,559

Table 1: Contributing areas of the spatial scales that are analysed.

	Temperature correction	Precipitation correction
Baseline case	none (RAW)	none (RAW)
Case 1	quantile-to-quantile transform (QQT)	quantile-to-quantile transform (QQT)
Case 1	linear regression (LIN)	logistic regression (LOG)

Table 2: Overview of cases.

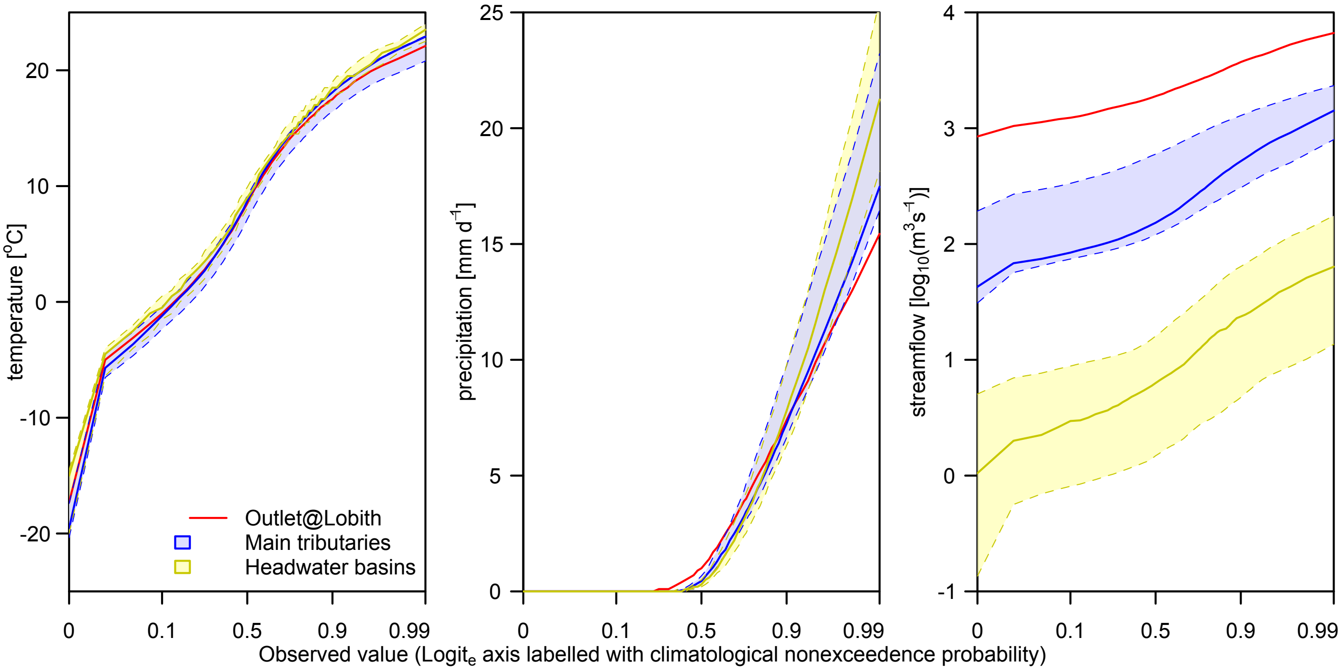

Figure 2: Distribution of daily averaged temperature, daily total precipitation and daily averaged streamflow in Rhine basins and stations. Three spatial scales are shown: 43 headwater basins, four large tributaries and the Rhine outlet at Lobith. For scales containing multiple locations, the median location is shown as a thick line and the 10th and 90th percentiles bound a shaded area.

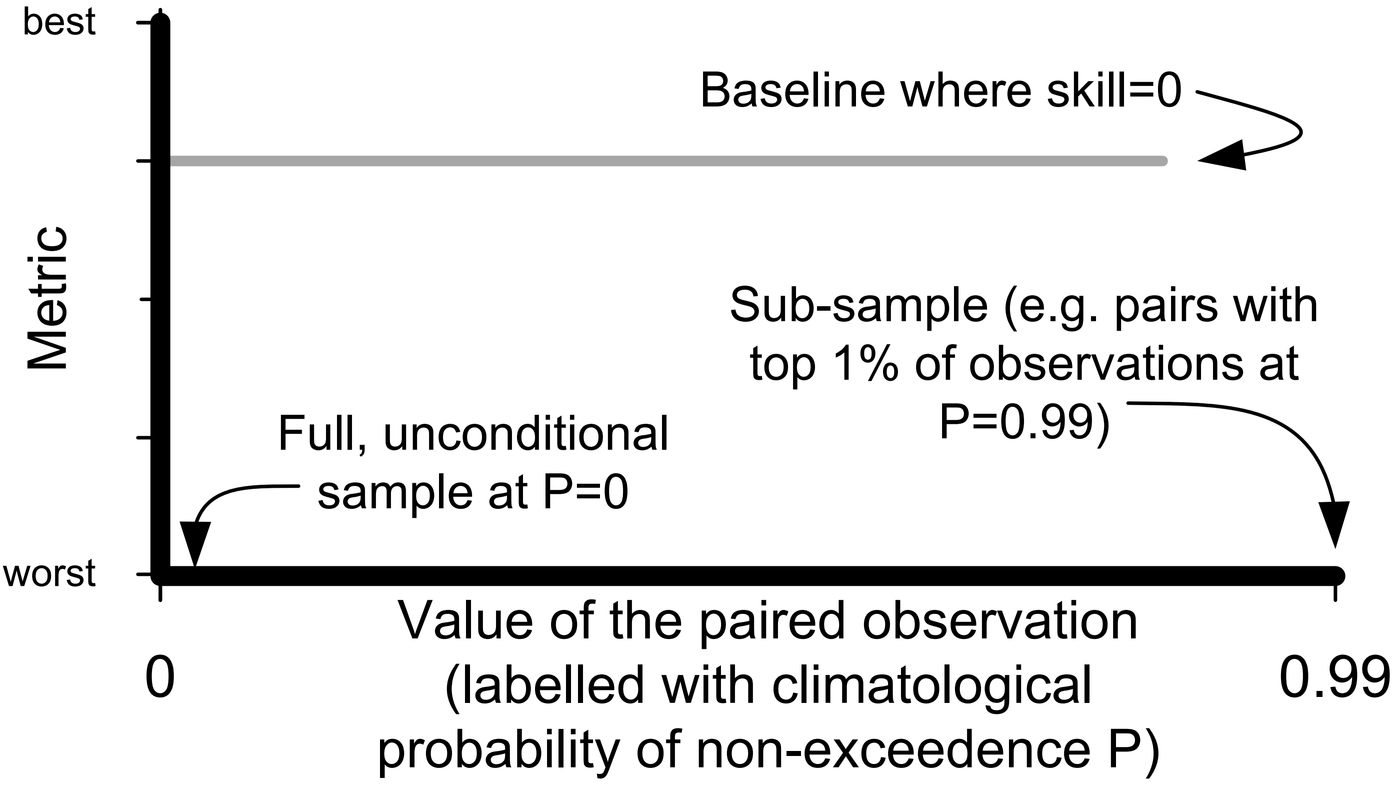
best
worst
Metric
Baseline where skill=0
Sub-sample (e.g. pairs with top 1% of observations at P=0.99)
Full, unconditional sample at P=0
0
0.99
Value of the paired observation (labelled with climatological probability of non-exceedence P)

Figure 3: Sample plot showing how the skill score plots are defined.

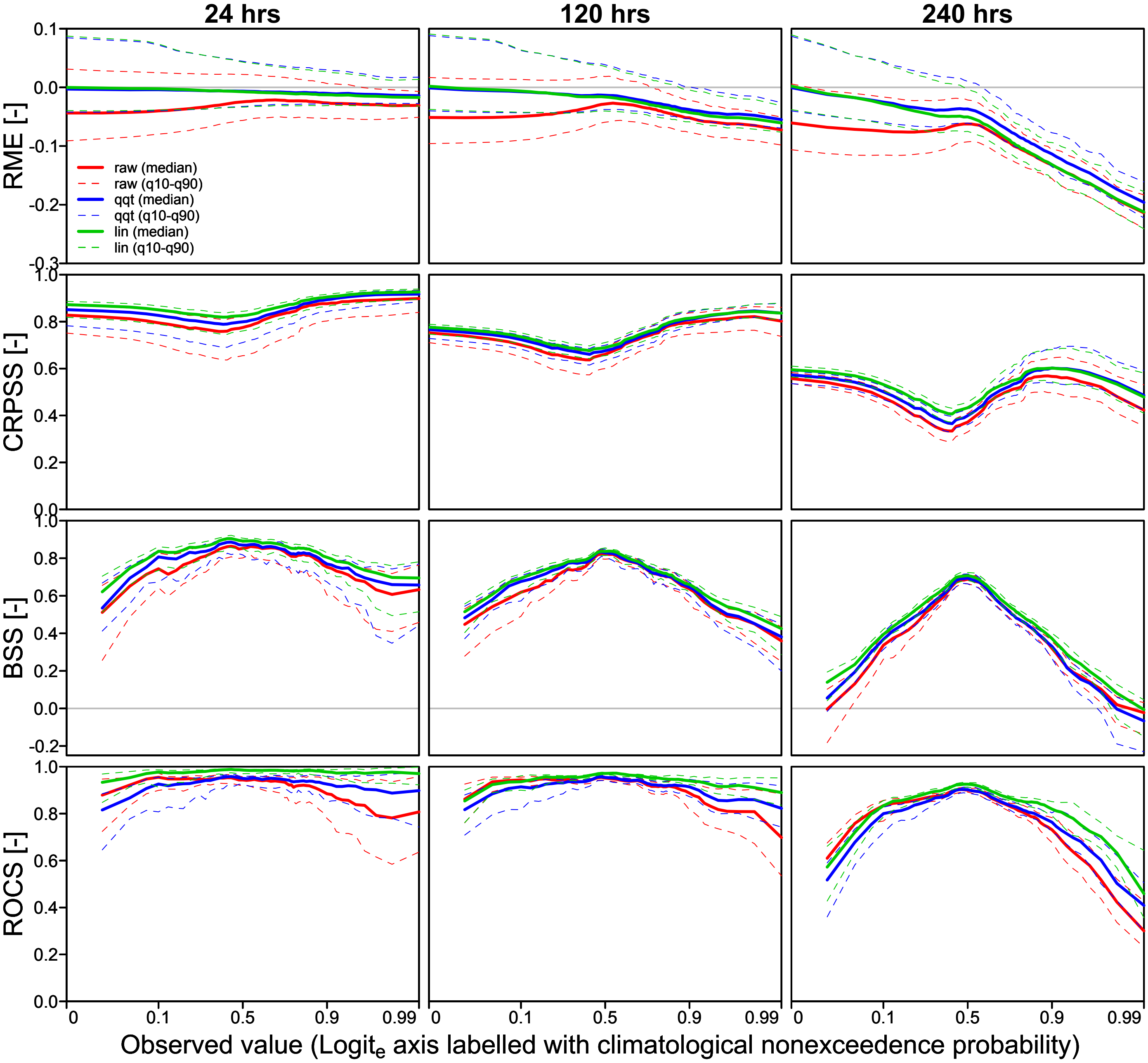

Figure 4: RME, CRPSS, BSS and ROCS for ensemble temperature forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

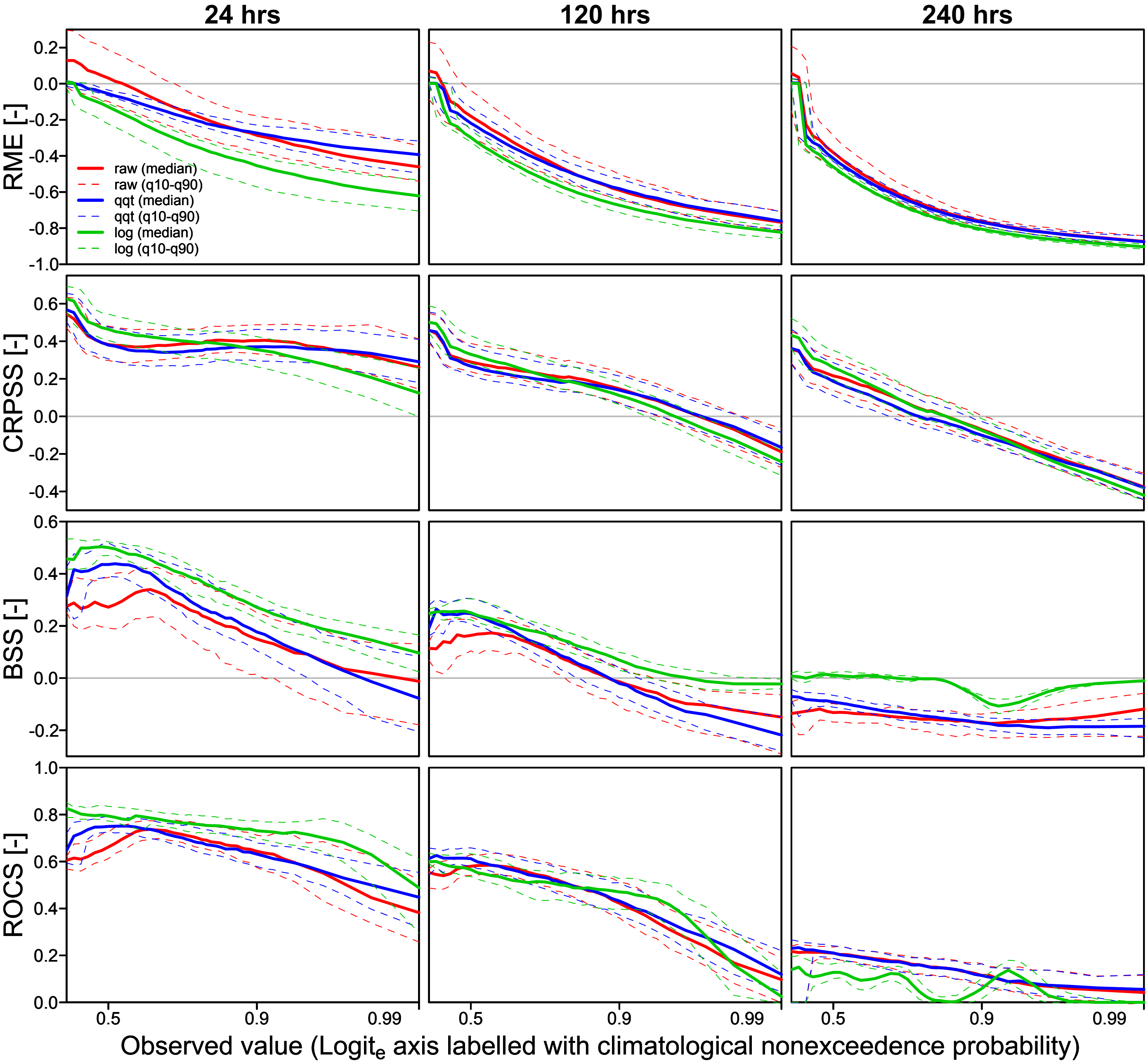

Figure 5: RME, CRPSS, BSS and ROCS for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

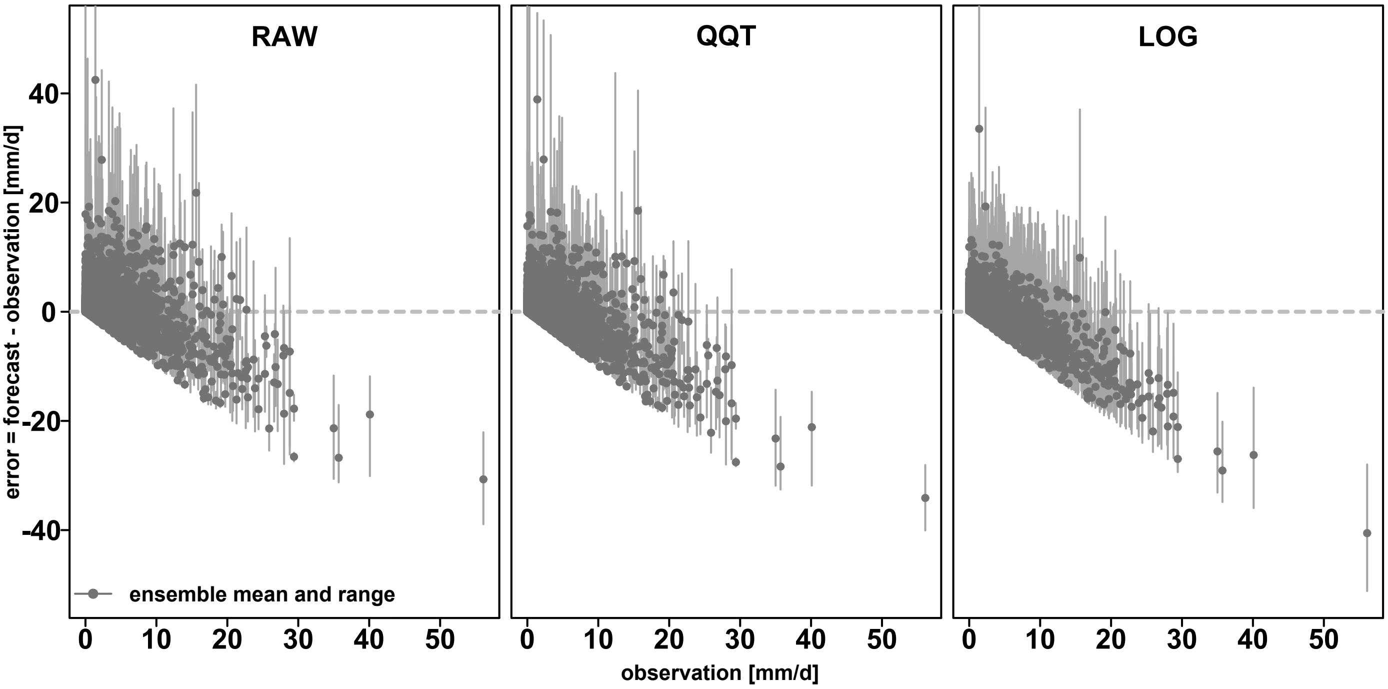

Figure 6: Forecast error versus observations for uncorrected (left), QQT-corrected (middle) and LOG-corrected precipitation ensembles for a single location (basin I-RN-0001, which is located in the Neckar sub-basin) at 120-hour lead time.

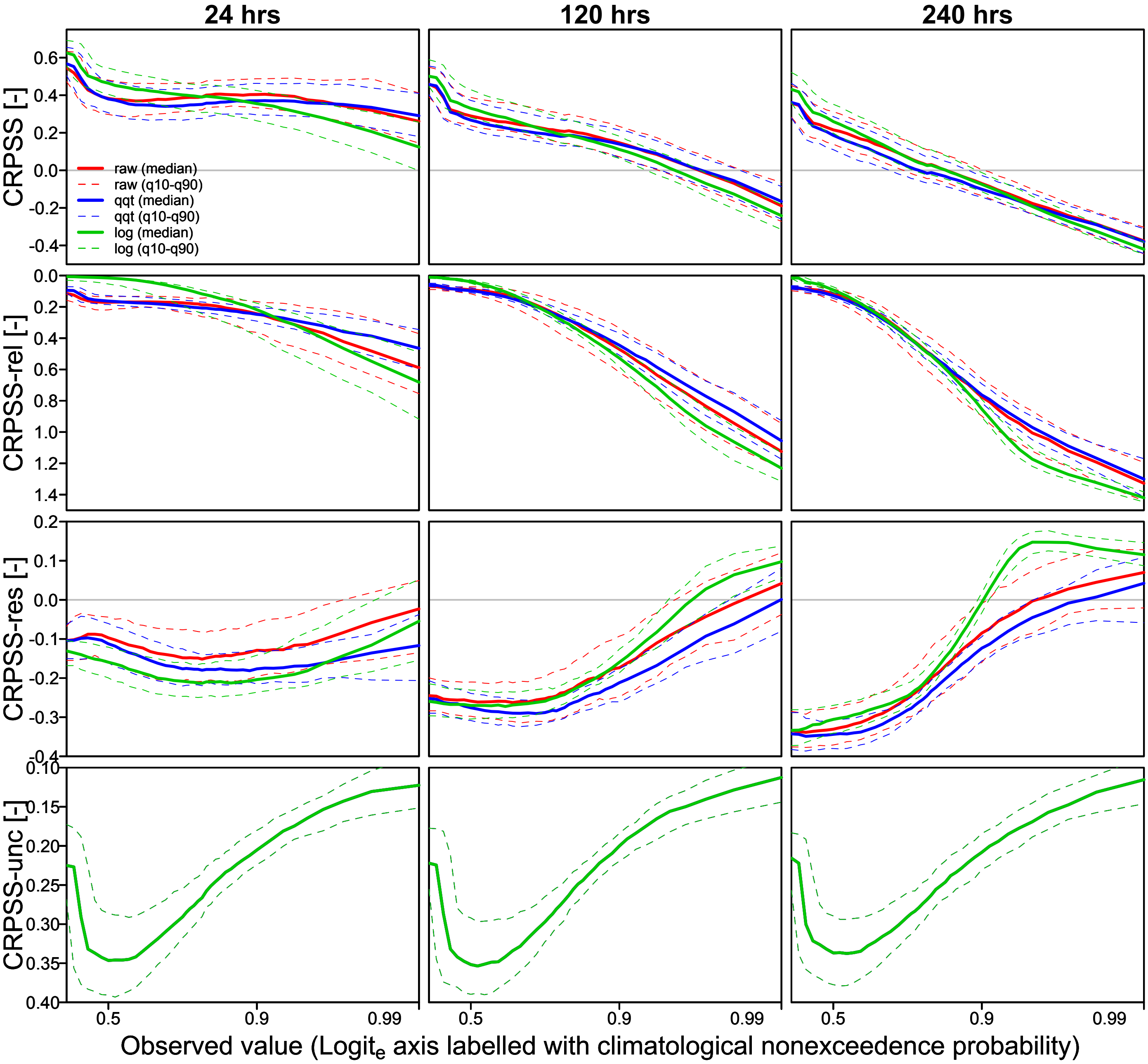

Figure 7: CRPSS calibration-refinement decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

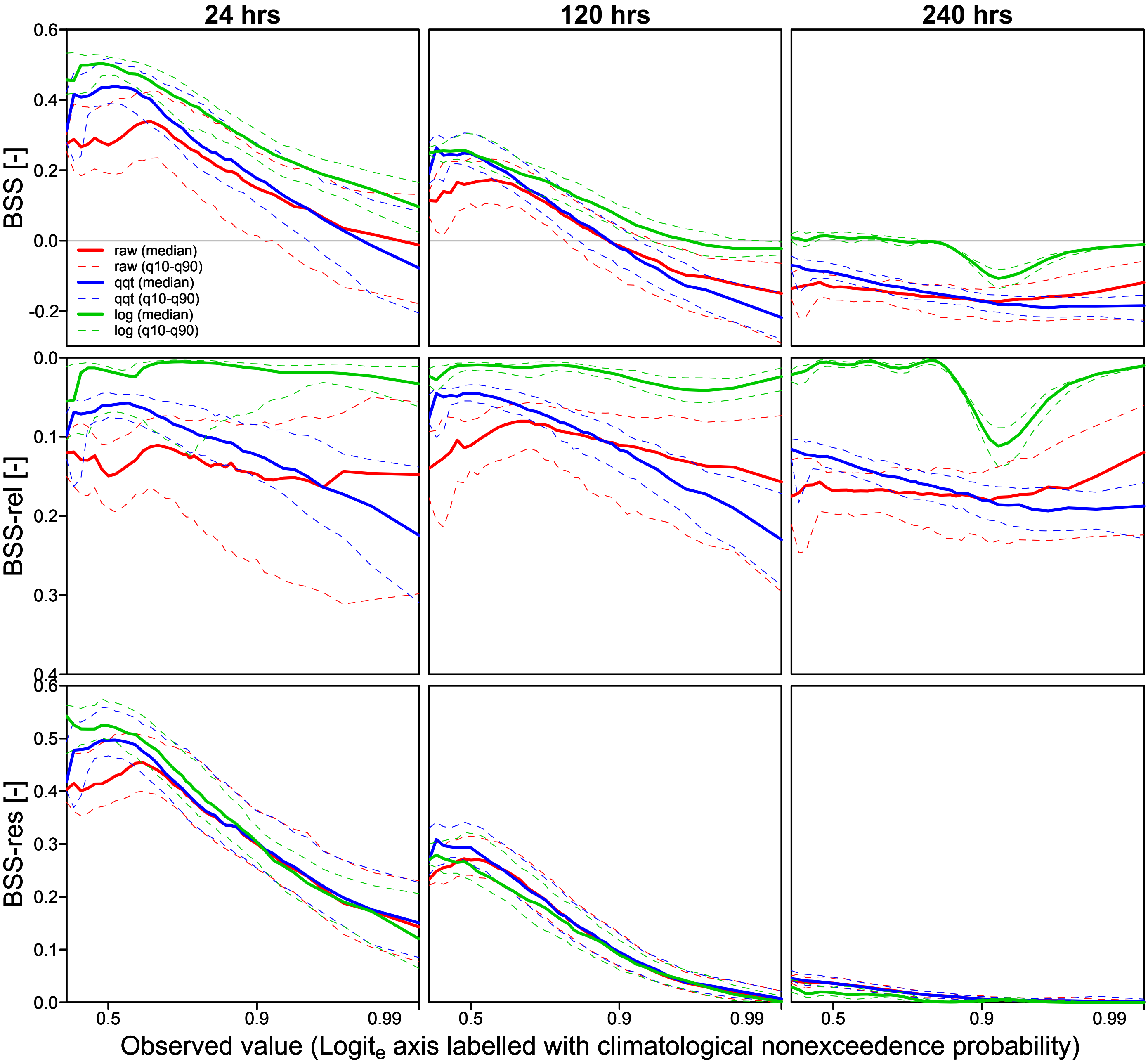

Figure 8: BSS calibration-refinement (Type I) decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

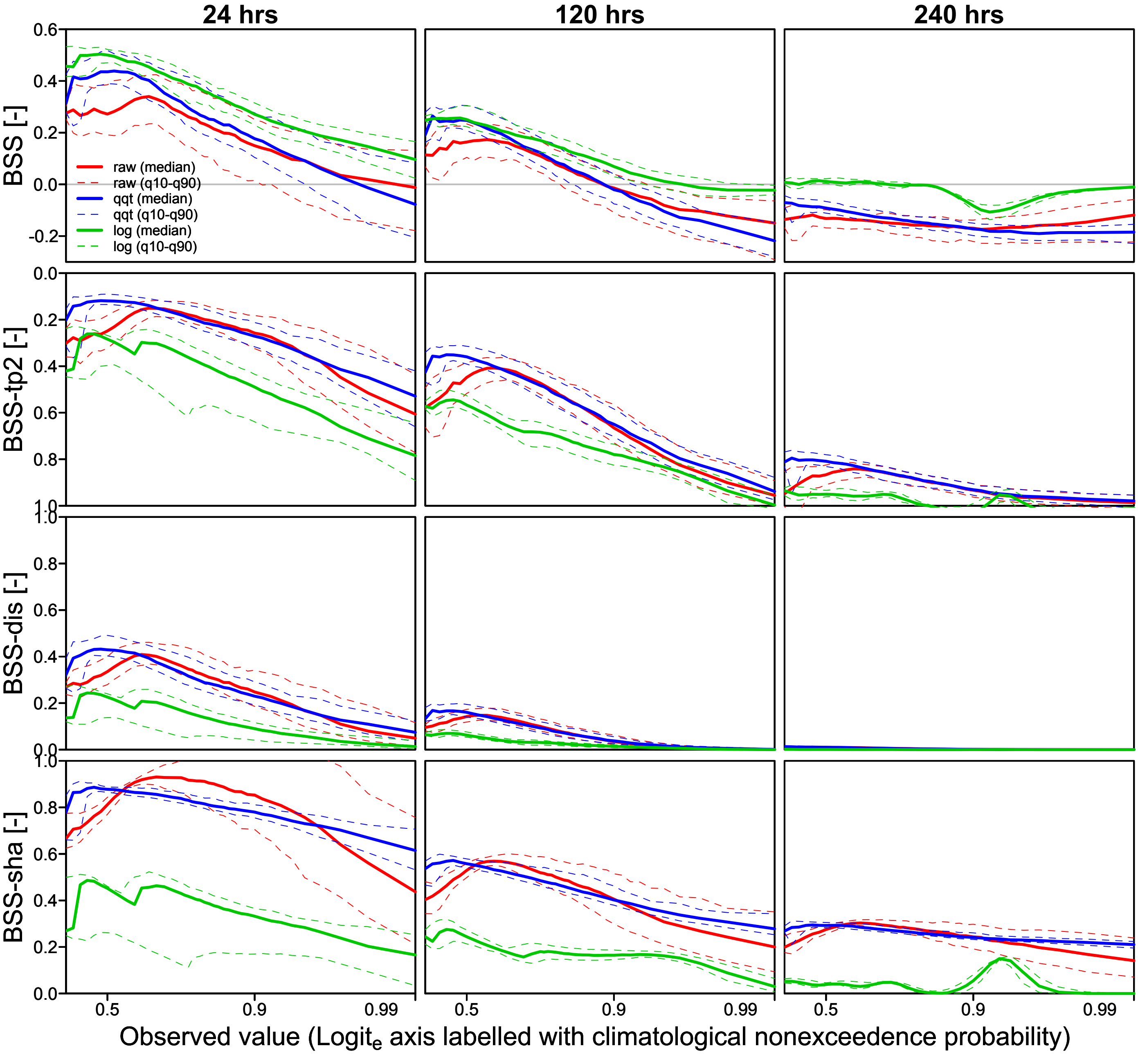

Figure 9: BSS likelihood-base rate (Type II) decomposition for ensemble precipitation forecasts at 24-hour, 120-hour and 240-hour ahead forecasts. The baseline is formed by sample climatology. The results pertain to 134 basins: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

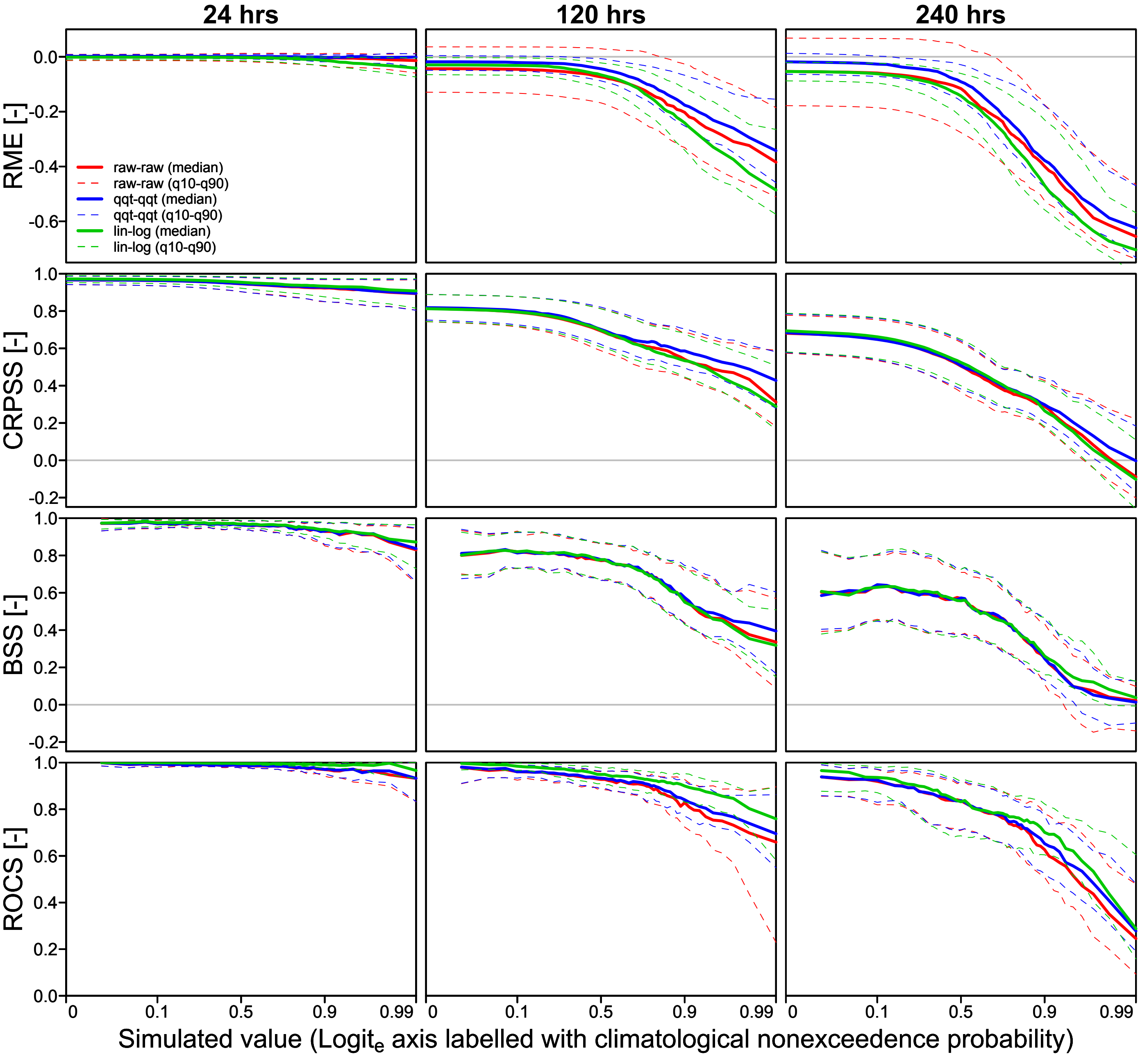

Figure 10: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

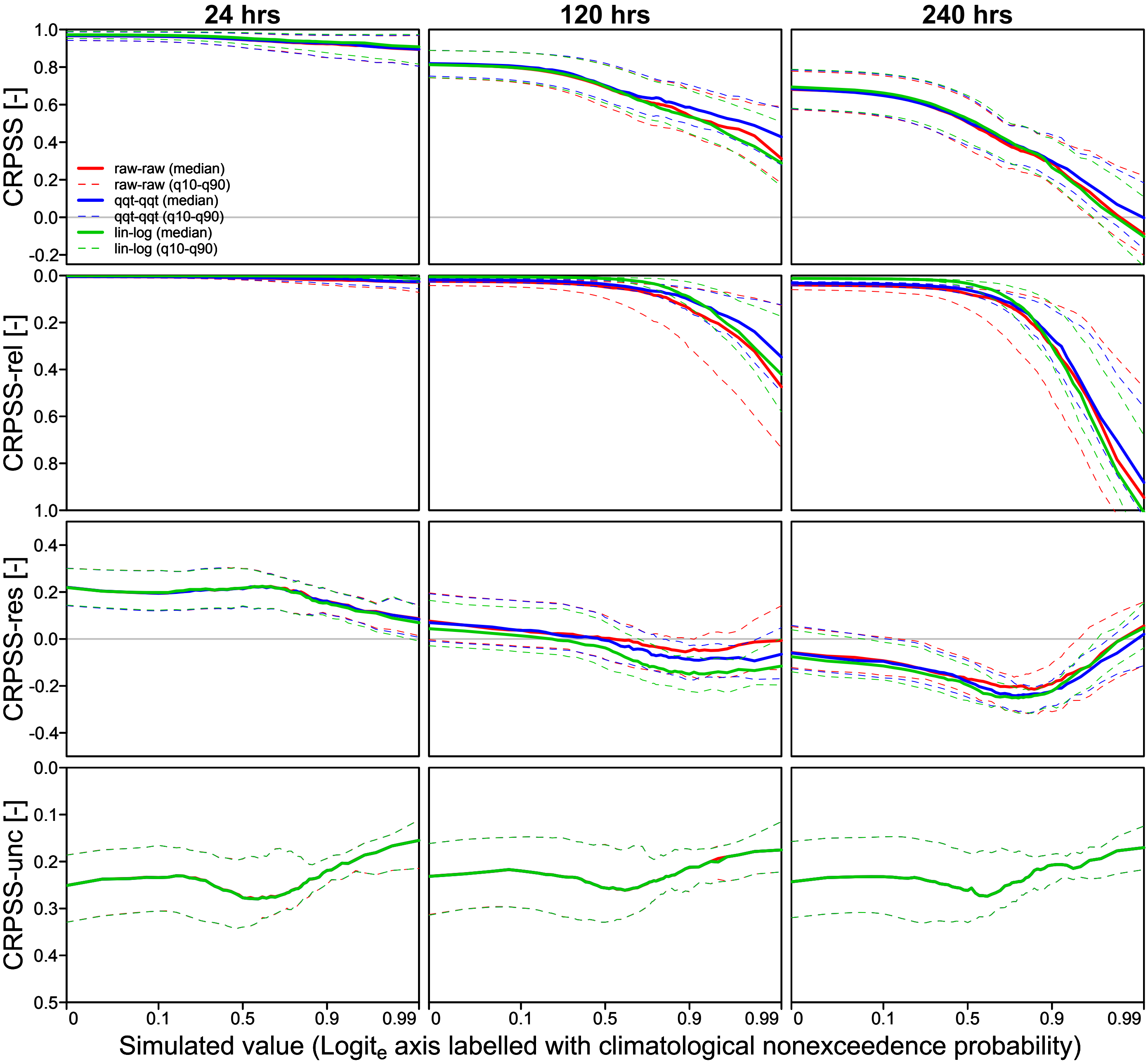

Figure 11: CRPSS calibration-refinement decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

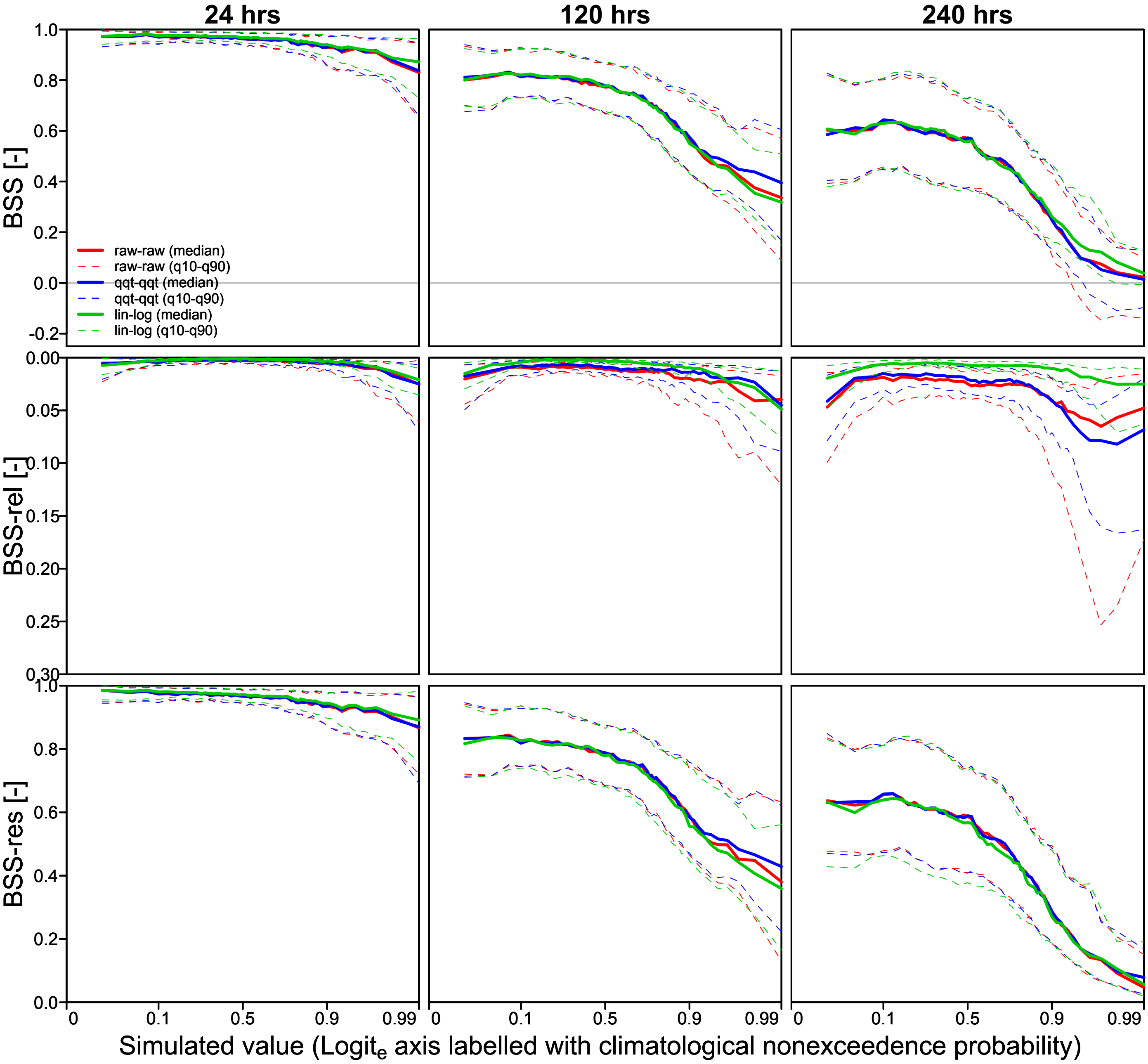

Figure 12: BSS calibration-refinement (Type I) decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

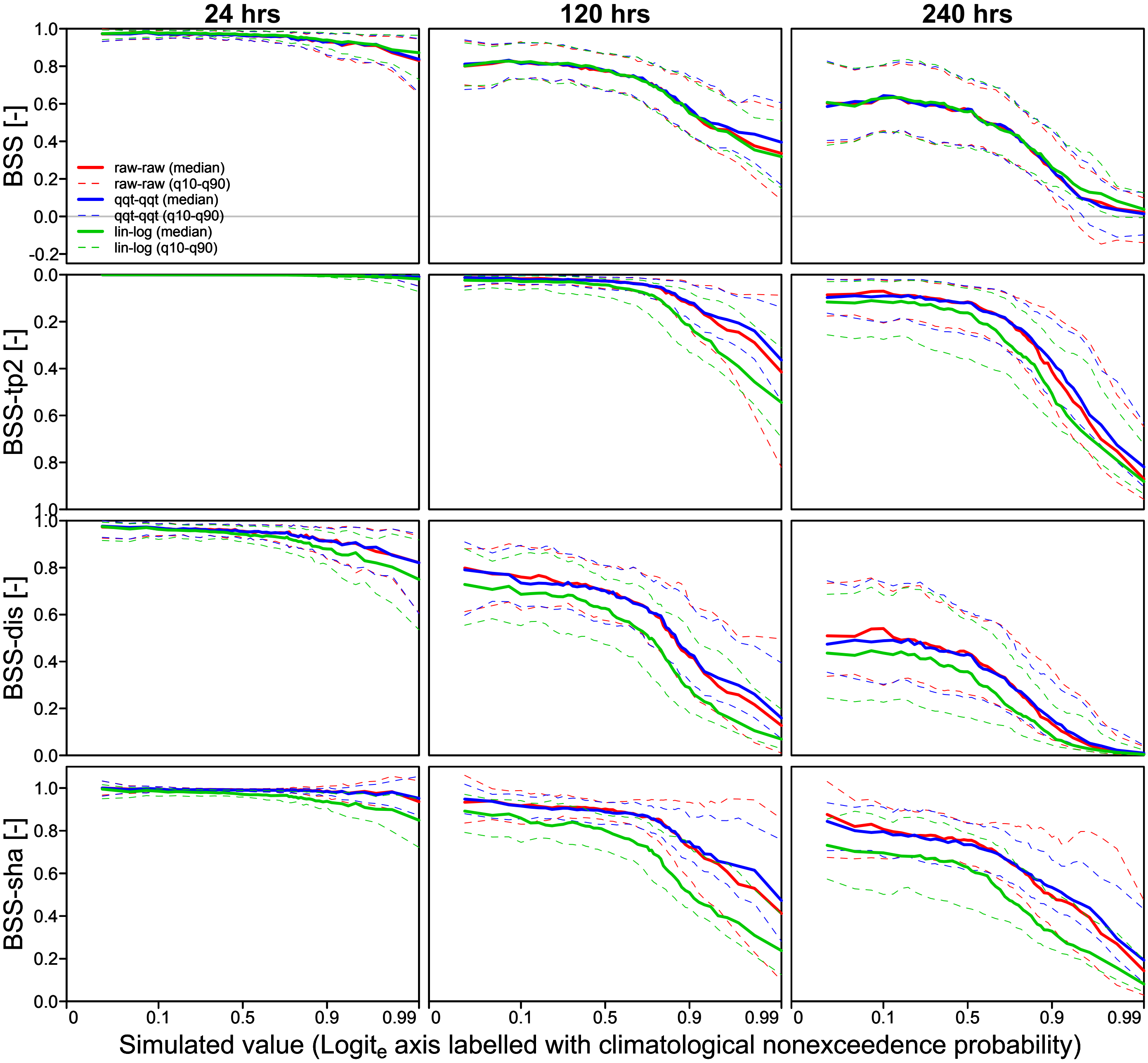

Figure 13: BSS likelihood-base rate (Type II) decomposition for ensemble streamflow forecasts for the headwater basins at 24-hour, 120-hour and 240-hour lead times. The baseline is formed by sample climatology. The results pertain to 43 locations: solid lines show the median value; dashed lines show the 0.10 and 0.90 quantiles.

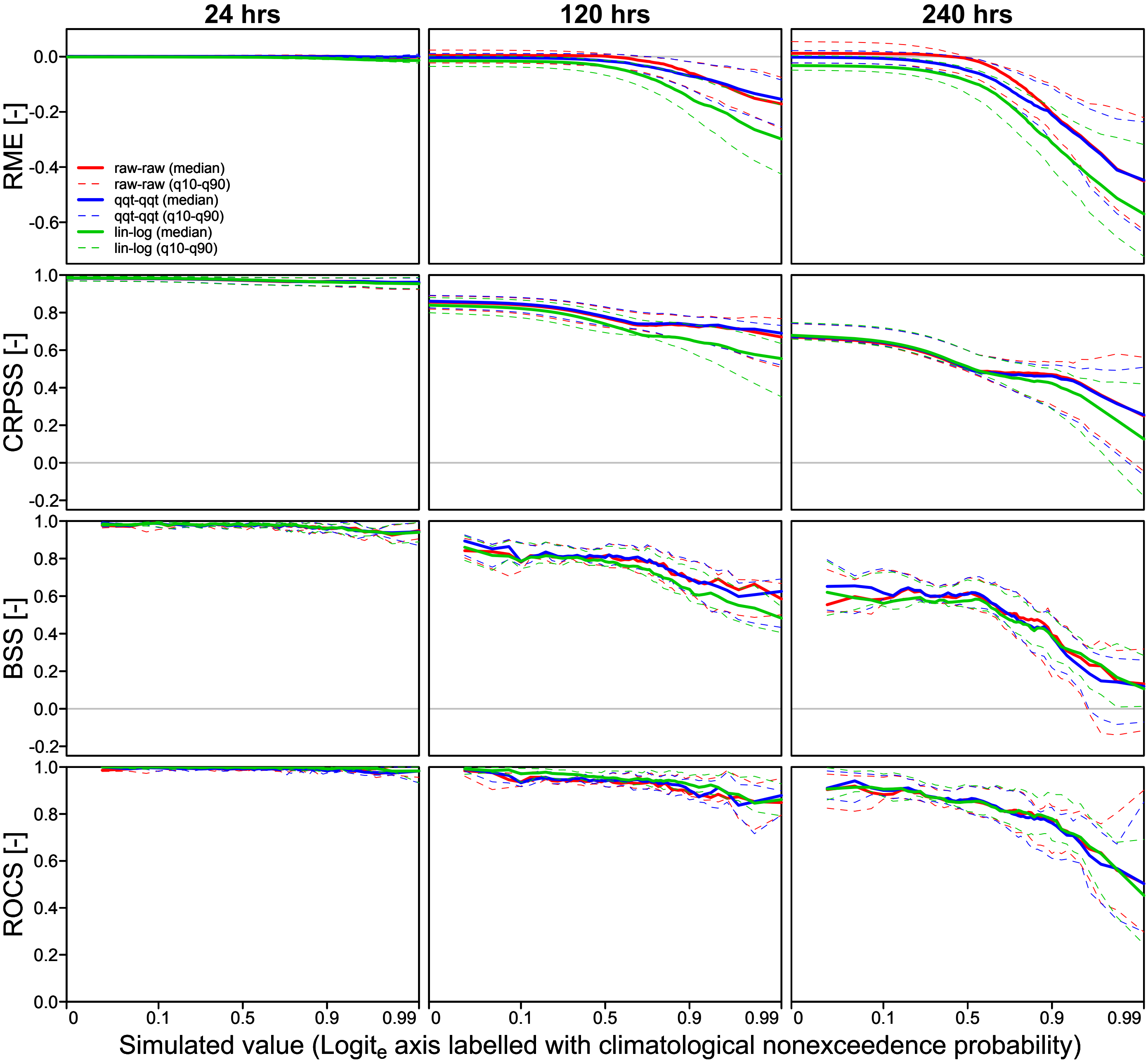

Figure 14: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the four main tributaries at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology. The results pertain to 4 locations: solid lines show the (interpolated) median value; dashed lines show the (interpolated) 0.10 and 0.90 quantiles.

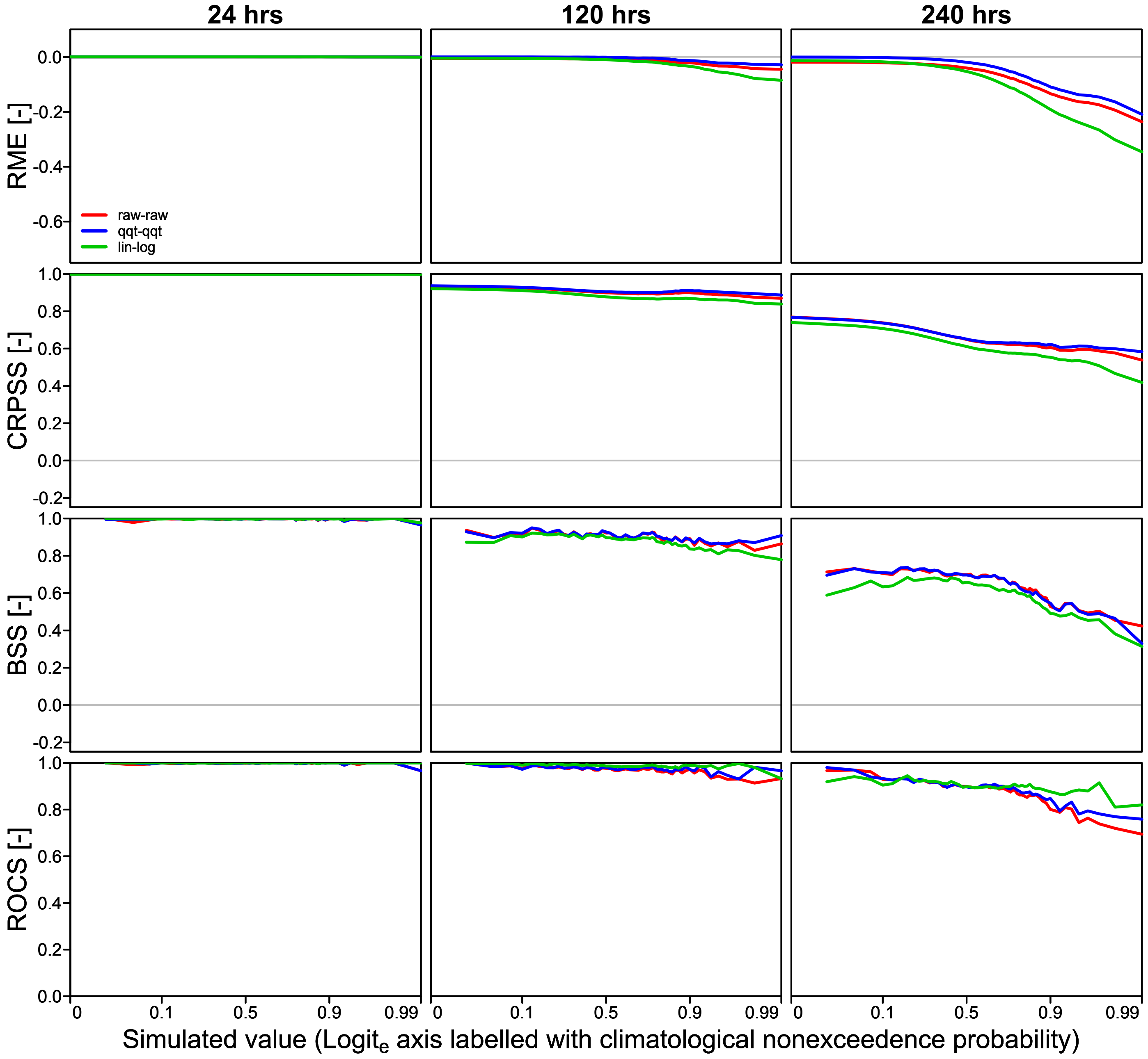

Figure 15: RME, CRPSS, BSS and ROCS for ensemble streamflow forecasts for the outlet at Lobith at 24-hour, 120-hour and 240-hour lead times. For CRPSS, BSS and ROCS, the baseline is formed by sample climatology.