Verification of temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service: an evaluation of the medium-range forecasts with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS) and a comparison to the frozen version of NCEP's Global Forecast System (GFS)

Revision number: Final

Prepared by Hydrologic Solutions Limited for the U.S. National Weather Service under Subcontract Agreement 2013-09 with LEN Technologies Incorporated (in fulfillment of Deliverable No. 2)

Dr. James Brown (james.brown@hydrosolved.com)

Friday, March 28, 2014

Abstract

Retrospective forecasts of temperature, precipitation, and streamflow were generated with the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service (NWS) for selected river basins in four NWS River Forecast Centers (RFCs), namely the Arkansas-Red Basin RFC (ABRFC), the Colorado Basin RFC (CBRFC), the California-Nevada RFC (CNRFC) and the Middle Atlantic RFC (MARFC). The meteorological hindcasts were produced with the HEFS Meteorological Ensemble Forecast Processor (MEFP). The MEFP was calibrated with forcing inputs from the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP). The streamflow hindcasts cover a ~15 year period from 1985-1999 with a forecast horizon of 1-14 days. Retrospective forecasts were also produced with the frozen (circa 1997) version of NCEP's Global Forecast System (GFS).The hindcasts were verified conditionally upon forecast lead time, magnitude of the observed and forecast variables, and season. Verification results are presented for the temperature and precipitation forecasts from the MEFP and for the streamflow forecasts before and after bias-correction with the HEFS Ensemble Postprocessor (EnsPost). This report presents the verification results, describes the expected performance and limitations of the HEFS for short- to medium-range forecasting with the GEFS, identifies the benefits of the GEFS when compared to the frozen GFS, and provides recommendations on future research and additional evaluation of the HEFS.

Document history

Action	Version	Person	Date
Complete first draft	1.0	James Brown	12/19/2013
Comments from Limin Wu	1.1	James Brown	12/27/2013
Comments from Haksu Lee	1.2	James Brown	12/27/2013
Minor edits	1.3	James Brown	12/27/2013
Added Table 3	1.4	James Brown	01/06/2014
Final edits based on OHD reviews	Final	James Brown	03/28/2014

Acknowledgements

This report was prepared by Hydrologic Solutions Limited under Subcontract Agreement 2013-09 with LEN Technologies Incorporated. The temperature, precipitation and streamflow hindcasts were prepared by the Office of Hydrologic Development (OHD), notably Kevin He and Xiaoshen Li (CHPS configurations and streamflow hindcasting), Limin Wu (MEFP calibration and hindcasting) and Satish Regonda (EnsPost calibration). The NWS RFCs, specifically AB-, CB-, CN- and MA-RFCs, provided guidance on the CHPS configurations, and some of the data required for hindcasting and verification. An earlier version of this report was reviewed by Limin Wu, Haksu Lee, Hank Herr and Mark Fresch.

CONTENTS

1. How to read this report

This report aims to: 1) provide a comprehensive scientific evaluation of the temperature, precipitation, and streamflow forecasts from the Hydrologic Ensemble Forecast Service Version 1 (HEFSv1) with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS); 2) benchmark the HEFSv1 with forcing inputs from the GEFS against a frozen (circa 1997) version of NCEP's Global Forecast System (GFS), in order to establish the benefits of the GEFS for operational hydrologic forecasting; and 3) communicate the strengths and weaknesses of the HEFSv1 and, where necessary, recommend specific enhancements or further studies. This section aims to guide readers with limited time or experience of ensemble forecasting or verification to the main results and conclusions. The following sections are particularly important:

I. Executive summary and recommendations. This describes the structure of the report and the strengths and weakness of the forecasts in non-technical terms;

II. Section 4.1. This provides a brief description of the study basins. Understanding the hydrology of the study basins is central to interpreting the quality of the HEFS forecasts and to applying the results more broadly (or understanding the risks of extrapolation);

III. Appendix C. This shows a selection of the paired streamflow forecasts and observations from which the verification results were derived. The plots comprise the bias-corrected streamflow forecasts with forcing inputs from the Meteorological Ensemble Forecast Processor (MEFP). The MEFP was calibrated with raw temperature and precipitation forecasts from the GEFS (MEFP-GEFS), the frozen GFS (MEFP-GFS) and a conditional or "resampled" climatology (MEFP-CLIM). The relative scatter of the observations within the ensemble forecast distribution provides some insight into the quality of the streamflow forecasts when using different forcing inputs. In general, the streamflow observations should fall randomly within the ensemble range. They should not fall consistently in one part of the ensemble forecast distribution or outside of the ensemble range;

IV. Section 4.4 and Appendix B. In order to understand the remainder of the report, it is necessary to consider the desirable attributes of ensemble forecasts and how they can be measured. Tutorials on forecast verification can be found in the documentation, presentations, and exercises that accompany recent training workshops on the HEFS and in the user's manual of the Ensemble Verification System (EVS). Important attributes of forecast quality are briefly described in Section 4.4, while Appendix B summarizes the verification measures used in this report; and

V. Section 5.3. The verification results are presented separately for the meteorological forecasts and the "raw" streamflow forecasts (which do not include streamflow post-processing). In Section 5.2, the raw streamflow forecasts are verified against simulated streamflows, in order to establish the potential benefits of the MEFP-GEFS forecasts without the impacts of hydrologic biases (the simulations and forecasts comprise the same hydrologic biases). In Section 5.3, the bias-corrected streamflow forecasts are verified against observed streamflows, in order to establish the actual benefits of the MEFP-GEFS forecasts in an operational context (where the streamflow bias-correction is often imperfect).

Some of the verification results are simpler to understand than others. Skill scores are generally simpler to understand and to compare between basins, partly because they are dimensionless. A skill score measures the fractional improvement of one forecasting system relative to another (0→1, although negative values are possible). For example, Figure 5 shows the fractional improvement of the MEFP-GEFS precipitation forecasts against sample climatology (an ensemble derived from the full, unconditional, sample of historical observations across all available dates). The results are also shown for the MEFP-GFS forecasts and for resampled climatology, MEFP-CLIM (an ensemble derived from a conditional sample of historical observations; that is, from sampling observations in a moving window around the forecast valid date across all historical years). Figure 6 shows the corresponding results for the MEFP temperature forecasts. Figures 25a/b show the skill of the bias-corrected streamflow forecasts with forcing inputs from the MEFP-GEFS and the MEFP-GFS. The baseline comprises the uncorrected streamflow forecasts with forcing inputs from resampled climatology, MEFP-CLIM. In addition to the overall skill, the contributions from the meteorological forcing and the hydrologic post-processing are shown separately. The improvement of one forecasting system over another can also be expressed as an average gain in forecast lead time (Figure 15). Using this approach, Figure 14 compares the MEFP-GEFS precipitation forecasts against the MEFP-GFS forecasts, while Figure 16 compares the MEFP-GEFS temperature forecasts against the MEFP-GFS forecasts. Figure 26 shows the corresponding results for the bias-corrected streamflow forecasts.

It is also important to understand the limitations of this study. First, it does not provide any guidance on the calibration or configuration of the HEFS. Such guidance would require hindcasting and verification for multiple calibration and configuration scenarios. Second, the report covers only a small fraction of the locations and scenarios under which the HEFS will be used operationally. It focuses on a pair of basins in each RFC, comprising one downstream basin with a single headwater. It does not consider the quality of the forecasts in regulated rivers, where the hydrologic errors are often difficult to model statistically. Similarly, it does not include rivers with multiple upstream contributions or lateral inflows, from which the uncertainties combine and propagate downstream. Finally, the report does not explicitly benchmark the HEFS against archived operational forecasts, such as the RFC single-valued forecasts, or operational forecast products, such as river flood warnings. Scientific evaluation of the HEFS is an ongoing activity. It requires an infrastructure for hindcasting, verification and archiving of data, as well as communicating verification concepts and results. This report provides an initial evaluation only. Further, targeted, evaluations should be conducted by the NWS RFCs in collaboration with the Office of Hydrologic Development (OHD).

2. Executive summary and recommendations

- Retrospective forecasts of temperature, precipitation, and streamflow were generated with the Hydrologic Ensemble Forecasts Service (HEFS) for a ~15-year period between 1985 and 1999. The hindcasts were produced for two basins in each of four River Forecast Centers (RFCs), namely the Arkansas-Red Basin RFC (ABRFC), the Colorado Basin RFC (CBRFC), the California-Nevada RFC (CNRFC) and the Middle Atlantic RFC (MARFC). The precipitation and temperature forecasts were generated with the HEFS Meteorological Ensemble Forecast Processor (MEFP). The MEFP produces ensemble forecasts of Mean Areal Temperature (MAT) and Mean Areal Precipitation (MAP), conditionally upon a raw, single-valued, forecast. Here, the MEFP was calibrated with the ensemble mean of the Global Ensemble Forecast System (GEFS). The GEFS uses Version 9.0.1 of the Global Forecast System (GFS), which comprises a horizontal resolution of T254 (~55km) for 1-8 days and T190 (~70km) for 9-16 days, together with a vertical resolution of L42 (42 vertical levels). In order to benchmark the forcing and streamflow hindcasts produced with the GEFS (denoted MEFP-GEFS), the MEFP was also calibrated with the frozen (circa 1997) version of the GFS (denoted MEFP-GFS), and a conditional or "resampled" climatology (denoted MEFP-CLIM). The latter involved resampling historical observations in a moving window around the forecast valid date. The streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS) and were bias-corrected with the Ensemble Post-processor (EnsPost). In all cases, the forecast time horizon was 1-14 days.
- The HEFS is being evaluated in several phases. The phased evaluation aims to: establish the expected performance and limitations of the HEFS; demonstrate that the outputs from the MEFP and the EnsPost are less biased and more skillful than the inputs; identify the key factors responsible for forecast error and skill in different situations; isolate the contributions from the meteorological and hydrologic uncertainties to the overall skill of the streamflow forecasts; establish a baseline for enhancements to the HEFS and, where appropriate, to recommend specific

enhancements or further studies; and illustrate how hindcasting and verification of the HEFS might be conducted in future. The temperature, precipitation, and streamflow forecasts were verified with the Ensemble Verification System. The results are presented by forecast lead time, season, and magnitude of the observed and forecast variables. The precipitation and temperature forecasts were verified against observed MAP and MAT, respectively. In order to establish the potential benefits of the meteorological forecasts separately from any hydrologic biases, the raw streamflow forecasts were verified against simulated flows. In addition, the bias-corrected streamflow forecasts were verified against observed flows. This allowed the actual benefits of the meteorological forecasts to be established in an operational context, where the hydrologic uncertainties may outweigh the meteorological uncertainties, and the EnsPost cannot be expected to remove all of the hydrologic biases.

- The correlations between the raw meteorological forecasts and observations are preserved or improved by the MEFP at all forecast lead times, in both seasons, and at all magnitudes of the observed variable. Also, the MEFP-GEFS forecasts consistently improve upon the MEFP-CLIM forecasts. Both are necessary, if not sufficient, attributes of reliable and skillful meteorological forecasts. Indeed, the MEFP aims to preserve the correlations in the raw, single-valued, forecasts and to produce ensemble forecasts that are reliable and *no less skillful* than resampled climatology.
- In general, the patterns of skill and bias in the MEFP-GFS forecasts are mirrored by the MEFP-GEFS forecasts. The MEFP-GEFS forecasts show much higher correlations and greater skill in CNRFC than in AB- or CB-RFCs. This is associated with the greater predictability of large storms in the California Coastal Ranges during the winter months. In MARFC, the MEFP-GEFS precipitation forecasts are highly skillful at early forecast lead times, particularly at moderate precipitation amounts, but the forecast skill declines more rapidly when compared to CNRFC. In keeping with the MEFP-GFS precipitation forecasts, the MEFP-GEFS forecasts are consistently less skillful in AB- and CB-RFCs than MA- and CN-RFCs (at least

during the first few days). This originates from a combination of reduced predictability in the southern plains and in the intermountain region of the western U.S., together with residual biases that were not removed by the MEFP. In general, both the MEFP-GEFS precipitation forecasts and the MEFP-GFS forecasts are unbiased and skillful during the first week, but show much lower skill and higher conditional biases during the second week.

- Despite the broad similarities between the MEFP-GEFS precipitation forecasts and the MEFP-GFS forecasts, the MEFP-GEFS forecasts show higher correlations and greater skill than the MEFP-GFS forecasts. In AB-, CB- and MA-RFCs, the MEFP-GEFS forecasts show higher correlations than the MEFP-GFS forecasts at all forecast lead times. In CNRFC, the improvements from the MEFP-GEFS are greatest after ~three days, as the raw GFS forecasts show similar correlations to the raw GEFS forecasts between 1-2 days (~0.8). During the first week, the MEFP-GEFS precipitation forecasts are consistently more skillful than the MEFP-GFS forecasts in AB-, CB- and MA-RFCs. However, after seven days, they are no more skillful than climatology, and, in CBRFC, the forecasts of light precipitation are somewhat less skillful than climatology. In CNRFC, the MEFP-GEFS precipitation forecasts are no more skillful than the MEFP-GFS forecasts between 1-2 days. However, during the second week, they are substantially more skillful than the MEFP-GFS forecasts, particularly at higher precipitation thresholds. When expressed as a net gain in forecast lead time over the period of skillful forcing, the MEFP-GEFS forecasts typically add 1-2 days in forecast lead time when compared to the MEFP-GFS forecasts.

- Both the MEFP-GEFS temperature forecasts and the MEFP-GFS temperature forecasts improve substantially upon resampled climatology. They also remain skillful for longer than the precipitation forecasts. Indeed, the MEFP-GEFS temperature forecasts remain skillful throughout the second week. However, in keeping with the precipitation forecasts, the benefits of the MEFP-GEFS temperature forecasts are generally more pronounced after the first 1-2 days. Thus, the errors saturate more quickly in the MEFP-GFS forecasts than the MEFP-

GEFS forecasts. For example, in the middle portion of the forecast horizon, the MEFP-GEFS forecasts show equivalent skill to the MEFP-GFS forecasts, but with 2-4 days of additional forecast lead time. The greatest improvements occur in CN-FTSC1 (Fort Seward, CA) and MA-CNNN6 (Cannonsville Reservoir, NY). In these basins, when measured against sample climatology, the MEFP-GEFS temperature forecasts are ~20% more skillful than the MEFP-GFS forecasts across all temperature thresholds. In general, the improvements are smaller at AB-BLKO2 (Blackwell, OK) and CB-DOLC2 (Dolores, CO). However, during the winter months (as well as the summer months at CB-DOLC2), the value added by the MEFP-GEFS increases when colder temperatures are included in the verification data. For example, during the winter months at CB-DOLC2, the mean Continuous Ranked Probability Skill Score (CRPSS) is ~0.6 when using the MEFP-GEFS to forecast temperatures above -8 °C (the 90% exceedence threshold) and ~0.4 when using the MEFP-GFS. While accurate forecasts of MAT are generally less important for hydrologic modeling than accurate forecasts of MAP, surface temperatures are important in determining the accumulation and melting of snow. Thus, in snow-dominated basins, such as CB-DOLC2, the additional skill of the MEFP-GEFS temperature forecasts may be important for hydrologic modeling.

- Both the MEFP-GEFS forecast and the MEFP-GFS forecasts comprise a range of conditional biases. In particular, there is a tendency for the precipitation forecasts to underestimate the Probability of Precipitation (PoP). This lack of reliability also effects the MEFP-CLIM forecasts. Indeed, the MEFP forecasts of PoP are substantially worse than unconditional climatology in some basins. In order to produce reliable forecasts of PoP at a daily accumulation, the forecasts must be reliable at a six-hourly accumulation. Furthermore, they must adequately capture the statistical dependencies between the six-hourly accumulations. In practice, the forecasts of PoP are unreliable at a daily accumulation, while the corresponding six-hourly forecasts are reliable, on average. This alludes to a problem with the modeling of precipitation intermittency at a six-hourly scale. More specifically, is alludes to a problem with the temporal variability of precipitation intermittency in the six-hourly MEFP forecasts.

Recommendation 1: Accurate modeling of the space-time covariability between precipitation and temperature is central to producing reliable meteorological forecast at aggregated scales. It is also important for hydrologic forecasting, as the outputs from hydrologic models are sensitive to the space-time covariability of the inputs. The MEFP uses the “Schaake Shuffle” to reproduce the historical space-time covariability of the observed MAT and MAP conditionally upon forecast valid date. It does not model these patterns conditionally upon the state of the atmosphere. Further investigation is warranted into the limitations of the Schaake Shuffle in reproducing the space-time covariability of precipitation and temperature, including whether other empirical structures (empirical copulas) can generate more realistic patterns. This investigation should consider a range of observed and forecast conditions, including moderate and intermittent precipitation, but also large and extreme events, where the space-time covariability may be substantially different than climatology.

- Alongside the underestimation of PoP, the MEFP precipitation forecasts systematically underestimate the largest observed precipitation amounts. This originates from a Type-II conditional bias in the precipitation forecasts (as distinct from a Type-I conditional bias or “lack of reliability”). Again, it is apparent in the MEFP-GEFS precipitation forecasts, as well as the MEFP-GFS forecasts. In general, the conditional bias increases as the forecast skill declines (i.e. approaches climatology); hence, it varies with location, season and forecast lead time, among other factors. This is understandable, because climatology is, by definition, conditionally biased with increasing amounts of observed precipitation. At early forecast lead times in CNRFC, the biases are sufficiently small, and the spread is sufficiently large, that the highest precipitation totals are generally forecast with some, non-zero, probability of occurrence. However, in other basins, and at longer forecast lead times, the largest precipitation totals are routinely underestimated by as much as the observed precipitation amount. While the MEFP-GEFS forecasts show similar conditional biases to the MEFP-GFS forecasts, they also comprise more spread in some cases. For example, at AB-BLKO2, the MEFP-GEFS forecasts are more likely to warn of the highest observed precipitation amounts, even if their central tendency is to underestimate. Currently, the MEFP is calibrated with the ensemble mean of the raw GEFS forecasts. Most of the skill in the frozen GFS is concentrated in the ensemble mean forecast. However, as atmospheric models become more skillful, post-processors may benefit from using higher moments, interactions, or even the individual ensemble members, providing the sampling uncertainties are reasonably small.

Recommendation 2: Future work should consider whether the raw temperature and precipitation forecasts used by the MEFP (notably the GEFS forecasts) contain valuable information in the ensemble spread and higher central moments, and how best to leverage this information. In this context, there is a trade-off between adding skillful predictors and the need to maintain a parsimonious description of the forecast errors. More generally, further work is needed on the limitations of statistical post-processing for large and extreme events. Here, the desire for unbiasedness must be weighed against the risk of obfuscating a weak, but potentially valuable, signal in the raw forecasts. The ability to calibrate the MEFP with reasonably small sampling uncertainty is important in this context. Thus, future work should leverage all of the available GEFS reforecasts and corresponding operational forecasts.

- In order to understand the benefits of the MEFP-GEFS forcing independently of any hydrologic biases, the raw streamflow forecasts were verified against simulated flows. In general, both the MEFP-GEFS streamflow forecasts and the MEFP-GFS forecasts are substantially more skillful than those with climatological forcing. Similarly, when compared to the MEFP-GFS streamflow forecasts, the MEFP-GEFS forecasts are consistently more correlated with the simulated streamflows and show higher skill. As the hydrologic models respond unevenly to meteorological forcing, depending on basin characteristics and antecedent conditions, the period over which the MEFP-GEFS forecasts improve upon the MEFP-GFS forecasts varies between basins. For example, at AB-BLKO2, the streamflow forecasts show a rapid decline in correlation with increasing forecast lead time. This originates from a lack of hydrologic persistence at AB-BLKO2 and the difficulty in forecasting precipitation beyond the short-range. In contrast, the basins in CBRFC are dominated by snow accumulation and melting. Here, much of the skill in the streamflow forecasts depends on the hydrologic uncertainties, specifically on the initial conditions in the hydrologic models. However, the timing and rate of snowmelt also depends on the accuracy of the temperature forecasts during the snowmelt period. When verifying the raw streamflow forecasts against simulated flows, the MEFP-GEFS forecasts are substantially more skillful than the equivalent MEFP-GFS forecasts. For example, at CB-DOLC2, the MEFP-GEFS forecasts contribute five or more days of additional forecast lead time in the medium-range alone. These improvements are greatest during the snowmelt period and originate from the increased accuracy of the MEFP-GEFS temperature forecasts. Significant improvements are also seen in MARFC, where the MEFP-GEFS streamflow forecasts contribute 2-4 days of additional forecast lead time. In general, however, these improvements are substantially lower when verifying the bias-corrected streamflow forecasts against observed flows (see below).

- In AB-, CN- and MA-RFCs, the raw streamflow forecasts are conditionally biased with increasing rates of simulated flow. These biases originate from a similar conditional bias in the MEFP precipitation forecasts and increase as the forecast skill declines. For example, in ABRFC, the conditional bias increases rapidly during the first week, as the precipitation forecasts show little skill beyond one week. In CNRFC, the conditional biases increase throughout the medium-range, as the forecasts remain skillful during the middle portion of the forecast horizon. In CBRFC, the streamflow forecasts are conditionally *unbiased* for most streamflow rates. This stems from the importance of snowmelt in generating large streamflows in CBRFC. Specifically, there is a weaker dependence of high streamflows on heavy precipitation and the conditional biases therein. In some basins, notably AB-BLKO2, the MEFP-GEFS forecasts partially compensate for the tendency to underestimate the highest flows with an increased spread and, thus, an increased chance of warning about the highest flows.

- The overall skill of the post-processed streamflow forecasts, as well as the relative contributions from the MEFP and the EnsPost, vary with basin, season, and forecast lead time. They also vary with the source of forcing used in the MEFP. In general, the post-processed MEFP-GEFS forecasts are substantially more skillful than the raw MEFP-CLIM forecasts. The overall skill is greatest in CBRFC and CNRFC, where the seasonal differences are also greatest. For example, during the wet season at CN-DOSC1 (Dos Rios, CA), the post-processed streamflow forecasts with MEFP-GEFS forcing are up to ~40% more skillful than the raw streamflow forecasts with MEFP-CLIM forcing. In the dry season, the overall skill increases to ~60% at the earliest forecast lead times. At low flows, a greater fraction of the total skill originates from streamflow post-processing, as the EnsPost benefits from hydrologic persistence. Also, in basins with a pronounced dry season, the meteorological forcing is more predictable during the summer months. For these reasons, the MEFP-GEFS streamflow forecasts do not substantially improve upon the MEFP-GFS forecasts at low flows.

- At moderate and higher streamflow thresholds, a greater fraction of the total skill in the post-processed streamflow forecasts originates from the MEFP. Thus, at higher flows, the MEFP-GEFS forecasts generally improve upon the MEFP-GFS forecasts. During the wet season in CB- and CN-RFCs, and throughout the year in MA- and AB-RFCs, the MEFP-GEFS forecasts typically show similar skill to the MEFP-GFS forecasts for 1-2 days longer. For example, at CN-FTSC1, the MEFP-GEFS forecasts detect streamflows above the 10% exceedence threshold with equivalent skill to the MEFP-GFS forecasts, but with an additional forecast lead time of ~2.5 days. However, when verifying the post-processed streamflow forecasts, the gains implied by the raw forecasts (against simulated flows) are not always realized by the EnsPost, particularly at high streamflow thresholds. Indeed, at early forecast lead times in AB- and MA-RFCs, and later forecast lead times in CB- and MA-RFCs, forecasts of moderate and high flows show a decline in CRPSS following streamflow post-processing. This may originate from a lack of stationarity in the hydrologic biases, among other things. For example, at CB-DOLC2, the hydrologic biases vary substantially between years, particularly during the snowmelt period. In practice, the hydrologic biases are often manifest as timing errors in the simulated flows, yet the EnsPost can only model these indirectly, as magnitude errors. In order to account for inter- and intra- annual variations in basin conditions, operational forecasters typically modify some combination of the inputs, parameters, and states of the hydrologic models at runtime. However, adjusted simulations are not consistently archived by the RFCs. This may lead to inconsistencies in the calibration and operational use of the EnsPost.

Recommendation 3: Data assimilation (DA) is the preferred approach to adjusting hydrologic model states. In principle, automated data assimilation would avoid inconsistencies between the calibration and operational use of the EnsPost caused by runtime modifications. Within an automated DA framework, adjustments to the hydrologic model states, and hence to the simulated flows, would be reproducible, both operationally and retrospectively. In addition, DA would increase the quality of the simulated flows, which are used by the EnsPost to quantify the hydrologic uncertainties and to eliminate any residual hydrologic biases. In this context, there is a need to better understand the limitations of the EnsPost, and of statistical post-processing more generally, for bias correcting forecasts of large and extreme events. In keeping with the meteorological forecasts, there is a risk of obfuscating a weak, but potentially valuable, signal in the raw streamflow reforecasts through statistical post-processing. Indeed, when the hydrologic uncertainties account for a large fraction of the total uncertainties in the streamflow forecasts, the benefits of improved meteorological forcing may be outweighed by residual hydrologic biases in the post-processed streamflow forecasts.

- In order to evaluate the quality of the HEFS and to establish a baseline for future enhancements, more comprehensive hindcasting and verification is needed. This should be conducted by all RFCs, in collaboration with the Office of Hydrologic Development, for a range of forcing inputs, and for a broader range of river basins, including regulated rivers and outlets. Further work is needed to compare the streamflow forecasts from the HEFS against the RFC operational forecasts. In addition, there is a need to evaluate decision support systems and other applications that rely on the HEFS, such as water quality, river navigation, and water supply. Such applications are necessary to demonstrate the wider, societal and economic, benefits of the HEFS and of ensemble forecasting more generally. In this context, there is a need for interdisciplinary and interagency collaborations on uncertainty and risk, as hydrologic forecasts are only one input to environmental decision making, and not necessarily the most important one.

3. Introduction

Uncertainties in the inputs to hydrologic models, combine with uncertainties in the model parameters, structures and initial conditions. These uncertainties propagate through the modeling system and lead to uncertainties about model predictions (Brown and Heuvelink, 2005; Schaake et al., 2006; Mattot et al., 2009; Cloke et al., 2013). Practical applications of hydrologic predictions frequently involve multiple variables, multiple space-time scales, and multiple interconnected systems, such as the atmosphere-surface, river-estuary, and surface-subsurface (Beven, 2008). They also involve interactions between physical, chemical, and biological processes, such as water quality and ecology. Thus, uncertainties from hydrologic modeling combine with other sources of uncertainty about environmental systems (Jakeman et al., 2006; Brown, 2010). As water resources generate social costs and benefits, uncertainties about hydrologic systems also involve risk. Indeed, hydrologic forecasts are frequently used in risk assessments, whether informally or as inputs to integrated decision support systems. Here, they combine with uncertainties about social and economic variables, decision frameworks and politics (Kahnemann et al., 1982; Hamlet et al., 2002; Filar and Haurie, 2010; Ramos et al., 2012; Demeritt et al., 2013). Thus, hydrologic forecasts are only one input to environmental decision making, and not necessarily the most important one. On the one hand, it is important to assess and communicate the uncertainties in hydrologic forecasts, as deterministic forecasts can lead to inadequate decisions or to persistent conflict and indecision (Beven, 2000; Handmer et al., 2001). On the other hand, simplicity and parsimony are also desirable, as the estimates of uncertainty must be useful for environmental decision making. Ultimately, the aim is to achieve an appropriate balance of scope and detail in accounting for these uncertainties. As water resources are central to many social, economic, and environmental issues, this requires interdisciplinary and interagency collaborations on uncertainty and risk (Brown, 2010; Demeritt et al., 2013).

Broadly, there are two approaches to quantifying the total uncertainty in hydrologic forecasts, namely disaggregated modeling of the individual sources of uncertainty ("bottom up") and aggregated modeling of the total uncertainty ("top down"). In terms of the former, the different sources of uncertainty are quantified with probability distributions and then propagated numerically, using ensemble techniques (Gneiting and Raftery, 2005; Helton et al., 2006; Cloke et al., 2013). In terms of the latter, the total uncertainty is modeled empirically, by estimating the probability distribution of the observed variable conditionally upon a “raw”, single-valued, forecast (Glahn and Lowery, 1972). A hybrid of these approaches involves statistical post-processing of ensemble forecasts (Gneiting et al., 2007; Montanari and Grossi, 2008; van Andel et al., 2013). Whether using source-based modeling, statistical modeling or some combination of the two, the quality of the meteorological forcing is important. However, top-down approaches, such as “model output statistics” (Regonda et al., 2013) use single-valued meteorological forecasts. In contrast, hydrologic Ensemble Prediction Systems (EPS) use forcing information from one or more meteorological EPS (Cloke et al., 2013).

In recent years, improvements in computing power and ensemble techniques, as well the underlying atmospheric models and data assimilation schemes, have increased the applicability of meteorological EPS to regional, as well as global scales (Buizza et al., 2005; Park and Xu, 2009; Warner, 2010; Hamill et al., 2013). This is an important motivation for developing hydrologic EPS. Indeed, the meteorological uncertainties may account for a large fraction of the total uncertainties in hydrologic forecasting. For short- to medium-range forecasting at a global scale, the development of meteorological EPS has been led by national and international forecasting agencies. Examples include the Global Atmospheric Model of the European Center for Medium Range Weather Forecasts (ECMWF; Hagedorn et al., 2012) and the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP; Hamill et al., 2013). At regional scales, universities and forecasting agencies have collaborated to develop limited area EPS, often using community tools, such as the Weather Research and Forecasting (WRF) model. Examples of limited area EPS include NCEP’s Short Range Ensemble Forecast System (SREF; Du et al., 2009; Brown et al., 2012), the University of Washington Mesoscale Ensemble (Grimit and Mass, 2002), and the Limited-area Ensemble Prediction System (LEPS) of the Consortium for Small-scale Modeling (COSMO; Marsigli et al., 2005). Elsewhere, limited area EPS have been nested into global EPS, in order to provide seamless forecasts across multiple spatial scales. Examples of nested models include the COSMO-LEPS, which uses selected members of the ECMWF GAM (Marsigli et al., 2005), and the UK Meteorological Office Global and Regional Ensemble Prediction System (MOGREPS; Schellekens et al., 2011). Alongside spatial nesting, meteorological EPS increasingly use multi-model combinations or "super-ensembles". Combinations range from the unweighted assembly of individual EPS to weighted aggregations based on statistical post-processing. By leveraging the unique or orthogonal information across several EPS, multi-model forecasts generally improve on the best performing single-model forecasts (e.g. Marsigli et al., 2013), although weaker models can reduce the skill of unweighted combinations (e.g. Hagedorn et al., 2012). Examples of multi-model EPS include the THORPEX Interactive Grand Global Ensemble (TIGGE; Bougeault et al., 2010) and NCEP's SREF (Du et al., 2009). In terms of the sources of uncertainty considered, multi-model forecasts may include multiple physics and multiple initial and boundary conditions, among others.

Increasingly, hydrologic EPS use a combination of meteorological EPS to account for the forcing uncertainties and biases and statistical post-processing to model the hydrologic uncertainties and biases (Pappenberger et al., 2005; Seo et al., 2006; Wood and Schaake, 2008; Montanari and Grossi, 2008; van Andel et al., 2013; Brown and Seo, 2013). However, there are important differences between hydrologic EPS (for a list of hydrologic EPS, see http://hepex.irstea.fr/operational-heps-systems-around-the-globe/, accessed 12/04/2013). Some hydrologic EPS use raw forcing from meteorological EPS without accounting for the hydrologic uncertainties and biases. This allows for the rapid integration of meteorological EPS into operational hydrologic forecasting. For example, the Meteorological Model-based Ensemble Forecast System (MMEFS) uses raw meteorological forecasts from the SREF, the GEFS and the North American Ensemble Forecast System (NAEFS). The MMEFS was developed by operational forecasters at the U.S. National Weather Service (NWS) to accelerate the implementation of meteorological EPS into hydrologic EPS (Philpott et al., 2012). While the meteorological forecasts are downscaled (interpolated) to basin-averaged quantities, the MMEFS does not correct for biases in the raw forcing or account for hydrologic uncertainties and biases. In practice, however, meteorological EPS generally contain biases that are important for hydrologic forecasting (Pappenberger and Buizza, 2009). Also, a significant fraction of the total uncertainty in streamflow forecasting may originate from the hydrologic uncertainties and biases (Philpott et al., 2012). In other hydrologic EPS, the residual biases in the meteorological forecasts are lumped together with the hydrologic biases are removed through streamflow post-processing. This involves calibrating a statistical post-processor on hydrologic forecasts. For example, the European Floods Awareness System (EFAS) uses ensemble forecasts of temperature and precipitation from the ECMWF GAM. The meteorological forecasts are input to the LISFLOOD-FP hydrologic model from which ensemble forecasts of streamflow are output (Thielen et al., 2009). The raw streamflow forecasts are then bias-corrected with a vector autoregressive model whose predictors comprise transforms of the raw forecasts and observations in wavelet space (Bogner and Pappenberger, 2011).

The Hydrologic Ensemble Forecast Service (HEFS) was developed by the Office of Hydrologic Development (OHD) of the NWS. The HEFS also uses a hybrid approach to quantifying and combining the meteorological and hydrologic uncertainties (Demargne et al., 2014). However, the meteorological uncertainties and biases are modeled separately from the hydrologic uncertainties and biases. In both cases, statistical post-processing is used to correct for biases in the raw forecasts. The meteorological uncertainties and biases are quantified with the Meteorological Ensemble Forecast Processor (MEFP). The MEFP produces ensemble forecasts of precipitation and temperature conditionally upon a raw, single-valued, forecast (Wu et al., 2011). The space-time covariability of precipitation and temperature is modeled with the Schaake Shuffle (Clark et al., 2004). For short- to medium-range forecasting, the raw forecasts used by the MEFP include the operational, single-valued, temperature and precipitation forecasts from the NWS River Forecast Centers (RFCs) and the ensemble mean of NCEP's GEFS. In removing the meteorological biases with the MEFP, the hydrologic uncertainties and biases can be modeled independently of the meteorological forcing (Seo et al., 2006; Demargne et al., 2014). For the same reason, they can be modeled independently of forecast lead time. The hydrologic uncertainties and biases are modeled in two stages. First, the meteorological forecasts from the MEFP are used to generate raw streamflow forecasts, which may contain hydrologic biases, but do not explicitly account for any hydrologic uncertainties. Secondly, the raw streamflow forecasts are post-processed with the Ensemble Postprocessor (EnsPost). The EnsPost models the hydrologic uncertainties and biases from the residuals between the observed and simulated streamflows (Seo et al., 2006). In the EnsPost, the observed streamflows are propagated forwards in time using an autoregressive AR(1,1) model with the simulated flow as an exogeneous predictor (Seo et al., 2006).

The HEFS is being implemented in several phases, with the initial version (HEFSv1) scheduled for operational use at all RFCs by the end of 2014. In order to establish a baseline for future enhancements, and to guide the operational use of the HEFSv1, several phases of hindcasting and verification are also underway. This involves retrospective forecasting of temperature, precipitation, and streamflow at selected RFCs and for selected sources of meteorological forcing. In an earlier phase of hindcasting, temperature, precipitation and streamflow forecasts were generated with the HEFSv1 using forcing inputs from the “frozen” version of NCEP’s GFS (Brown, 2013). The frozen GFS employs a horizontal resolution of T62 or ~250km. In this report, the MEFP is calibrated with NCEP’s operational GEFS. The GEFS uses Version 9.0.1 of the GFS, which comprises a horizontal resolution of T254 (~55km) for 1-8 days and T190 (~70km) for 9-16 days, and a vertical resolution of L42 or 42 levels (Wei et al. 2008; Hamill et al. 2011; Hamill et al. 2013). The hindcasts are produced for selected river basins in four NWS RFCs, namely the Arkansas-Red Basin RFC (ABRFC), the Colorado Basin RFC (CBRFC), the California-Nevada RFC (CNRFC) and the Middle Atlantic RFC (MARFC). The hindcasts are verified conditionally upon forecast lead time, season, and magnitude of the observed and forecast variables. Limited combinations of these attributes are also considered. In order to establish the benefits of the GEFS separately from any hydrologic biases, the raw streamflow forecasts are verified against simulated streamflows. In addition, the post-processed streamflow forecasts are verified against observed flows.

The report is separated into three parts. It begins with the Material and Methods section, comprising an overview of the study basins and datasets, the HEFS methodology, and the verification strategy (Section 4). The results are then presented separately for the meteorological forecasts (Section 5.1), the raw streamflow forecasts (Section 5.2) and the bias-corrected streamflow forecasts (Section 5.3). Finally, the Discussion and Conclusions (Section 6) lead to guidance on the expected performance and limitations of the HEFSv1 for medium-range forecasting with the GEFS, together with recommendations on future enhancements.

4. Materials and methods

4.1 Study area

Four pairs of basins were used to evaluate the HEFSv1, each comprising one headwater and one immediately downstream basin. Figure 1 and Table 1 show the location of each basin, its average elevation, area, and the location of the nearest grid node in the GFS and the GEFS. Table 1 also shows the annual precipitation, the fraction of precipitation that generates runoff (the runoff coefficient), and the ratio of precipitation to potential evaporation (the climate index). The drainage areas range from 275 square kilometers (DRRC2) to 5457 square kilometers (FTSC1) and the runoff coefficients vary from 0.12 (CBNK1) to 0.58 (CNNN6). The basins were chosen for a combination of practical and hydrological reasons. First, they all originate from RFCs for which testing of the HEFSv1 is currently underway, namely AB-, CB-, CN-, and MA-RFCs. Indeed, the same basins were used to verify the precipitation, temperature and streamflow forecasts with forcing inputs from the frozen GFS (Brown, 2013). Second, as the uncertainties and biases propagate from upstream to downstream locations, it is important to understand the quality of the HEFSv1 in headwater basins. Third, headwater basins are important for operational forecasting of water quantity and quality, including flood warning and reservoir operations. Further downstream, the HEFS will be impacted by additional sources of bias and uncertainty, of which some are inherently difficult to quantify (e.g. the downstream effects of river regulations, simplified hydraulic routing and composite timing errors; see Raff et al., 2013). As part of the phased evaluation of the HEFS, more complex regimes, as well as additional sources of forcing, will be considered in future.

Figure 2 shows the daily means of temperature and precipitation across each pair of basins, together with the daily mean runoff for the headwater and downstream basins separately. The averages are shown by calendar month and were derived from gauged temperature, precipitation, and streamflow over a 20 year period between 1979 and 1999 (see Section 4.3). Nominally, two seasons are identified for each RFC, namely a “wet season" and a "dry season." These seasons are used in the calibration of the EnsPost (Appendix A) and in the verification of the forcing and streamflow forecasts (Section 5).

As indicated in Figure 2, there are marked differences in the seasonality and covariability of precipitation and runoff among basins and among RFCs. The strongest seasonality occurs in CNRFC, where precipitation quickly translates into runoff. In CBRFC and, to a lesser extent, in MARFC, snow accumulates during the cool season and leads to runoff during the late spring and early summer. In ABRFC, the relationship between precipitation and runoff is modulated by the shallow terrain and the high vegetation cover in these basin, as well as increased evapotranspiration during the summer months.

The basins in ABRFC comprise the Chikaskia River at Corbin, Kansas (CBNK1), and the Chikaskia River near Blackwell, Oklahoma (BLKO2). These basins experience a warm, and humid, summer climate. During the late spring and early summer, cool air from Canada and the Rocky Mountains combines with moist air from the Gulf of Mexico and hot air from the Sonoran Desert, leading to intense thunderstorms and tornados in Kansas and Oklahoma.

The basins in CBRFC are located on the Dolores River in Colorado, with the headwater near Rico (DRRC2) and the downstream basin in Dolores (DOLC2). The Dolores River is a tributary of the Colorado River and occupies a narrow valley incised into the sandstone of the San Juan Mountains. Precipitation is reasonably constant throughout the year, but falls primarily as snow during the winter months and in the higher elevations of DRRC2. The snowpack melts in the late spring and early summer, which leads to a sharp increase in runoff between April and July (Figure 2). Of the pairs of basins considered, DRRC2 and DOLC2 show the greatest differences in streamflow climatology between the headwater and downstream basins. For the purposes of hydrologic modeling, DRRC2 is separated into two sub-basins, while DOLC2 is separated into three sub-basins, in order to accommodate the varied elevations there. The lower sub-basin of DRRC2 accounts for 77% of the total area of DRRC2 while, in DOLC2, the lower middle and upper sub-basins account for 17%, 61% and 22% of the total area, respectively.

The basins in CNRFC comprise the Middle Fork of the Eel River at Dos Rios (DOSC1) and the Eel River at Fort Seward (FTSC1). These basins are located on the windward slopes of the North Coast Ranges in northern California (Figure 1). During the late summer and early autumn, the upper reaches of the Eel River experience little or no precipitation and streamflow. Low flows are accentuated by diversions to the Russian River for use in the Potter Valley Hydro-Electric Project. In late autumn, cooler temperatures are accompanied by rapidly increasing precipitation, to which the streamflows respond through November and continue increasing until January (Figure 2). During the winter months, the predictability of heavy precipitation is increased by the onshore movement of weather fronts from the Pacific coast and their orographic lifting in the North Coast Ranges. The coastal mountains of northern California and the Pacific Northwest are also susceptible to "atmospheric rivers", which carry moisture in narrow bands from the tropical oceans to the mid-latitudes. Atmospheric rivers can lead to persistent, heavy, precipitation and extreme flooding in the North Coast Ranges and further inland (Smith et al., 2010). For the purposes of hydrologic modeling, both DOSC1 and FTSC1 are separated into two sub-basins, with the lower sub-basins accounting for, respectively, 77% and 97% of the total area of each basin.

The basins in MARFC comprise the West Branch of the Delaware River at Walton in Pennsylvania (WALN6) and the inflow to Cannonsville Reservoir in New York State (CNNN6). The West Branch of the Delaware rises near Mount Jefferson in Schoharie County, NY, and flows through Delaware County until it reaches the Cannonsville Reservoir, approximately 15 miles downstream of Walton. In terms of daily precipitation amounts, the climatology is relatively constant throughout the year (Figure 2). However, during the winter months and in the higher elevations, the majority of precipitation falls as snow. With increasing rainfall and rising temperatures, the snowpack melts during the late spring, with streamflow peaking between March and April before declining rapidly in the summer months. Owing to the proximity of WALN6 and CNNN6, their runoff patterns are very similar throughout the year. The Cannonsville Reservoir is operated by the New York City Department of Environmental Protection (NYCDEP). It is one of three reservoirs in the Delaware River Basin and nineteen reservoirs overall that supply NYC with drinking water. In order to improve the management of water supply from these reservoirs, the NYCDEP are currently evaluating streamflow forecasts from the HEFSv1.

4.2 The Hydrologic Ensemble Forecast Service (HEFS) methodology

Further details on the HEFS methodology can be found in Appendix A. The HEFS models the total uncertainty in streamflow at some future times, $\mathbf{q}_f$, conditionally upon the observed streamflow up to, and including, the current time, $\mathbf{q}_c$. The total predictive uncertainty is factored into two main sources of uncertainty, namely the "hydrologic uncertainties" and the "meteorological uncertainties". The meteorological uncertainties are included in the raw streamflow forecast and the hydrologic uncertainties are modeled in an adjusted streamflow forecast. Omitting notation of the random variables for simplicity,

$$\underbrace{f_1(\mathbf{q}_f\,|\,\mathbf{q}_c)}_{\text{Total}} = \int \underbrace{f_2(\mathbf{q}_f\,|\,\mathbf{q}_c, \mathbf{q}_r)}_{\text{Adjusted}}\ \underbrace{f_3(\mathbf{q}_r\,|\,\mathbf{q}_c)}_{\text{Raw}}\ d\mathbf{q}_r, \tag{1}$$

where $\mathbf{q}_r$ denotes the raw streamflow forecast. The raw streamflow forecast is estimated with the Hydrologic Ensemble Processor (HEP). The HEP integrates a finite number of "equally likely" forecasts of precipitation and temperature through the hydrologic models. These forecasts include the meteorological uncertainties, which are modeled explicitly

$$\underbrace{f_3(\mathbf{q}_r\,|\,\mathbf{q}_c)}_{\text{Raw}} = \int \underbrace{f_4(\mathbf{q}_r\,|\,\mathbf{q}_c, \mathbf{m}_f)}_{\text{Raw}\,|\,\text{Forcing}}\ \underbrace{f_5(\mathbf{m}_f)}_{\text{Forcing}}\ d\mathbf{m}_f, \tag{2}$$

where $\mathbf{m}_f$ denotes the future forcing. The meteorological uncertainties are quantified with the Meteorological Ensemble Forecast Processor (MEFP). The MEFP models the future forcing conditionally upon a raw forecast, $\mathbf{r}_f$; that is, by estimating the conditional distribution, $f_6(\mathbf{m}_f\,|\,\mathbf{r}_f)$

$$f_5(\mathbf{m}_f) = f_6(\mathbf{m}_f\,|\,\mathbf{r}_f). \tag{3}$$

In this study, the raw forcing comprises the ensemble mean of NCEP's GEFS and the ensemble mean of the frozen GFS. However, other sources of forcing are supported by the MEFP, including the RFC single-valued quantitative precipitation forecasts (Schaake et al., 2007; Wu et al., 2011). The HEFS does not currently isolate the contributions from other sources of uncertainty, such as the initial conditions or parameters of the hydrologic models (Appendix A). Rather, the overall effects of these hydrologic uncertainties are modeled in the adjusted streamflow forecast using the Ensemble Post-processor (EnsPost; Seo et al., 2006). In all cases, the parameters of future quantities are estimated from subsets of the historical data, for which a degree of stationarity is assumed.

4.3 Datasets

Hindcasts of mean areal temperature (MAT) and mean areal precipitation (MAP) were generated with the MEFP for a ~15 year period between 1985 and 1999. The hindcasts of MAP and MAT were produced at 12Z each day. Each forecast comprised ~50 ensemble members, with lead times varying from 6 to 336 hours in six-hour increments. Inputs to the MEFP comprised raw precipitation and temperature forecasts from the latest version of NCEP's operational GEFS (Hamill et al., 2013). In order to evaluate the skill of the MEFP-GEFS forecasts, forcing inputs were also generated with the frozen (circa 1997) version of NCEP's GFS (Hamill et al., 2006) and with a conditional or "resampled" climatology. The latter involved resampling the historical observations of MAP and MAT in a moving window of, respectively, 61 days and 31 days around the forecast valid date. The MEFP hindcasts with forcing inputs from the GEFS, GFS and resampled climatology are denoted MEFP-GEFS, MEFP-GFS and MEFP-CLIM, respectively. In keeping with the recommended operational practice, the parameters of the MEFP were estimated from all available data (Table 2). The parameters of the bivariate relationship between the forecasts and observations are estimated from the historical pairs of forecasts and observations. The hindcasts from the operational GEFS comprise a ~25 year period from 1985-2010, while the hindcasts from the frozen GFS comprise a ~27 year period from 1979-2006 (Table 3). As indicated above, the MEFP also models the space-time covariability of precipitation and temperature. The Schaake Shuffle is trained with the historical observations of MAP and MAT alone. Alongside the covariability of temperature and precipitation, the Schaake Shuffle determines the number of ensemble members in the MEFP forecasts (one per historical year).

Raw streamflow hindcasts were generated with the Community Hydrologic Prediction System (CHPS) using the precipitation and temperature forecasts from the MEFP. The hindcasts were produced with the hydrologic models and parameter settings used operationally at each RFC. In AB-, CB- and CN-RFCs, the Snow Accumulation and Ablation Model (SNOW-17; Anderson, 1973) is used together with the Sacramento Soil Moisture Accounting Model (SAC-SMA; Burnash, 1995). In MARFC, the SAC-SMA is substituted with an empirical model, based on the Antecedent Precipitation Index (API), but adapted for continuous simulations (the so-called "Continuous API" model; Sittner et al., 1969). The models are integrated with a six-hourly timestep in AB- and MA-RFCs and an hourly timestep in CB- and CN-RFCs. Routing from the headwater to the downstream basin is conducted with Lag/K using constant or variable lag and attenuation (e.g. WALN6 to CNNN6 uses a constant lag with no attenuation). In most RFCs, an ADJUST-Q operation is used to blend the recently observed streamflow into the operational forecast. However, ADJUST-Q was omitted from the streamflow hindcasting, as the EnsPost employs a weighted combination of the recently observed and forecast streamflows (Appendix A). In order to calibrate the EnsPost, and to establish the relative importance of the meteorological and hydrologic uncertainties, simulated streamflows were generated for each basin and used to verify the streamflow forecasts (see below). The parameters of the EnsPost were estimated from the historical pairs of observed and simulated streamflows in each basin (Table 2).

Observations of precipitation and temperature were obtained from each RFC and comprised areal averages (MAP, MAT) of the gauged precipitation and temperature in each basin. The data comprise six-hourly observations at {0Z,6Z,12Z,18Z} between ~1950-1999. Streamflow observations were obtained from the United States Geological Survey (USGS) and comprise daily mean streamflows at the outlet of each basin. The averages were determined from observations of river stage, beginning at midnight in local time, and converted to streamflow using a measured stage-discharge relation (Kennedy, 1983). Subsequently, they were converted to runoff values (mm/day) for ease of comparison between basins. While stage observations were available at the outflow of the Cannonsville Reservoir (CNNN6), the NYCDEP use inflow forecasts to manage the reservoir levels. Thus, the HEFS was calibrated and verified at the inflow to Cannonsville Reservoir. The inflows were estimated by NYCDEP using gauged reservoir levels and outflows. The outflows comprise all diversions, spills and releases, but evaporation is not considered. During the dry season, this can lead to approximation errors for low streamflows, which are assigned zero if the inflow estimates are negative. There are short periods of missing data in several RFCs. In particular, the streamflow observations are missing between 1st October 1996 and 1st October 1998 in DRRC2 and between 1st January 1999 and 31st December 1999 in CNNN6.

As indicated above, the HEFS forecasts are issued at 12Z each day, while precipitation, temperature and streamflow are all observed in local time. In order to pair the meteorological observations and forecasts, the observed values were chosen from the nearest available time in {0Z, 6Z, 12Z, 18Z}. This introduced a timing error into the observations of +1 hours, 0 hours, -1 hours and -2 hours for MARFC, ABRFC, CBRFC and CNRFC, respectively. As the forecasts were verified at an aggregated support of one day or larger (see below), this timing error was deemed acceptable. However, pairing of the observed and forecast streamflows was complicated by the daily scale of the streamflow observations. Ultimately, any fractional downscaling of the observed streamflows to match the forecast day of 12Z-12Z would require a model of the temporal dependencies at the downscaled support. This could introduce significant biases, as the forecasts begin ~12 hours after the observations. Instead, the first ~12 hours of forecasts were ignored. This eliminated all timing errors associated with pairing in CBRFC and CNRFC, where the forecasts are issued hourly, and in ABRFC, where the six-hourly forecasts are offset from UTC by 6 hours. In MARFC, it introduced a one-hour timing error, as the observed day (5Z-5Z) is offset from the nearest available six-hourly forecast (6Z) by one hour. Pairing of the streamflow forecasts and simulations was straightforward, and daily averages were formed from 12Z to12Z each day.

4.4 Verification strategy

Verification was conducted with the NWS Ensemble Verification System (EVS; Brown et al., 2010). The forecasts were verified conditionally upon season, forecast lead time, and magnitude of the observed and forecast variables. Limited combinations of these attributes were also considered, but were often constrained by the sampling uncertainties of the verification metrics. In evaluating the quality of the HEFS forecasts, unconditional bias and skill are important, as the HEFS is an operational forecasting system for which many applications, with varying sensitivities to streamflow amount, are anticipated. However, "average conditions", particularly the ensemble mean, generally favor dryer weather and lower flows, as precipitation and streamflow are both skewed variables. Thus, conditional verification is also important. The MEFP forecasts were verified against observed temperature and precipitation. The streamflow forecasts were verified in two stages. First, in order to understand the benefits of the MEFP-GEFS forcing separately from any hydrologic biases, the raw streamflow forecasts were verified against simulated streamflows. Any differences between the hydrologic forecasts and simulations reflect the contribution of meteorological uncertainty to the streamflow forecasts, independently of any hydrologic uncertainties (but notwithstanding errors in the meteorological observations). Second, the post-processed streamflow forecasts were verified against observed streamflows. This allowed the benefits of the MEFP-GEFS forcing to be established in an operational context, where the EnsPost cannot be expected to remove all hydrologic biases.

As indicated in Table 2, the periods used to calibrate and validate the HEFS are neither completely dependent nor independent. While statistical models generally perform better under dependent than independent validation, the HEFS was designed with a minimum number of parameters to estimate. Not surprisingly, therefore, experiments with the MEFP (e.g. Wu et al., 2011) and with the EnsPost (e.g. Seo et al., 2006) have shown negligible differences between dependent and cross-validation when using a calibration period of 20+ years. In calibrating the HEFS, the recommended operational practice was followed; that is, to use all available data, subject to a “reasonable” degree of stationarity in the forcing and streamflow observations (Seo et al., 2006; Wu et al., 2011).

When verifying forecasts of continuous random variables, such as precipitation and streamflow, verification is often performed both unconditionally and conditionally upon particular events (Wilks, 2006; Demargne et al., 2010; Jolliffe and Stephenson, 2011). In order to compare the verification results between basins and seasons, for different forecast lead times and valid times, and for specific aggregation periods, common events were identified for each basin. Specifically, for each verifying dataset, v, aggregation period, a, and basin, b, a climatological distribution function, $\hat{F}_{n,v,a,b}(x)$, was computed from the n values of the hydrometeorological variable, x, between 1985 and 1999. Real-valued thresholds were then determined for $k \approx 100$ climatological exceedence probabilities, c_p, $\hat{F}^{-1}_{n,v,a,b}(c_p)$, where $c_p \in [0,1]$ and $p = 1,\ldots,k$. Verification measures that depend continuously on threshold value, such as the mean error, were derived from the conditional sample in which the observed value exceeded the threshold. For consistency, exceedence thresholds are used throughout; for continuous measures, this implies greater emphasis on high streamflows. Measures defined for discrete events, such as the Brier Score, were computed from the observed and forecast probabilities of exceeding the threshold. When verifying the raw streamflow forecasts, $\hat{F}_{n,v,a,b}(x)$ was derived separately for the streamflow observations and simulations. While the sampling uncertainties were not quantified here (see Brown and Seo, 2013 for an example), the verification results are only considered for sample sizes of 30 or more. For continuous measures, the sample size is determined by the number of verification pairs. For discrete measures, it comprises the smaller of the number of occurrences and non-occurrences.

Key attributes of forecast quality are obtained by examining the joint probability distribution of the observed variable, Y, and the forecast variable, X, $f_{XY}(x, y)$. The joint distribution can be factored into $f_{XY}(x, y) = f_{Y|X}(y \mid x)\ f_X(x)$, which is known as the “calibration-refinement” (CR) factorization and $f_{XY}(x, y) = f_{X|Y}(x \mid y)\ f_Y(y)$, which is known as the “likelihood-base rate” (LBR) factorization (Murphy and Winkler, 1987). The conditional distribution, $f_{Y|X}(y|x)$, reflects the Type-I conditional bias or **reliability** of the forecast probabilities when compared to $f_X(x)$ and **resolution** when only its sensitivity to X is considered. For a given level of reliability, sharp forecasts (i.e. forecasts with smaller spread or a greater deviation from climatology) are sometimes preferred over unsharp ones, as they contribute less uncertainty to decision making (Gneiting et al., 2007). Put differently, as the **sharpness** increases, other attributes of forecast quality must also increase to maintain a given level of forecast skill. The conditional distribution, $f_{X|Y}(x|y)$, reflects the **Type-II conditional bias** of the forecasts when compared to $f_Y(y)$ and **discrimination** when only its sensitivity to Y is considered. If Y is assumed certain, i.e. the forecast probability distribution is given by the Dirac delta function, $f_Y(y)=\delta(y)$, the forecasts must be perfectly sharp (deterministic) and perfectly accurate to have no Type-II conditional bias. In practice, no single metric provides a complete description of forecast quality (Hersbach, 2000; Bradley et al., 2004). Appendix B describes the metrics used in this paper.

5. Results and analysis

5.1 Quality of the precipitation and temperature forecasts

The precipitation and temperature forecasts from the MEFP were verified against observed MAP and MAT, respectively. The results are presented by forecast lead time, season, and magnitude of the forcing variable.

5.1.1 Forecast lead time

Figure 3 shows the correlation between the ensemble mean forecast and observed precipitation for the upstream and downstream basin in each RFC. The results are shown for the raw GEFS forecasts and the MEFP outputs, which include MEFP-CLIM, MEFP-GFS and MEFP-GEFS. In general, the highest correlations occur in MARFC and CNRFC, where the MEFP forecasts benefit from the regulating effects of the Atlantic Ocean and the Pacific Ocean, respectively. However, the forecast skill declines more rapidly in MARFC than CNRFC. For example, the MEFP-GEFS forecasts show correlations of ~0.8

in both CNRFC and MARFC at a forecast lead time of one day, and then decline to ~0.3 in MARFC after six days, while remaining at ~0.6 in CNRFC. The forecasts in CBRFC show the lowest overall correlations and the largest differences between basins. This reflects the mountainous terrain surrounding the Dolores River, where the average elevation exceeds 2,500m in the headwater basin, CB-DRRC2.

The MEFP aims to increase the skill of the raw forecasts by reducing bias, while preserving the information content in the raw forecasts; that is, by minimizing sampling uncertainty and other statistical artifacts. The MEFP employs a linear regression in normal space, with the observed variable as the predictand and the raw forecast as the predictor (Wu et al., 2011), for which the correlation coefficient is an important measure. As indicated in Figure 3, the MEFP preserves or increases the correlations between the ensemble mean of the raw GEFS forecasts and the observed variable. In preserving these correlations, the MEFP benefits from the improvements in the GEFS when compared to the GFS. Indeed, the MEFP-GEFS precipitation forecasts show higher correlations than the MEFP-GFS forecasts in all basins and at most forecast lead times, with increases in correlation of 0.1-0.2 during the short- to medium-range (Figure 3). In AB-, CB- and MA-RFCs, the MEFP-GEFS forecasts show higher correlations than the MEFP-GFS forecasts at all forecast lead times. In CNRFC, the improvements from the MEFP-GEFS are greatest after ~3 days, as the GEFS and GFS forecasts are both highly correlated with the observations at earlier forecast lead times (~0.8). When expressed as a gain in forecast lead time over the period of skillful forcing, the MEFP-GEFS forecasts typically add 1-2 days in forecast lead time to the MEFP-GFS forecasts (see Section 5.1.3 also).

Figure 4 shows the relative mean error (RME) of the MEFP-CLIM, MEFP-GFS and MEFP-GEFS precipitation forecasts. In general, the forecasts underestimate the observed precipitation amount by ~5-15%, with the smallest biases in CNRFC and the largest biases in CBRFC. However, these biases are small in absolute terms, amounting to less than 0.5mm/day of accumulation in CBRFC. Also, there is a slight underforecasting bias in the MEFP-CLIM forecasts in AB-, CB- and MA-RFCs. In principle, the MEFP-CLIM forecasts should be unconditionally unbiased, as they are resampled from the same historical observations of MAP and MAT used to verify the MEFP forecasts. However, different periods of record were used to calibrate the MEFP and to conduct verification (Table 2). Thus, some of the differences between forcing sources, as well as the slight underforecasting bias in the MEFP-CLIM forecasts, may be related to climatological variability.

Figures 5 shows the mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-CLIM, MEFP-GFS and MEFP-GEFS precipitation forecasts against sample climatology. Sample climatology comprises the unconditional probability distribution of the observed MAP between 1985 and 1999. In keeping with the correlation results, the MEFP precipitation forecasts show the greatest skill in CNRFC, where the atmospheric predictability is greatest, followed by MARFC and ABRFC. In CBRFC, the precipitation forecasts show limited or negative skill and, while the MEFP-GEFS forecasts improve upon the MEFP-GFS forecasts, neither improve upon the MEFP-CLIM forecasts beyond ~5 days. Moreover, the MEFP-CLIM forecasts are somewhat unskillful when compared to unconditional climatology. As indicated above, this may originate from differences in the climatology of the calibration and verification periods, with a longer period used to calibrate the MEFP-CLIM in CB-DRRC2 and CB-DOLC2 (1979-2005) than conduct verification (1985-1999). Notwithstanding the limited skill of the MEFP precipitation forecasts in CBRFC, the MEFP-GEFS forecasts are consistently more skillful than the MEFP-GFS forecasts. As with the correlation results, for a given amount of skill, the MEFP-GEFS forecasts typically add 1-2 days in forecast lead time to the MEFP-GFS forecasts (Figure 5).

Figures 6 shows the mean CRPSS of the MEFP-CLIM, MEFP-GFS and MEFP-GEFS temperature forecasts against sample climatology. Unlike the MEFP precipitation forecasts, the MEFP temperature forecasts consistently improve upon sample climatology. This reflects the strong seasonality in temperature when compared to precipitation (Figure 2) and the resulting advantage of a conditional (resampled) climatology over an unconditional one. For the same reason, the differences between the MEFP-CLIM, MEFP-GFS and MEFP-GEFS forecasts are more important than the absolute skill of the temperature forecasts. As indicated in Figure 6, the MEFP-GFS and the MEFP-GEFS temperature forecasts both improve substantially upon resampled climatology and remain skillful for longer than the equivalent precipitation forecasts, owing to the relative predictability of temperature versus precipitation. However, in keeping with the precipitation forecasts, the benefits of the MEFP-GEFS are generally more pronounced after the first 1-2 days. In other words, as the errors begin to saturate in the MEFP-GFS forecasts, the MEFP-GEFS forecasts remain skillful for longer. Indeed, in the middle portion of the forecast horizon, the MEFP-GEFS adds 2-4 days of forecast lead time, for an equivalent CRPSS, when compared to the MEFP-GFS. While hydrologic models are generally more sensitive to precipitation forcing than temperature, accurate forecasts of surface temperature are important for predicting the timing of snowmelt in snow-dominated basins, such as CB-DOLC2 and CB-DRCC2.

5.1.2 Magnitude of the forcing variable

Figure 7 shows the correlation of the ensemble mean forecast and observed variable for increasing amounts of observed precipitation. The correlations are shown for the raw GEFS inputs, as well as the MEFP outputs. The results are plotted against climatological exceedence probability on a probit scale, but labeled with actual probability. For example, 0.01 represents a daily total precipitation amount that is exceeded, on average, only once in every 100 days (Table 1). As indicated in Section 4.4, verification scores that depend continuously on the data, such as the correlation coefficient and CRPSS, were derived from the subsample of verification pairs whose observed value exceeded the threshold. Thus, Figure 7 does not include zero precipitation amounts and the origin of each curve corresponds to the observed probability of precipitation (PoP). The results are shown for the upstream and downstream basins in each RFC and at selected forecast lead times. Figure 8 shows the CRPSS of the MEFP-GEFS and the MEFP-GFS forecasts for increasing amounts of observed precipitation, where skill is measured against sample climatology. The results are plotted for the downstream basin in each RFC and at selected forecast lead times. Alongside the mean CRPSS, the skill is decomposed into selected attributes of forecast quality, namely the relative reliability, $\text{CRPSS}_{\text{REL}}$, the relative resolution, $\text{CRPSS}_{\text{RES}}$, and the relative uncertainty $\text{CRPSS}_{\text{UNC}}$.

As shown in Appendix B, smaller values of the $\text{CRPSS}_{\text{REL}}$, and larger values of the $\text{CRPSS}_{\text{RES}}$, both contribute to increased skill.

As indicated in Figure 7, the correlations between the raw GEFS precipitation forecasts and observations are consistently preserved by the MEFP, with similar or higher correlations in the MEFP-GEFS forecasts at a range of precipitation thresholds and forecast lead times (notwithstanding some sampling noise). Furthermore, the MEFP-GEFS forecasts consistently preserve or improve upon the correlations between the MEFP-GFS forecasts and observations. The greatest improvements occur in MARFC, where the MEFP-GEFS forecasts show substantially higher correlations at all precipitation thresholds and at all forecast lead times. In ABRFC, the correlations are improved at early forecast lead times, particularly for smaller precipitation thresholds. However, the benefits of the GEFS inputs decline with increasing forecast lead time and increasing precipitation amount. More generally, the effects of declining predictability and increasing sampling uncertainty (i.e. lower background correlations, higher precipitation thresholds) are to obscure any differences between the MEFP-GEFS and the MEFP-GFS forecasts until they become indistinguishable from climatology. In CBRFC, the raw GEFS forecasts show similar correlations to the MEFP-GFS forecasts, but the MEFP-GEFS forecasts apparently show higher correlations. In principle, the correlations may be improved by the MEFP, as the raw forecasts are translated into multiple predictors or “canonical events”, which aim to capture the skill in the raw forecasts at multiple temporal scales. However, some caution is necessary, as the raw GEFS forecasts exhibit only weak correlations and similar improvements are not seen in other RFCs. Unlike most basins, where the improvements from the MEFP-GEFS forecasts typically decline with increasing forecast lead time, the improvements in CNRFC are greatest during the middle portion of the forecast horizon. For example, at CN-FTSC1, the MEFP-GEFS and the MEFP-GFS forecasts show similar correlations at all precipitation thresholds after one day, but the MEFP-GEFS forecasts show higher correlations after seven days (Figure 7).

During the first week, the MEFP-GEFS precipitation forecasts are consistently more skillful than the MEFP-GFS forecasts at AB-BLKO2, CB-DOLC2 and MA-CNNN6 (Figure 8). However, during the second week, the MEFP-GEFS forecasts are no more skillful than climatology and, in CB-DOLC2, the forecasts of light precipitation are somewhat less skillful than climatology. Overall, the greatest improvements from the GEFS occur in MA-CNNN6, where the MEFP-GEFS forecasts are 15-30% more skillful than climatology, while the MEFP-GFS forecasts are only 5-15% more skillful. At early forecast lead times, the MEFP-GEFS forecasts show similar skill to the MEFP-GFS forecasts in CN-FTSC1. However, they are substantially more skillful after seven days, particularly at higher precipitation thresholds. At lower precipitation thresholds, the residual skill of the MEFP-GEFS forecasts mainly originates from improved resolution. In contrast, at higher thresholds, it mainly originates from improved reliability. The reliability component of the mean Continuous Ranked Probability Score (CRPS) is closely related to the "flatness" of the verification rank histogram (Hersbach, 2000). In other words, small $\mathrm{CRPSS}_{\mathrm{REL}}$ implies that any two ranked ensemble members capture the observation with equal probability. The relative uncertainty is a property of the observed sample alone and subtracts from the overall skill (Appendix B). By definition, a climatological probability forecast is perfectly reliable (0.0) and has no resolution (0.0); thus, the CRPS of a climatological probability forecast is determined by the uncertainty component alone. In that case, the $\mathrm{CRPSS}_{\mathrm{UNC}}$ is greatest (1.0) when the conditional subsample of observations is equal to the full sample, i.e. to sample climatology. Since Figure 8 originates at the PoP, the $\mathrm{CRPSS}_{\mathrm{UNC}}$ is always less than 1.0 and it declines with increasing precipitation amount.

Figure 9 shows the reliability diagram for the MEFP-GEFS precipitation forecasts at the downstream basin in each RFC. The results are shown at a forecast lead time of 1-2 days (24-48 hours) and for selected precipitation events, which are expressed as climatological exceedence probabilities. In addition to the 1-2 day PoP, Figure 9 shows the average reliability across the four, six-hourly sub-periods from which the 1-2 day accumulation was derived. In order to produce reliable forecasts at aggregated scales, a forecasting system must: 1) produce reliable forecast at disaggregated scales; and 2) adequately capture the statistical dependencies between the disaggregated scales. In the context of PoP, this implies adequate modeling of any statistical dependencies in precipitation intermittency at the six-hourly scale. As indicated in Appendix B, the reliability diagram compares the average observed and forecast probabilities across several categories of forecast probability, together with the sample size or “sharpness” of the forecasts in each category. For a discrete event, such as flooding, an ensemble forecasting system is reliable if the event is observed with the same relative frequency as the forecast probability implies. Reliability is an important attribute of an operational forecasting system, as decisions are based on attributes of the forecasts, rather than the observed outcomes. Nevertheless, attributes of forecast quality that depend on observed outcomes (e.g. when flooding is observed to occur) are important for guiding model development and operational practice. Figure 10 shows the Relative Operating Characteristic (ROC) for both the MEFP-GEFS precipitation forecasts and the MEFP-GFS forecasts at the downstream basin in each RFC. For a discrete event, such as flooding, an ensemble forecasting system is discriminatory if it correctly forecasts the event with a probability higher than chance (i.e. higher than the climatological probability) and correctly forecasts its non-occurrence with a probability lower than chance. The results are shown for a forecast lead time of 1-2 days and for increasing precipitation amounts. The ROC curves were fitted under an assumption of bivariate normality between the Probability of Detection (PoD) and the Probability of False Detection (PoFD) and are shown together with the empirical pairs of PoD and PoFD for each exceedence threshold (Appendix B). As the ROC is insensitive to reliability, it is often interpreted alongside the reliability diagram.

As indicated in Figure 9, the MEFP-GEFS forecast are generally reliable for moderate and high precipitation amounts, particularly in AB-BLKO2, MA-CNNN6 and CN-FTSC1. In DOLC2, the forecasts of moderate precipitation (C_p=0.1 or >6.5 mm/day) are somewhat over-confident at high forecast probabilities, but the sample sizes are also relatively small (~10-100 occurrences). However, in CB-DOLC2, CN-FTSC1 and MA-CNNN6, the forecasts of PoP are consistently too low. For example, in MA-CNNN6, when precipitation was forecast with an average probability of 0.4, it was observed with a probability of 0.6. In this context, PoP is defined as the probability of exceeding the smallest measurable precipitation amount. The latter was nominally defined as 0.25mm (for both the forecasts and observations). However, the results were found to be insensitive to the precise definition of PoP, and similar biases were observed for “light” precipitation amounts. Underestimation of PoP and light precipitation was also observed for the MEFP-GFS forecasts (see Brown, 2013 also). Again, this suggests a problem in the modeling, estimation, or implementation of the MEFP for PoP and light precipitation, whether using the GEFS or the GFS. However, while the forecasts of PoP are generally unreliable at the daily scale, the corresponding six-hourly forecasts are reliable, on average. As indicated above, temporal aggregation relies on accurate modeling of the statistical dependencies between the disaggregated quantities. By implication, the lack of reliability in the accumulated PoP may originate from inadequate modeling of the temporal dependencies at the six-hourly scale; that is, from the handling of precipitation intermittency by the Schaake Shuffle.

For moderate and high precipitation amounts, both the MEFP-GEFS forecasts and the MEFP-GFS forecasts are discriminatory, as well as reliable (Figure 10). An ensemble forecasting system is discriminatory if it correctly forecasts an event with a probability higher than chance and correctly forecasts its non-occurrence with a probability lower than chance. At a forecast lead time of 1-2 days, the MEFP-GEFS precipitation forecasts can discriminate between the exceedence and non-exceedence of selected precipitation thresholds in all basins (Figure 10). While the MEFP forecasts are more discriminatory than sample climatology at all precipitation thresholds, they are generally most discriminatory at moderate and high thresholds. The greatest discrimination occurs in CN-FTSC1, where the atmospheric predictability is highest (e.g. Figure 3). The lowest discrimination occurs in CB-DOLC2, where the correlations and skill are also low. While the MEFP-GEFS forecasts are highly discriminatory in FTSC1, they are no more discriminatory than the MEFP-GFS forecasts. This reflects the small additional skill in the MEFP-GEFS forecasts during the first few days (e.g. Figure 8). In all other basins, the MEFP-GEFS forecasts are more discriminatory than the MEFP-GFS forecasts, at least for some precipitation thresholds. The greatest improvements occur in MA-CNNN6, where the discrimination is uniformly higher across all precipitation thresholds. By way of example, a decision maker may accept a PoFD of 5% in forecasting a one-day precipitation total greater than 14.75mm (C_p=0.05). In that case, the MEFP would forecast the event correctly on 55% of occasions when using the GFS and on 62% of occasions when using the GEFS. In AB-BLKO2, the additional discrimination is greatest at moderate and high precipitation thresholds (C_p=0.1 and C_p=0.05), whereas in DOLC2, the GEFS adds more discrimination for PoP and light precipitation amounts (C_p=0.25).

While the MEFP-GEFS precipitation forecasts are generally reliable or unbiased conditionally upon the forecast variable, they are not necessarily unbiased conditionally upon the observed variable. Indeed, in a separate study by Brown (2013), the MEFP-GFS forecasts systematically underestimated the highest precipitation totals, particularly at longer forecast lead times, where the forecasts were no more skillful than climatology (climatology is, by definition, conditionally biased). This originated from the inability of the frozen GFS to accurately detect the largest precipitation amounts. Figure 11 shows box plots of errors in the MEFP-GFS and the MEFP-GEFS precipitation forecasts for the downstream basin in each RFC. Each box represents one ensemble forecast from the period 2-3 days (72-96 hours). Selected quantiles of the forecast error are plotted together with the median error and range (extreme residuals) as whiskers. The boxes are arranged by increasing amount of observed precipitation. At the highest precipitation amounts, the MEFP-GEFS forecasts show similar conditional biases to the MEFP-GFS forecasts (Figure 11), and the forecast median generally underestimates the observed precipitation by 50-100%. However, the MEFP-GEFS forecasts typically contain greater spread, both in terms of the interquartile range and the overall range. In other words, the MEFP-GEFS forecasts are more likely to predict the highest observed precipitation amounts (or similarly large amounts) with some, non-zero, probability of occurrence, even if their central tendency is to underestimate. Currently, the MEFP is calibrated with the ensemble mean of the raw forecasts (Appendix A). For the purposes of statistical post-processing, most of the information content in the frozen GFS is concentrated in the ensemble mean forecast (Wilks and Hamill, 2007; Wu et al., 2011). However, as atmospheric models and EPS become more skillful, statistical post-processors may benefit from using higher moments, interactions, or even the individual ensemble members, providing the sampling uncertainties are reasonably small. Future work should consider whether the raw GEFS forecasts contain valuable information beyond the ensemble mean and how best to leverage this information in the MEFP.

5.1.3 Season

Seasonal verification was performed for the “wet” and “dry” seasons in each RFC. As indicated in Figure 2, the relationship between temperature, precipitation and streamflow varies between RFC. In MA- and CN-RFCs, the dry season coincides with the summer, whereas in AB- and CB-RFCs, it coincides with the winter. The forecasts were verified at increasingly high thresholds of the observed variable. These thresholds are expressed in terms of climatological exceedence probabilities. In order to compare the verification results between seasons, the thresholds were derived from the overall observed sample, ensuring fixed amounts of precipitation and temperature throughout the year. Figure 12 shows the CRPSS of the MEFP-GEFS precipitation forecasts, together with the MEFP-GFS forecasts, for increasing amounts of observed precipitation. Figure 13 shows the corresponding results for the MEFP temperature forecasts. The CRPSS is based on sample climatology, which comprises the climatology of observations in the paired sample, i.e. in the seasonal subsample. The results are plotted for the downstream basin in each RFC and for the wet and dry seasons, as well as the overall period. Three forecast lead times are considered, namely one, seven and 14 days.

As indicated in Figure 12, the skill of the MEFP precipitation forecasts varies widely between the four basins considered, with the greatest CRPSS in CN-FTSC1, followed by MA-CNNN6, AB-BLKO2 and CB-DOLC2. In general, the MEFP forecasts are more skillful during the winter months. The winter occurs during the “wet” season in CN- and MA-RFCs and during the “dry” season in AB- and CB-RFCs. The seasonal differences in CRPSS typically amount to ~10-15% at early forecast lead times, with smaller differences as the overall skill declines. For example, in MA-CNNN6, the maximum CRPSS of the MEFP-GEFS forecasts is ~60% during the wet season at a forecast lead time of one day and ~46% during the dry season. At a forecast lead time of seven days, this declines to ~11% during the wet season and ~6% during the dry season. The MEFP precipitation forecasts are generally skillful in AB-BLKO2, CN-FTSC1 and MA-CNNN6 during both the wet and dry seasons. However, in CBRFC, the MEFP-GFS and the MEFP-GEFS forecasts are less skillful than climatology for some precipitation amounts and forecast lead times, but particularly during the wet season and at longer forecast lead times. As indicated in

Section 5.1.1, this may originate from differences in the climatology of the calibration and verification periods, as the latter comprises only a subset of years from the former. Despite the differences between basins, the MEFP-GEFS forecasts consistently improve upon the MEFP-GFS forecasts; that is, for all basins, in both seasons, and at most forecast lead times and precipitation amounts. In general, the greatest improvements occur in the earlier portion of the forecast horizon, but not always on the first day (Figure 12). Indeed, in CN-FTSC1, the MEFP-GEFS precipitation forecasts are no more skillful than the MEFP-GFS forecasts during the first 1-2 days, but are substantially more skillful after seven days, particularly during the wet season and for high precipitation amounts (by ~10-15% at $C_p \approx 0.01$, relative to the climatological baseline). In MA-CNNN6, the greatest overall improvements occur during the summer months at a forecast lead time of ~four days. Here, the maximum CRPSS is achieved at moderately high precipitation amounts ($C_p \approx 0.05$), where the MEFP-GEFS forecasts are ~30% more skillful than climatology and the MEFP-GFS forecasts are only ~10% more skillful.

For the temperature forecasts (Figure 13), the seasonal variations in CRPSS are controlled by three factors. First, the seasonality is reversed in AB- and CB-RFCs when compared to CN- and MA-RFCs. Thus, the two groups show similar patterns of skill in opposite seasons. Second, the CRPSS measures the *relative* quality of the forecasts. Under warm (summer) conditions, the MEFP forecasts are most skillful at the lowest and highest temperatures, where sample climatology is, by definition, conditionally biased. They are less skillful at moderate temperatures, where sample climatology performs reasonable well. Under cool (winter) conditions, the MEFP forecasts are more skillful at relatively warmer temperatures, again because sample climatology does not predict conditionally upon season (and warm temperatures are unusual in cool months). Third, the MEFP forecasts are conditionally biased at the coldest observed temperatures. As forecast skill partly depends on these conditional biases, the MEFP forecasts are less skillful during the cold season at the lowest observed temperatures. Nevertheless, the MEFP forecasts are considerably more skillful than sample climatology at all forecast lead times and for all observed temperatures. In keeping with the precipitation forecasts, the MEFP-GEFS temperature forecasts consistently improve upon the MEFP-GFS forecasts (Figure 13). The greatest improvements occur in CN-FTSC1 and MA-CNNN6 at a forecast lead time of ~four days. Here, the CRPSS of the MEFP-GEFS forecasts is ~20% higher across all temperature thresholds. In general, the improvements are smaller in AB-BLKO2 and CB-DOLC2. However, during the winter months (as well as the summer months in CB-DOLC2), the value added by the MEFP-GEFS increases when colder temperatures are included in the verification sample. For example, during the winter months in CB-DOLC2, the CRPSS is ~0.6 when using the MEFP-GEFS to forecast temperatures that are greater than -8 degrees C (C_p=0.9) and ~0.4 when using the MEFP-GFS. In CBRFC, the additional skill in the MEFP-GEFS temperature forecasts may be important for hydrologic forecasting, as the timing and magnitude of the snowmelt are sensitively dependent on temperature during the snowmelt period (Figure 2).

Figure 14 shows the net gain in forecast lead time associated with the MEFP-GEFS precipitation forecasts when compared to the MEFP-GFS forecasts. As illustrated in Figure 15, the net gain in forecast lead time, Δt, comprises the difference in forecast lead time between the MEFP-GEFS forecasts and the MEFP-GFS forecasts when they achieve the same value of a verification score, m. Figure 14 shows the net gain in forecast lead time for three verification scores, namely the correlation coefficient, the mean Continuous Ranked Probability Skill Score (CRPSS) and the Brier Skill Score (BSS). The correlation coefficient and the CRPSS were determined from the unconditional sample of verification pairs. The BSS was determined for a daily precipitation total that is exceeded, on average, once every ten days (C_p=0.1). In computing the CRPSS and BSS, sample climatology was used as the reference forecast. In order to reduce the sampling uncertainty of Δt, the net gain was averaged over the first seven days of the forecast horizon, denoted $\overline{\Delta t}$. The average comprised only finite values of Δt; that is, instances where both sources of forcing achieved the same values of m at some point during the 14-day forecast horizon. The average net gain, $\overline{\Delta t}$, is negative when the MEFP-GEFS forecasts achieve a lower score, on average, than the MEFP-GFS forecasts and positive when the MEFP-GEFS forecasts achieve a higher score, on average, than the MEFP-GFS forecasts. The average net gain is zero when the forecasts have equivalent scores at all forecast lead times or the positive and negative values of Δt balance over the averaging period. Figure 16 shows the equivalent plot for the MEFP temperature forecasts, where the BSS comprises a daily mean temperature that is exceeded with 90% probability (C_p=0.9) or, equivalently in terms of the BSS, a temperature that is not exceeded with 10% probability.

As indicated in Figure 14, the MEFP-GEFS precipitation forecasts show a net gain in forecast lead time when compared to the MEFP-GFS forecasts. Put differently, the MEFP-GEFS forecasts show equivalent correlation, CRPSS and BSS at longer forecast lead times than the MEFP-GFS forecasts. When averaged over the first seven days, the MEFP-GEFS forecasts show equivalent correlation and skill for 1-2 days longer than the MEFP-GFS forecasts. For each RFC, these improvements are broadly consistent between seasons, metrics and basins (Figure 14). Between RFCs, the improvements are somewhat larger in CBRFC than other RFCs. However, in CBRFC, the apparent gains from the MEFP-GEFS should be treated with caution. For example, as shown in Figure 8, the precipitation forecasts show little or no improvement on sample climatology (i.e. the curves are flat). In this context, the sampling uncertainties are larger and a gain in forecast lead time may not be physically meaningful. For temperature, the benefits of the MEFP-GEFS forecasts are generally more pronounced (Figure 16). Here, the MEFP-GEFS forecasts show equivalent correlation and skill for 2-4 days longer than the MEFP-GFS forecasts. This is apparent across all RFCs, both seasons and for most verification scores. For example, in AB-BLKO2, the MEFP-GEFS produces temperature forecasts for the 10th percentile of the climatological distribution with equivalent BSS, but with 4.5 days of additional forecast lead time, to those produced by the MEFP-GFS. Elsewhere, in MA-CNNN6, the MEFP-GEFS temperature forecasts show equivalent correlations ~four days later than the MEFP-GFS forecasts during the dry season and ~three days later during the wet season (Figure 16).

5.2 Quality of the raw streamflow forecasts

In order to understand the benefits of the MEFP-GEFS forcing separately from any hydrologic biases, the raw streamflow forecasts were verified against simulated streamflows. The verification results are presented by forecast lead time and season and amount of streamflow.

5.2.1 Forecast lead time and season

Figure 17 shows the relative mean error (RME) of the streamflow forecasts with forcing inputs from the MEFP. The results are shown separately for the “wet” and “dry” seasons in each RFC, as well as the overall period (see Figure 2 also). The RME is plotted against forecast lead time for each basin and for each source of forcing used in the MEFP, namely MEFP-CLIM, MEFP-GFS and MEFP-GEFS. When verified against simulated flows, errors in the streamflow forecasts originate from errors in the MEFP forcing and in the observed forcing from which the simulations are produced (these are assumed to be negligible). Furthermore, when the forcing and streamflow climatologies are non-stationary, discrepancies in the calibration and verification periods can also be important. In this study, the HEFS was applied to the operational hydrologic models, which are calibrated for different periods in each basin. In general, these periods are not aligned with the calibration of the MEFP or the EnsPost. They also differ from the verification period, i.e. 1985-1999. Thus, biases in the raw streamflow forecasts could originate from multiple sources, including biases in the MEFP forcing.

As indicated in Figure 17, the ensemble mean forecast underestimates the simulated flows in AB- and CN-RFCs and overestimates the simulated flows in CBRFC. In MARFC, the ensemble mean is relatively unbiased during the wet season and for the overall period, but shows an underforecasting bias during the dry season, which increases with forecast lead time. In general, the RME of the overall period is dominated by the wet season, as the mean error is sensitive to high flows (in contrast to the mean relative error, for example). In ABRFC, the ensemble mean forecast underestimates the simulated streamflow by 5-10% (Figure 17). This is broadly consistent with the precipitation forecasts, which underestimate the observed precipitation by 5-10%, on average (Figure 4). However, in CBRFC, the simulated flows are overestimated by the ensemble mean forecast, while the precipitation forecasts underestimate the observed precipitation. This discrepancy is understandable because CB-DRRC2 and CB-DOLC2 are snow-dominated basins. In these basins, errors in precipitation are not immediately reflected in streamflow and the quality of the streamflow forecasts also depends strongly on temperature during the snowmelt period. The greatest biases occur in CNRFC, where the MEFP-GEFS streamflow forecasts underestimate the simulated streamflows by 5-15%. Here, the MEFP-GFS and MEFP-GEFS precipitation forecasts over- and under-estimate the observed precipitation, respectively, while the MEFP-CLIM forecasts are relatively unbiased (Figure 4). This ordering is broadly reflected in the raw streamflow forecasts, with smaller negative biases in the MEFP-GFS forecasts than the MEFP-GEFS forecasts. As indicated above, differences between the hydrologic forecasts and simulations may originate from several sources of error other than the MEFP forecasts.

Figure 18 shows the correlation of the ensemble mean forecast and simulated streamflow by forecast lead time. The results are shown for the upstream and downstream basin in each RFC and for each source of forcing in the MEFP, namely resampled climatology, MEFP-GFS and MEFP-GEFS. Again, the results are shown separately for the wet and dry seasons, as well as the overall period. Figure 19 shows the CRPSS of the streamflow forecasts with forcing inputs from the MEFP-GFS and the MEFP-GEFS. Here, skill is measured against the streamflow forecasts with forcing inputs from resampled climatology, MEFP-CLIM.

As indicated in Figure 18, the correlations are substantially higher for the MEFP-GEFS streamflow forecasts than resampled climatology. The MEFP-GEFS streamflow forecasts are also substantially more skillful than the MEFP-CLIM forecasts (Figure 19). Similarly, when compared to the MEFP-GFS streamflow forecasts, the MEFP-GEFS forecasts are consistently more correlated with the simulated streamflows and show higher CRPSS. At early forecast lead times, the correlations are similar for all sources of forcing (Figure 18) and the CRPSS is lower (Figure 19). This is understandable because the initial conditions are common to all streamflow forecasts. Also, the hydrologic models respond unevenly to meteorological forcing, depending on basin characteristics and antecedent conditions. For example, in AB-CBNK1 and AB-BLKO2, the streamflow forecasts show limited skill (Figure 19) and a rapid decline in correlations with increasing forecast lead time (Figure 18). This originates from the lack of hydrologic persistence in these basins and the difficulty in forecasting precipitation beyond the short range. Indeed, while the MEFP-GEFS precipitation forecasts improve substantially upon the MEFP-GFS forecasts, they are no more skillful than the MEFP-CLIM forecasts after ~one week

(Figure 8). In contrast, CB-DRRC2 and CB-DOLC2 are snow-dominated basins, where forecast skill is driven by the hydrologic uncertainties and the streamflow correlations decline only gradually over time (Figure 18). The MEFP-GEFS streamflow forecasts are substantially more skillful than the MEFP-GFS forecasts in these basins (Figure 19), contributing five or more days of additional forecast lead time in the medium-range alone. However, when verified against observed streamflows, this additional skill is not fully translated into the post-processed streamflow forecasts (see Section 5.3.1).

In keeping with the quality of meteorological forecasts in MA- and CN-RFCs (e.g. Figure 7 and Figure 8), and the relative importance of the meteorological uncertainties in these basins (Section 5.3), the streamflow forecasts show high correlations and good skill in MA- and CN-RFCs. Here, the benefits of the MEFP-GEFS forecasts are greatest during the wet season, particularly in MA-WALN6 and MA-CNNN6, where the MEFP-GEFS streamflow forecasts show equivalent CRPSS ~2 days ahead of the MEFP-GFS forecasts (Figure 19). In these basins, the MEFP-GFS streamflow forecasts are indistinguishable from climatology after ~ten days, while the MEFP-GEFS forecasts remain skillful throughout the medium-range. During the summer months, the streamflows in CNRFC are dominated by hydrologic persistence, with little flow in the upper reaches of the Eel River (Figure 2). Under these conditions, the MEFP-GEFS forecasts do not significantly improve on the MEFP-GFS forecasts (Figure 19).

5.2.2 *Magnitude of streamflow*

The raw streamflow forecasts were verified conditionally upon the amount of streamflow at each forecast lead time. Figure 20 shows the RME of the ensemble mean forecast at selected forecast lead times, while Figure 21 shows the correlation of the ensemble mean forecast and simulated streamflow. The results are shown for each source of forcing in the MEFP, namely resampled climatology, MEFP-GFS and MEFP-GEFS. The scores are plotted against climatological exceedence probability on a probit scale, but are labeled with actual probability. As indicated in Section 4.4, continuous measures, such as the RME, were derived from the subset of verification pairs whose simulated value exceeded the threshold. Thus, flows denoted by a climatological probability of C_p=0.1 comprise the 10% of flows that exceed this threshold and not the 90% of flows that fall below it. Figure 22 shows the BSS of the MEFP-GEFS and MEFP-GFS streamflow forecasts, where the baseline comprises the streamflow forecasts with climatological forcing. For discrete measures, such as the Brier Score, the forecast event and its complement have the same error in absolute terms. Thus, the BSS is equivalent for a streamflow rate that exceeds a threshold of C_p=0.1 or does not exceed this threshold. The hydrologic initial conditions do not contribute to the BSS, as the MEFP-CLIM forecasts were initialized from the same (warm) states as the MEFP-GFS and MEFP-GEFS forecasts. This is important because the initial conditions contribute a significant fraction of the total skill in some basins, notably in CB-DRRC2 and CB-DOLC2.

In AB-, CN- and MA-RFCs, the raw streamflow forecasts are conditionally biased in the ensemble mean with increasing streamflow amount (Figure 20) and they are increasingly biased at longer forecast lead times. Indeed, the conditional biases increase in proportion to the decline of forecast skill over time, as climatology is, by definition, conditionally biased. For example, in ABRFC, the conditional biases increase rapidly over the first week, as the forecasts show little skill beyond one week. In CNRFC, the conditional biases are much smaller at ~seven days than ~14 days, as the forecasts remain skillful during the middle portion of the forecast horizon. These conditional biases originate from the precipitation forecasts rather than the hydrologic modeling (Figure 11). Consequently, they are sensitive to the degree of conditional bias in the precipitation forecasts, as well as the sensitivity of the streamflow forecasts to the meteorological forcing. For the same reason, the streamflow forecasts are conditionally unbiased for most streamflow rates in CB-DRRC2 and CB-DOLC2, despite the large conditional biases in the precipitation forecasts (Figure 11). This stems from the importance of snowmelt in generating large streamflows in these basins (Figure 2). As snow accumulation involves a time integral over the accumulation period, there is a weaker dependence of high streamflows on heavy precipitation. In general, the raw streamflow forecasts show similar conditional biases for the MEFP-GFS and MEFP-GEFS forcing, with the latter contributing slightly higher (negative) biases in CNRFC, particularly under dry conditions, and slightly lower (negative) biases in ABRFC.

In order to illustrate the practical impacts of these conditional biases on forecasting error and the capacity to warn about unusually high streamflows, box plots were computed from the raw streamflow forecasts. The results are shown in Figure 23 for the downstream basin in each RFC and at a forecast lead time of five days (recalling that the conditional biases increase over time). The box plots are organized by increasing amount of simulated streamflow. Each box represents one ensemble forecast from the period 4-5 days. Selected quantiles of the forecasting errors are plotted together with the median error and range (extreme residuals) as whiskers. In keeping with the precipitation forecasts (Figure 11), the streamflow forecasts with MEFP-GEFS forcing show similar conditional biases to those with MEFP-GFS forcing, but increased spread in some basins. In CB-DOLC2, CN-FTSC1 and MA-CNNN6, both the MEFP-GFS and the MEFP-GEFS forecasts show sufficiently large spread, and sufficiently low biases, to provide some warning of the highest simulated flows (Figure 23). In AB-BLKO2, the MEFP-GEFS forecasts partially compensate for a large conditional bias with increased spread. Here, the MEFP-GEFS forecasts consistently warn of the highest simulated flows (Figure 23), whereas the MEFP-GFS forecast are unable to provide a warning in most cases.

While the streamflow forecasts with MEFP-GEFS forcing show similar conditional biases to the MEFP-GFS forecasts, the MEFP-GEFS forecasts show higher correlations and improved skill in most basins, at most streamflow rates, and for most forecast lead times (Figure 21 and Figure 22). These improvements are sometimes more pronounced during the middle portion of the forecast horizon than earlier forecast lead times. For example, in CNRFC, the MEFP-GEFS forecasts show similar correlations to the MEFP-GFS forecasts after two days, but higher correlations after seven days (Figure 21). This partly reflects the lagged response of the hydrologic models to forcing. It also reflects the characteristics of the MEFP-GEFS forcing (e.g. Figure 7 and Figure 8), where the marginal improvements over the MEFP-GFS were greater during the middle portion of the forecast horizon (e.g. in CN-FTSC1). As with the unconditional CRPSS (Figure 19), the greatest improvements in BSS occur in CB-DRRC2 and CB-DOLC2 (Figure 22). For example, the MEFP-GFS forecasts are 10% more skillful than the MEFP-CLIM forecasts at predicting streamflow rates that are above (or below) the median streamflow, while the MEFP-GEFS forecasts are 50% more skillful at predicting these rates (Figure 22).

5.3 Quality of the bias-corrected streamflow forecasts

In order to establish the benefits of the MEFP-GEFS forcing in an operational context, the post-processed streamflow forecasts were verified against observed flows. The results are presented by forecast lead time and season, and amount of streamflow. The overall skill is determined relative to the raw streamflow forecasts with climatological forcing, and the contributions from the MEFP and the EnsPost are factored out. Alongside the verification results, a selection of the paired forecasts and observations is provided in Appendix C. By plotting the ensemble mean and range against the observed streamflow amounts, the strengths and weaknesses of the HEFS forecasts can be evaluated (albeit subjectively) for specific hydrologic events, providing some insight into timing and amplitude errors before and after streamflow post-processing, both for the MEFP-GFS and MEFP-GEFS forcing.

5.3.1 Forecast lead time and season

Figure 24 shows the RME of the bias-corrected streamflow forecasts with forcing inputs from the MEFP-GEFS and the MEFP-GFS. The RME of the raw streamflow forecasts is also shown for the MEFP-GEFS forcing. The ability of the EnsPost to produce reliable and skillful hydrologic forecasts will depend on several factors, including any residual biases in the MEFP forcing, the length and quality of the calibration data, whether the assumptions of the EnsPost are met, and the skill of the predictors. The hydrologic uncertainties and biases are estimated from the residuals between the observed and simulated streamflows (Appendix A). These residuals include errors in the observed forcing, from which the simulated flows are produced, and in the observed streamflows (both of which are assumed to be reasonably unbiased). They do not include errors in the meteorological forecasts, which are modeled by the MEFP. Thus, any residual biases from the MEFP will propagate into the streamflow forecasts. Alongside the various sources of bias in the streamflow forecasts, there are various measures of unconditional and conditional bias. These measures have different sensitivities and practical implications. For example, the ensemble mean, as well as the mean error, are sensitive to outliers, particularly for skewed variables, such as precipitation and streamflow.

Nevertheless, the mean error of the ensemble mean forecast is an important measure of unconditional unbiasedness, and unconditional unbiasedness is an important attribute of an operational forecasting system. When estimating the parameters of the EnsPost, the estimates are not guaranteed to produce forecasts that are unconditionally unbiased (Seo et al., 2006). Rather, the predictions aim to minimize the mean CRPS of the post-processed streamflow forecasts. The unconditional bias is only one component of the CRPS, specifically a contribution to the reliability term (Hersbach, 2000).

Following streamflow post-processing, the MEFP-GEFS streamflow forecasts show similar unconditional biases to the MEFP-GFS forecasts (Figure 24). In general, the EnsPost reduces the RME of the ensemble mean forecast. However, these improvements are not uniform. For example, in AB-BLKO2 and AB-CBNK1, the unconditional biases are increased during the dry season while, in MA-CNNN6, they are increased during the wet season. Elsewhere, in CN-DOSC1 and CN-FTSC1, there is an underforecasting bias, particularly during the wet season. This is not improved by the EnsPost and is greater for the MEFP-GEFS forecasts than the equivalent MEFP-GFS forecasts. In practice, the meteorological uncertainties and biases interact with the hydrologic uncertainties and biases and lead to complex behaviors. Depending on the relative magnitude and direction of these biases, and the sensitivities of the hydrologic models, the meteorological biases may be exaggerated or modulated by streamflow post-processing. For example, in CNRFC, the meteorological uncertainties account for a significant fraction of the total uncertainties (see below). In these basins, the raw MEFP-GEFS streamflow forecasts show similar RME to the bias-corrected forecasts (Figure 24). Both underestimate the observed streamflow, when compared to the MEFP-GFS forecasts, because the MEFP-GEFS precipitation forecasts are drier, on average, than the MEFP-GFS precipitation forecasts (Figure 4). In contrast, during the wet season (which dominates the overall period), the post-processed streamflow forecasts are relatively unbiased in CBRFC. Superficially, this appears to be inconsistent with the underforecasting bias in the MEFP-GEFS precipitation forecasts (Figure 4) and the over-forecasting bias in the raw streamflow forecasts (Figure 24). However, this reflects the greater importance of the hydrologic uncertainties in CBRFC (see below), as well as the complex relationship between the meteorological and hydrologic uncertainties in these two basins. Here, the RME (among other unconditional statistics) is particularly sensitive to the high streamflows that occur during the snowmelt period. Snowmelt is sensitively dependent upon temperature, but relatively insensitive to biases in precipitation. Consequently, the underforecasting biases in the MEFP-GEFS precipitation forecasts (Figure 4) are not reflected in the MEFP-GEFS streamflow forecasts (Figure 24).

Figure 25a and Figure 25b show the overall CRPSS of the post-processed streamflow forecasts for the upstream and downstream basins in each RFC, respectively. The results are shown for the streamflow forecasts with forcing inputs from the MEFP-GEFS, as well as the MEFP-GFS. The baseline in the CRPSS comprises the raw streamflow forecasts with forcing inputs from MEFP-CLIM. In addition to the overall skill, the CRPSS is factored into contributions from the MEFP and the EnsPost

$$\underbrace{\frac{CRPS_{CLIM}-CRPS_{GEFSPOST}}{CRPS_{CLIM}}}_{\text{Total skill}}=\underbrace{\frac{CRPS_{CLIM}-CRPS_{GEFS}}{CRPS_{CLIM}}}_{\text{MEFP - GEFS skill}}+\underbrace{\frac{CRPS_{GEFS}-CRPS_{GEFSPOST}}{CRPS_{CLIM}}}_{\text{EnsPost skill}}, \qquad (4)$$

where the subscripts denote the source of forcing used in the MEFP. As indicated above, the streamflow forecasts were all initialized from the same (warm) states. Thus, the skill from the initial conditions is factored out of the CRPSS.

The overall skill of the post-processed streamflow forecasts, as well as the relative contributions from the MEFP and the EnsPost, vary with basin, season, and forecast lead time. They also vary with the source of forcing used in the MEFP. In general, the bias-corrected MEFP-GEFS forecasts are substantially more skillful than the raw MEFP-CLIM forecasts. The overall skill is greatest in CBRFC and CNRFC, where the seasonal differences are also greatest. For example, during the wet season in CN-DOSC1, the post-processed MEFP-GEFS forecasts are up to ~40% more skillful than the raw MEFP-CLIM forecasts (Figure 25a). In the dry season, the overall skill increases to ~60% at the earliest forecast lead times (Figure 25a). However, the origins of this skill are quite different, depending on the season considered. In both RFCs, the dry season is dominated by low flows and hydrologic persistence. During prolonged dry periods, the meteorological forecasts are relatively unimportant (indeed, the MEFP contributes negatively in CBRFC), while the EnsPost benefits from hydrologic persistence. In contrast, during the wet season, the overall skill is driven by the meteorological forcing in CNRFC, while the hydrologic uncertainties remain important in CBRFC. This is understandable, because CB-DRRC2 and CB-DOLC2 are snow-dominated basins, where the EnsPost also benefits from hydrologic persistence under snowmelt conditions. The hydrologic uncertainties are less important in CN-DOSC1 and CN-FTSC1, including at high flows. Here, predictability is greatly enhanced during the winter months by the onshore movement of weather fronts from the Pacific coast and their orographic lifting in the North Coast Ranges. In AB-CBNK1 and AB-BLKO2, the overall CRPSS is lower, as the meteorological forecasts are less skillful (e.g. Figure 5), and the hydrologic persistence is also lower. In both AB- and MA-RFCs, the relative contributions from the MEFP and the EnsPost vary with forecast lead time (Figure 25a and Figure 25b). At the earliest forecast lead times (~1-2 days), much of the skill originates from the EnsPost. After ~two days, the MEFP contributes a significant fraction of the total CRPSS.

As indicated in Figure 25a and Figure 25b, the MEFP-GEFS streamflow forecasts are consistently more skillful than the MEFP-GFS forecasts. The fraction of skill contributed by the MEFP also increases when using the MEFP-GEFS forcing, particularly during the wet season in CB-DRRC2, CB-DOLC2, MA-WALN6 and MA-CNNN6. During the middle portion of the forecast horizon, the MEFP-GFS forcing contributes little or no skill in these basins (even negative skill in CB-DRRC2 and CB-DOLC2). In contrast, the MEFP-GEFS forcing accounts for most or all of the skill in the MEFP-GEFS streamflow forecasts at corresponding forecast lead times (not including any background skill from the hydrologic initial conditions). However, neither the MEFP-GEFS forcing nor the MEFP-GFS forcing contribute valuable skill during the dry season in CBRFC (Figure 25a and Figure 25b). Rather, in CB-DOLC2, both the MEFP-GEFS precipitation forecasts and the MEFP-GFS forecasts are somewhat less skillful than the MEFP-CLIM forecasts (Figures 25b; Figure 4). In practice, however, the MEFP forcing does not translate into streamflow during the winter months, as most of the precipitation in CB-DRRC2 and CB-DOLC2 falls as snow.

Figure 26 shows the net gain in forecast lead time associated with the MEFP-GEFS streamflow forecasts when compared to the MEFP-GFS forecasts. As illustrated in Figure 15, the net gain in forecast lead time, Δt, comprises the difference in forecast lead time between the MEFP-GEFS forecasts and the MEFP-GFS forecasts achieving the same value of a verification score, m. Figure 26 shows the net gain in forecast lead time for three verification scores, namely the correlation coefficient, the CRPSS and the BSS. The correlation coefficient and the CRPSS were determined from the unconditional sample of verification pairs. The BSS was determined for a daily streamflow rate that is exceeded, on average, once every ten days (C_p=0.1). In order to reduce the sampling uncertainty of Δt, an average was computed over the first seven days of the forecast horizon, denoted $\overline{\Delta t}$. The average comprised only finite values of Δt; that is, instances where both forecasts achieved the same values of m at some point during the 14-day forecast horizon. As indicated in Figure 26, the MEFP-GEFS forecasts generally show equivalent correlation, CRPSS and BSS at longer forecast lead times than the MEFP-GFS forecasts. Indeed, during the wet season in CB- and CN-RFCs, and throughout the year in MA- and AB-RFCs, the MEFP-GEFS typically adds 1-2 days in forecast lead time for all verification measures. For example, in CN-FTSC1, the MEFP-GEFS forecasts show equivalent skill to the MEFP-GFS forecasts in predicting streamflow rates above C_p=0.1, but with an additional forecast lead time of ~2.5 days (Figure 26). In CB- and CN-RFCs, the dry season is characterized by persistent low flows (Figure 2). Here, the benefits of the MEFP-GEFS forcing are, understandably, smaller, as the meteorological and hydrologic uncertainties are low.

5.3.2 Magnitude of streamflow

The post-processed streamflow forecasts were verified for increasing amounts of observed streamflow. Figure 27 shows the Brier Skill Score (BSS) at a forecast lead time of ~five days (~90-114 hours: see Section 4.3) for the upstream and downstream basins in each RFC. As indicated above, the benefits of the MEFP-GEFS forecasts are typically greatest during the middle portion of the forecast horizon. The BSS is plotted against climatological exceedence probability, C_p. For example, C_p=0.1 denotes the daily mean streamflow that is exceeded, on average, once every ten days. The BSS measures the gain in skill (or reduction in BS) of the post-processed streamflow forecasts with MEFP-GEFS forcing, and with MEFP-GFS forcing, relative to those with climatological forcing. In addition, the BSS is shown for the raw streamflow forecasts with MEFP-GEFS forcing. By conditioning on the observed and forecast variables, the BSS can be factored into more detailed attributes of forecast quality (Appendix B). When conditioning on the forecast variable, these comprise the "relative reliability" and "relative resolution", which are also shown in Figure 27. For each forecast probability issued, the reliability component of the BS measures the extent to which the average forecast probability differs from the average observed probability (Appendix B). As statistical post-processing focuses on the same conditional biases, much of the skill contributed by the EnsPost may originate from improvements in reliability (Seo et al., 2006; Brown and Seo, 2013). The resolution measures the sensitivity of the observed outcomes when grouping by forecast probability; that is, whether the forecast probability is closely related to an event occurring or not occurring, after factoring out any conditional bias (Appendix B). As indicated above, all components of the BSS are relative to the streamflow forecasts with MEFP-CLIM forcing, which removes any contribution from the hydrologic initial conditions.

At low streamflow thresholds, the post-processed MEFP-GEFS forecasts show similar skill to the post-processed MEFP-GFS forecasts (Figure 26). Here, the majority of skill originates from the EnsPost, rather than the MEFP. Under low flow conditions, the streamflow forecasts benefits from hydrologic persistence; that is, from the prior observed streamflow, which is used as an auxiliary predictor in the EnsPost (Appendix A). At these thresholds, the bias-corrected streamflow forecasts are substantially more reliable, and somewhat more resolved, than the raw forecasts (Figure 26). In AB-, CB- and CN-RFCs, the upstream and downstream basins show similar patterns of skill, with similar (non-equal) contributions from reliability and resolution. However, in MARFC, the forecast probabilities of exceeding low flows are substantially less skillful at Cannonsville (MA-CNNN6) than Walton (WALN6). This reflects a combination of lower resolution and higher conditional bias at CNNN6 (Figure 26), which may originate from the truncation of estimated inflows to the Cannonsville Reservoir under dry conditions.

Figure 28 shows the reliability diagrams for the bias-corrected streamflow forecasts with MEFP-GEFS forcing at a forecast lead time of ~2 days (~18-42 hours). The results are shown for the downstream basin in each RFC and for selected streamflow thresholds. By comparing the average observed and forecast probabilities for each group (bin) of forecast probabilities, the reliability of the forecasts can be established in terms of absolute probabilistic error. As indicated in Figure 28, the bias-corrected MEFP-GEFS forecasts are both reasonably sharp (confident) and reliable (not over-confident) at low and moderate streamflow thresholds. In MA-CNNN6, the forecasts of low flows (C_p=0.9) are truncated at low probability thresholds. Figure 29 shows the Relative Operating Characteristic (ROC) for the bias-corrected streamflow forecasts with MEFP-GEFS forcing. The results are shown for the downstream basin in each RFC, for selected thresholds, and at a forecast lead time of ~2 days (~18-42 hours). The ROC curves were fitted under an assumption of bivariate normality between the Probability of Detection (PoD) and the Probability of False Detection (PoFD) and are shown together with the empirical pairs of PoD and PoFD for each threshold (Appendix B). The streamflow forecasts are most discriminatory in CB-DOLC2 and CN-FTSC1. In BLKO2, the forecasts are less discriminatory at moderate streamflow thresholds and, in CNNN6, they are less discriminatory at all streamflow thresholds, but particularly at low flows (C_p=0.9). However, the bias-corrected forecasts are substantially more discriminatory than sample climatology in all basins and at all streamflow thresholds (the diagonal line in Figure 29).

At moderate and higher streamflow thresholds, a greater fraction of the total skill in the streamflow forecasts originates from the MEFP than the EnsPost. Following streamflow post-processing, the MEFP-GEFS forecasts generally improve upon the MEFP-GFS forecasts, particularly in CBRFC (Figure 26). However, these improvements are modest and are subject to sampling uncertainties. Also, at high streamflow thresholds, the benefits of the EnsPost are less clear. Indeed, while the MEFP-GEFS forecasts are no less skillful than the MEFP-GFS forecasts, they are not improved by hydrologic post-processing. Figure 30 shows the CRPSS of the MEFP-GEFS forecasts before and after streamflow post-processing, together with the post-processed MEFP-GFS forecasts. Again, following streamflow post-processing, the MEFP-GEFS forecasts are generally more skillful than the MEFP-GFS forecasts. However, except in CNRFC where the

EnsPost contributes little to the overall skill under wet conditions (e.g. see Figure 25a and Figure 25b for the wet season), the MEFP-GEFS forecasts show some *loss* of skill from streamflow post-processing. This also occurs for the MEFP-GFS forecasts (results not shown). Indeed, at early forecast lead times in AB- and MA-RFCs, and later forecast lead times in CB- and MA-RFCs, the post-processed MEFP-GEFS forecasts are less skillful than the raw MEFP-GEFS forecasts at moderate and higher flows (C_p<0.1) (Figure 30). Nevertheless, at early forecast lead times, the post-processed MEFP-GEFS forecasts are substantially more skillful than the MEFP-CLIM forecasts, both at high and low flows (Figure 30). They are also more discriminatory than a climatological probability forecast (Figure 29).

6. Discussion and conclusions

Retrospective forecasts of temperature, precipitation and streamflow were generated with the Hydrologic Ensemble Forecasts Service (HEFS) for selected river basins in four NWS River Forecast Centers (RFCs), namely the Arkansas-Red Basin RFC (ABRFC), the Colorado Basin RFC (CBRFC), the California-Nevada RFC (CNRFC) and the Middle Atlantic RFC (MARFC). The precipitation and temperature forecasts were generated with the HEFS Meteorological Ensemble Forecast Processor (MEFP). The MEFP produces ensemble forecasts of Mean Areal Temperature (MAT) and Mean Areal Precipitation (MAP), conditionally upon a raw, single-valued, forecast. Here, the single-valued forecast comprised the ensemble mean from NCEP's Global Ensemble Forecast System (GEFS). Until recently, the MEFP used the frozen (circa 1997) version of NCEP's Global Forecast System (GFS; see Brown 2013). The frozen GFS employs a horizontal resolution of T62 or ~250km. The operational GEFS uses Version 9.0.1 of the GFS, which comprises a horizontal resolution of T254 (~55km) for 1-8 days and T190 (~70km) for 9-16 days, and a vertical resolution of L42 or 42 levels (Wei et al. 2008; Hamill et al. 2011). A new reforecast dataset was recently completed by NCEP (Hamill et al. 2013) and was used to calibrate the MEFP. The hindcasts of temperature, precipitation and streamflow were generated for a 15 year period between 1985 and 1999. While the GEFS reforecasts were available until 2010, the observations of MAT, MAP and streamflow were not consistently available in all basins after 1999. In CB- and CN-RFCs, the upstream and downstream basins were separated into multiple sub-basins, in order to accommodate the varied elevations there. The forecast time horizon was 1-14 days and the timestep of the hydrologic models was six-hourly in MA- and AB-RFCs and hourly in CB- and CN-RFCs. Streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS). In AB-, CB- and CN-RFCs, the hydrologic models included the Sacramento Soil Moisture Accounting model (SAC-SMA) and the Snow Accumulation and Ablation Model (SNOW-17). In MARFC, the SAC-SMA was substituted with an empirical model, based on the Antecedent Precipitation Index (API). The raw streamflow forecasts were post-processed with the Ensemble Postprocessor (EnsPost). The EnsPost accounts for the hydrologic uncertainties and reduces any hydrologic biases (Seo et al., 2006).

The precipitation, temperature and streamflow forecasts were verified with the Ensemble Verification System (Brown et al., 2010). The results are presented by forecast lead time, season, and magnitude of the observed and forecast variables. In order to establish the benefits of the MEFP-GEFS for operational hydrologic forecasting, hindcasts of temperature, precipitation and streamflow were also generated with the frozen version of NCEP’s GFS (MEFP-GFS). Both the MEFP-GEFS forecasts and the MEFP-GFS forecasts were compared to a “resampled” or conditional climatology (MEFP-CLIM). This involved resampling the historical observations of MAP and MAT in a moving window of, respectively, 61 days and 31 days around the forecast valid date. By conditioning separately on the observed and forecast variables, the forecast errors were factored into several attributes of forecast quality, including independent measures, such as reliability and discrimination. The precipitation and temperature forecasts were verified against observed MAP and MAT, respectively. In order to establish the benefits of the MEFP-GEFS forecasts separately from any hydrologic biases, the raw streamflow forecasts were verified against simulated flows. In addition, the bias-corrected streamflow forecasts were verified against observed flows. This allowed the benefits of the MEFP-GEFS forcing to be established in an operational context, where the hydrologic uncertainties may outweigh the meteorological uncertainties, and the EnsPost cannot be expected to remove all of the hydrologic biases.

A detailed comparison between the raw GEFS reforecasts and the MEFP-GEFS forecasts was hampered by the different spatial scales of inputs and outputs (i.e. a single grid node versus a basin average). Nevertheless, the correlations between the raw forecasts and observations are preserved or improved by the MEFP at all forecast lead times, in both seasons, and at all magnitudes of the observed variable. Also, the MEFP-GEFS forecasts consistently improve upon the MEFP-CLIM forecasts. Both are important attributes of reliable and skillful meteorological forecasts. Indeed, the MEFP aims to preserve the correlations in the raw, single-valued, forecasts and to produce ensemble forecasts that are reliable and no less skillful than resampled climatology.

In general, the patterns of skill and bias in the MEFP-GFS forecasts (Brown, 2013) are mirrored by the MEFP-GEFS forecasts. For example, the MEFP-GEFS forecasts show much higher correlations and greater skill in CNRFC than in AB- or CB-RFCs. This is associated with the greater predictability of large storms in the North Coast Ranges during the winter months. However, when compared to the MEFP-GFS forecasts, the seasonal variations in forecast skill are less pronounced in the MEFP-GEFS forecasts. For example, in CN- and MA-RFCs, the MEFP-GEFS precipitation forecasts show similar overall skill in the wet and dry seasons, while the MEFP-GFS forecasts are much more skillful during the wet season. In MARFC, the MEFP-GEFS precipitation forecasts are highly skillful at early forecast lead times, particularly at moderate precipitation amounts, but the forecast skill declines more rapidly when compared to CNRFC. In keeping with the MEFP-GFS precipitation forecasts, the MEFP-GEFS forecasts are consistently less skillful in AB- and CB-RFCs than MA- and CN-RFCs (at least during the first few days). This originates from a combination of reduced predictability in the southern plains and in the intermountain region of the western U.S., together with residual biases that were not removed by the MEFP. In general, both the MEFP-GEFS precipitation forecasts and the MEFP-GFS forecasts are unbiased and skillful during the first week, but show much lower skill and higher conditional biases during the second week.

Despite the broad similarities between the MEFP-GEFS forecasts and the MEFP-GFS forecasts, the MEFP-GEFS forecasts show higher correlations and greater skill than the MEFP-GFS forecasts. In AB-, CB- and MA-RFCs, the MEFP-GEFS precipitation forecasts show higher correlations than the MEFP-GFS forecasts at all forecast lead times. In CNRFC, the improvements from the MEFP-GEFS are greatest after ~three days, as the raw GFS forecasts show similar correlations during the first 1-2 days (~0.8). In keeping with the correlation results, the MEFP-GEFS precipitation forecasts are consistently more skillful than the MEFP-GFS forecasts in AB-, CB- and MA-RFCs. However, after seven days, the MEFP-GEFS forecasts are no more skillful than climatology and, in CB-DOLC2, the forecasts of light precipitation are somewhat less skillful than climatology. In principle, the MEFP-CLIM forecasts should not be less skillful than sample climatology. However, this may originate from differences in the climatology of the calibration and verification periods, as the latter comprises only a subset of years from the former. Overall, the MEFP-GEFS precipitation forecasts show the greatest benefits in MA-CNNN6, where they are 15-30% more skillful than climatology, while the MEFP-GFS forecasts are only 5-15% more skillful. As indicated above, at early forecast lead times, the MEFP-GEFS forecasts show similar skill to the MEFP-GFS forecasts in CNRFC. However, after seven days, the MEFP-GEFS forecasts are substantially more skillful than the MEFP-GFS forecasts, particularly at higher precipitation thresholds. At lower precipitation thresholds, the residual skill of the MEFP-GEFS forecasts mainly originates from improved resolution whereas, at higher thresholds, it mainly originates from improved reliability. However, reliable forecasts may be conditionally biased given the observed variable. Indeed, the MEFP-GEFS forecasts systematically underestimate the highest observed precipitation amounts (see below). While the MEFP-GEFS forecasts are more discriminatory than the sample climatology at all precipitation thresholds, they are generally most discriminatory at moderate and high thresholds. The greatest discrimination occurs in CN-FTSC1, where the atmospheric predictability is highest. The lowest discrimination occurs in CB-DOLC2, where the correlations and skill are also low. When expressed as a net gain in forecast lead time over the period of skillful forcing, the MEFP-GEFS precipitation forecasts typically add 1-2 days in forecast lead time when compared to the MEFP-GFS forecasts.

Both the MEFP-GEFS temperature forecasts and the MEFP-GFS forecasts improve substantially upon resampled climatology. They also remain skillful for longer than the precipitation forecasts. Indeed, the MEFP-GEFS temperature forecasts remain skillful throughout the second week. However, in keeping with the precipitation forecasts, the benefits of the MEFP-GEFS are generally more pronounced after the first 1-2 days. Thus, the errors saturate more quickly in the MEFP-GFS temperature forecasts than the MEFP-GEFS forecasts. For example, in the middle portion of the forecast horizon, the MEFP-GEFS forecasts show equivalent CRPSS to the MEFP-GFS forecasts, but with 2-4 days of additional forecast lead time. The greatest improvements occur in CN-FTSC1 and MA-CNNN6. In these basins, when measured against sample climatology, the MEFP-GEFS forecasts are ~20% more skillful than the MEFP-GFS forecasts across all temperature thresholds. In general, the improvements are smaller in AB-BLKO2 and CB-DOLC2. However, during the winter months (as well as the summer months in CB-DOLC2), the value added by the MEFP-GEFS increases when colder temperatures are included in the verification data. For example, during the winter months in CB-DOLC2, the CRPSS is ~0.6 when using the MEFP-GEFS to forecast temperatures above -8 °C (C_p=0.9) and ~0.4 when using the MEFP-GFS. While accurate forecasts of MAT are generally less important for hydrologic modeling than accurate forecasts of MAP, surface temperatures are important in determining the accumulation and melting of snow. Thus, in snow-dominated basins, such as CB-DRRC2 and CB-DOLC2, the additional skill of the MEFP-GEFS temperature forecasts may be important for hydrologic modeling.

As described by Hamill et al. (2013), there is a software bug in Version 9.0.1 of the GFS, whereby incorrect look-up tables are used in the land surface parameterization. This is known to introduce biases into the raw forecasts of near-surface temperatures. These biases were not apparent in the MEFP temperature forecasts. However, they pose a challenge for continuity of service with the HEFS, as the MEFP relies on operational forecasts whose statistical properties are reasonably stable and consistent with those of the hindcasts.

Both the MEFP-GEFS forecast and the MEFP-GFS forecasts comprise a range of conditional biases. In particular, there is a tendency for the precipitation forecasts to underestimate the Probability of Precipitation (PoP). This lack of reliability also effects the MEFP-CLIM forecast. Indeed, the MEFP forecasts of PoP are substantially worse than unconditional climatology in some basins. In order to produce reliable forecasts of PoP at a daily accumulation, the forecasts must be reliable at a six-hourly accumulation. Furthermore, they must adequately capture the statistical dependencies between the six-hourly accumulations. In practice, the forecasts of PoP are unreliable at a daily accumulation, while the corresponding six-hourly forecasts are reliable, on average. This alludes to a problem with the modeling of precipitation intermittency at a six-hourly accumulation. More specifically, is alludes to a problem with the temporal variability of precipitation intermittency in the MEFP forecasts. The space-time covariability of precipitation and temperature is modeled with the Schaake Shuffle (Clark et al., 2004). The Schaake Shuffle reproduces the historical space-time covariability of the observed MAT and MAP conditionally upon forecast valid date. It does not model these patterns conditionally upon the state of the atmosphere. Thus, it cannot account for any differences in covariability under dry versus wet conditions (beyond the unconditional probability of matching those conditions). Further investigation is warranted into the limitations of the Schaake Shuffle, particularly for extreme events, and whether other empirical structures, such as high-resolution forecasts or conditional climatologies, can better reproduce the space-time covariability of precipitation and temperature (e.g. Shefzik et al., 2013).

Alongside the underestimation of PoP, the precipitation forecasts are conditionally biased with increasing amounts of observed precipitation. This originates from a Type-II conditional bias in the MEFP precipitation forecasts. Again, it is apparent in the MEFP-GEFS precipitation forecasts, as well as the MEFP-GFS forecasts, and resampled climatology. While climatology is, by definition, perfectly reliable, it is conditionally biased given the observed variable. For this reason, the Type-II conditional bias increases as the forecast skill declines; hence, it varies with location, season and forecast lead time, among other factors. At early forecast lead times in CNRFC, the biases are sufficiently small, and the spread is sufficiently large, that the highest precipitation totals are generally forecast with some, non-zero, probability of occurrence. However, in other basins, and at longer forecast lead times, the largest precipitation totals are routinely underestimated by as much as the observed precipitation amount. While the MEFP-GEFS forecasts show similar conditional biases to the MEFP-GFS forecasts, they also comprise more spread in some cases. For example, in AB-BLKO2, the MEFP-GEFS forecasts are more likely to warn of the highest observed precipitation amounts, even if their central tendency is to underestimate.

Currently, the MEFP is calibrated with the ensemble mean of the raw GEFS forecasts (Appendix A). Most of the skill in the frozen GFS is concentrated in the ensemble mean forecast (Wilks and Hamill, 2007; Wu et al., 2011). However, as atmospheric models and EPS become more skillful, post-processors may benefit from using higher moments, interactions, or even the individual ensemble members. Thus, future work should consider, first, whether the GEFS contains useful information beyond the ensemble mean and, second, how best to leverage this information, while maintaining a parsimonious description of the forecast errors. More generally, further work is needed on the limitations of statistical post-processing for large and extreme events. Here, the desire for unbiasedness must be weighed against the risk of obfuscating a weak, but potentially valuable, signal in the raw forecasts. The ability to calibrate the MEFP with reasonably small sampling uncertainty is important in this context. Thus, future work should leverage all of the available GEFS reforecasts and corresponding operational forecasts.

In order to understand the benefits of the MEFP-GEFS forcing independently of any hydrologic biases, the raw streamflow forecasts were verified against simulated flows. In general, both the MEFP-GEFS streamflow forecasts and the MEFP-GFS forecasts are substantially more skillful than those with climatological forcing. Similarly, when compared to the MEFP-GFS streamflow forecasts, the MEFP-GEFS forecasts are consistently more correlated with the simulated streamflows and show higher CRPSS. At a forecast lead time of one day, the correlations are similar for all sources of forcing and the CRPSS is lower. This is understandable, because the streamflow forecasts with MEFP-CLIM forcing comprise the same initial conditions as those with MEFP-GEFS and MEFP-GFS forcing. As the hydrologic models respond unevenly to meteorological forcing, depending on basin characteristics and antecedent conditions, the period over which the MEFP-GEFS forecasts improve upon the MEFP-GFS forecasts varies between basins. For example, in AB-CBNK1 and AB-BLKO2, the streamflow forecasts show a rapid decline in correlation with increasing forecast lead time. This originates from a lack of hydrologic persistence in ABRFC and the difficulty in forecasting precipitation beyond the short-range, but particularly heavy precipitation. Indeed, while the MEFP-GEFS precipitation forecasts are more skillful than the equivalent MEFP-GFS forecasts, they are no more skillful than the MEFP-CLIM forecasts after ~one week. In contrast, the basins in CBRFC are dominated by snow accumulation and melting. Here, much of the skill in the streamflow forecasts depends on the hydrologic uncertainties, specifically on the initial conditions in the hydrologic models. However, the timing and rate of snowmelt also depends on the accuracy of the temperature forecasts during the snowmelt period. In CB-DRRC2 and CB-DOLC2, the highest streamflows occur during the snowmelt period, and integral measures of error and skill, such as the CRPSS, are sensitive to these flows. When verifying the raw streamflow forecasts against simulated flows, the MEFP-GEFS forecasts are substantially more skillful than the equivalent MEFP-GFS forecasts. For example, in CB-DOLC2, the MEFP-GEFS forecasts contribute five or more days of additional forecast lead time in the medium-range alone. These improvements are greatest during the snowmelt period and originate from the increased accuracy of the MEFP-GEFS temperature forecasts, as the MEFP-GEFS precipitation forecasts are unskillful beyond ~3-6 days. Substantial improvements are also seen in MA-WALN6 and MA-CNNN6, where the MEFP-GEFS streamflow forecasts contribute 2-4 days of additional forecast lead time. In general, however, the benefits of the MEFP-GEFS are less pronounced when verifying the bias-corrected streamflow forecasts against observed flows (see below).

In AB-, CN- and MA-RFCs, the raw streamflow forecasts are conditionally biased with increasing rates of simulated flow. These biases originate from a similar conditional bias in the MEFP precipitation forecasts (to which the hydrologic models are more sensitive in AB-, CN- and MA-RFCs). By definition, climatology is conditionally biased for large values of the observed (and simulated) variable. Thus, the degree of conditional bias increases as the forecast skill declines. For example, in ABRFC, the conditional bias increases rapidly over the first week, as the forecasts show little skill beyond one week. In CNRFC, the conditional biases increase throughout the medium-range, as the forecasts remain skillful during the middle portion of the forecast horizon. In CBRFC, the streamflow forecasts are conditionally *unbiased* for most streamflow rates. This stems from the importance of snowmelt in generating large streamflows in these basins. Indeed, while the precipitation forecasts show large conditional biases in CB-DRRC2 and CB-DOLC2, snow accumulation is a time integral over the accumulation period. This implies a weaker dependence of high streamflows on heavy precipitation and the conditional biases therein. In practice, while the MEFP-GEFS streamflow forecasts are consistently more skillful than the MEFP-GFS forecasts in all basins, both the MEFP-GEFS forecasts and the MEFP-GFS forecasts show large Type-II conditional biases. Nevertheless, in some basins (notably AB-BLKO2), the MEFP-GEFS forecasts partially compensate for the tendency to underestimate the highest flows with an increased spread and, thus, an increased chance of warning about the highest flows.

The overall skill of the post-processed streamflow forecasts, as well as the relative contributions from the MEFP and the EnsPost, vary with basin, season, and forecast lead time. They also vary with the source of forcing used in the MEFP. In general, the post-processed MEFP-GEFS forecasts are substantially more skillful than the raw MEFP-CLIM forecasts. The overall skill is greatest in CBRFC and CNRFC, where the seasonal differences are also greatest. For example, during the wet season in CN-DOSC1, the post-processed MEFP-GEFS forecasts are up to ~40% more skillful than the raw MEFP-CLIM forecasts. In the dry season, the overall skill increases to ~60% at the earliest forecast lead times. At low flows, a greater fraction of the total skill originates from streamflow post-processing, as the EnsPost benefits from hydrologic persistence. Also, in basins with a pronounced dry season, the meteorological forcing is more predictable during the summer months. For these reasons, the MEFP-GEFS forecasts do not substantially improve upon the MEFP-GFS forecasts at low flows.

At moderate and higher streamflow thresholds, a greater fraction of the total skill in the post-processed streamflow forecasts originates from the MEFP. Thus, at higher flows, the MEFP-GEFS forecasts generally improve upon the MEFP-GFS forecasts, particularly in CBRFC. During the wet season in CB- and CN-RFCs, and throughout the year in MA- and AB-RFCs, the MEFP-GEFS forecasts typically show similar skill to the MEFP-GFS forecasts for 1-2 days longer. For example, in CN-FTSC1, the MEFP-GEFS forecasts can detect streamflows above C_p=0.1 with equivalent skill to the MEFP-GFS forecasts, but with an additional forecast lead time of ~2.5 days. However, when verifying the post-processed streamflow forecasts (against observed flows), the gains implied by the raw forecasts (against simulated flows) are not always realized by the EnsPost, particularly at high streamflow thresholds. Indeed, at early forecast lead times in AB- and MA-RFCs, and later forecast lead times in CB- and MA-RFCs, forecasts of moderate and high flows (C_p<0.1) show a decline in CRPSS following streamflow post-processing. This may originate from a lack of stationarity in the hydrologic biases. For example, in CB-DOLC2, the hydrologic biases vary substantially between years, particularly during the snowmelt period (Appendix C). In some years, there are large discrepancies between the observed and simulated flows (e.g. 1986; Figure C05) while, in other years, the simulated flows closely match the observed flows (e.g. 1998; Figure C08). In practice, the hydrologic biases are often manifest as timing errors in the simulated flows, yet the EnsPost can only model these indirectly, as magnitude errors. In order to account for inter- and intra-annual variations in basin conditions, operational forecasters typically modify ("mod") some combination of the inputs, parameters and states of the hydrologic models. However, adjusted simulations are not consistently archived by the RFCs. Also, they cannot be used operationally when the EnsPost is calibrated with raw simulations. Data assimilation is the preferred approach to adjusting model states (Liu et al., 2012). In principle, automated data assimilation would avoid these inconsistencies between the calibration and operational use of the EnsPost. It would also reduce the hydrologic biases in the raw simulations and account for non-stationarities in the hydrologic errors.

In order to evaluate the quality of the HEFS and to establish a baseline for future enhancements, more comprehensive hindcasting and verification is needed. This should be conducted by all RFCs, for a range of forcing inputs, and for a broader range of river basins, including regulated rivers and outlets. Further work is needed to compare the streamflow forecasts from the HEFS against the RFC operational forecasts. While such comparisons are not straightforward (e.g. because the raw forcing data used by the HEFS is not used for operational forecasting), they are necessary to benchmark the HEFS and to show that, overall, the forecasts improve on existing products. In addition, there is a need to evaluate decision support systems and other applications that rely on the HEFS, such as water quality, ecology, river navigation, and water supply. Such applications will show varying sensitivities to the HEFS forecasts and are necessary to demonstrate the wider, societal and economic, benefits of the HEFS and of ensemble forecasting more generally. In this context, there is a need for interdisciplinary and interagency collaborations on uncertainty and risk, as hydrologic forecasts are only one input to environmental decision making, and not necessarily the most important one.

7. Glossary of terms and acronyms

ADJUST-Q – A procedure implemented within the CHPS to “blend” an operational streamflow forecast with the most recent streamflow observation. A rudimentary form of Data Assimilation that relies on hydrologic persistence

Aggregation and Disaggregation – forming larger or smaller control volumes, respectively

Bias – A systematic difference between an estimate of some quantity and its “true” (generally meaning observed) value

BS – Brier Score. The average squared deviation between the predicted probabilities that a discrete event occurs (such as flooding) and the corresponding observed outcome (0 or 1)

BSS – Brier Skill Score. The fractional reduction in the BS of one forecasting system relative to another. A value of 1 denotes perfect skill, 0 indicates that the forecasting systems are equivalent, and a negative value denotes a loss of skill

Calibration – A process of estimating model parameters based on observations and corresponding (raw) predictions. In post-processing and verification, calibration has a second meaning, namely to correct for biases in ensemble forecasts by increasing their reliability. See Calibration-refinement

Calibration-refinement – One factorization of the joint probability distribution of the forecasts and observations, obtained by conditioning on the forecast variable. Calibration is also known as reliability or Type-I conditional bias. See Likelihood-base-rate

Canonical Event – a partitioning of time scales in order to account for the varying information content of the different forcing inputs to MEFP (e.g., RFC QPF/QTF, GFS, and CFSv2)

CFSv2 – Climate Forecast System. A fully coupled model representing the interaction between the Earth's oceans, land and atmosphere that generates forecasts from 1-270 days. See also: http://cfs.ncep.noaa.gov/

CHPS – The Community Hydrologic Prediction System (pronounced “chips”)

Climatology – The science that deals with average weather conditions over long periods. Climatology also refers the historical record of observations (e.g. mean areal averages of actual temperature and precipitation) used to drive a model

Conditional bias – A bias in the forecasts over a subsample of the verification pairs. The subsample may originate from the application of one or more conditions to the paired data, such as observed values that exceed a given threshold. See Bias

Continuous API – Continuous Antecedent Precipitation Index. An empirical hydrologic model used by the Middle Atlantic RFC

Correlation coefficient – Pearson product-moment correlation coefficient. The covariance of two variables divided by the product of their standard deviations. A degree of linear association between two variables, with -1 and 1 denoting perfect negative and positive association, respectively, and 0 denoting the absence of a linear association (but not necessarily a non-linear association)

CRPS – Continuous ranked probability score. The integral square difference between a forecast probability distribution and the observed outcome. It is typically averaged over many such cases (known as the “mean CRPS”)

CRPSS – The continuous ranked probability skill score. The fractional reduction in CRPS of one forecasting system when compared to another (the reference or baseline). A value of 1 denotes perfect skill, 0 indicates that the forecasting systems are equivalent, and a negative value denotes a reduction in skill

DA – Data Assimilation. A procedure for updating model states (and possibly other variables) with recent observations, thereby improving forecasts

Disaggregation – (see aggregation/disaggregation)

Discrimination – Discrimination is an attribute of forecast quality that measures the sensitivity of the forecast probabilities to different observed outcomes. A forecasting system is discriminatory if its forecast probabilities vary for different observed outcomes. Discrimination is insensitive to conditional bias, i.e. a forecasting system may be discriminatory but have large Type-II conditional biases. A component of the Likelihood-base-rate factorization

Ensemble Forecast – A collection of equally likely predictions of the future states of the atmosphere or hydrologic system, based on sampling of the different sources of uncertainty and propagating them through a modeling system (such as CHPS). An "ensemble trace" comprises two or more forecast lead times

EnsPost – Ensemble Post-processor. A software tool and a statistical technique that accounts for hydrologic uncertainties and biases separately from the forcing uncertainties and biases

ESP – Ensemble Streamflow Prediction. In NWS operations, this has the specific meaning of forcing the NWS River Forecast System with a sample of observations from the same dates in previous years, i.e. climatological forcing. Some RFCs have augmented the original ESP algorithms to account for additional information

EVS – Ensemble Verification System. A software tool for verifying ensemble forecasts

Forcings – The model inputs (e.g., precipitation and temperature) that drive or "force" a hydrologic model

Forecast Issue Time – The date/time at which a forecast is issued, also known as "T0." This differs from the Forecast Valid Time

Forecast Lead time – The difference between the Forecast Valid Time and the Forecast Issue Time

Forecast Valid Time – The time at which a forecast is valid

GEFS - Global Ensemble Forecast system – An ensemble forecasting system that uses an enhanced version of the GFS

GFS – Global Forecast System. An operational NWP model developed by NCEP. The operational GFS is run four times daily, with forecasts out to 384 hours. The GFS was also "frozen" in 1997 (the "frozen GFS") and used to generate hindcasts beginning in 1979, which are used to calibrate the MEFP. The frozen GFS is a legacy model and operational forecasts will end in 2013. See GEFS also

HEFS – Hydrologic Ensemble Forecast Service. Also, HEFSv1, the first version of the HEFS

HEP – Hydrologic Ensemble Processor. A component of the HEFS implemented within the CHPS. The HEPS integrates a finite number of "equally likely" traces of precipitation and temperature through the NWS hydrologic models

HEPS – Hydrologic Ensemble Prediction System. The general approach of which the HEFS is one example

Hindcast – A retrospective forecast or reforecast. A forecast begins on each of several historical days. Reforecast is a term frequently used for weather models

Lag/K – A simple technique for routing an inflow hydrograph downstream, originally developed as a graphical routing procedure. The outflow hydrograph comprises one or both of a time lag and attenuation (K) of the input hydrograph

Likelihood-base-rate – The second of two factorizations of the joint probability distribution of the forecasts and observations, obtained by conditioning on the observed variable. See Calibration-refinement

Long-range – The latter portion of the forecast time horizon, generally interpreted as more than ~14 days, where the forecast skill is lowest. See short-range and medium-range also.

MAP – Mean Areal Precipitation over a basin/watershed

MAT – Mean Areal Temperature over a basin/watershed

Medium-range – The middle portion of the forecast time horizon, generally interpreted as ~5-14 days. See short-range and long-range also

MEFP – Meteorological Ensemble Forecast Processor. A software tool and statistical technique that produces ensemble forecasts of temperature and precipitation using (single-valued) operational forecasts from NWP models. The forecast spread is derived from historical information about forecast errors

MOS – Model Output Statistics. A statistical technique for bias-correcting weather and water forecasts (e.g. Hydrologic MOS or HMOS)

NQT – Normal Quantile Transform. A transformation made to a data sample so that it follows a normal probability distribution (i.e. so that the histogram of values would appear normal)

NWP – Numerical Weather Prediction

NWSRFS – National Weather Service River Forecast System. Replaced by CHPS

NYCDEP – New York City Department of Environmental Protection

PoD – Probability of Detection. The probability that a discrete event is detected by an ensemble forecasting system. An event is detected when the forecast probability exceeds a pre-defined threshold and the event occurs. In general, a high threshold will reduce the PoFD, but may also reduce the PoD. Hence, the PoD and PoFD are typically compared in a ROC diagram

PoFD – Probability of False Detection. The probability that a discrete event is incorrectly detected by an ensemble forecasting system. An event is incorrectly detected when the forecast probability exceeds a pre-defined threshold and the event does not occur. In general, a low threshold will increase the PoD, but may also increase the PoFD. Hence, the PoD and PoFD are typically compared in a ROC diagram

PoP – Probability of precipitation. The probability that a non-zero precipitation amount will occur

Probit – A non-linear plotting scale. The probit function is the quantile function (inverse of the cumulative distribution function) associated with the standard normal probability distribution

Reforecast – See Hindcast. Commonly used in the atmospheric sciences

Reliability (Type-I conditional bias or calibration) – A flood forecasting system is "reliable" if flooding occurs with the same relative frequency as the forecast probabilities imply. For example, flooding should occur 20% of the time when the forecast probability is 0.2. An attribute of forecast quality and a component of the Calibration-refinement factorization

Resampled climatology – A procedure for generating an ensemble of precipitation and temperature forecasts from the MEFP using historical observations. The observations are resampled in a moving window either side of the forecast valid date across all historical years. A smooth probability distribution is then fitted to the resampled observations and ensemble members are derived from the fitted distribution. See sample climatology also

Resolution – Should not be confused with spatial or temporal resolution. Resolution is an attribute of forecast quality that measures the sensitivity of the observed outcomes to differences in the forecast probabilities of those outcomes. Resolution is insensitive to conditional bias, i.e. a forecasting system may be resolved but unreliable. A component of the Calibration-refinement factorization

RME – Relative Mean Error. The average fractional bias of the ensemble mean forecast or the mean error of the ensemble mean, divided by the mean observed value. Positive, zero, and negative values denote a positive, zero, and negative bias, respectively

ROC – The Relative Operating Characteristic. Measures the ability of a forecasting system to correctly predict (or "discriminate") the occurrence of an event (PoD) while avoiding too many incorrect forecasts when it does not occur (PoFD)

SAC-SMA – The Sacramento Soil Moisture Accounting Model. A conceptual hydrologic model used in CHPS

Sample climatology – an unconditional probability distribution, comprising all historical observations at a given space-time scale within a period of interest (e.g. 1985-1999). Unlike resampled climatology, the observations are not sub-sampled within a local (e.g. seasonal) window or smoothed by fitting a parametric distribution to the observations. In general, for variables that show a strong seasonality, resampled climatology is much more skillful than sample climatology

Sharpness – Sharpness is an attribute of the forecast variable used in verifying ensemble forecasts. Specifically, it refers to the variability (e.g. measured by the variance) of the forecast probabilities. Sharpness may be considered desirable insofar as decisions may be hampered if a forecast lacks sharpness (i.e. comprises a larger range of possibilities), but sharpness is not desirable at the expense of other attributes of forecast quality, such as reliability. A component of the Likelihood-base-rate factorization

Short-range – The early part of the forecast time horizon, generally interpreted as ~1-5 days or less, where the forecast skill is highest. See medium-range and long-range also

Simulation – A hydrologic prediction based on observed temperature and precipitation (as distinct from a forecast, which comprises forecast inputs)

Skill – The fractional improvement of one forecasting system relative to a baseline. The measure used for skill could vary (e.g. the Brier Skill Score uses the Brier Score).

SNOW-17 – Snow Accumulation and Ablation Model 17. A conceptual hydrologic model for snow processes, incorporated in the CHPS

SREF – Short-Range Ensemble Forecast (SREF) system. An NCEP model that issues short-range ensemble forecasts

Support – Synonymous with scale. The temporal or spatial control volume.

T0 – Forecast issue (System/Basis) Time. The time at which a forecast is produced

Type-II conditional bias – A bias in the ensemble forecasts when viewed conditionally upon the observed variable. For example, a bias in the forecast ensemble mean when the observations exceed a given threshold. An attribute of forecast quality and a component of the Likelihood-base-rate factorization

Uncertainty – An attribute of the Calibration-refinement factorization, not to be confused with the more general concept of “uncertainty.” Specifically, it refers to the variability (e.g. measured by the variance) of the observations

UTC – Coordinated Universal Time, also known as Zulu (Z) time and synonymous with Greenwich Mean Time (GMT). Forecasts from the HEFSv1 are issued daily at 12Z

WPC – Weather Prediction Center, formerly the Hydrometeorological Prediction Center

XEFS – Experimental Ensemble Forecast System. The experimental precursor to the HEFS

8. References

Anderson, E.A. 1973. National Weather Service River Forecast System-Snow Accumulation and Ablation Model, NOAA Technical Memorandum: NWS Hydro-17, U.S. National Weather Service.

Beven, K.J. 2000. On model uncertainty, risk and decision making. *Hydrological Processes* **14**, 2605-2606.

Beven, K.J. 2008: *Environmental Modelling: An Uncertain Future?* Routledge: London, 328 pp.

Bogner, K., and Pappenberger, F. 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resources Research* **47**, W07524. doi:10.1029/2010WR009137.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society* **91**, 1059–1072. doi: http://dx.doi.org/10.1175/2010BAMS2853.1

Bradley, A.A., Schwartz, S.S. and Hashino, T. 2004. Distributions-oriented verification of ensemble streamflow predictions. *Journal of Hydrometeorology* **5**(3), 532-545.

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

Brown, J. D., Demargne, J., Seo, D-J, and Liu, Y. 2010b. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling and Softw*are **25**, 854-872.

Brown, J.D. 2010a. Prospects for the open treatment of uncertainty in environmental research. *Progress in Physical Geography* **34**, 75-100, doi:10.1177/0309133309357000.

Brown, J.D, Seo, D.-J. and Du, J. 2012. Verification of precipitation forecasts from NCEP's Short Range Ensemble Forecast (SREF) system with reference to ensemble streamflow prediction using lumped hydrologic models. *Journal of Hydrometeorogy*, **13**(3), 808-836.

Brown, J.D. 2013. *Verification of temperature, precipitation and streamflow forecasts from the NWS Hydrologic Ensemble Forecast Service (HEFS): medium-range forecasts* *with forcing inputs from the frozen version of NCEP's Global Forecast System*. Technical Report prepared by Hydrologic Solutions Limited for the U.S. National Weather Service, Office of Hydrologic Development [Available at: http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/Contract_2012-04-HEFS_Deliverable_02_Phase_I_report_FINAL.pdf, accessed 11th September 2013], 133pp.

Brown, J.D. and Heuvelink, G. 2005. Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In: Anderson, M. (Ed.) *The Encyclopedia of Hydrological Sciences*, Chichester: John Wiley and Sons, 1181–1195.

Brown, J.D., and 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, doi: 10.1002/hyp.9263.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M. 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review* **133**, 1076–1097. doi: http://dx.doi.org/10.1175/MWR2905.1

Burnash, R.J.C. 1995. The NWS river forecast system—catchment modeling. In: Singh, V.P. (Ed.), *Computer Models of Watershed Hydrology*. Water Resources Publications, Littleton, Colorado, 311–366.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. 2004. The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields. *Journal of Hydrometeorology* **5**, 243–262.

Cloke, H.L., Pappenberger, F., van Andel, S-J, Schaake, J., Thielen, J., Ramos, M-H 2013. Hydrological ensemble prediction systems. *Hydrological Processes* **27**, 1–4. doi: 10.1002/hyp.9679.

Demargne, J., Brown, J. D., Liu, Y., Seo, D-J, Wu, L., Toth, Z., and Zhu, Y. 2010. Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters* **11**(2), 114-122.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D-J., Hartman, R., Herr, H.D. Fresch, M., Schaake, J. and Zhu, Y. 2014. The science of NOAA's operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society*, in press.

Demeritt, D., Nobert, S., Cloke, H. L. and Pappenberger, F. 2013. The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrological Processes* **27**, 147–157. doi: 10.1002/hyp.9419.

Du, J., DiMego, G., Toth, Z., Jovic, D., Zhou, B., Zhu, J., Chuang, H., Wang, J., Juang, H., Rogers, E. and Lin, Y. 2009. NCEP short-range ensemble forecast (SREF) system upgrade in 2009. *19th Conference on Numerical Weather Prediction and 23rd Conference on Weather Analysis and Forecasting*, *Omaha, Nebraska*, American Meteorological Society. [Available at: http://www.emc.ncep.noaa.gov/mmb/SREF/reference.html, accessed 12/06/13]

Filar, J.A. and Haurie, A. (eds) 2010. *Uncertainty and environmental Decision Making: A Handbook of Research and Best Practice*. International Series in Operations Research and Management Science. Springer. 338 pp.

Glahn, H. and Lowry, D. 1972. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology* **11**(8), 1203-1211.

Gneiting, T. and Raftery, A.E. 2005. Weather forecasting with ensemble methods. *Science* **310**(5746), 248-249.

Gneiting, T., Balabdaoui, F., and Raftery, A.E. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**(2), 243–268.

Green, D.M., and Swets, J.M. 1966. *Signal detection theory and psychophysics.* John Wiley and Sons: New York, 455pp.

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

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M. and Palmer, T. N. 2012. Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 138: 1814–1827. doi: 10.1002/qj.1895.

Hamill, T.M., Whitaker, J. S., Fiorino, M. and Benjamin, S. G. 2011. Global ensemble predictions of 2009's tropical cyclones initialized with an ensemble Kalman filter. *Monthly Weather Review* **139**, 668–688.

Hamill, T.M., Bates, G.T., Whitaker, J. S., Murray, D.R., Fiorino, M., Galarneau Jr., T., Zhu, Y., and Lapenta, W. 2013. NOAA's second-generation global medium-range ensemble reforecast data set. *Bulletin of the American Meteorological Society*, in press.

Hamill, T.M., Whitaker, J. S. and Mullen, S. L. 2006. Reforecasts: an important data set for improving weather predictions. *Bulletin of the American Meteorological Society* **87**(1), 33-46.

Hamlet, A. F., Huppert, D. and Lettenmaier, D. P. 2002. Economic value of long-lead streamflow forecasts for Columbia River hydropower. *Journal of Water Resources Planning and Management* **128**, 91-101.

Handmer, J., Norton, T. and Dovers, S. (eds) 2001. *Uncertainty, Ecology and Policy: Managing Ecosystems for Sustainability*. Prentice-Hall: Harlow.

Hanley, J. 1988. The robustness of the ''binormal'' assumptions used in fitting ROC curves. *Medical Decision Making* **8**, 197–203.

Helton, J.C., Johnson, J.D., Salaberry, C.J. and Storlie, C.B. 2006. Survey of sampling based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety* **91**, 1175–1209.

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

Jakeman, A.J., Letcher, R.A. and Norton, J.P. 2006. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software* **21**, 602-614.

Jolliffe, I.T., and Stephenson, D.B. (eds). 2011. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley and Sons: Chichester.

Kahnemann, D., Slovic, P. and Tversky, A. 1982: *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press. 544 pp.

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

Kelly, K.S., and Krzysztofowicz, R. 1997. A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrology and Hydraulics* **11**, 17–31.

Kennedy, E.J. 1983. *Techniques of Water-Resources Investigations of the United States Geological Survey, Book 3. Chapter A13: Computation of Continuous Records of Streamflow*. US Government Printing Office, 52pp. [Available at: http://pubs.usgs.gov/twri/twri3-a13/pdf/TWRI_3-A13.pdf, accessed 02/01/13].

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P. 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences* **16**, 3863–3887.

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

Marsigli, C. Montani, A. and Paccagnella, T. 2013. Perturbation of initial and boundary conditions for a limited-area ensemble: multi-model versus single-model approach. *Quarterly Journal of the Royal Meteorological Society*, in press.

Matott, L.S., Babendreier, J.E., and Parucker, S.T. 2009. Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resources Research* **45**, WO6421, doi:10.1029/2008WR007301.

Metz, C. E., and Pan, X. 1999. ‘‘Proper’’ binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology* **43**, 1–33.

Montanari, A., and Grossi, G. 2008. Estimating the uncertainty of hydrological forecasts: A statistical approach. *Water Resources Research* **44**, W00B08, doi:10.1029/2008WR006897.

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

Pappenberger, F., Beven, K. J., Hunter, N., Bates, P. D., Couweleeuw, B. T., Thielen J. and de Roo, A. P. J. 2005. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Sciences* **9**, 381-393.

Pappenberger, F., and Buizza, R. 2009. The skill of ECMWF precipitation and temperature predictions in the Danube basin as forcings of hydrological models. Weather and Forecasting **24**, 749–766.

Park, S.K. and Xu, L. 2009. *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*. Springer-Verlag: Berlin.

Philpott, A.W., Wnek, P. and Brown, J.D. 2012. Verification of ensembles at the Middle Atlantic River Forecast Center. *92nd American Meteorological Society Annual Meeting*, January 22-26, 2012, New Orleans, LA [Available at: https://ams.confex.com/ams/92Annual/webprogram/Paper199532.html, accessed 02/02/13].

Raff, D., Brekke, L., Werner, K., Wood, A. and White, K. 2013. *Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information*. A report of the U.S. Army Corps of Engineers (USACE), Bureau of Reclamation, and the National Oceanic and Atmospheric Administration (NOAA), CWTS 2013-1 [Available at http://www.ccawwg.us/docs/Short-Term_Water_Management_Decisions_Final_3_Jan_2013.pdf, accessed 04/04/13].

Ramos, M. H., van Andel, S. J., and Pappenberger, F. 2012. Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences Discussions* **9**, 13569-13607, doi:10.5194/hessd-9-13569-2012.

Regonda, S.K., Seo, D-J., Lawrence, B., Brown, J.D., and Demargne, J. 2013. Short-term ensemble streamflow forecasting using operationally produced single-valued streamflow forecasts – A Hydrologic Model Output Statistics (HMOS) approach. *Journal of Hydrology* **497**, 80-96.

Schaake, J., Franz, K., Bradley, A., and Buizza, R. 2006. The Hydrologic Ensemble Prediction EXperiment (HEPEX). *Hydrology and Earth Systems Sciences Discussion* **3**, 3321–3332.

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X. and Seo, D.J. 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth Systems Sciences* **4**, 655-717.

Schefzik, R., Thorarinsdottir, T. L. and Gneiting, T. 2013. Uncertainty quantification in complex simulation models using ensemble copula coupling. Preprint, arXiv:1302.7149v1.

Schellekens, J., Weerts, A.H., Moore, R.J., Pierce, C.E., and 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.

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

Seo, D-J., Demargne, J., Wu, L., Liu, Y., Brown, J. D., Regonda, S. and Lee, H. 2010. *Hydrologic Ensemble Prediction for Risk-Based Water Resources Management and Hazard Mitigation*. 4th Federal Interagency Hydrologic Modeling Conference, June 27-July 1, 2010, Las Vegas, NV.

Sittner, W., Schauss, C., and Monro, J. 1969. Continuous Hydrograph Synthesis with an API-Type Hydrologic Model. *Water Resources Research*, **5**(5), 1007-1022.

Smith, B.L., Yuter, S.E., Neiman, P.J. and Kingsmill, D.E. 2010. Water Vapor Fluxes and Orographic Precipitation over Northern California Associated with a Landfalling Atmospheric River. *Monthly Weather Review* **138**, 74–100. doi: http://dx.doi.org/10.1175/2009MWR2939.1

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

van Andel, S. J., Weerts, A., Schaake, J. and Bogner, K. 2013. Post-processing hydrological ensemble predictions intercomparison experiment. Hydrological Processes **27**, 158–161. doi: 10.1002/hyp.9595.

Warner, T.T. 2010. *Numerical Weather and Climate Prediction*. Cambridge University Press: Cambridge. 548pp.

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

Wilks, D.S. 2006. *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Elsevier: San Diego.

Wilks, D. S. and Hamill, T.M. 2007. Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review* **135**, 2379–2390. doi: http://dx.doi.org/10.1175/MWR3402.1.

Wood, A.W. and Schaake, J.C. 2008. Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology* **9**, 132-148.

Wu, L., Seo, D.-J., Demargne, J., Brown, J.D., Cong, S. and Schaake, J. 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast via meta-Gaussian distribution models. *Journal of Hydr*ology, **399**(3-4), 281-298.

9. Tables

Table 1: characteristics of the study basins

Characteristic	ABRFC		CBRFC		CNRFC		MARFC	
	CBNK1	BLKO2	DRRC2	DOLC2	DOSC1	FTSC1	WALN6	CNNN6
Lat (outlet)	37.1292	36.8086	37.6389	37.4725	39.71	40.22	42.1661	42.0628
Long (outlet)	-97.6017	-97.2775	-108.06	-108.497	-123.32	-123.63	-75.1403	-75.3747
Lat (GFS)	37.5	37.5	37.5	37.5	40.0	40.0	42.0	42.0
Long (GFS)	-97.5	-97.5	-107.5	-107.5	-122.5	-122.5	-75.0	-75.0
Lat (GEFS, week 1)	37.2171	36.7490	37.6853	37.6853	39.5578	40.026	42.3667	41.8985
Long (GEFS, week 1)	-97.5	-97.5	-108.2812	-108.2812	-123.2812	-123.75	-75.0	-75.0
Lat (GEFS, week 2)	37.123	36.4991	37.7469	37.7469	39.6186	40.2426	42.1143	42.1143
Long (GEFS, week 2)	-97.5	-97.5	-108.125	-108.75	-123.125	-123.75	-75.0	-75.0
Area (total, km^2)	2057	4815	275	1305	1930	5457	860	1175
Mean elev. (m)	115	140	2567	2115	340	247	180	157
Annual P (mm)	935.68	1017.4	961.94	805.95	1682.36	1438.92	1038.27	1053.22
C_p[P]=0.1 (mm)	7.96	8.61	7.38	5.74	13.87	13.42	9.09	9.15
C_p[P]=0.05 (mm)	16.33	17.25	12.37	9.4	25.58	25.17	14.18	14.42
C_p[P]=0.01 (mm)	37.12	41.23	24.79	19.73	54.33	51.74	29.97	29.12
Runoff coefficient	0.12	0.14	0.45	0.42	0.42	0.53	0.57	0.58
P/PE	0.74	0.78	0.93	0.78	1.92	2.17	1.49	1.51
Q_{action} (mm/d)	3.403	7.789	N/A	9.872	N/A	N/A	8.763	N/A
C_p[Q>Q_{action}]	0.0117	0.00602	N/A	0.00073	N/A	N/A	~0	N/A
Q_{flood} (mm/d)	5.924	10.585	N/A	14.789	N/A	N/A	17.612	N/A
C_p[Q>Q_{flood}]	0.00484	0.00315	N/A	~0	N/A	N/A	~0	N/A
C_p[Q]=0.1 (mm/d)	0.031	0.024	0.133	0.094	0.017	0.015	0.15	0.106
C_p[Q]=0.75 (mm/d)	0.255	0.224	1.248	0.916	2.182	1.842	1.959	1.936
C_p[Q]=0.9 (mm/d)	0.554	0.615	3.946	2.92	4.87	5.537	3.716	3.654

P = precipitation
C_p = climatological probability
PE = potential evaporation
Q = streamflow
Q_{action} = action stage in millimeters per day (mm/d)
Q_{flood} = flood stage in mm/d

Table 2: period of record used to calibrate the HEFS

RFC/Basin	MEFP-CLIM	MEFP-GFS		MEFP-GEFS		EnsPost
		Bivariate[1]	Shuffle[2]	Bivariate	Shuffle	
AB-BLKO2	1979-1999	1979-1999	1951-1999	1985-1999	1951-1999	1979-1999
AB-CBNK1	1979-1999	1979-1999	1951-1999	1985-1999	1951-1999	1979-1999
CB-DRRC2	1979-2003	1979-2005	1961-2003	1985-2005	1961-2003	1979-1999
CB-DOLC2	1979-2003	1979-2005	1961-2003	1985-2005	1961-2003	1979-1999
CN-DOSC1	1979-1998	1979-2005	1961-1998	1985-2009	1961-1998	1979-1999
CN-FTSC1	1979-1998	1979-2005	1961-1998	1985-2009	1961-1998	1979-1999
MA-WALN6	1979-1998	1979-1998	1950-1998	1985-1998	1950-1998	1979-1999
MA-CNNN6	1979-1998	1979-1998	1950-1998	1985-1998	1950-1998	1979-1999

[1] The bivariate probability distribution of the raw forecasts and observations, whose parameters are estimated from the paired data
[2] The space-time co-variability of precipitation and temperature is estimated with the Schaake Shuffle, based on observed MAP/MAT

Table 3: important characteristics of the first and second generation reforecasts

Characteristic	First generation (frozen GFS)	Second generation (GEFS)
Variables	Total accumulated precipitation, min/max temperature 2m above the surface	Total accumulated precipitation, min/max temperature 2m above the surface
Ensemble members	15 members (1 control, 14 perturbed)	11 members (1 control, 10 perturbed)
Horizontal resolution	T62 (~200 km)	T254 (~55 km) for 1-8 days, T190 (~70 km) for 8 to 16 days
Vertical resolution	L28 (28 levels)	L42 (42 levels)
Forecast horizon	15 days	16 days
Forecast time step	12 hours	6 hours
Initialization cycle	00 UTC	00 UTC
Period of record	1979 – 2006 (~27 yrs)	1985 – 2010 (~25 yrs)
GFS version	~1997	9.0.1
Availability	NOAA/ESRL	NOAA/ESRL
Reference	Hamill et al. (2006)	Hamill et al. (2013)

10. Figures

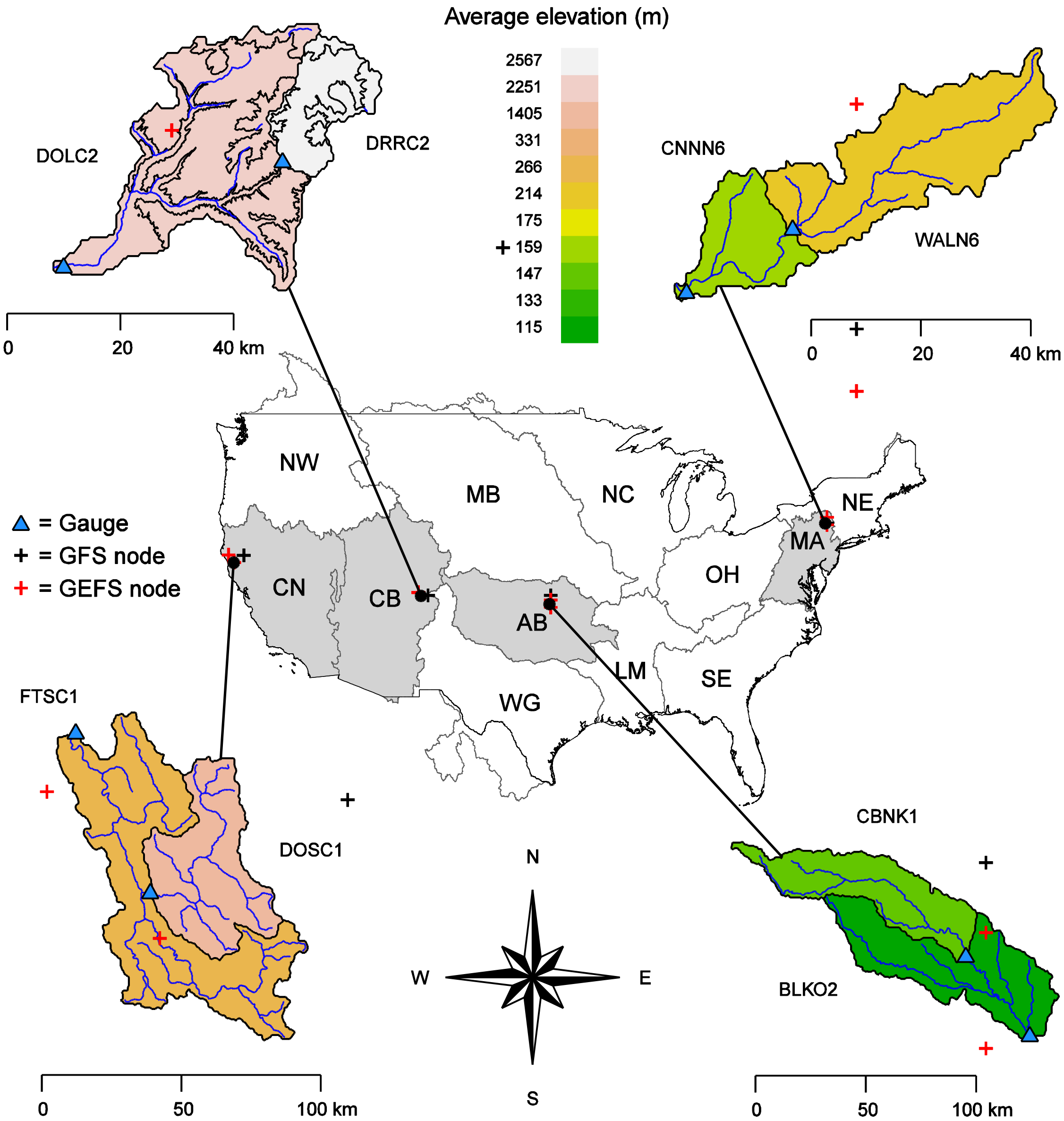

Figure 1: The eight study basins, comprising one upstream and one downstream basin in each of AB-, CB-, CN- and MA-RFCs.

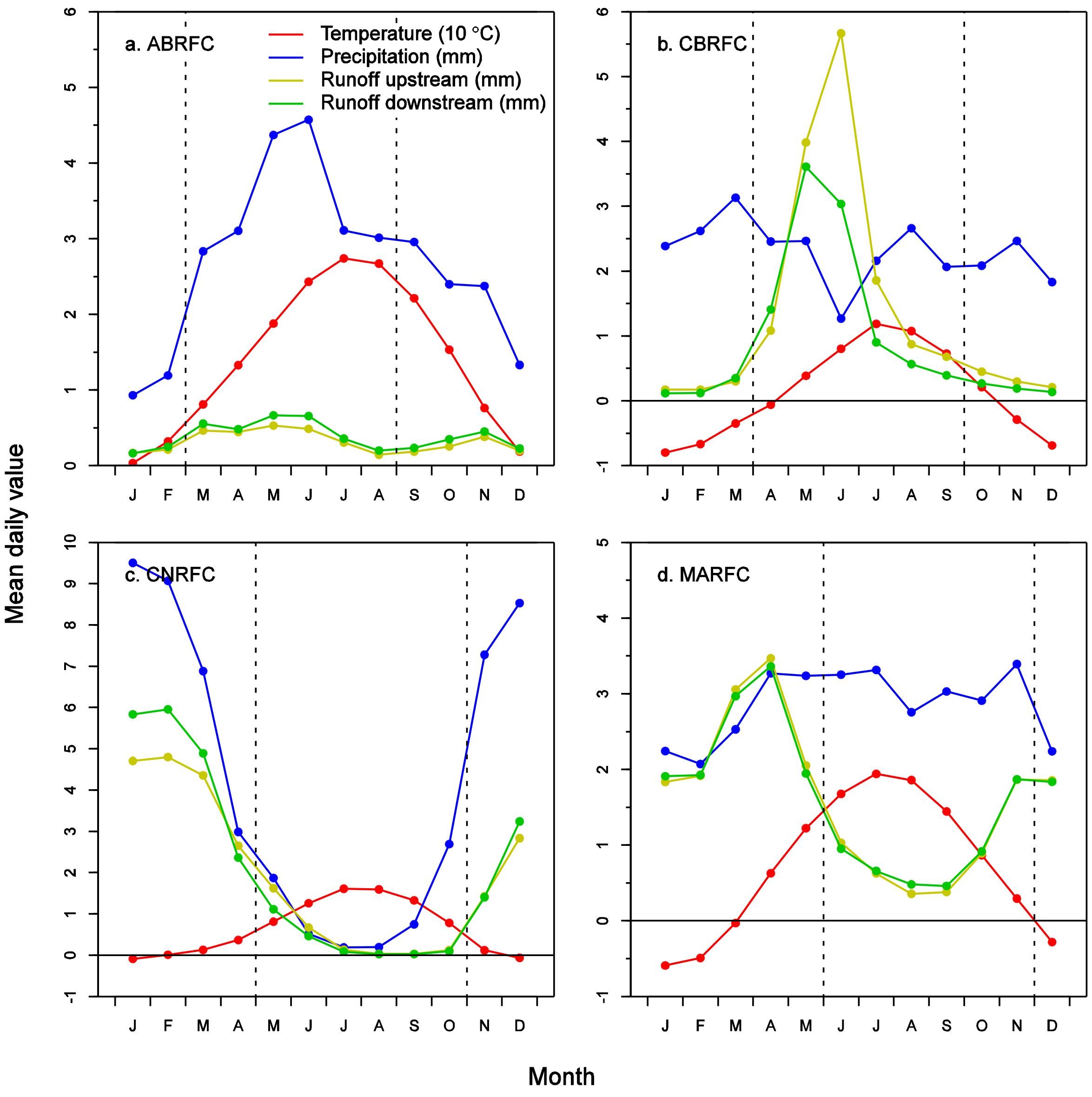

Figure 2: Daily averages of temperature, precipitation and runoff by calendar month for each study basin. The meteorological variables are averaged over the upstream and downstream basins (weighed by basin area).

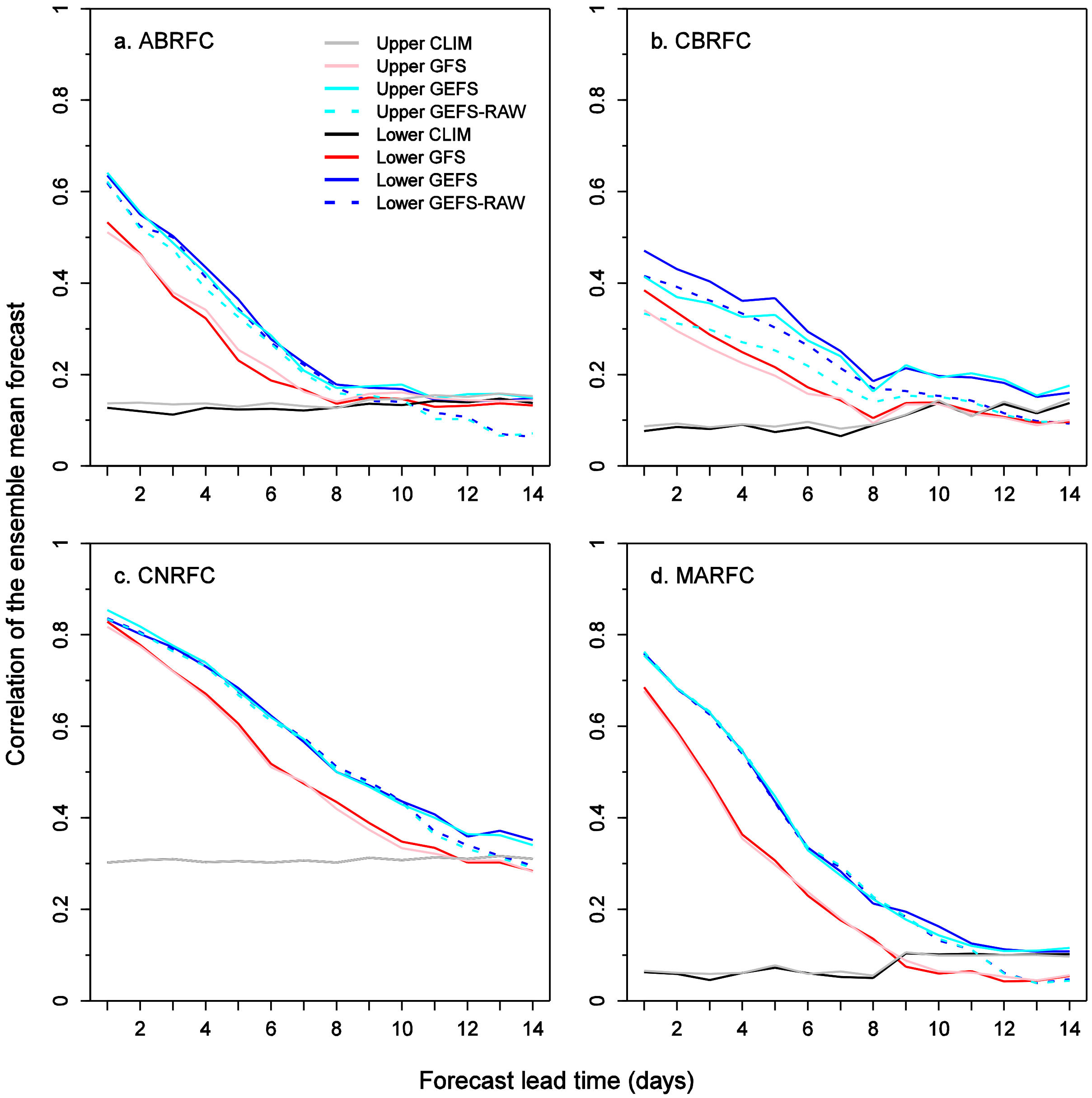

Figure 3: Correlation between the ensemble mean forecast and observed precipitation amounts by forecast lead time. The results are shown for the upstream and downstream basin in each RFC and comprise the raw GEFS reforecasts (GEFS-RAW) and the MEFP outputs with forcing inputs from the GFS, GEFS and resampled climatology (CLIM).

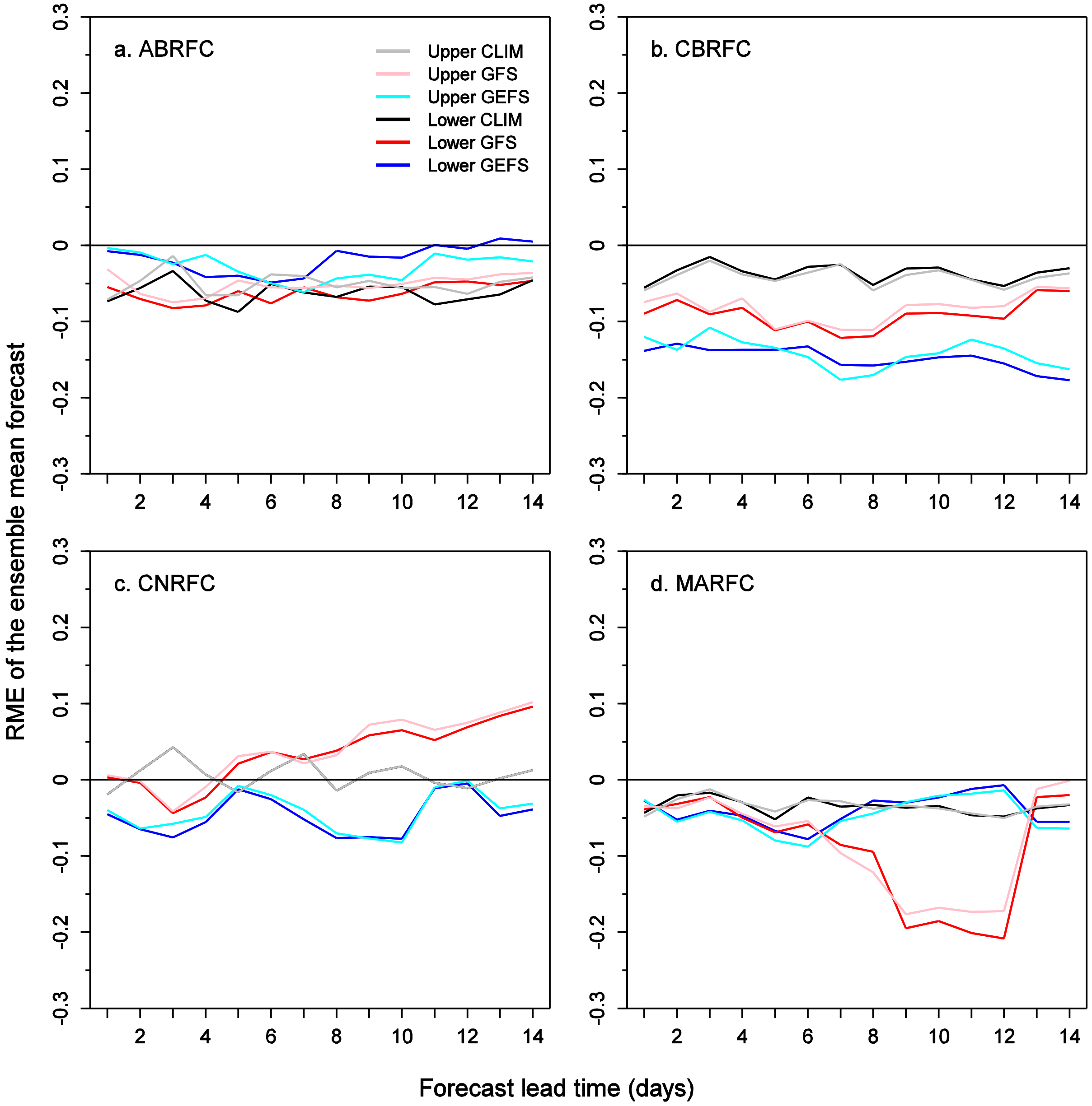

Figure 4: Relative mean error (RME) of the MEFP precipitation forecasts with forcing inputs from the GFS, GEFS and resampled climatology (CLIM). The results are shown by forecast lead time for the upstream and downstream basin in each RFC.

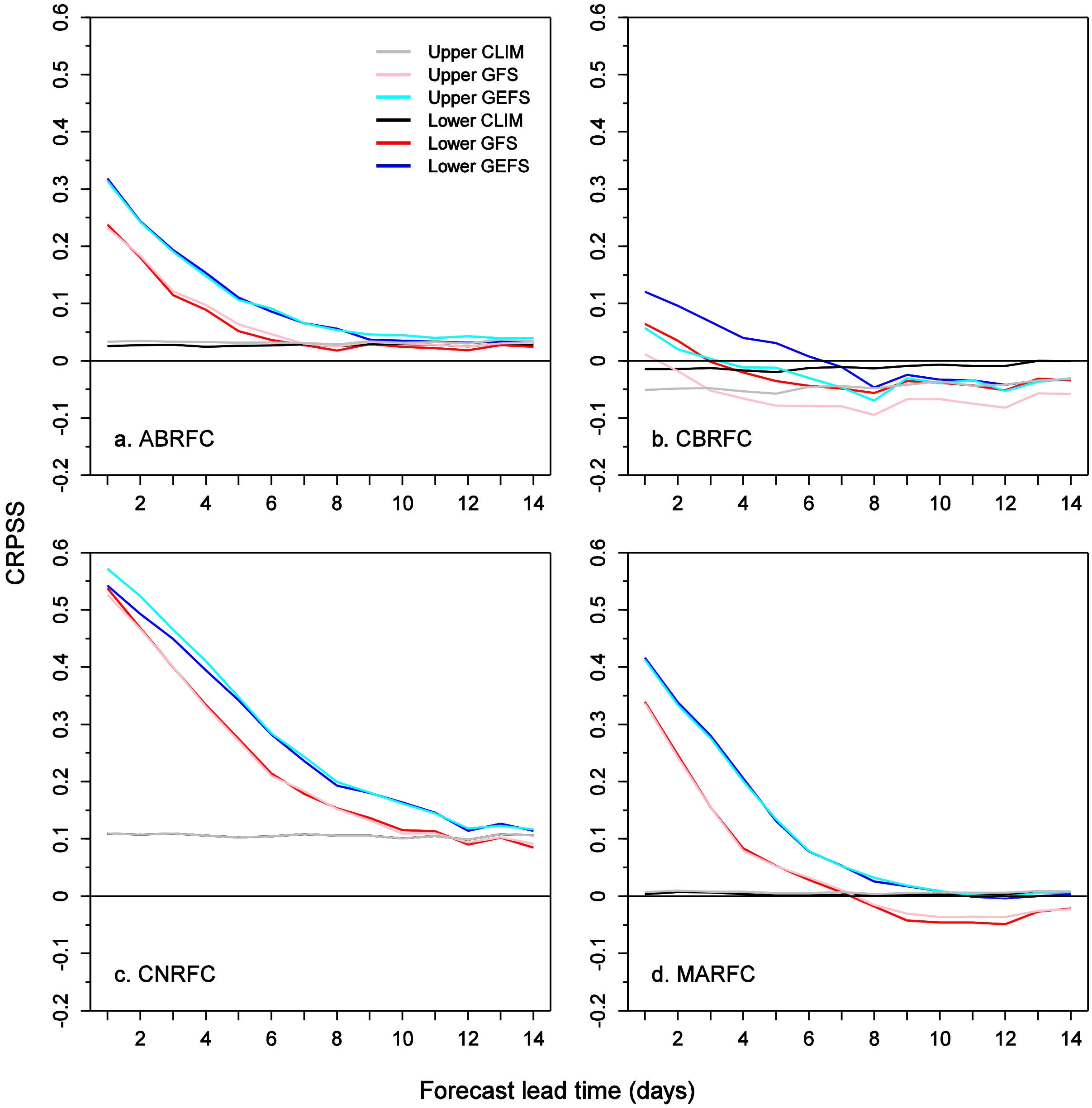

Figure 5: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP precipitation forecasts with forcing inputs from the GFS, GEFS and resampled climatology (CLIM). The results are shown by forecast lead time for the upstream and downstream basin in each RFC. The reference forecast is sample climatology.

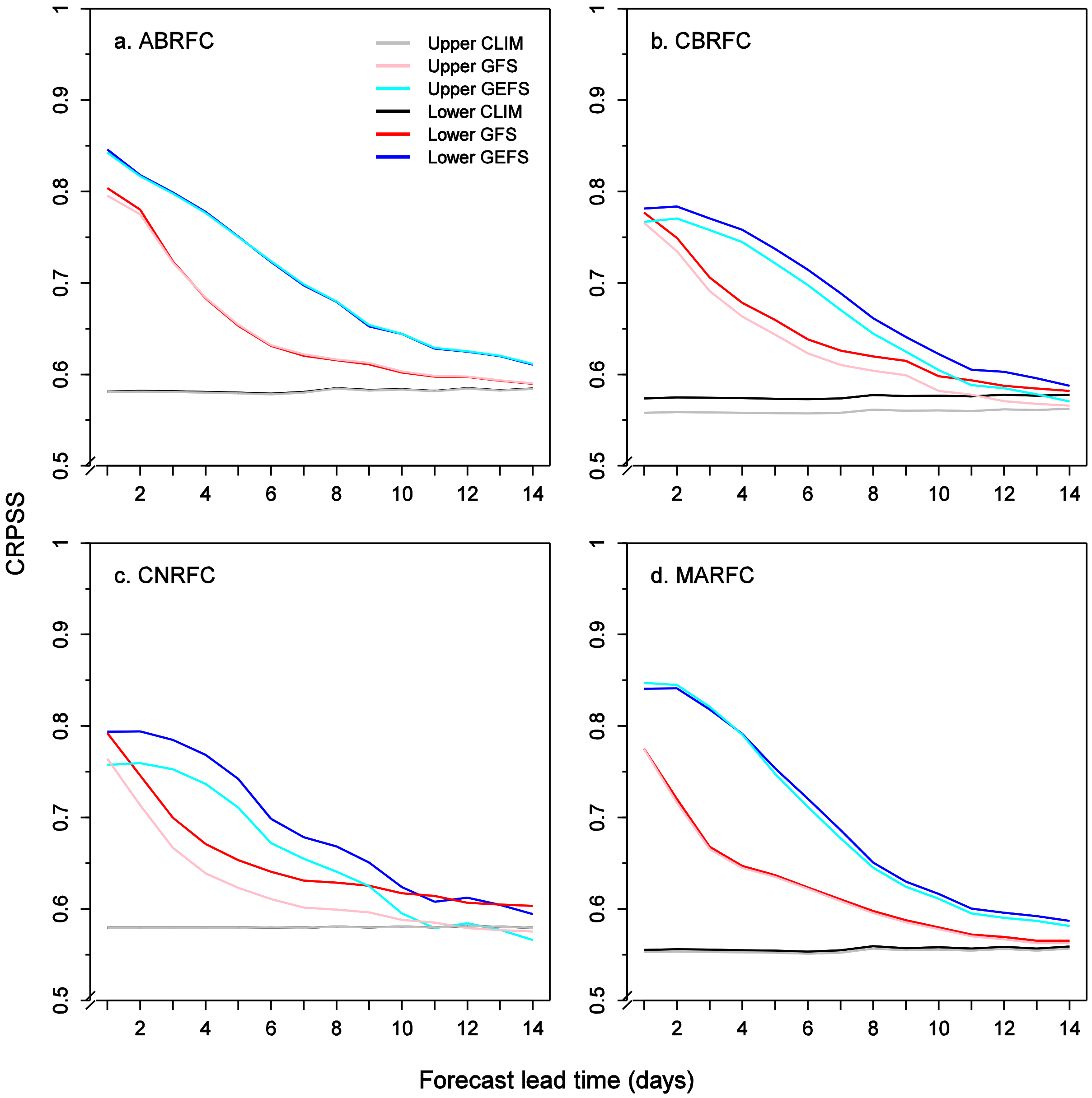

Figure 6: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP temperature forecasts with forcing inputs from the GFS, GEFS and resampled climatology (CLIM). The results are shown by forecast lead time for the upstream and downstream basin in each RFC. The reference forecast is sample climatology. Note the discontinuity on the range axis.

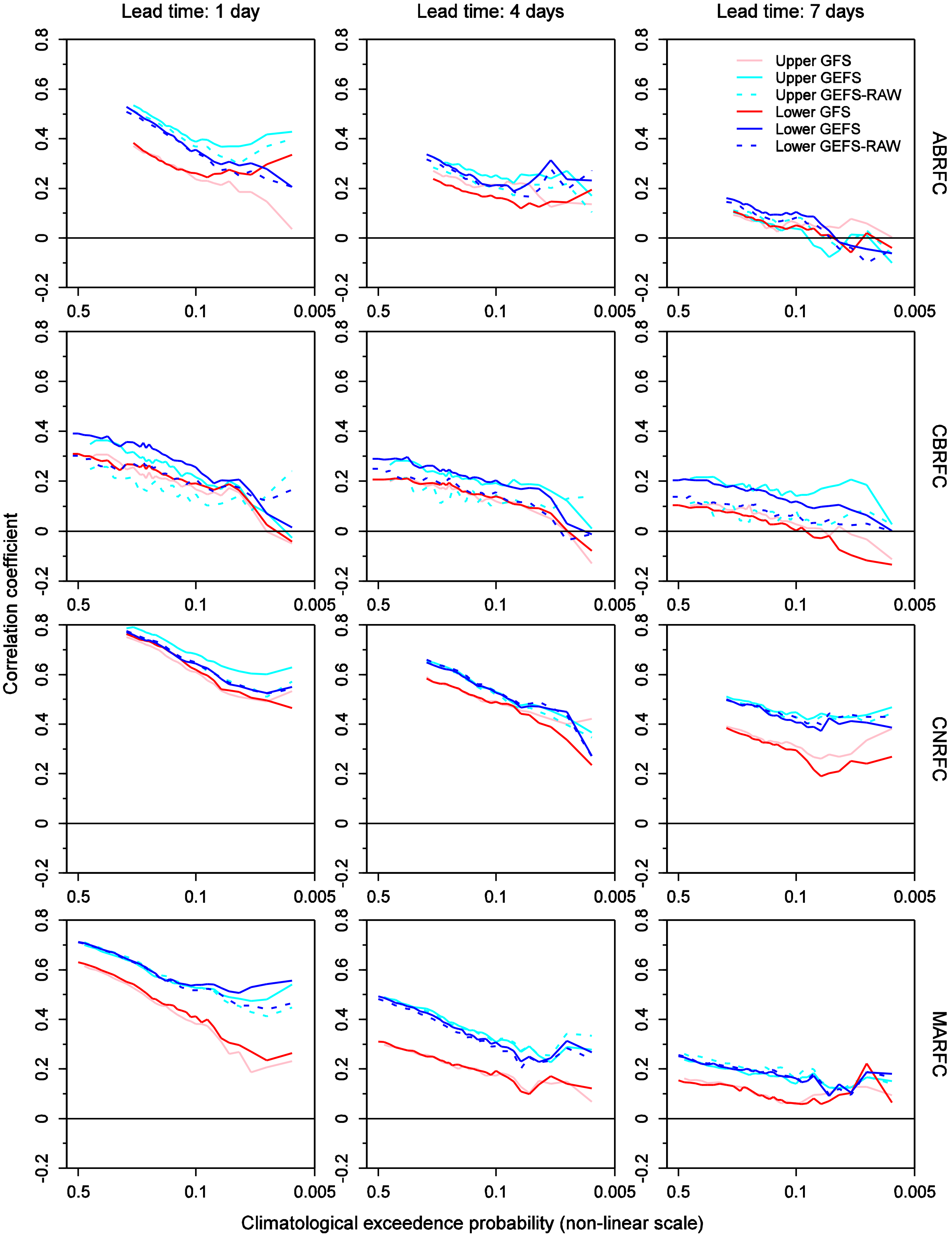

Figure 7: Correlation between the ensemble mean forecast and observed precipitation amount at increasing thresholds of observed precipitation. The results are shown for the raw GEFS reforecasts, the MEFP-GFS forecasts and the MEFP-GEFS forecasts. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

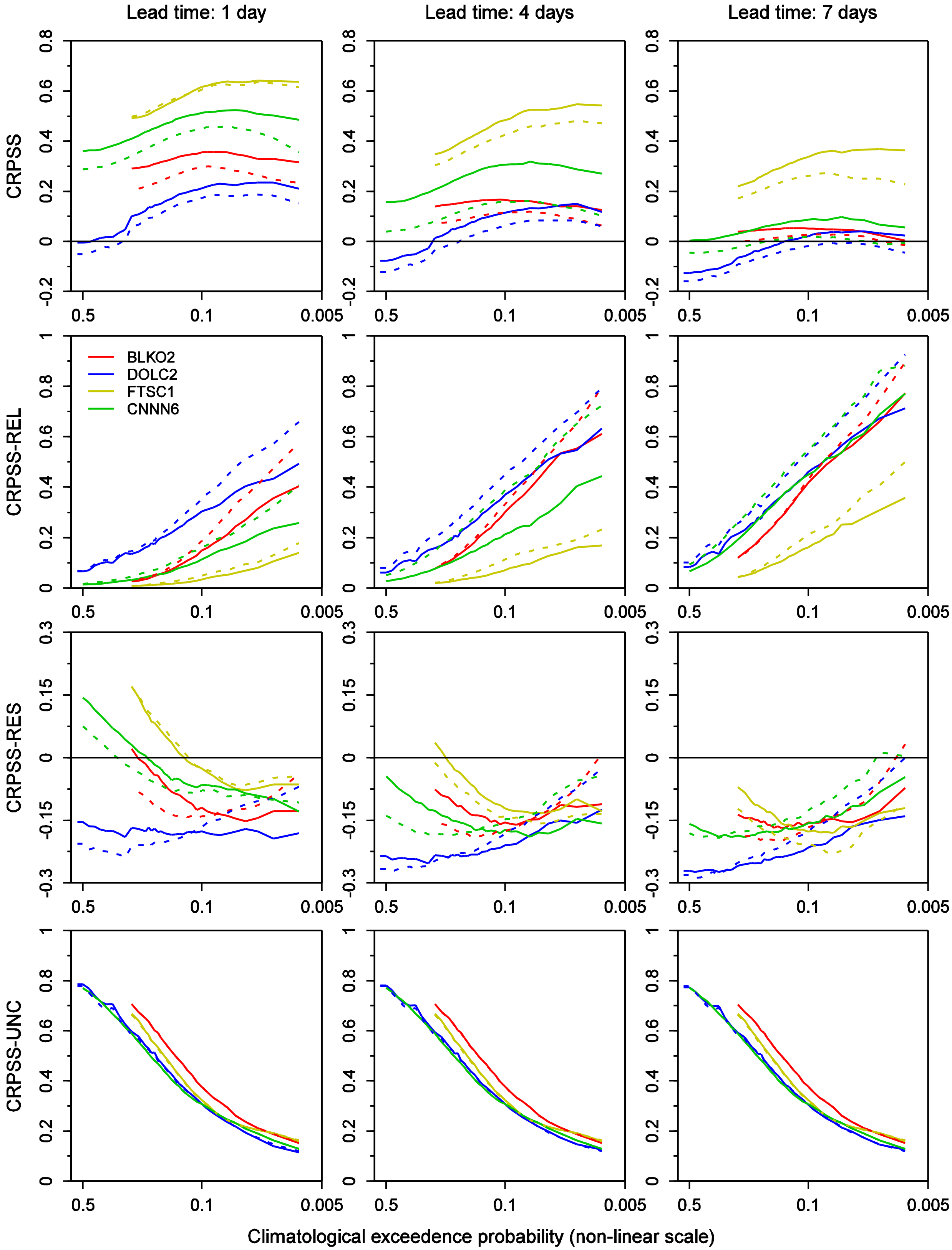

Figure 8: Calibration-refinement factorization of the CRPSS for the MEFP-GEFS precipitation forecasts and the corresponding MEFP-GFS forecasts (dashed). The results are shown for the downstream basins and the reference forecast is sample climatology. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

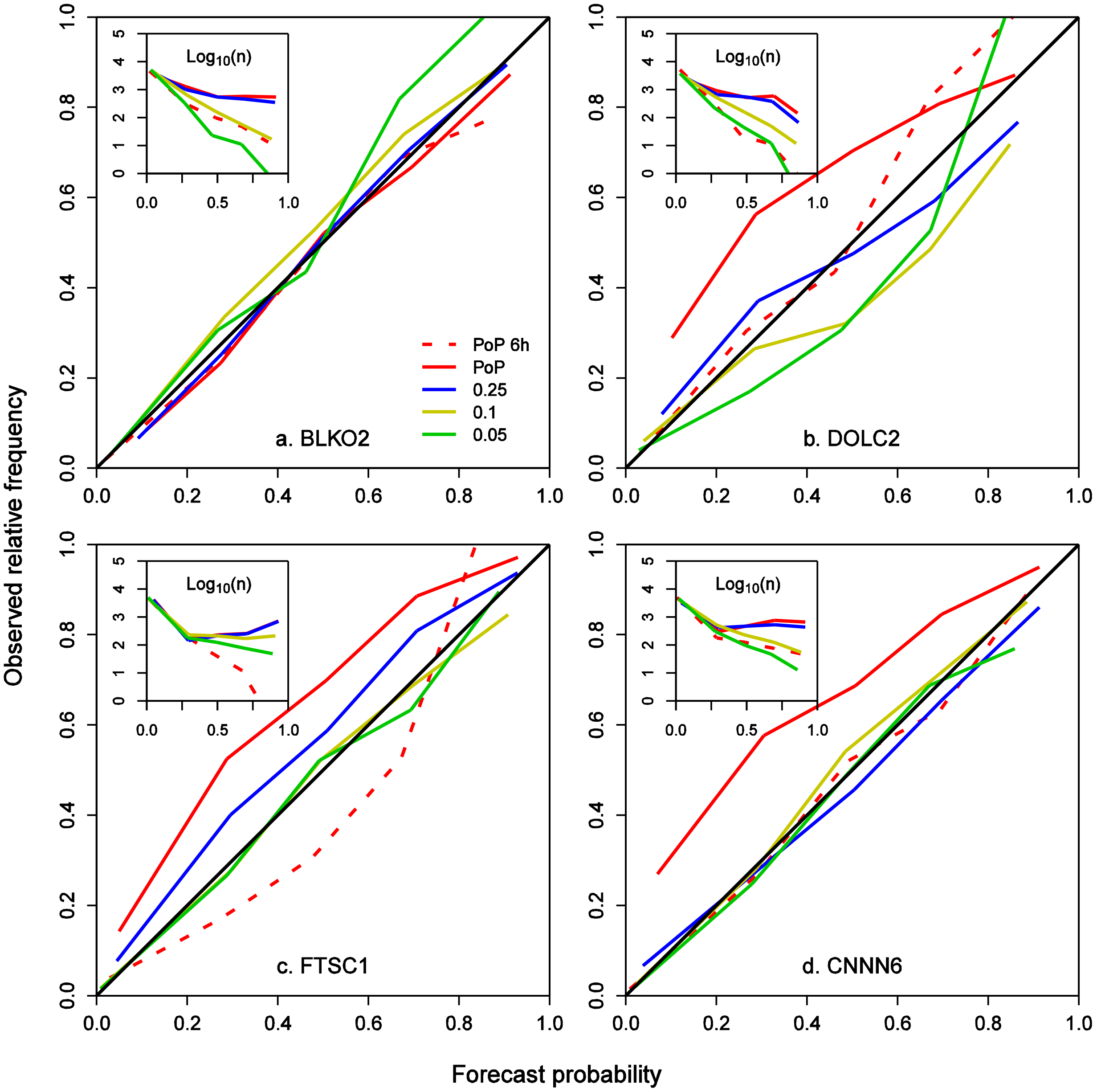

Figure 9: Reliability diagrams for the MEFP-GEFS precipitation forecasts at the downstream basin in each RFC. The results are shown for selected precipitation thresholds, which are expressed as climatological exceedence probabilities. The solid lines comprise the daily accumulation with a forecast lead time of 1-2 days. The dashed lines comprise the average reliability of the four, six-hourly, accumulations from 1-2 days at the probability of precipitation (PoP) threshold.

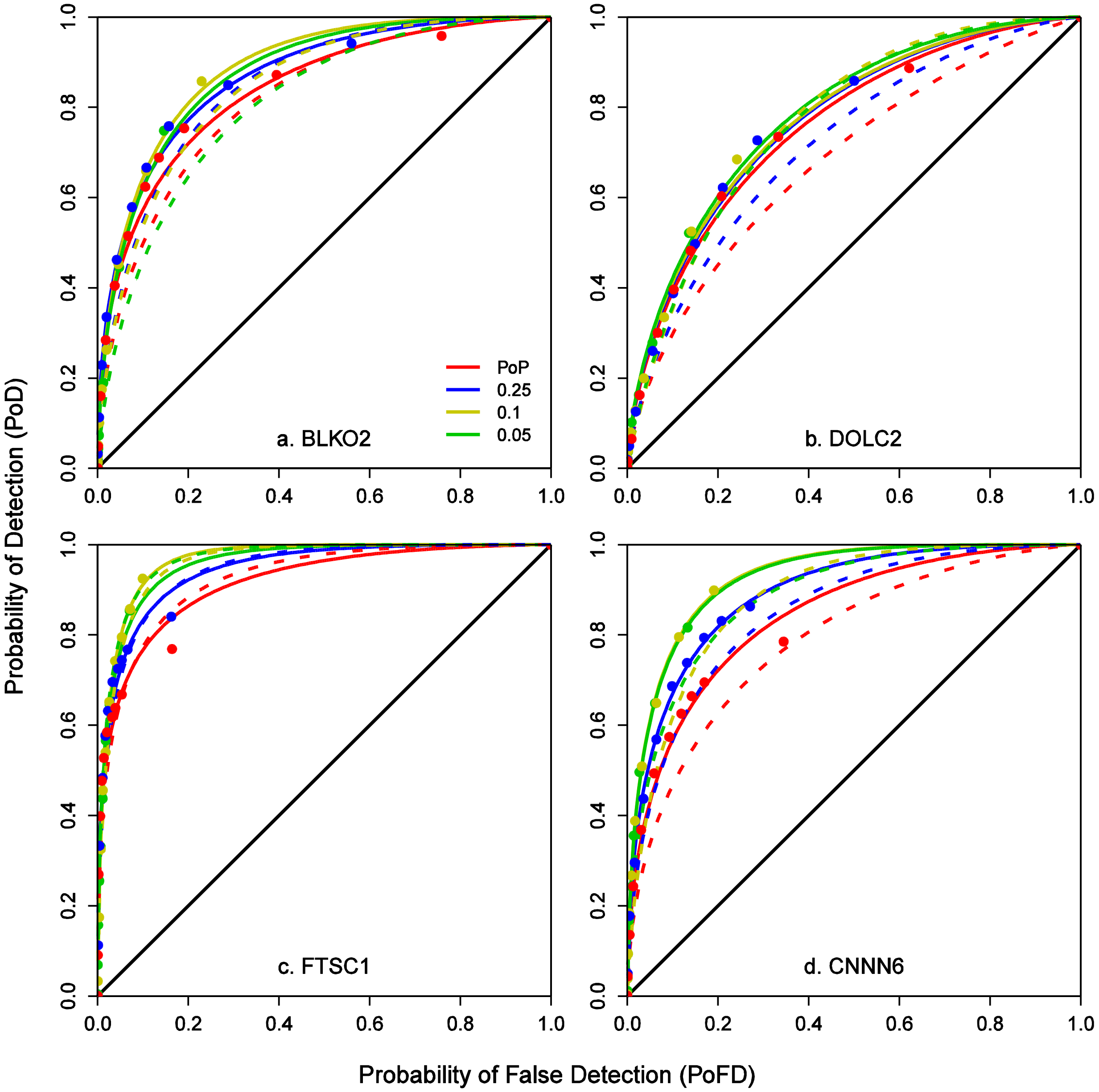

Figure 10: Relative Operating Characteristic (ROC) curves for the MEFP-GEFS precipitation forecasts and the corresponding MEFP-GFS forecasts (dashed). The results are shown at a forecast lead time of 1-2 days for the downstream basin in each RFC and for selected precipitation thresholds, which are expressed as climatological exceedence probabilities. The ROC curves were fitted under an assumption of bivariate normality between the empirical pairs of PoD and PoFD (shown as dots).

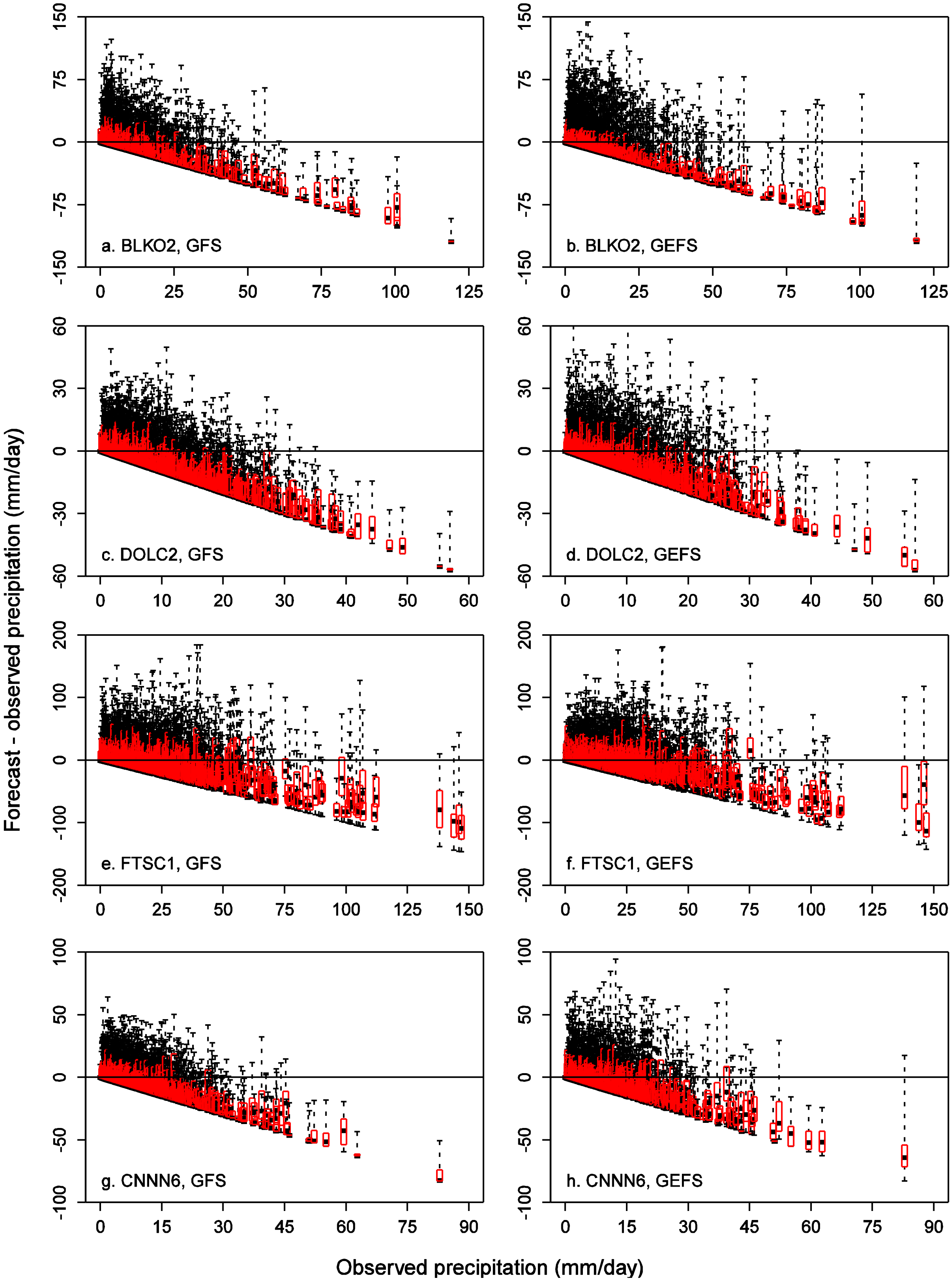

Figure 11: Box plots of errors (forecast - observed) in the MEFP-GFS precipitation forecasts (left column) and the corresponding MEFP-GEFS forecasts (right column). The results are shown at a forecast lead time of 3-4 days for the downstream basin in each RFC.

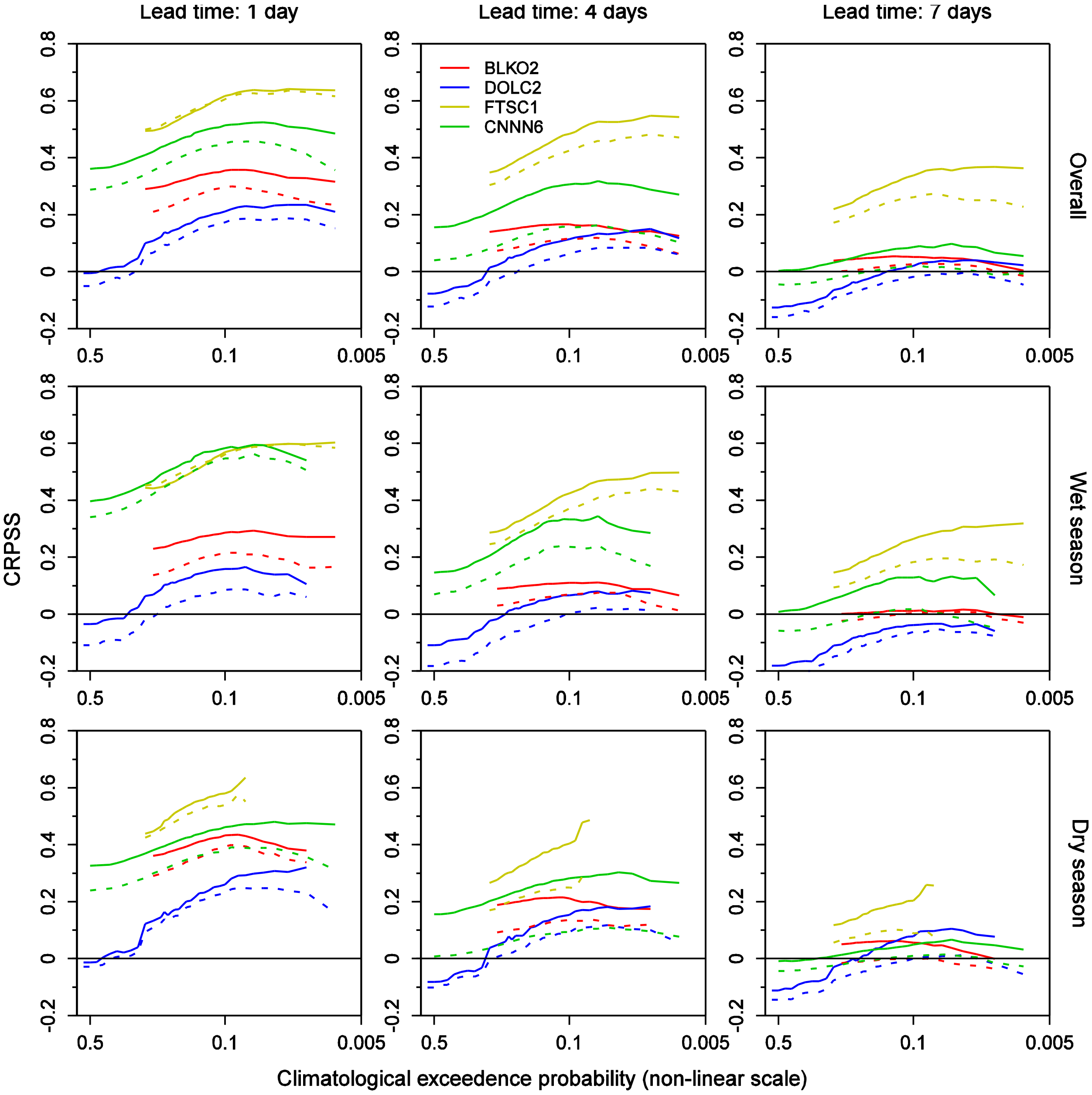

Figure 12: Mean CRPSS of the MEFP-GEFS precipitation forecasts and the corresponding MEFP-GFS forecasts (dashed) at selected forecast lead times. The reference forecast is sample climatology. The results are shown for the downstream basin in each RFC and for the wet and dry seasons, as well as the overall period. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

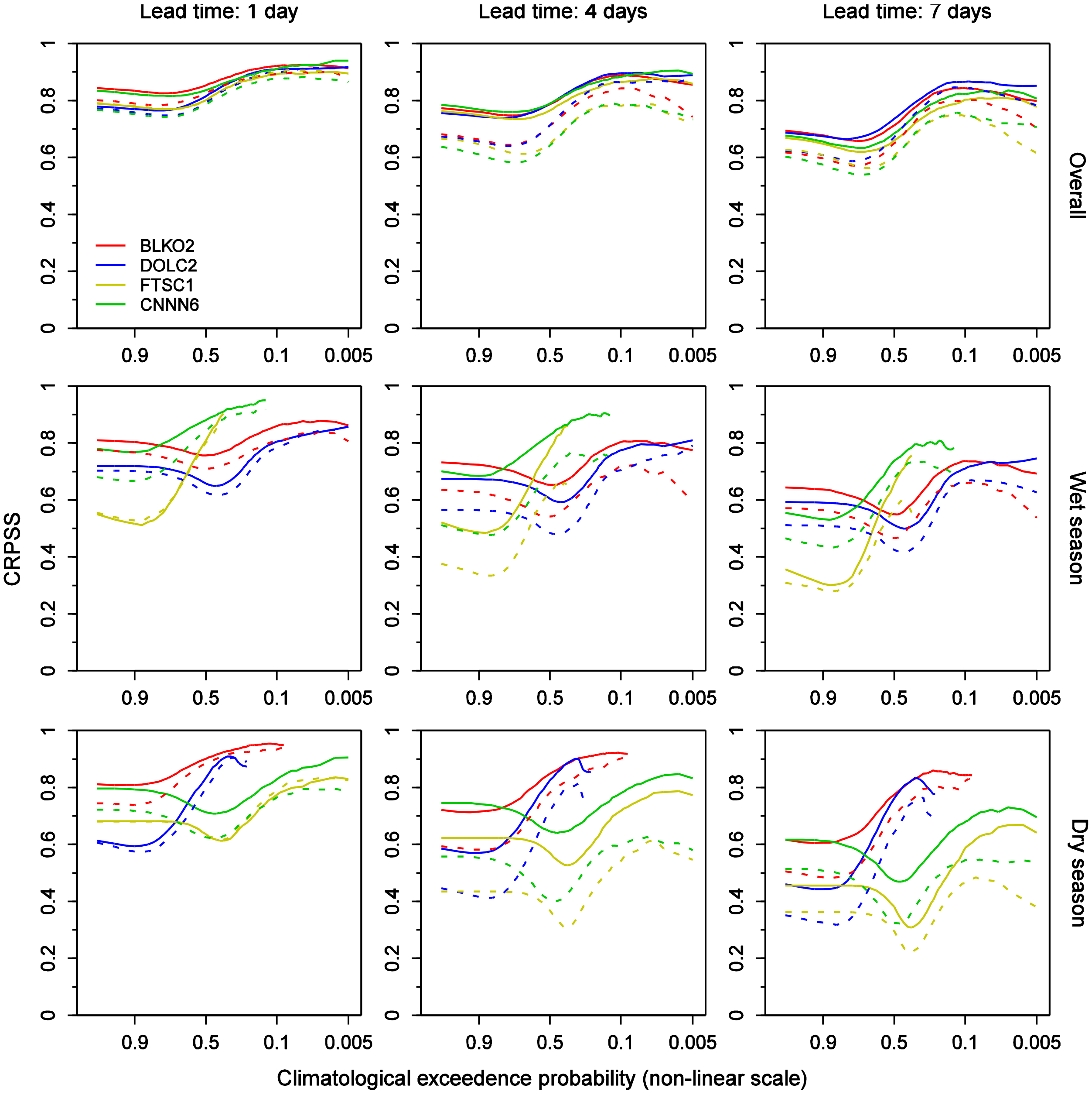

Figure 13: Mean CRPSS of the MEFP-GEFS temperature forecasts and the corresponding MEFP-GFS forecasts (dashed) at selected forecast lead times. The reference forecast is sample climatology. The results are shown for the downstream basin in each RFC and for the wet and dry seasons, as well as the overall period. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

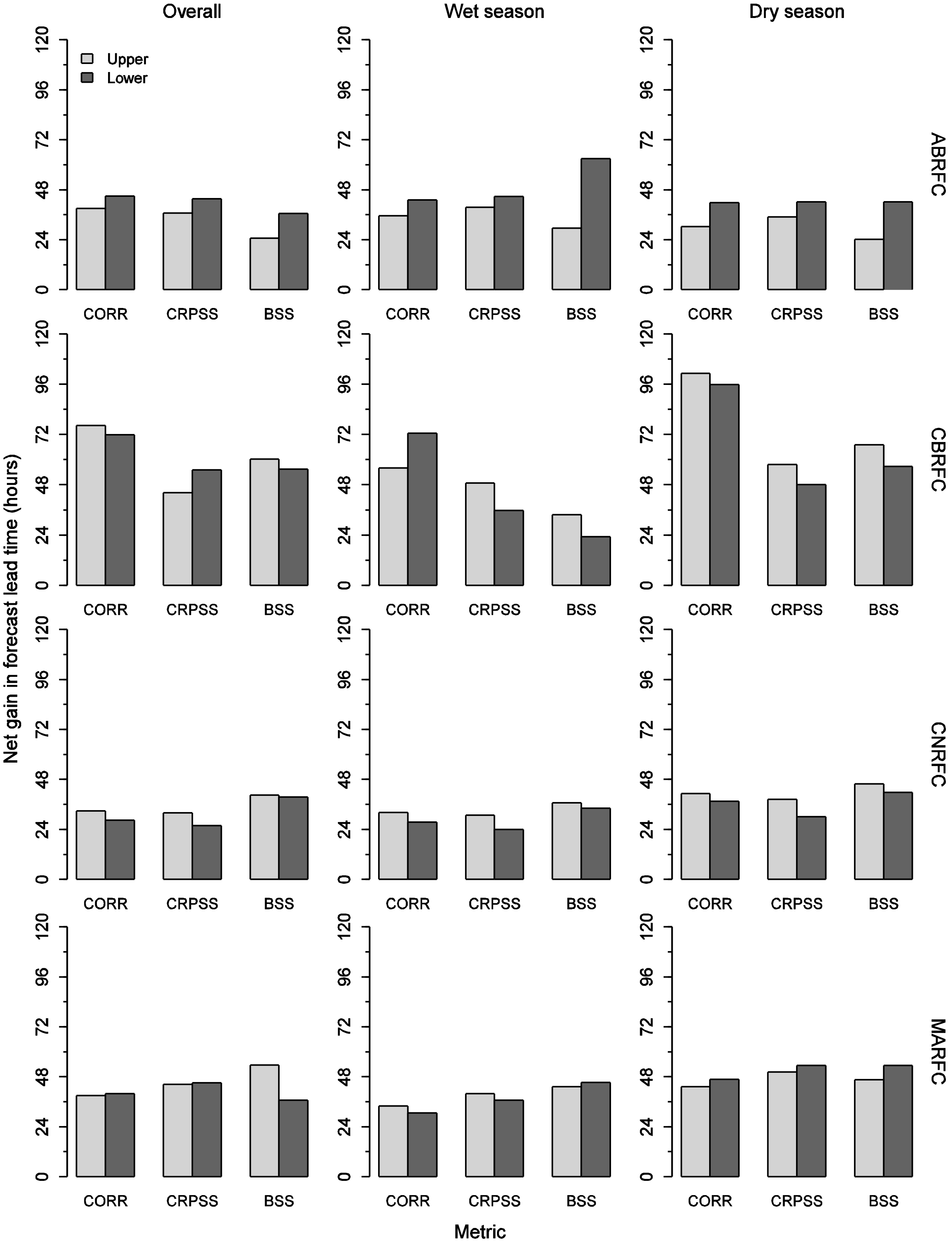

Figure 14: Average net gain in forecast lead time for the MEFP-GEFS precipitation forecasts when compared to the MEFP-GFS forecasts. The results are shown for the wet and dry seasons, as well as the overall period. The net gain in forecast lead time was averaged across lead times of 1-7 days and is shown for three verification metrics. The BSS comprises a threshold of C_p=0.1.

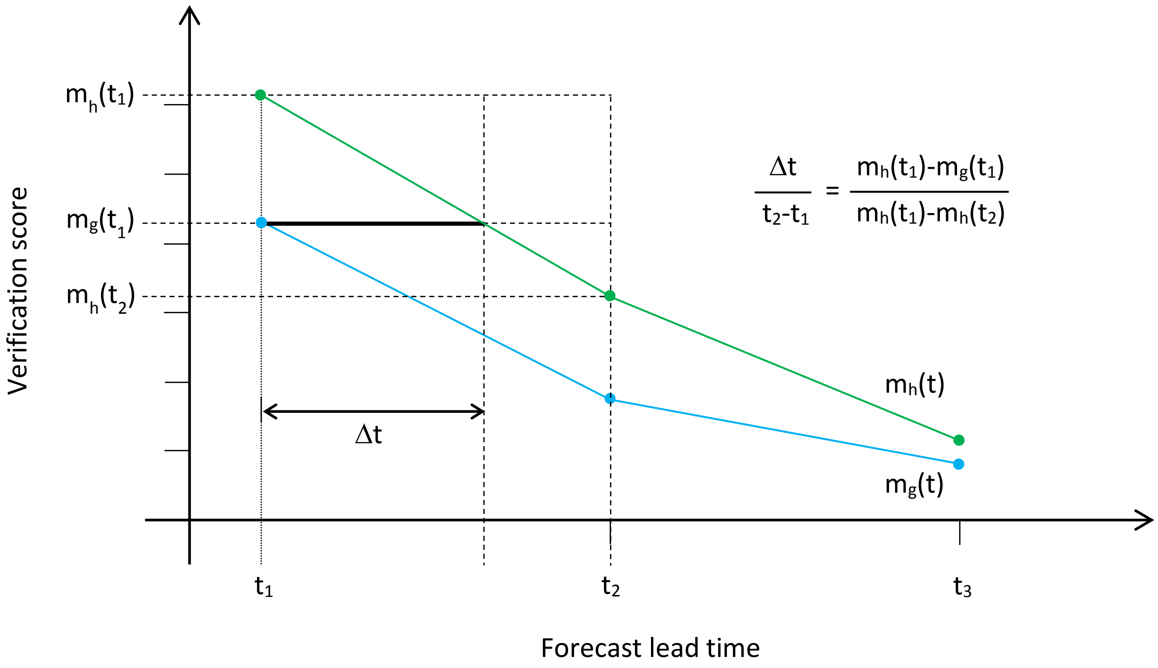

Figure 15: An illustration of the gain in forecast lead time, Δt, for one verification time-series, $m_h(t)$, relative to another, $m_g(t)$. The gain in lead time is defined as the difference in time, Δt, from $m_g(t_1)$ until an equivalent value is reached on time-series, $m_h(t)$. In practice, Δt may comprise several whole time increments in addition to a partial increment. A similar procedure is used to define a reduction in forecast lead time. The average net gain is defined as the average Δt across several start times $(t_1,t_2,\ldots,t_n)$.

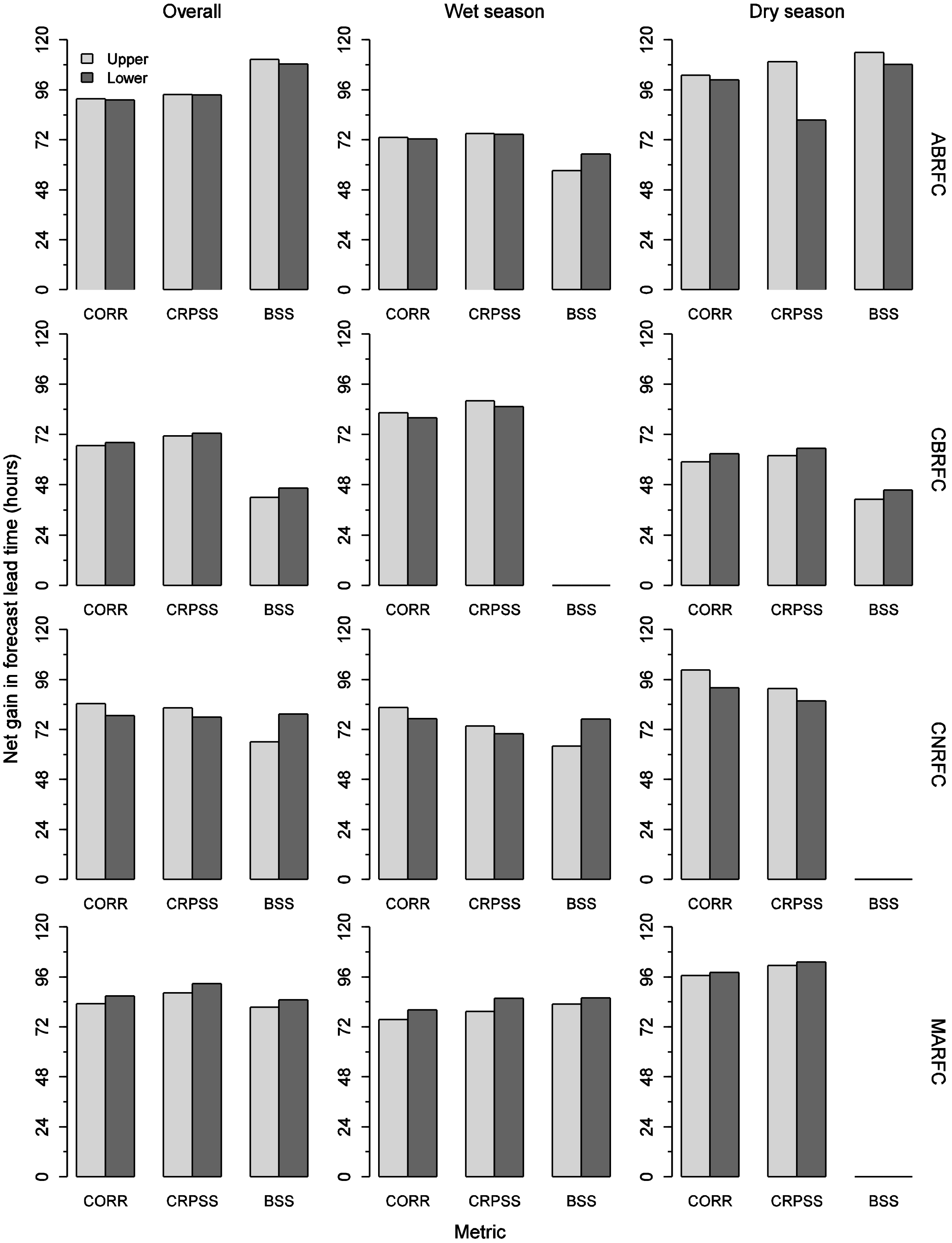

Figure 16: Average net gain in forecast lead time for the MEFP-GEFS temperature forecasts when compared to the MEFP-GFS forecasts. The results are shown for three verification metrics and for the wet and dry seasons, as well as the overall period. The net gain in forecast lead time was averaged across lead times of 1-7 days. The BSS comprises a threshold of C_p=0.9.

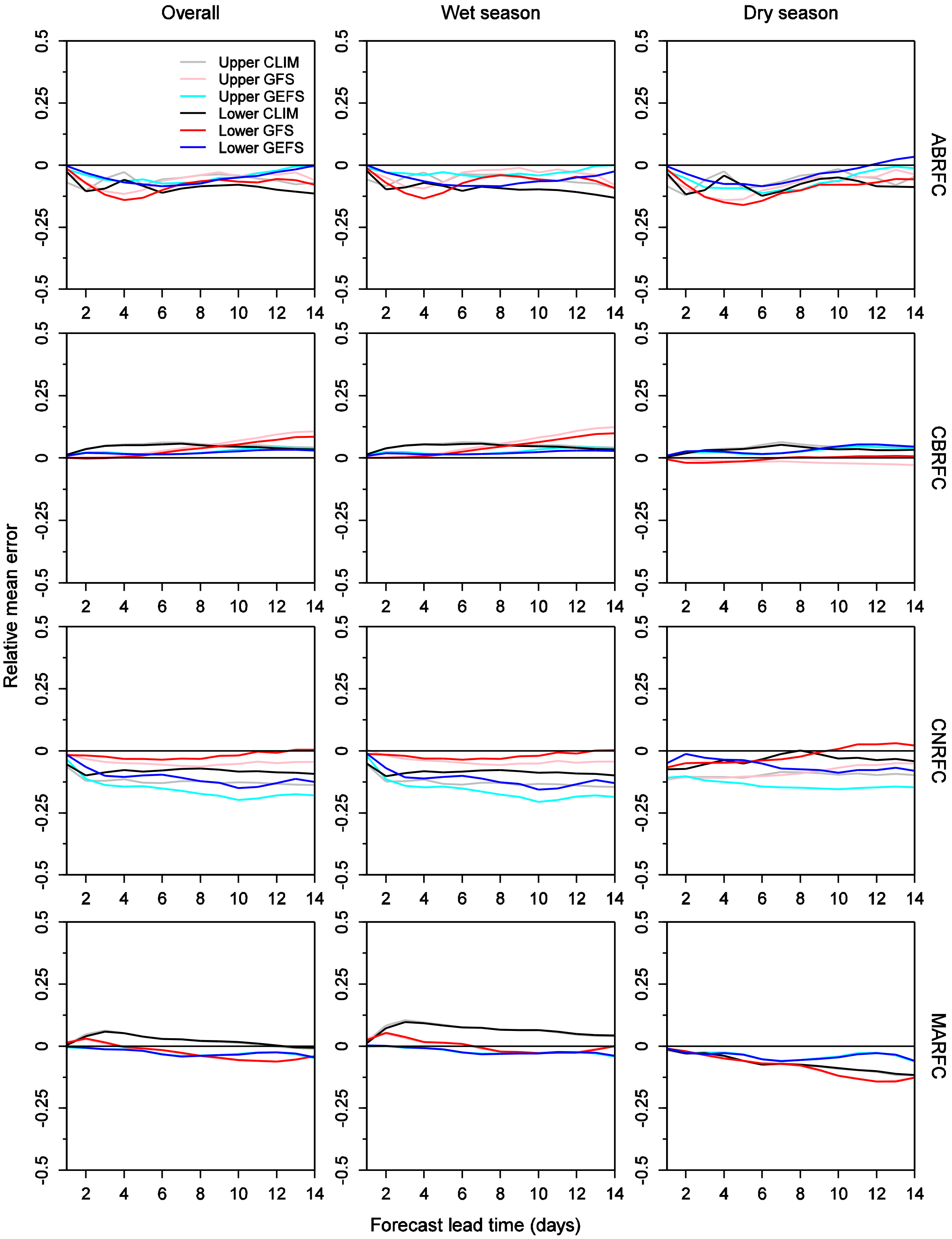

Figure 17: Relative mean error (RME) of the MEFP-GEFS, MEFP-GFS and MEFP-CLIM streamflow forecasts against simulated streamflow. The results are shown for the upstream and downstream basin in each RFC and for the wet and dry seasons, as well as the overall period.

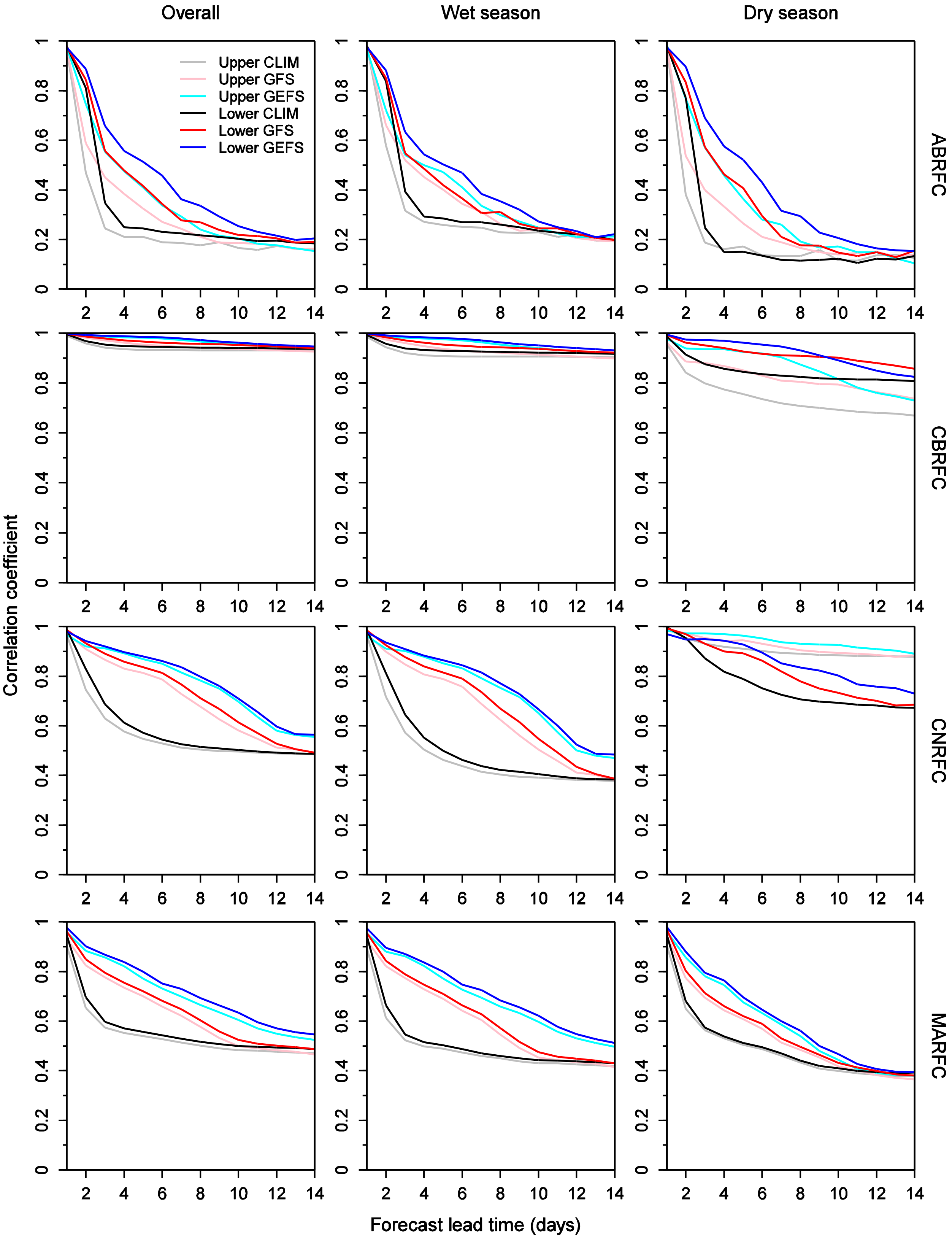

Figure 18: Correlation of the ensemble mean forecast and simulation for the MEFP-GEFS, MEFP-GFS and MEFP-CLIM streamflow forecasts. The results are shown for the upstream and downstream basin in each RFC and for the wet and dry seasons, as well as the overall period.

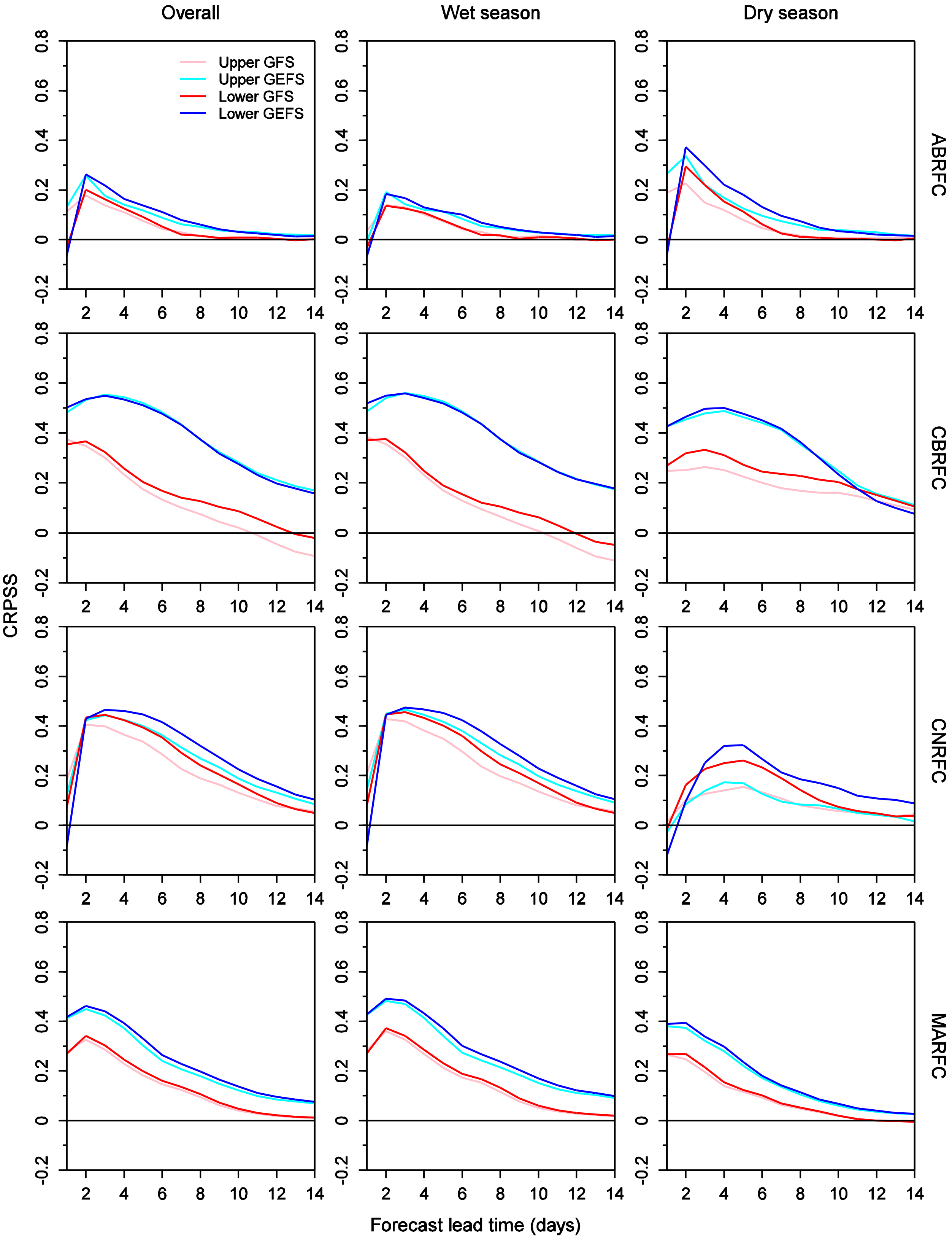

Figure 19: Mean CRPSS of the MEFP-GFS and the MEFP-GEFS streamflow forecasts versus the MEFP-CLIM forecasts. The results are shown for the upstream and downstream basin in each RFC and for the wet and dry seasons, as well as the overall period.

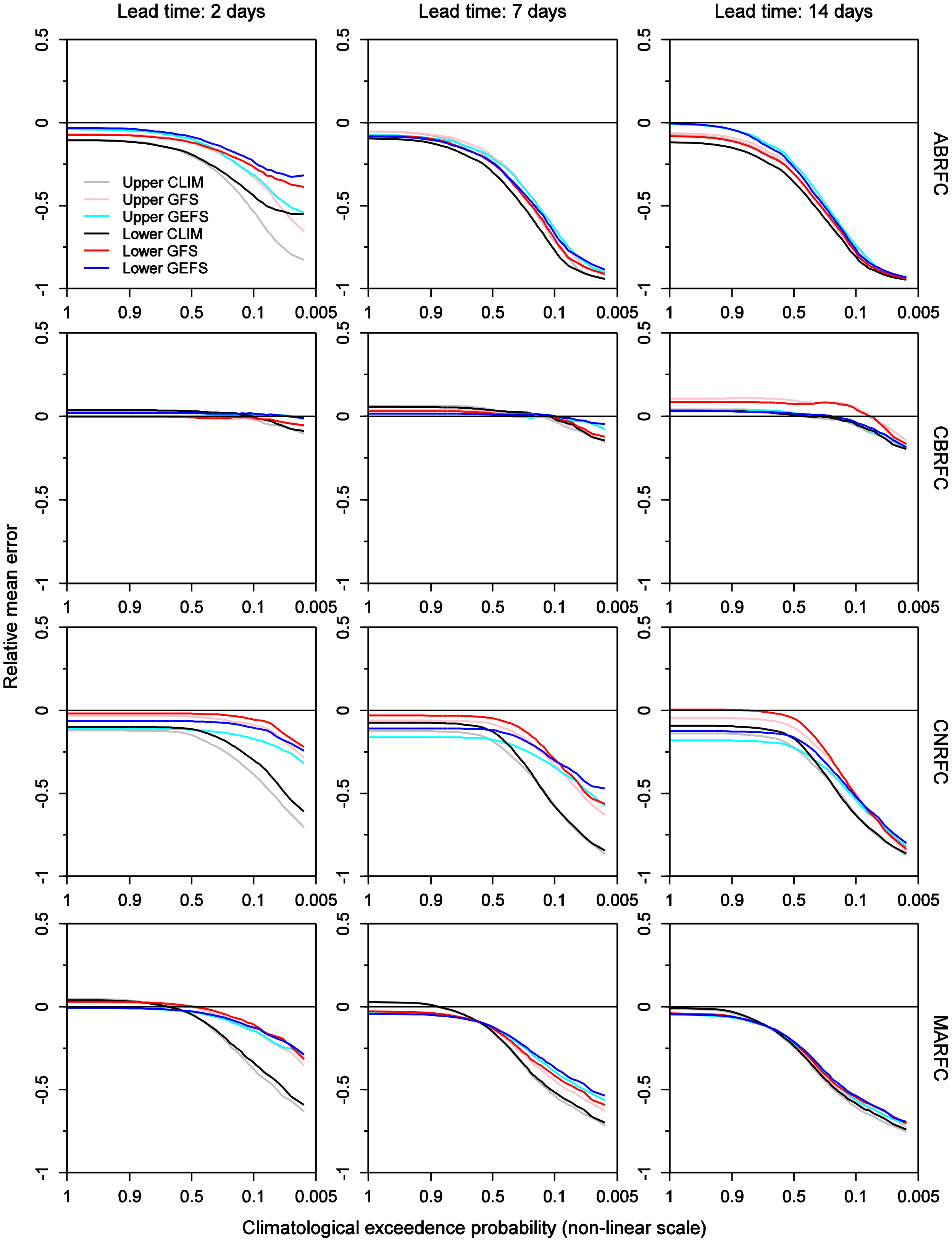

Figure 20: Relative mean error (RME) of the MEFP-GEFS, MEFP-GFS and MEFP-CLIM streamflow forecasts against simulated streamflow. The results are shown for the upstream and downstream basin in each RFC and at selected forecast lead times. The streamflow thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

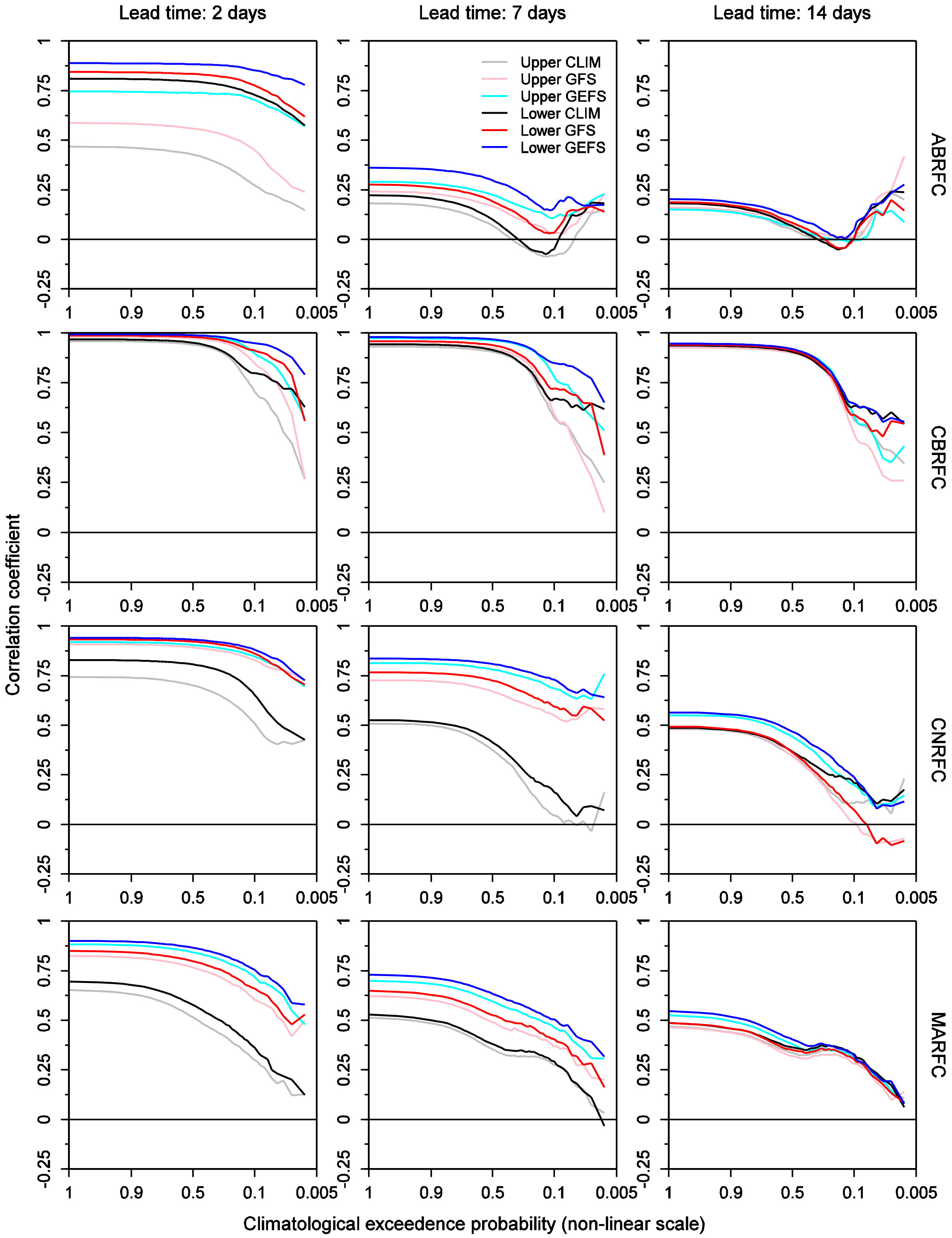

Figure 21: Correlation of the ensemble mean forecast and simulated streamflow for the MEFP-GEFS, MEFP-GFS and MEFP-CLIM forecasts. The results are shown for the upstream and downstream basin in each RFC and at selected forecast lead times. The streamflow thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

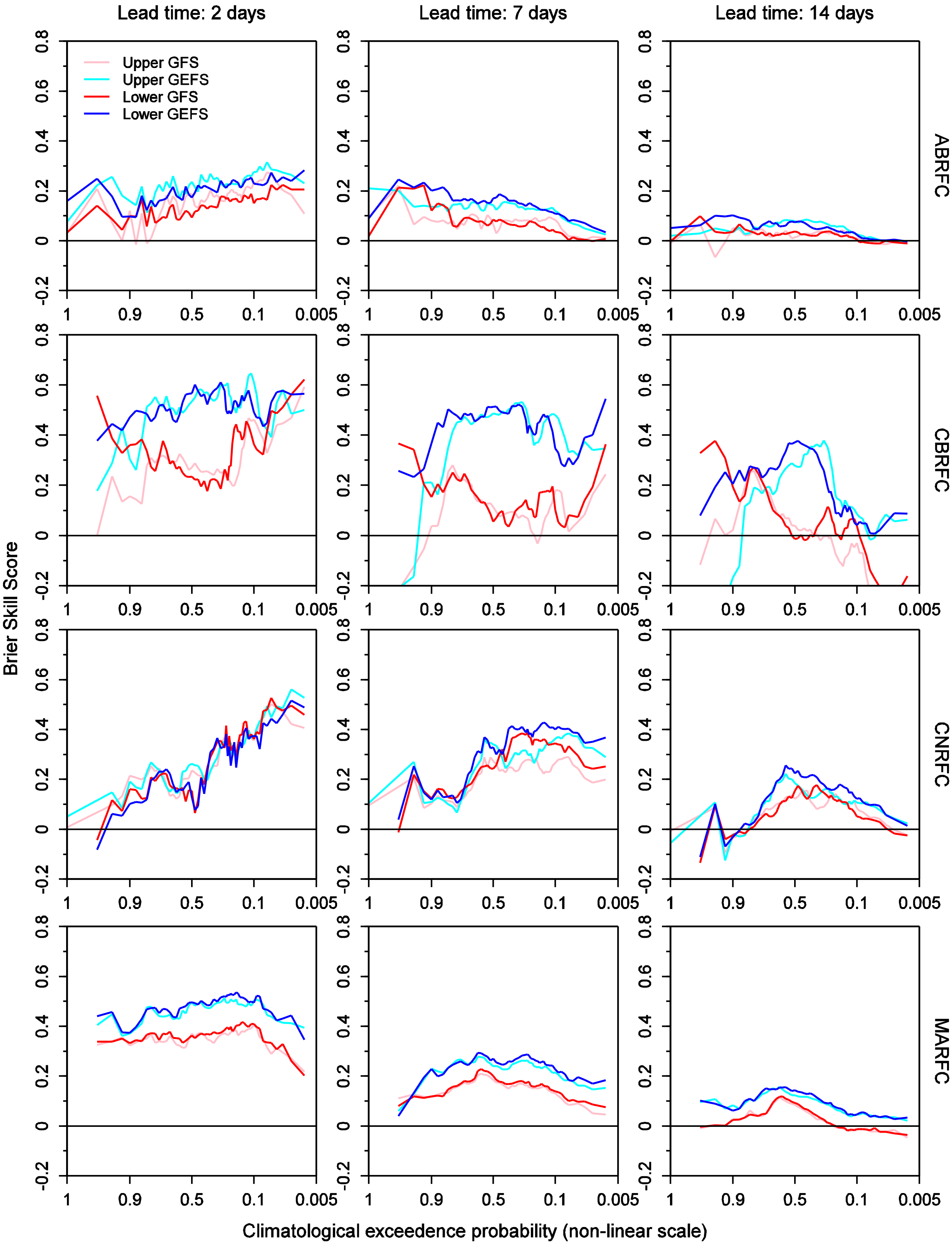

Figure 22: Brier Skill Score (BSS) of the MEFP-GEFS and MEFP-GFS streamflow forecasts versus the MEFP-CLIM forecasts. The results are shown for the upstream and downstream basin in each RFC and at selected forecast lead times. The streamflow thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

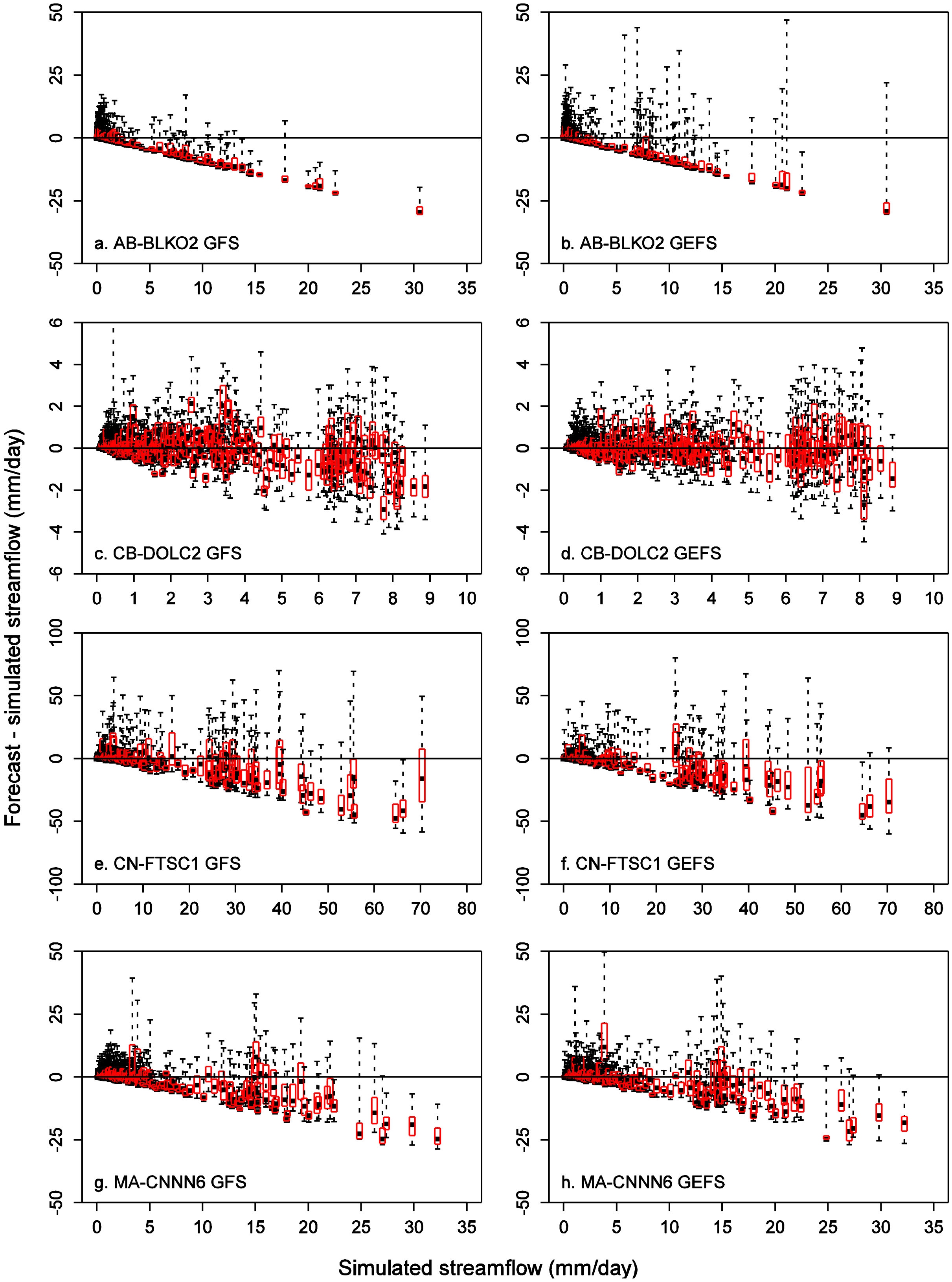

Figure 23: Box plots of errors (forecast - simulated) in the MEFP-GFS streamflow forecasts (left column) and the corresponding MEFP-GEFS forecasts (right column). The results are shown at a forecast lead time of 4-5 days for the downstream basin in each RFC.

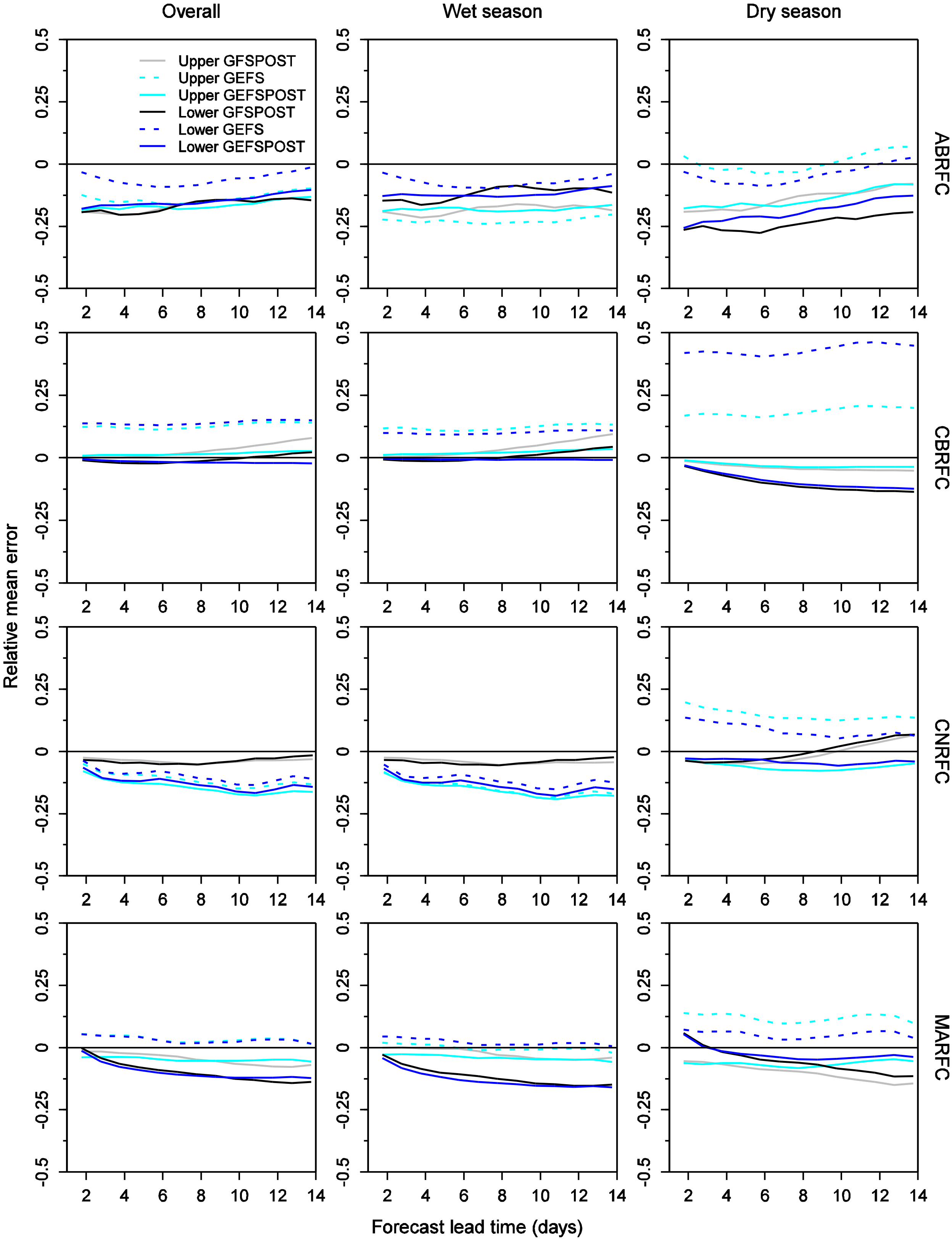

Figure 24: Relative mean error (RME) of the raw streamflow forecasts with MEFP-GEFS forcing (GEFS) and the corresponding post-processed streamflow forecasts (GEFSPOST). The results are also shown for the post-processed streamflow forecasts with MEFP-GFS forcing (GFSPOST).

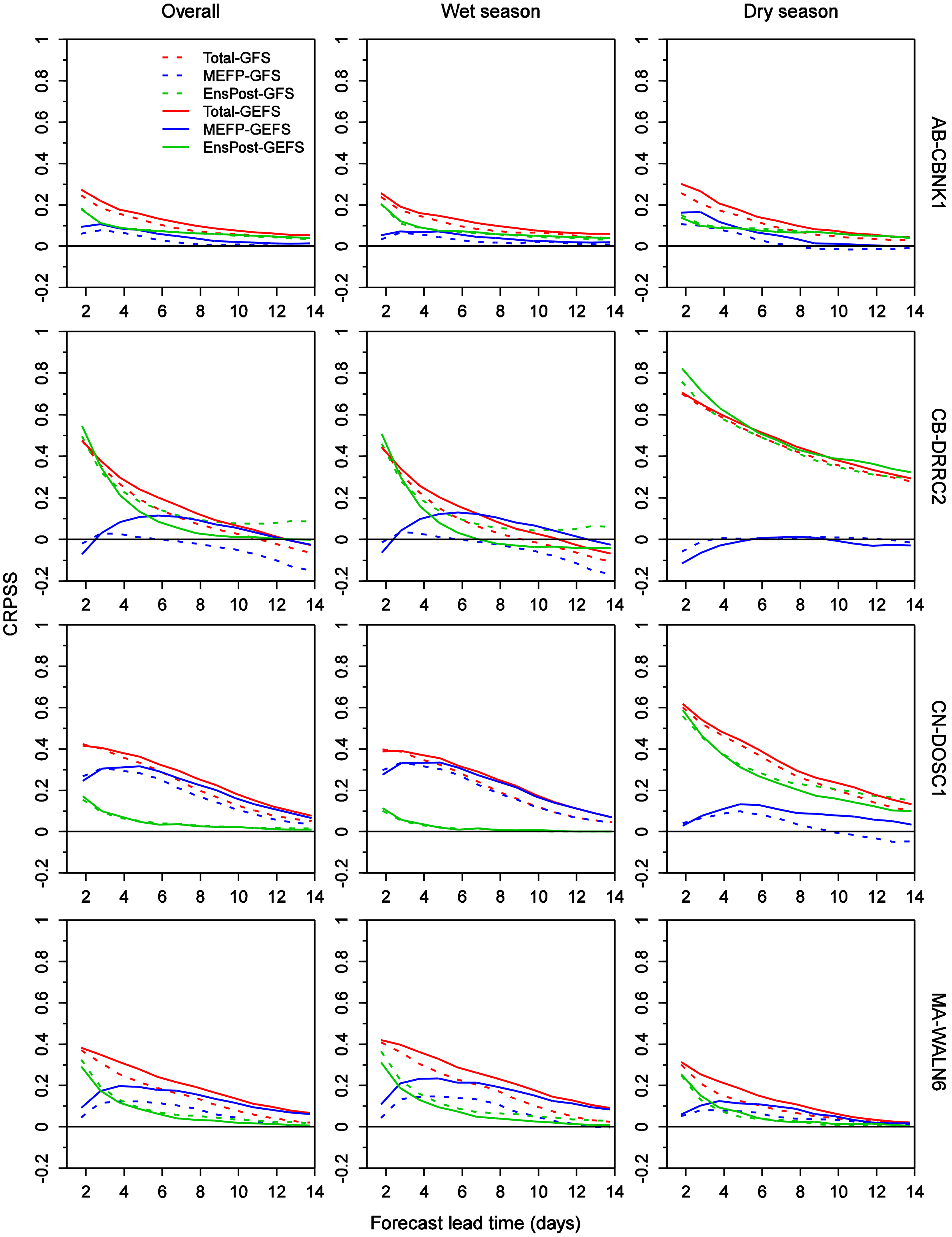

Figure 25a: Mean CRPSS of the bias-corrected MEFP-GEFS and MEFP-GFS streamflow forecasts against the raw streamflow forecasts with forcing inputs from MEFP-CLIM. The CRPSS is decomposed into contributions from the meteorological forcing and the EnsPost. The results are shown for the upstream basin in each RFC and for the dry and wet seasons, as well as the overall period.

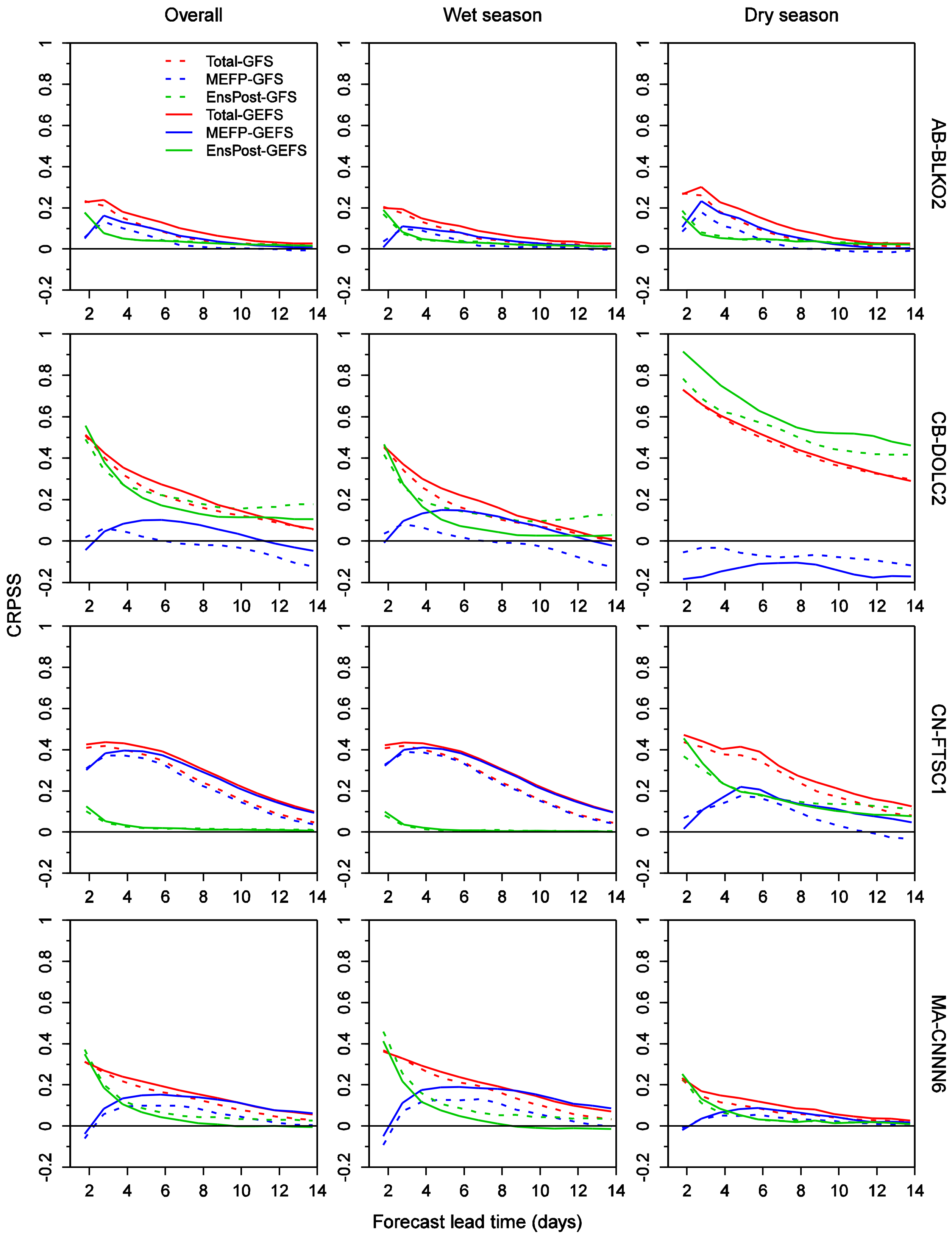

Figure 25b: Mean CRPSS of the bias-corrected MEFP-GEFS and MEFP-GFS streamflow forecasts against the raw streamflow forecasts with forcing inputs from MEFP-CLIM. The CRPSS is decomposed into contributions from the meteorological forcing and the EnsPost. The results are shown for the downstream basin in each RFC and for the dry and wet seasons, as well as the overall period.

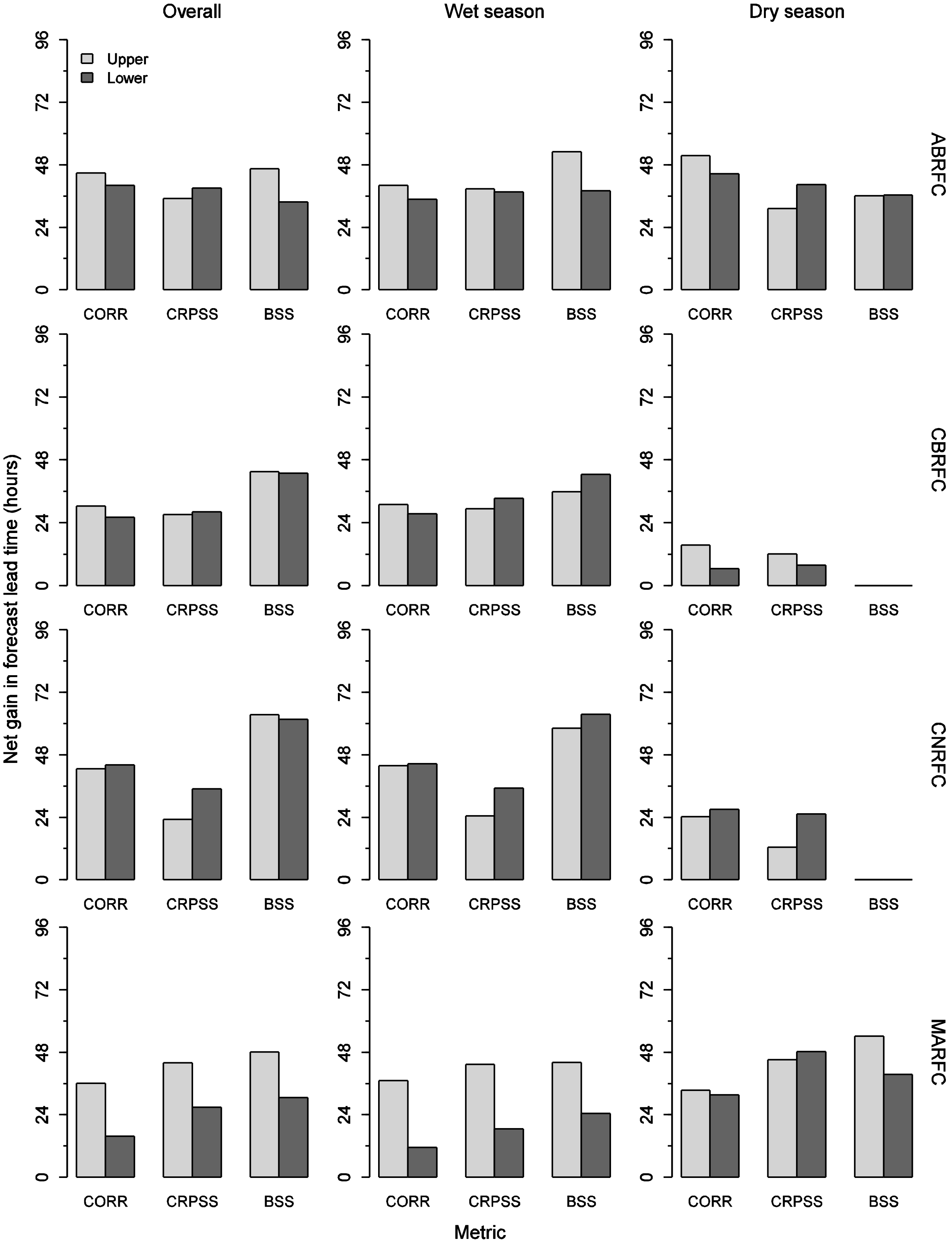

Figure 26: Average net gain in forecast lead time for the bias-corrected streamflow forecasts with forcing inputs from the MEFP-GEFS versus the MEFP-GFS. The results are shown for three verification metrics and for the wet and dry seasons, as well as the overall period. The net gain in forecast lead time was averaged across lead times of 1-7 days. The BSS comprises a threshold of C_p=0.1.

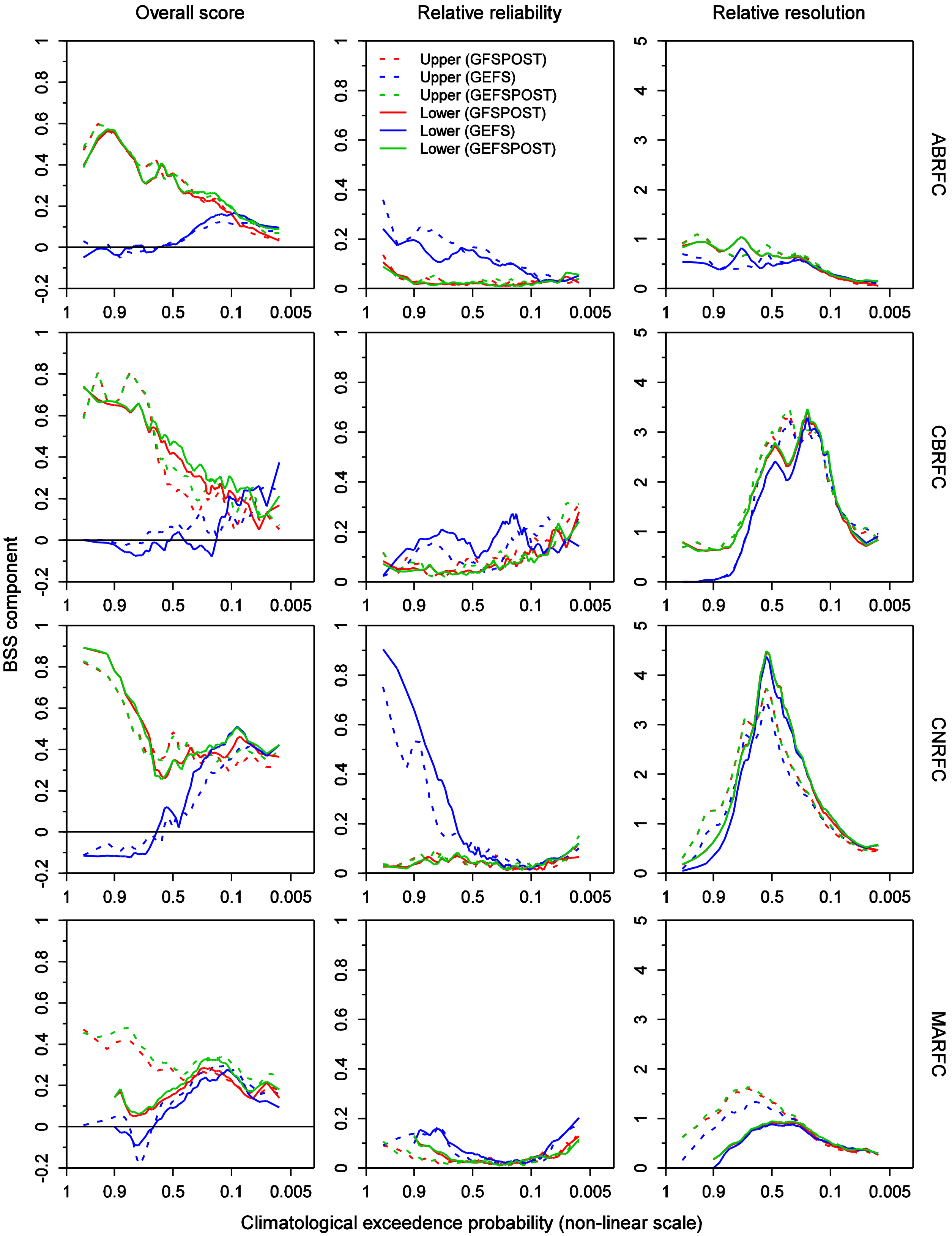

Figure 27: Calibration-refinement factorization of the BSS for the raw streamflow forecasts with GEFS forcing (GEFS) and for the post-processed streamflow forecasts with MEFP-GFS forcing (GFSPOST) and MEFP-GEFS forcing (GEFSPOST). The reference forecast is MEFP-CLIM. The results are shown for the downstream basin in each RFC at a forecast lead time of ~5 days.

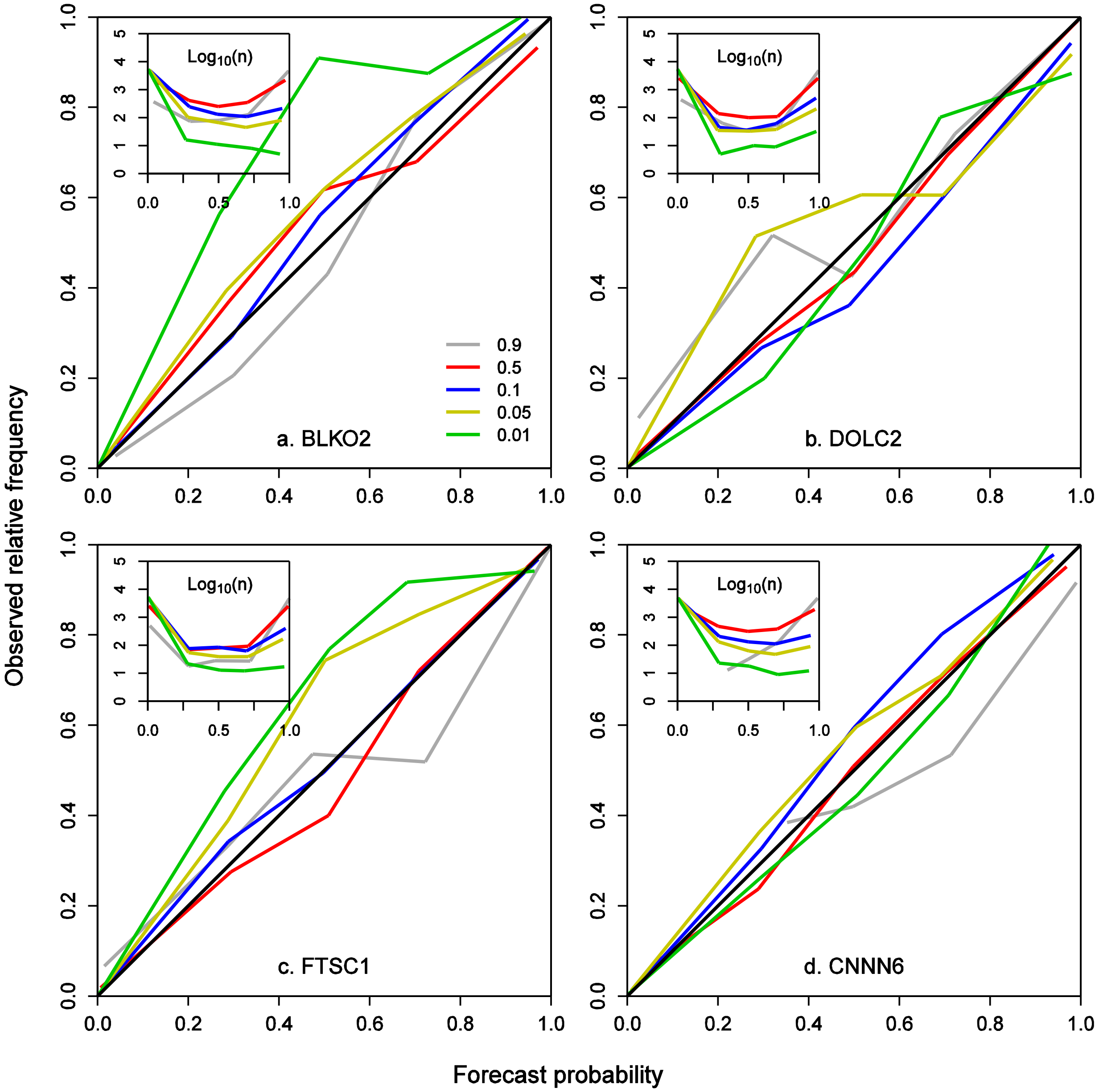

Figure 28: Reliability diagrams for the post-processed MEFP-GEFS streamflow forecasts at the downstream basin in each RFC. The results are shown for selected streamflow thresholds, which are expressed as climatological exceedence probabilities, and for a forecast lead time of ~2 days.

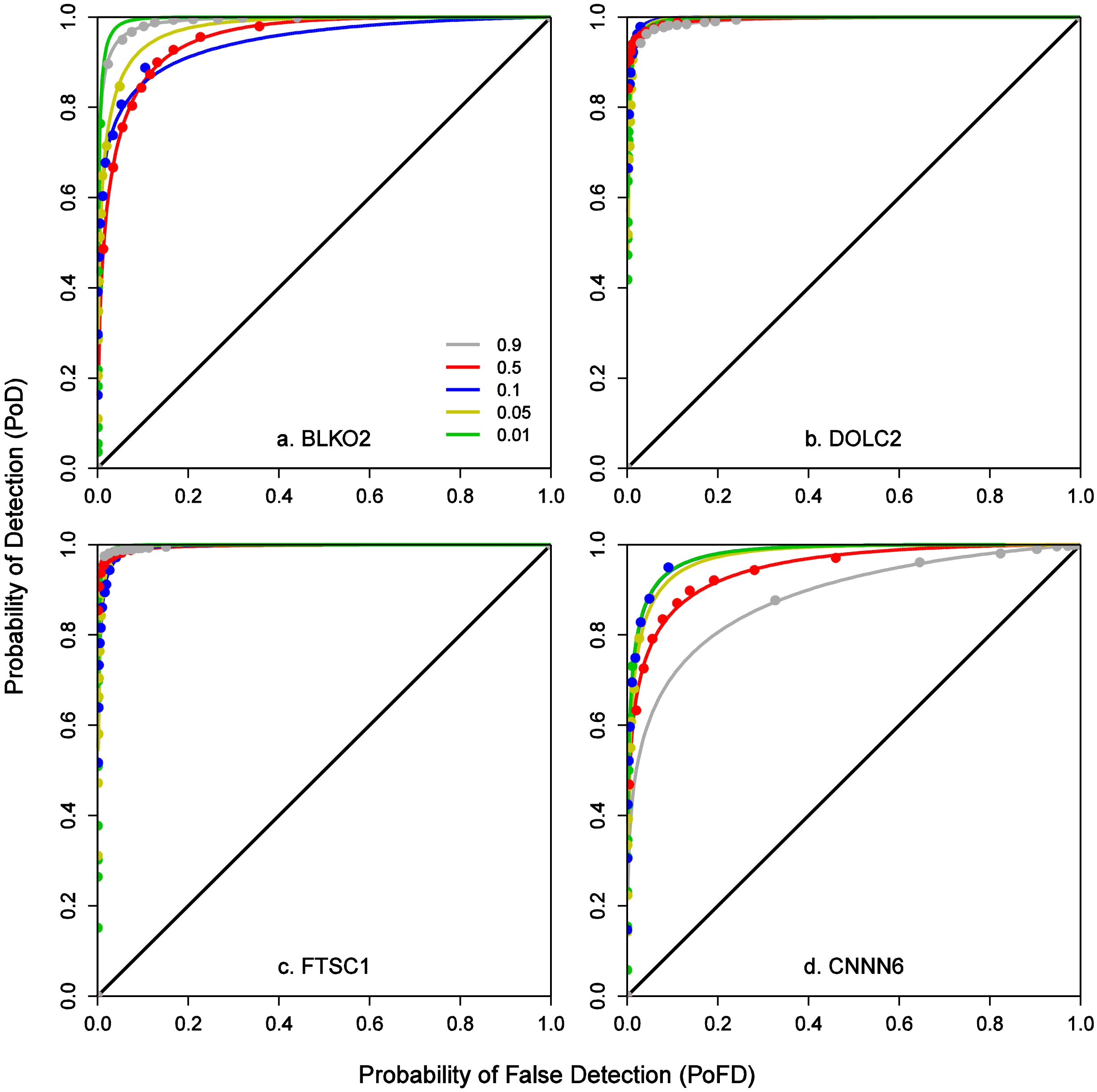

Figure 29: Relative Operating Characteristic (ROC) curves for the post-processed MEFP-GEFS streamflow forecasts. The results are shown at a forecast lead time of ~2 days for the downstream basin in each RFC and for selected streamflow thresholds, which are expressed as climatological exceedence probabilities. The ROC curves were fitted under an assumption of bivariate normality between the empirical pairs of PoD and PoFD (shown as dots).

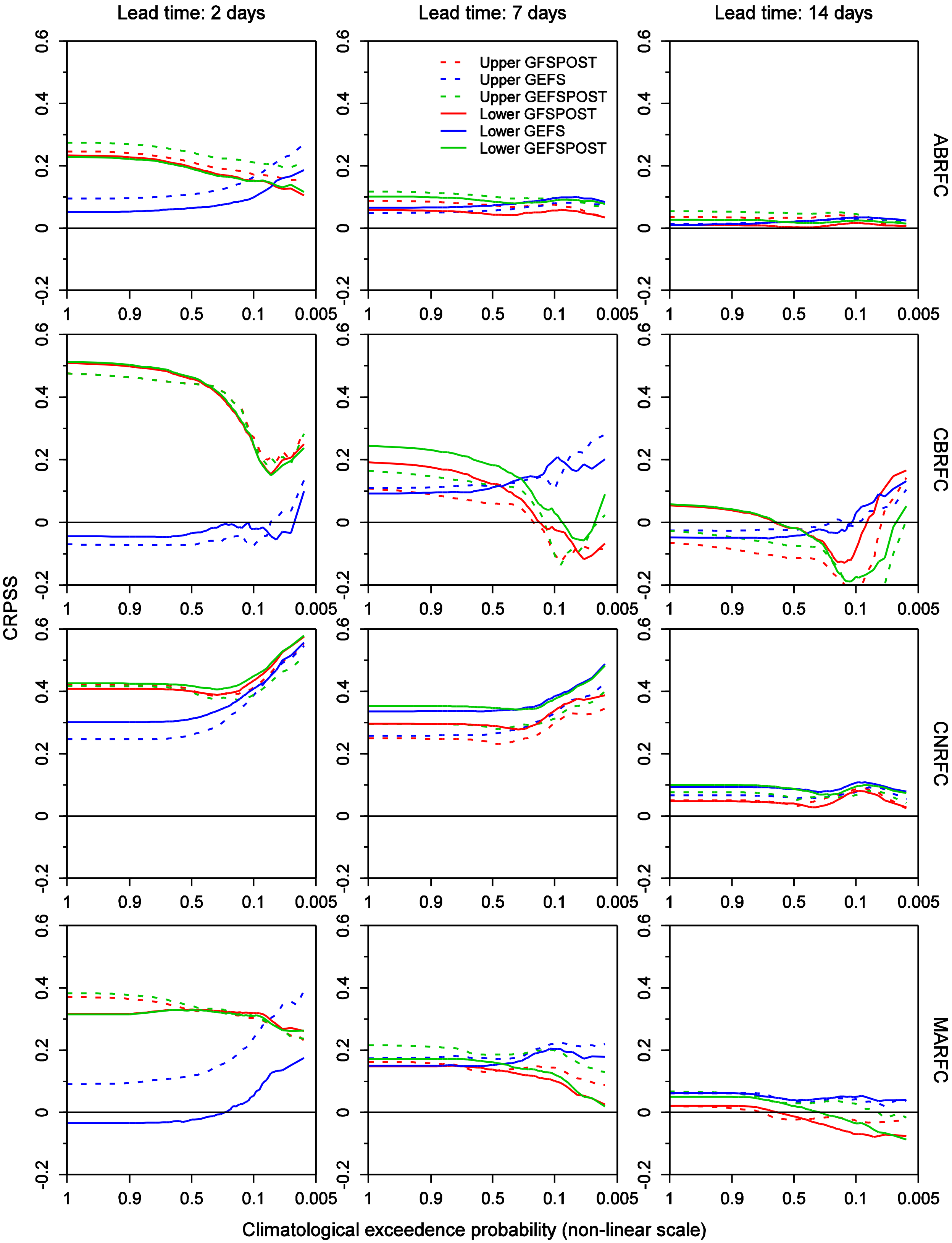

Figure 30: Mean CRPSS of the raw streamflow forecasts with GEFS forcing (GEFS) and for the post-processed streamflow forecasts with MEFP-GFS forcing (GFSPOST) and MEFP-GEFS forcing (GEFSPOST). The reference forecast is MEFP-CLIM. The results are shown for the downstream basin in each RFC and at selected forecast lead times.

APPENDIX A: The Hydrologic Ensemble Forecast Service (HEFS)

A detailed description of the Hydrologic Ensemble Forecast Service (HEFS) can be found in Seo et al. (2010) and Demargne et al. (2014), and only a brief outline is provided here. Let $\mathbf{q}_f$ denote the observed streamflow at some future times and $\mathbf{q}_c$ denote the observed streamflow up to the current time. Omitting the random variables for simplicity, the conditional distribution, $f_1(\mathbf{q}_f \,|\, \mathbf{q}_c)$, may be factored into a “raw” streamflow forecast, $f_3(\mathbf{q}_r \,|\, \mathbf{q}_c)$, and an “adjusted” streamflow forecast, given the raw forecast, $f_2(\mathbf{q}_f \,|\, \mathbf{q}_c, \mathbf{q}_r)$

$$\underbrace{f_1(\mathbf{q}_f \,|\, \mathbf{q}_c)}_{\text{Total}} = \int \underbrace{f_2(\mathbf{q}_f \,|\, \mathbf{q}_c, \mathbf{q}_r)}_{\text{Adjusted}} \; \underbrace{f_3(\mathbf{q}_r \,|\, \mathbf{q}_c)}_{\text{Raw}} \, d\mathbf{q}_r, \tag{A1}$$

where $\mathbf{q}_r$ denotes the raw model forecast (or the simulated streamflow if the adjustment can be made independently of forecast lead time). The future (observed) streamflow is then estimated by factoring out the raw forecast from the adjusted forecast. The raw forecast, $f_3(\mathbf{q}_r \,|\, \mathbf{q}_c)$, may be further separated into specific sources of uncertainty in the hydrologic modeling,

$$f_3(\mathbf{q}_r \,|\, \mathbf{q}_c) = \iiint f_4(\mathbf{q}_r \,|\, \mathbf{m}_f, \mathbf{i}, \mathbf{p}, \mathbf{q}_c) \;\; f_5(\mathbf{m}_f \,|\, \mathbf{i}, \mathbf{p}, \mathbf{q}_c) \;\; f_6(\mathbf{p} \,|\, \mathbf{i}_f, \mathbf{q}_c) \;\; f_7(\mathbf{i}_f \,|\, \mathbf{q}_c) \; d\mathbf{m}_f d\mathbf{i} \; d\mathbf{p}, \tag{A2}$$

where $\mathbf{i}$ denotes the initial conditions, $\mathbf{p}$ denotes the model parameters and $\mathbf{m}_f$ denotes the meteorological forcing. Although updating with streamflow and other observations (e.g. soil moisture) may be desirable (Liu et al, 2012), this is not currently supported by the HEFS.

The conditional distribution, $f_4(\mathbf{q}_r \,|\, \mathbf{m}_f, \mathbf{i}, \mathbf{p}, \mathbf{q}_c)$, is estimated with the HEP, which integrates the adjusted forcing from the MEFP through the hydrologic models. The MEFP generates precipitation and temperature forcing conditionally upon a raw forecast (Wu et al., 2011). The raw forcing may comprise the RFCs operational quantitative precipitation and temperature forecasts or the ensemble mean of NCEP’s GFS, among others. For gridded meteorological forecasts, the MEFP uses the raw forecast whose grid node is nearest to the basin centroid. In forming predictors from the raw forecasts, the MEFP separates the forecast horizon into multiple temporal scales. At each scale, the predictors are aggregated into time periods or "canonical events" that reflect the underlying skill in the raw forecasts at different aggregation periods. Thus, while short-range forecasts may be skillful at hourly or daily aggregations, long-range forecasts may benefit from predictors formed at larger (e.g. monthly) aggregations. By separately factoring precipitation occurrence and amount, the MEFP allows for a highly parsimonious model of $\mathbf{m}_f$ (Wu et al., 2011). The space-time covariances in $\mathbf{m}_f$ are modeled with the Schaake Shuffle, which re-orders the ensemble members to match the rank ordering of observations from similar dates in the past (see Clark et al., 2004 and Wu et al., 2011 for details). Currently, the uncertainties in the initial conditions and parameters of the hydrologic model are not modeled separately (see below).

The raw streamflow forecast is then adjusted by the EnsPost to account for any "residual" hydrologic uncertainty, not included in the raw forecast (Seo et al., 2006). This adjustment is factored into the conditional distribution, $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$. The structure and modeling of the adjusted forecast will depend on the sources of uncertainty that are addressed in the raw forecast. For example, without factoring any sources of uncertainty into $f_3(\mathbf{q}_r \mid \mathbf{q}_c)$, the adjusted forecast, $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$ may be approximated with a simple model of the total uncertainty, such that the contributions from $(\mathbf{i}, \mathbf{p}, \mathbf{m}_f)$ are lumped into $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$. Regonda et al. (2013) describe one approach to lumped modeling of $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$, known as "Hydrologic Model Output Statistics" (HMOS). Conversely, $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$ would be structureless if the hydrologic uncertainties were properly accounted for in $f_3(\mathbf{q}_r \mid \mathbf{q}_c)$. In practice, a compromise is sought in the HEFS whereby the hydrologic uncertainties $(\mathbf{i}, \mathbf{p})$ are lumped into the adjusted forecast, $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$, but the critically important meteorological uncertainties, $(\mathbf{m}_f)$, are modeled separately by the MEFP,

$$\underbrace{f_3(\mathbf{q}_r\,|\,\mathbf{q}_c)}_{\text{Raw}} = \int \underbrace{f_4(\mathbf{q}_r\,|\,\mathbf{q}_c,\mathbf{m}_f)}_{\text{Raw}\,|\,\text{Forcing}}\ \underbrace{f_5(\mathbf{m}_f)}_{\text{Forcing}}\ d\mathbf{m}_f. \tag{A3}$$

Thus, while the hydrologic uncertainties are not factored into specific contributions, their aggregate effects on $f_2(\mathbf{q}_f\,|\,\mathbf{q}_c,\mathbf{q}_r)$ are modeled by the EnsPost in a highly simplified way (Seo et al., 2006). Here, the model predicted and observed streamflows are transformed using the Normal Quantile transform (NQT; Kelly and Krzysztofowicz, 1997) and their joint distribution modeled as bivariate normal. In order to account for the temporal dependencies, future streamflows are assumed conditionally independent of past streamflows, given the present (Markov property) and an AR(1,1) structure used to model these dependencies (Seo et al., 2006). In modeling the residual uncertainty, the EnsPost assumes that the forcing ensembles are unconditionally and conditionally unbiased and that the hydrologic biases and uncertainty are independent of forecast lead time. Specifically, the model predicted streamflow, $\mathbf{q}_r$, in Eqn. A1 is substituted with simulated streamflow. This is reasonable in the context of the HEP, but implies that any residual biases in the meteorological forcing will also factor in the post-processed streamflow.

While the HEFS distinguishes between the meteorological and hydrologic uncertainties, further lumping of these uncertainties is not *necessarily* undesirable. Rather, modeling of $f_7(\mathbf{m}_f)$ is complicated by the “mixed” nature of precipitation, both in terms of precipitation occurrence and amount and liquid versus solid precipitation. It is also complicated by the sensitivity of streamflow to the correct modeling of space-time and cross-variable relationships in the forcing. The Schaake Shuffle is often used to capture these dependencies (Clark et al., 2004; Kang et al., 2010; Wu et al., 2011), but has several limitations. An intermediate solution between lumped modeling of the forcing contribution in $f_2(\mathbf{q}_f\,|\,\mathbf{q}_c,\mathbf{q}_r)$ and posterior modeling of $f_5(\mathbf{m}_f)$ may involve an *a priori* estimate of $f_5(\mathbf{m}_f)$ with a raw ensemble of meteorological forcing, together with a posterior adjustment to the streamflow for any residual forcing bias and uncertainty; that is, by substituting the raw forcing for $\mathbf{m}_f$ in Eqn. A3. This approach is used operationally by the European Floods Awareness System (EFAS; Thielen et al., 2009) and is currently being evaluated by the NWS Eastern Region as part of their Meteorological Model Ensemble Forecast System (MMEFS; Philpott et al., 2012).

The total uncertainty in Eqn. A1 is approximated, numerically, by integrating a finite number of “equally likely” ensemble members through the operational forecasting system. The HEFS is embedded within the Community Hydrologic Prediction System (CHPS), which provides the operational forecasting environment. A phased implementation of the HEFS is currently underway, with the first version (HEFSv1) due to be implemented across all RFCs by 2014. In support of this phased implementation, hindcasting and verification is being conducted at ~30 river basins in five RFCs (partly described here). The hindcasts are also being used by the NYCDEP in their Operational Support Tool (OST) for managing water supply to NYC.

APPENDIX B: Verification metrics

a. Relative mean error

The relative mean error (RME), or relative bias, measures the average difference between a set of forecasts and corresponding observations as a fraction of the average observation. Here, it measures the average difference between the ensemble mean forecast, $\overline{y}$, and the corresponding observation, x, over n pairs of forecasts and observations

$$RME = {\sum_{i=1}^{n}(\overline{y}_i - x_i)} \Big/ {\sum_{i=1}^{n} x_i} . \tag{B1}$$

The RME provides a measure of relative bias in the ensemble mean forecast, and may be positive, zero, or negative. A positive RME denotes overforecasting and a negative RME denotes underforecasting (insofar as the ensemble mean should equal the observed value).

b. Brier Score and Brier Skill Score

The Brier Score (BS; Brier, 1950) quantifies the mean square error of n forecast probabilities that Q exceeds q

$$BS = {1 \big/ n} \sum_{i=1}^{n} \left\{ F_{X_i}(q) - F_{Y_i}(q) \right\}^2, \ where\ F_{X_i}(q) = Pr\left[X_i > q\right] and\ F_{Y_i}(q) = \begin{cases} 1, Y_i > q; \\ 0,\ otherwise, \end{cases} \tag{B2}$$

where $F_{Y_i}(q)$ and $F_{X_i}(q)$ denote the ith observed and forecast probabilities that Q exceeds q, respectively. By conditioning on the forecast probability, and partitioning over J categories, the BS is decomposed into the calibration-refinement measures of Type-I conditional bias (CB) or ‘reliability’ (REL), resolution (RES), and uncertainty (UNC) (see Bradley et al., 2004 also)

$$BS = \underbrace{{}^{1}\!/_{n}\sum_{j=1}^{J} N_j \left\{F_{X_j}(q) - \bar{F}_{Y_j}(q)\right\}^2}_{\text{REL}} - \underbrace{{}^{1}\!/_{n}\sum_{j=1}^{J} N_j \left\{\bar{F}_{Y_j}(q) - \bar{F}_{Y}(q)\right\}^2}_{\text{RES}} + \underbrace{\sigma_Y^2(q)}_{\text{UNC}} . \tag{B3}$$

Here, $\bar{F}_Y(q)$ represents the average relative frequency (ARF) with which the observation exceeds q. The term $\bar{F}_{Y_j}(q)$ represents the conditional observed ARF, given that the forecast probability falls within the *j*th category, which occurs N_j times. Normalizing by the climatological variance, $\sigma_Y^2(q)$, leads to the Brier Skill Score (BSS)

$$BSS = 1 - \frac{BS}{\sigma_Y^2(q)} = \frac{RES}{\sigma_Y^2(q)} - \frac{REL}{\sigma_Y^2(q)}. \tag{B4}$$

By conditioning on the K=2 two possible observed outcomes, {0,1}, the BS is decomposed into the likelihood-base-rate measures of Type-II CB (T2), discrimination (DIS), and sharpness (SHA),

$$BS = \underbrace{{}^{1}\!/_{n}\sum_{k=1}^{K} N_k \left\{\bar{F}_{X_k}(q) - \bar{F}_{Y_k}(q)\right\}^2}_{\text{T2}} - \underbrace{{}^{1}\!/_{n}\sum_{k=1}^{K} N_k \left\{\bar{F}_{X_k}(q) - \bar{F}_{X}(q)\right\}^2}_{\text{DIS}} + \underbrace{\sigma_X^2(q)}_{\text{SHA}} . \tag{B5}$$

where $\bar{F}_{X_k}(q)$ denotes the conditional ARF that X is forecast to exceed q given that Y is observed to exceed q (k=1) or observed to not exceed q (k=2), where N_k is the conditional sample size for each case, and $\bar{F}_X(q)$ denotes the unconditional ARF. Here, $\bar{F}_{Y_k}(q)$ denotes the conditional ARF that Y is observed to exceed q. Since $\bar{F}_{Y_k}(q)$ is either zero or one, the Type-II CB can only be zero if the forecasts are perfectly sharp. Conditionally upon the observed outcome, the BSS is given by,

$$BSS = 1 - \frac{SHA}{\sigma_Y^2(q)} + \frac{DIS}{\sigma_Y^2(q)} - \frac{T2}{\sigma_Y^2(q)}. \tag{B6}$$

c. Continuous Ranked Probability Score and skill score

The Continuous Ranked Probability Score (CRPS) measures the integral square difference between the cumulative distribution functions of the observed and predicted variables

$$CRPS = \int \left\{ F_X(q) - F_Y(q) \right\}^2 dq. \quad \text{(B7)}$$

The mean CRPS comprises the CRPS averaged across n pairs of forecasts and observations. As in Eqn. B3, although with a somewhat different interpretation, the CRPS can be factored into a combination of reliability, resolution and uncertainty (see Hersbach, 2000). The Continuous Ranked Probability Skill Score (CRPSS) is a ratio of the mean CRPS of the main prediction system, $\overline{CRPS}$, and a reference system, $\overline{CRPS}_{REF}$

$$CRPSS = \frac{\overline{CRPS}_{REF} - \overline{CRPS}}{\overline{CRPS}_{REF}}. \quad \text{(B8)}$$

d. Relative Operating Characteristic

The Relative Operating Characteristic (ROC; Green and Swets, 1966) measures the ability of a forecasting system to correctly predict the occurrence of an event (Probability of Detection or PoD) while avoiding too many incorrect forecasts when it does not occur (Probability of False Detection or PoFD). For probability forecasts, this trade-off is expressed as a probability threshold, d, at which the forecast triggers a decision. 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

$$PoD = \sum_{i=0}^{n} I_{X_i}\left(F_{X_i}(q) > d \mid Y_i > q\right) \Big/ \sum_{i=0}^{n} I_{Y_i}(Y_i > q). \quad \text{(B9)}$$

where I denotes the indicator function. The empirical PoFD is

$$PoFD = \left. \sum_{i=0}^{n} I_{X_i}\left(F_{X_i}(q) > d \mid Y_i \le q\right) \middle/ \sum_{i=0}^{n} I_{Y_i}(Y_i \le q) \right. . \tag{B10}$$

Here, the relationship between the PoD and PoFD is assumed bivariate normal (Hanley, 1988; Metz and Pan, 1999)

$$PoD = \Phi\left\{a + b\Phi^{-1}\left(PoFD\right)\right\} \; where \; a = \frac{\mu_{PoD} - \mu_{PoFD}}{\sigma_{PoD}} \; and \; b = \frac{\sigma_{PoFD}}{\sigma_{PoD}}, \tag{B11}$$

and Φ is the cumulative distribution function of the standard normal distribution. The means of the PoD and PoFD are μ_{PoD} and μ_{PoFD}, respectively, and their corresponding standard deviations are σ_{PoD} and σ_{PoFD}. Calculation of the fitted ROC amounts to estimating the parameters, a and b, of the linear relationship between the PoD and the PoFD in normal space, for which Ordinary Least Squares regression was used.

APPENDIX C: Event-based analysis of the streamflow forecasts

Paired streamflow forecasts and observations are presented for selected years in the downstream basin of each RFC. The results comprise the bias-corrected streamflow forecasts with forcing inputs from the MEFP-GFS and MEFP-GEFS. The results are also shown for the raw streamflow forecasts with climatological forcing. The plots include the single-valued streamflow observations and simulations, together with the ensemble range of the corresponding streamflow forecast (maximum – minimum value) on each forecast valid date during one calendar year. The results are shown at forecast lead times of ~18-42 hours, ~42-66 hours, ~162-186 hours and ~306-330 hours and for calendar years 1986, 1990, 1994, and 1998. The plots support visual inspection of the HEFS streamflow forecasts, including timing and amplitude errors for specific hydrologic events and in different portions of the streamflow hydrographs. However, some care (and subjective interpretation) is needed in separating between random and systematic behaviors over a small number of hydrologic events. Thus, the plots should only be viewed as supplementary to the verification results presented in Section 5 of this report.

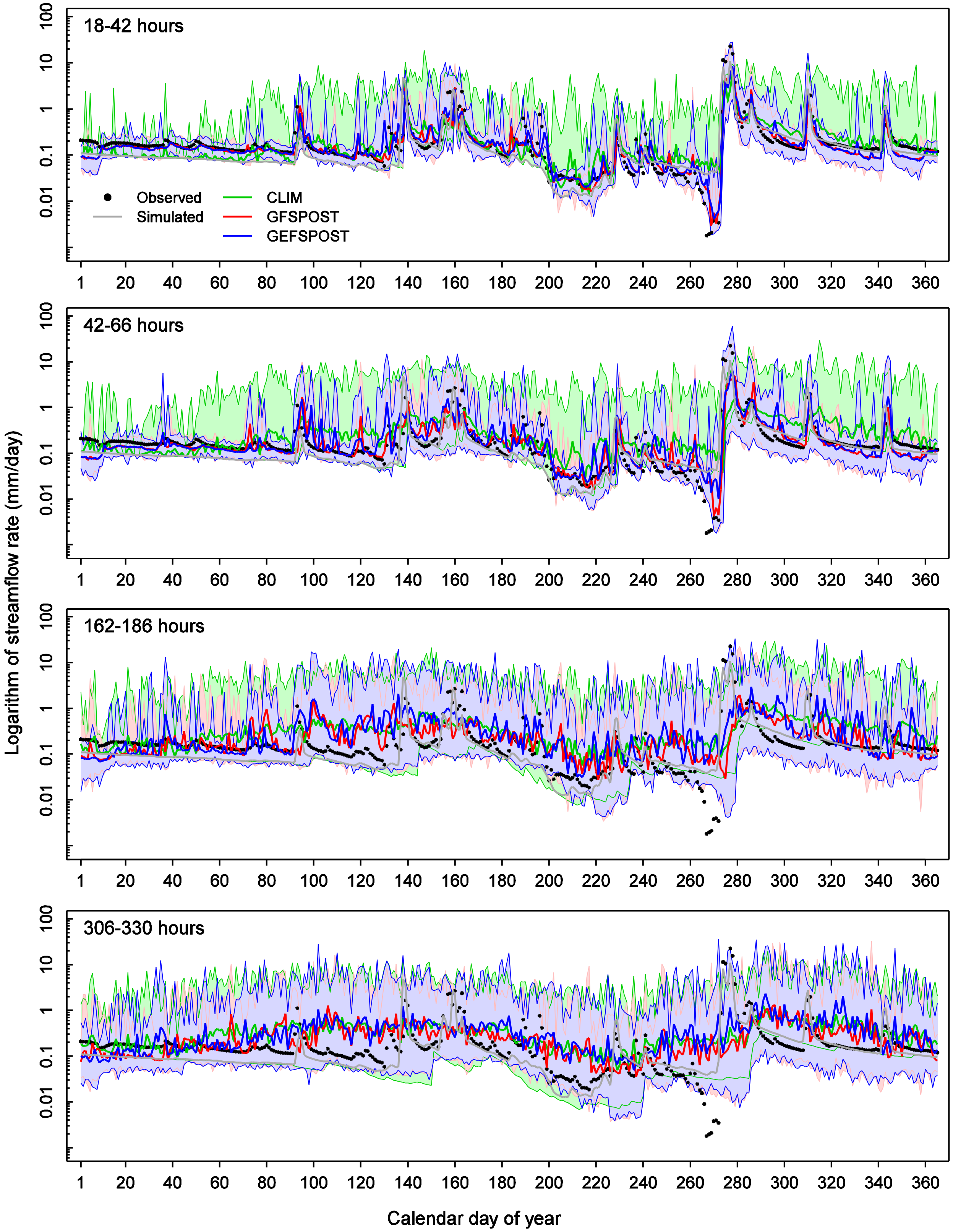

Figure C01: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

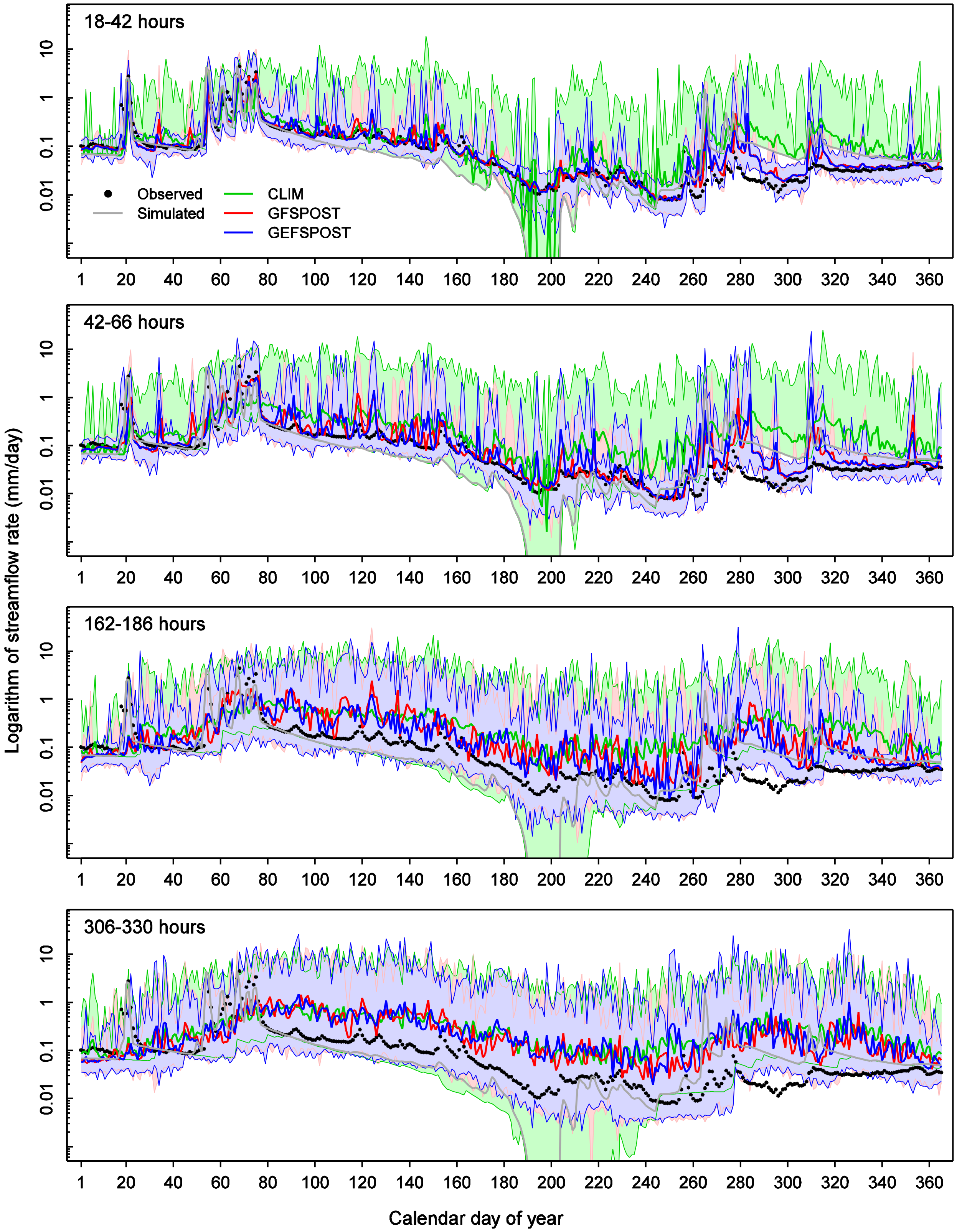

Figure C02: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

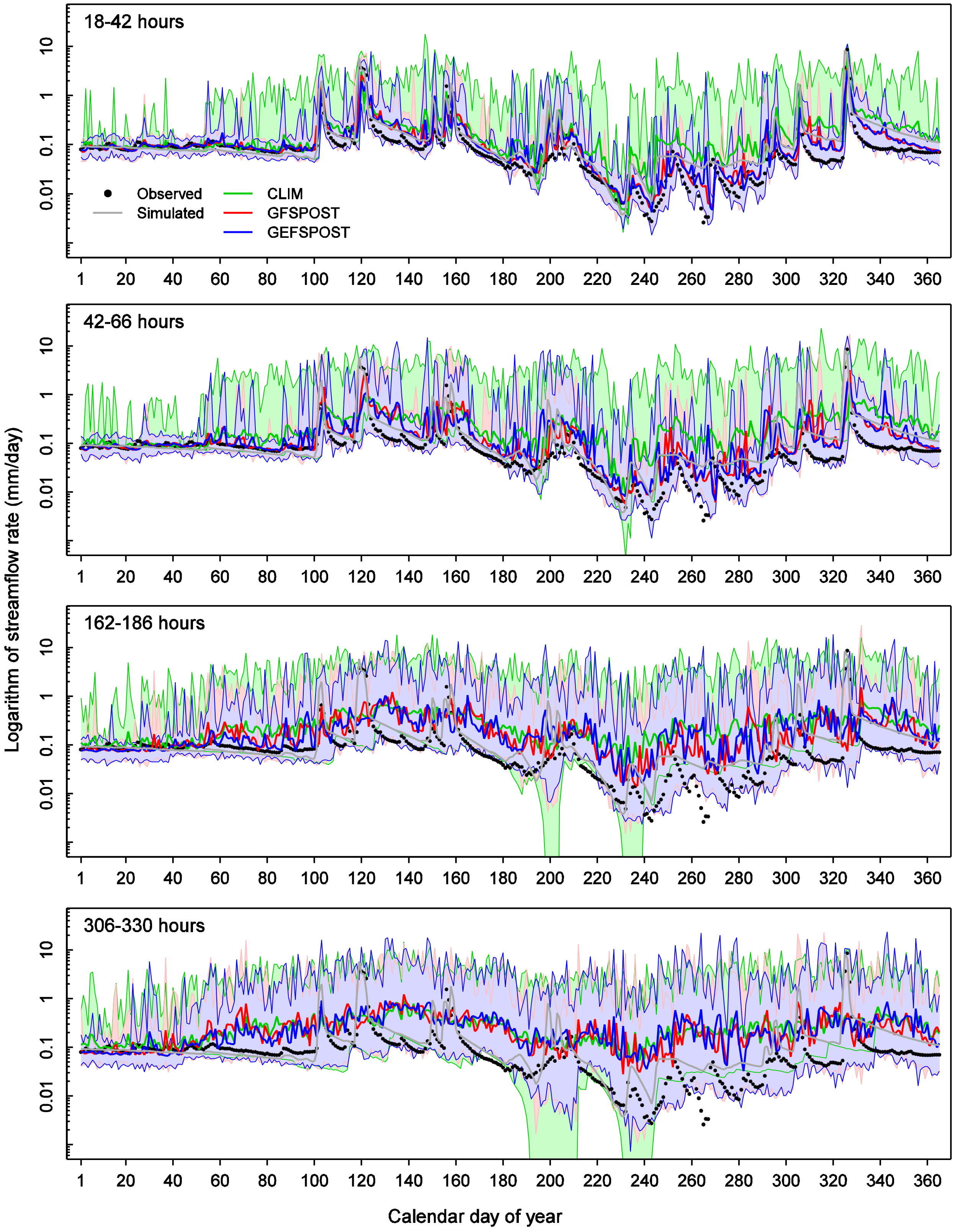

Figure C03: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1994 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

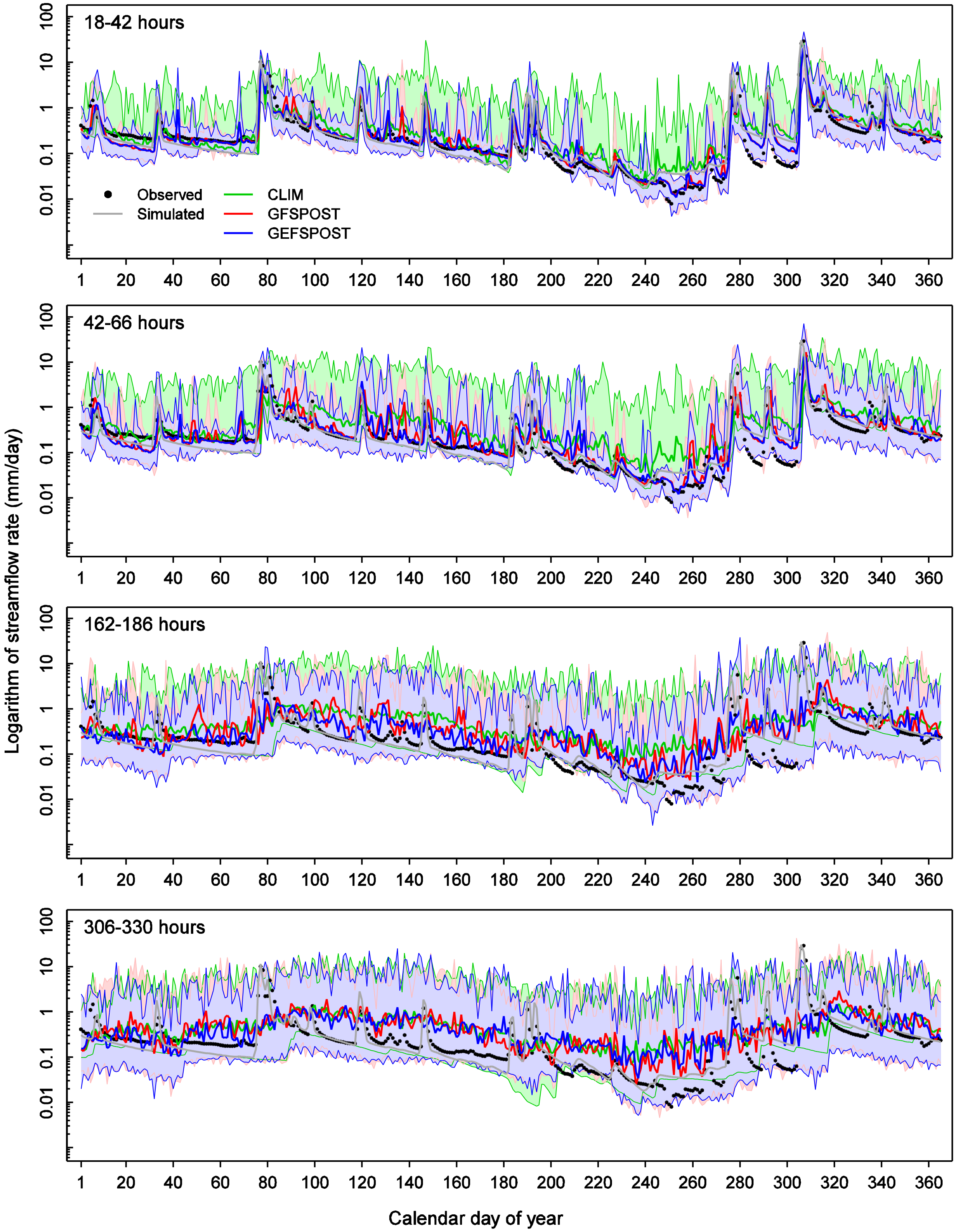

Figure C04: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1998 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

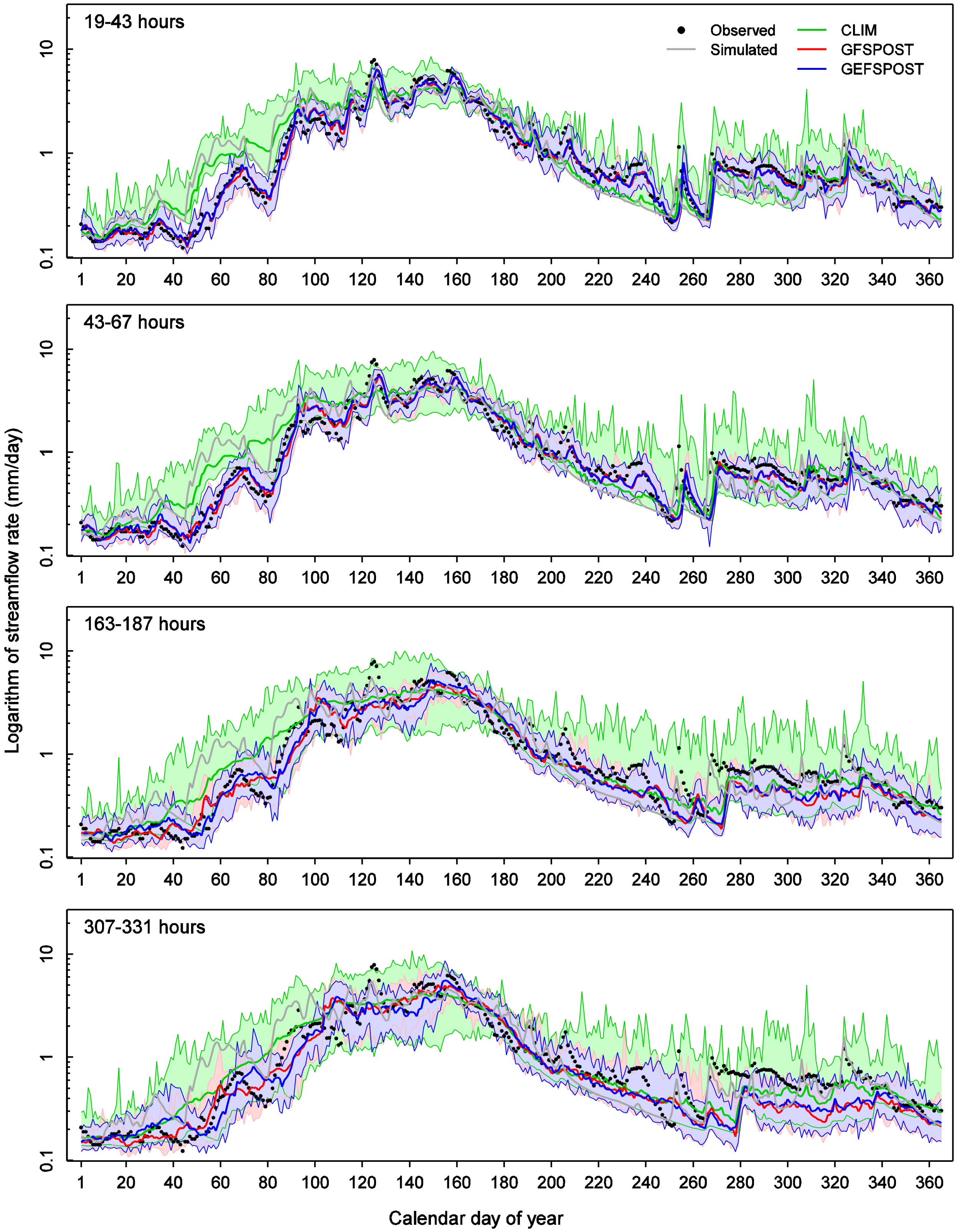

Figure C05: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

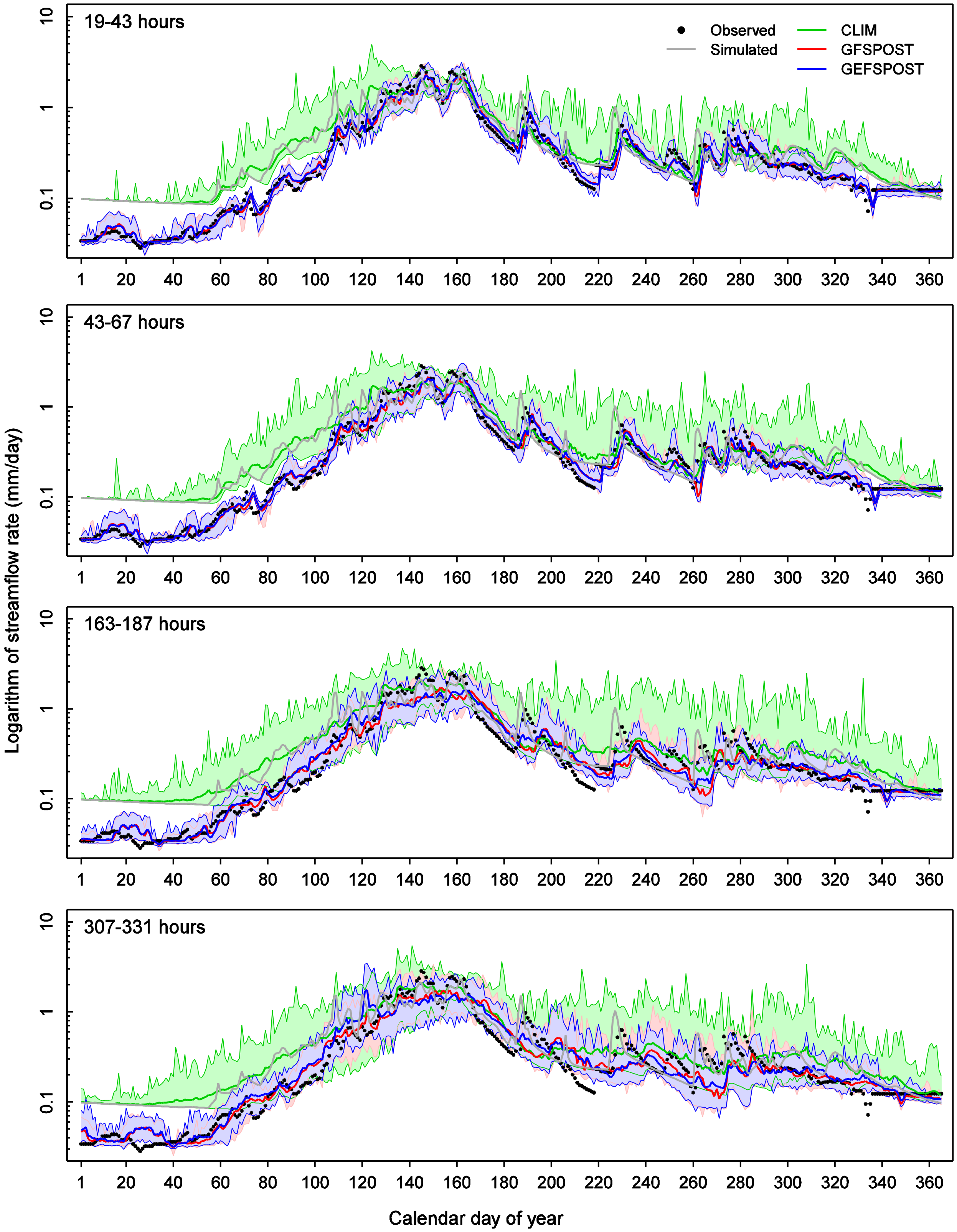

Figure C06: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

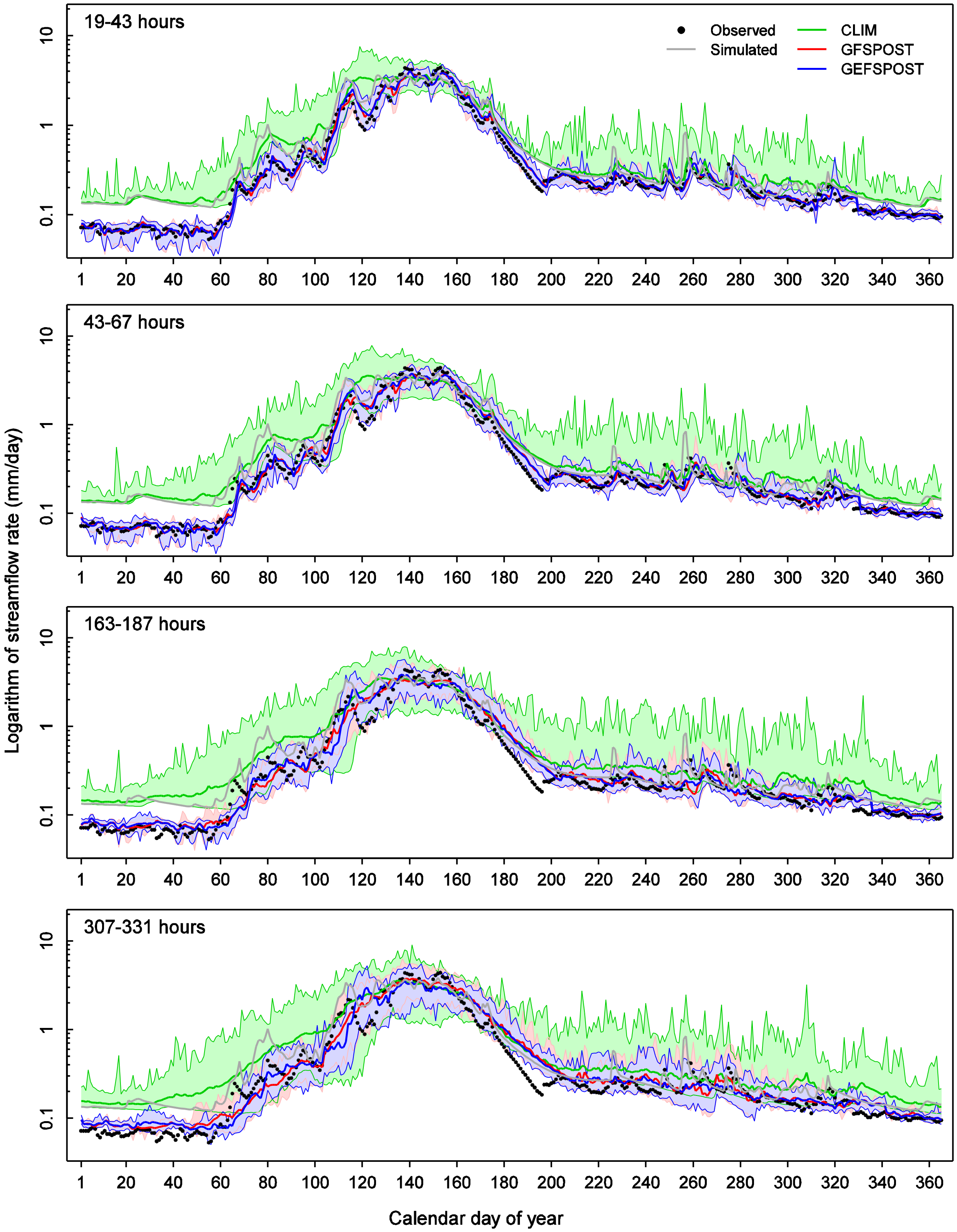

Figure C07: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1994 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

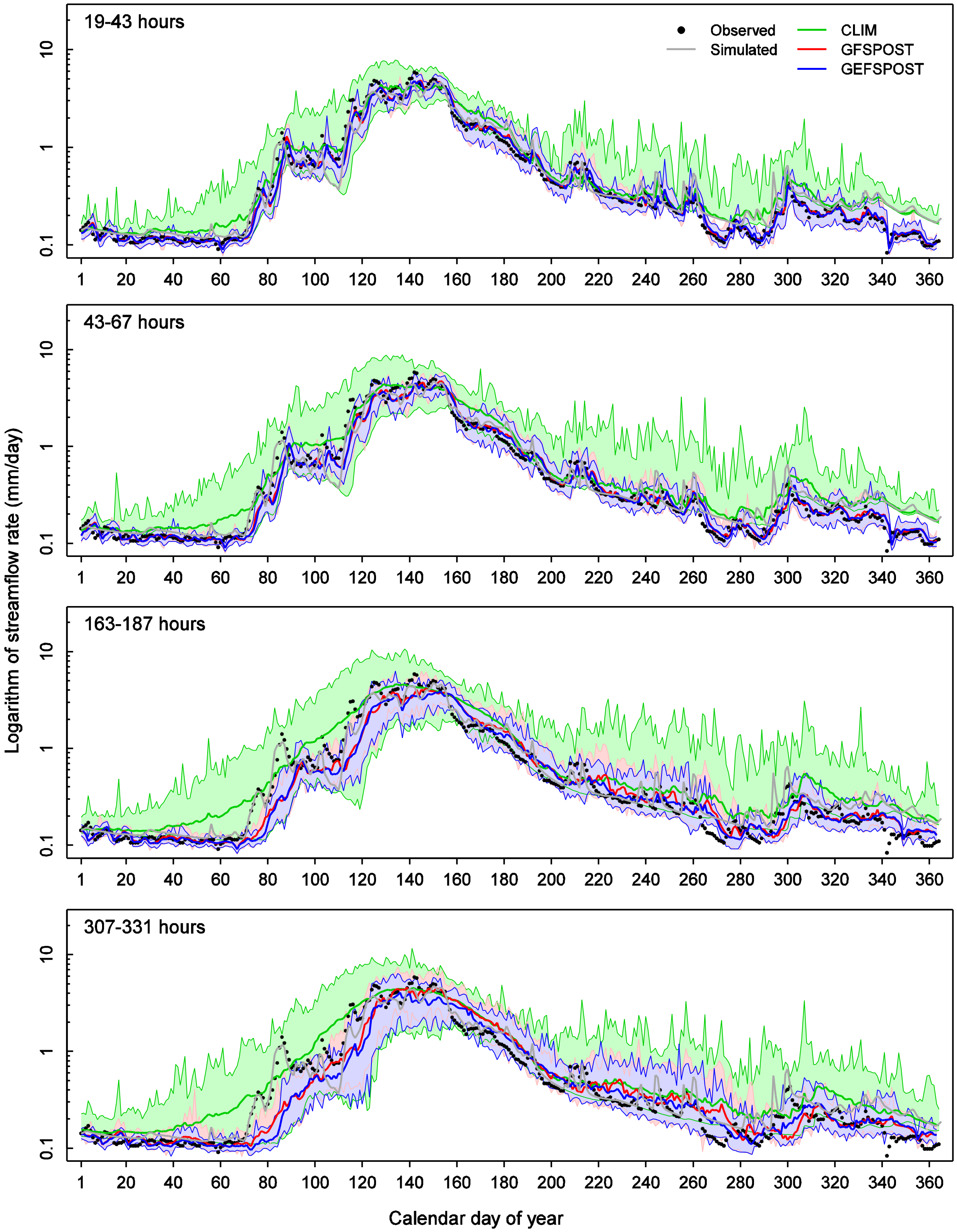

Figure C08: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1998 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

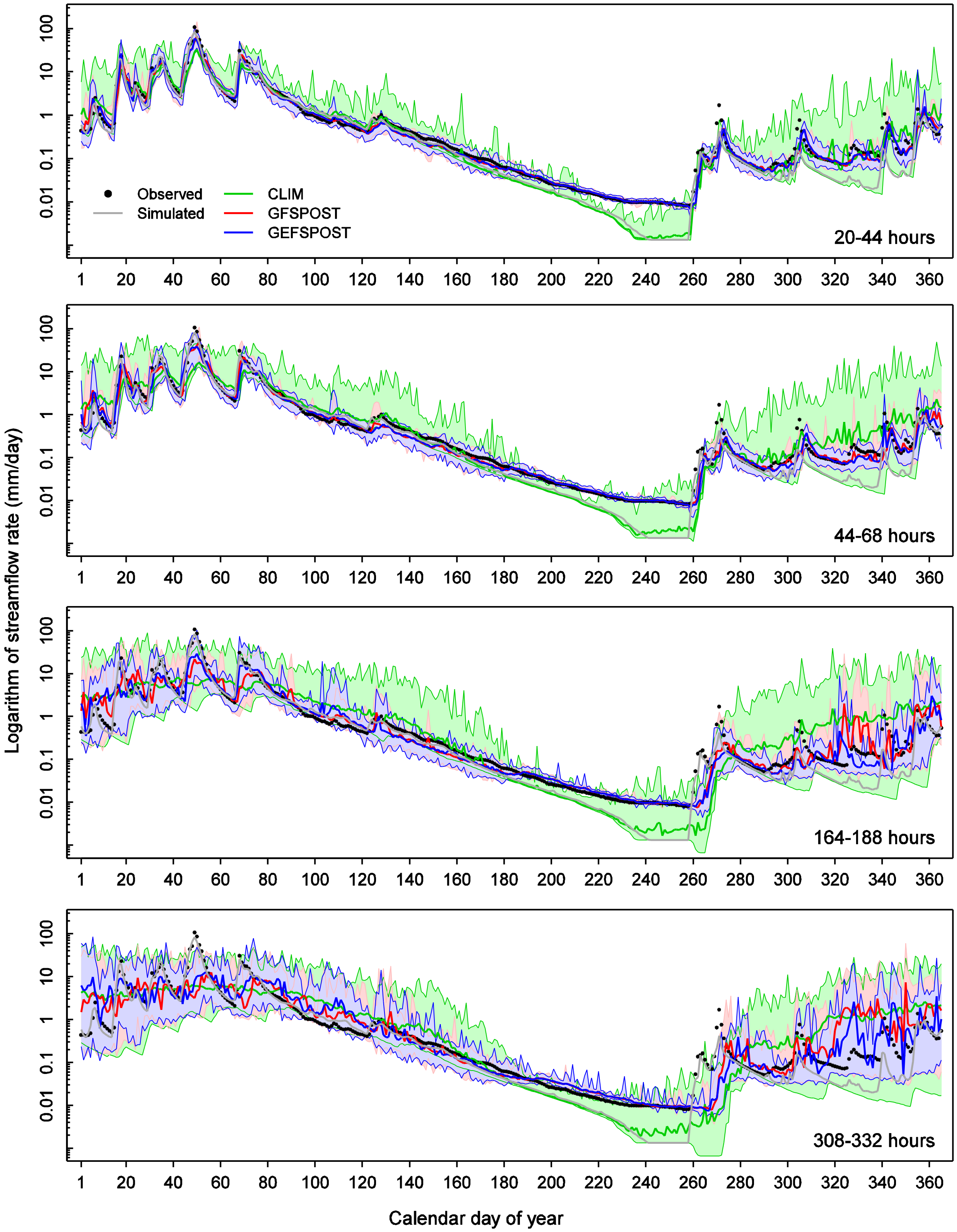

Figure C09: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

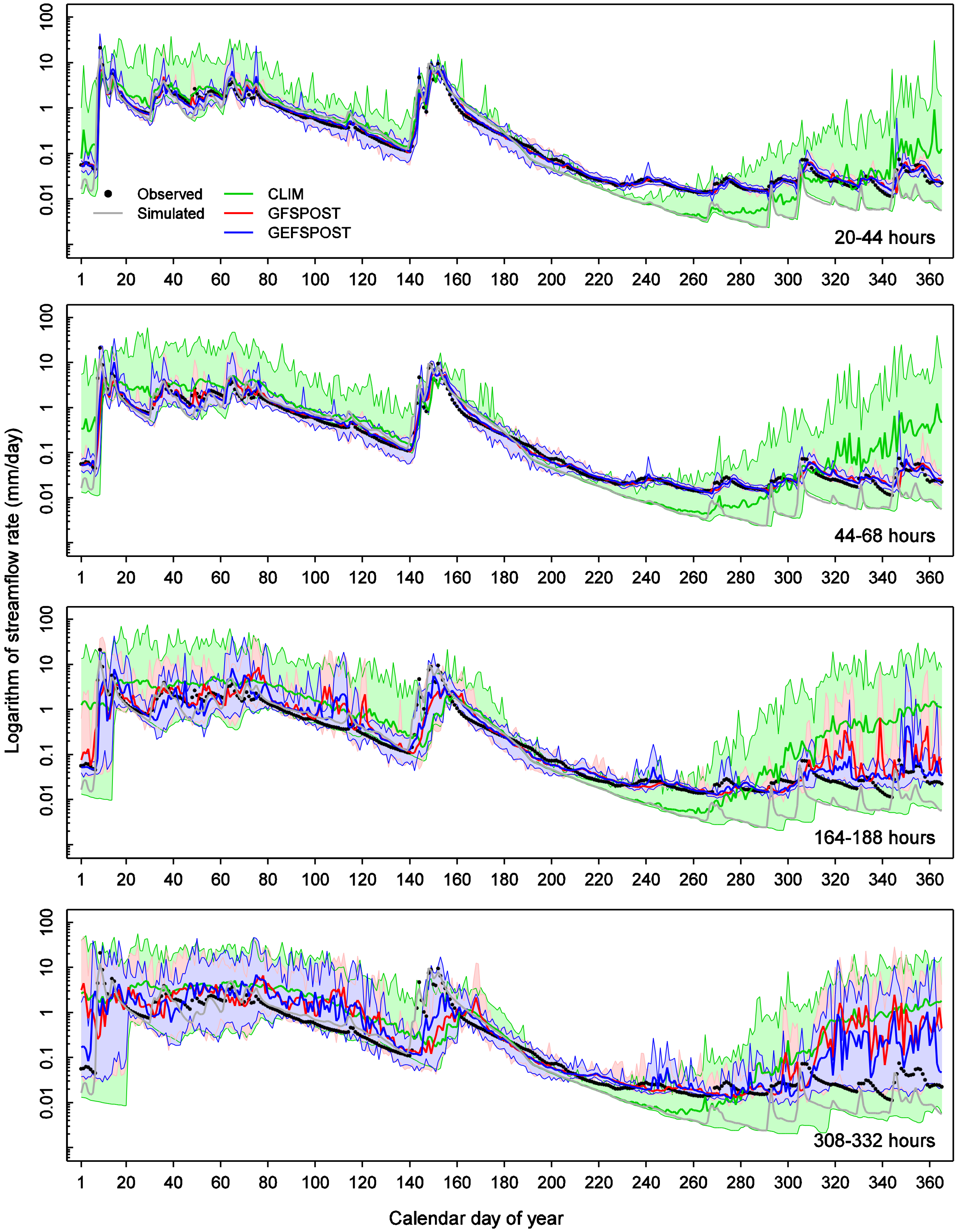

Figure C10: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

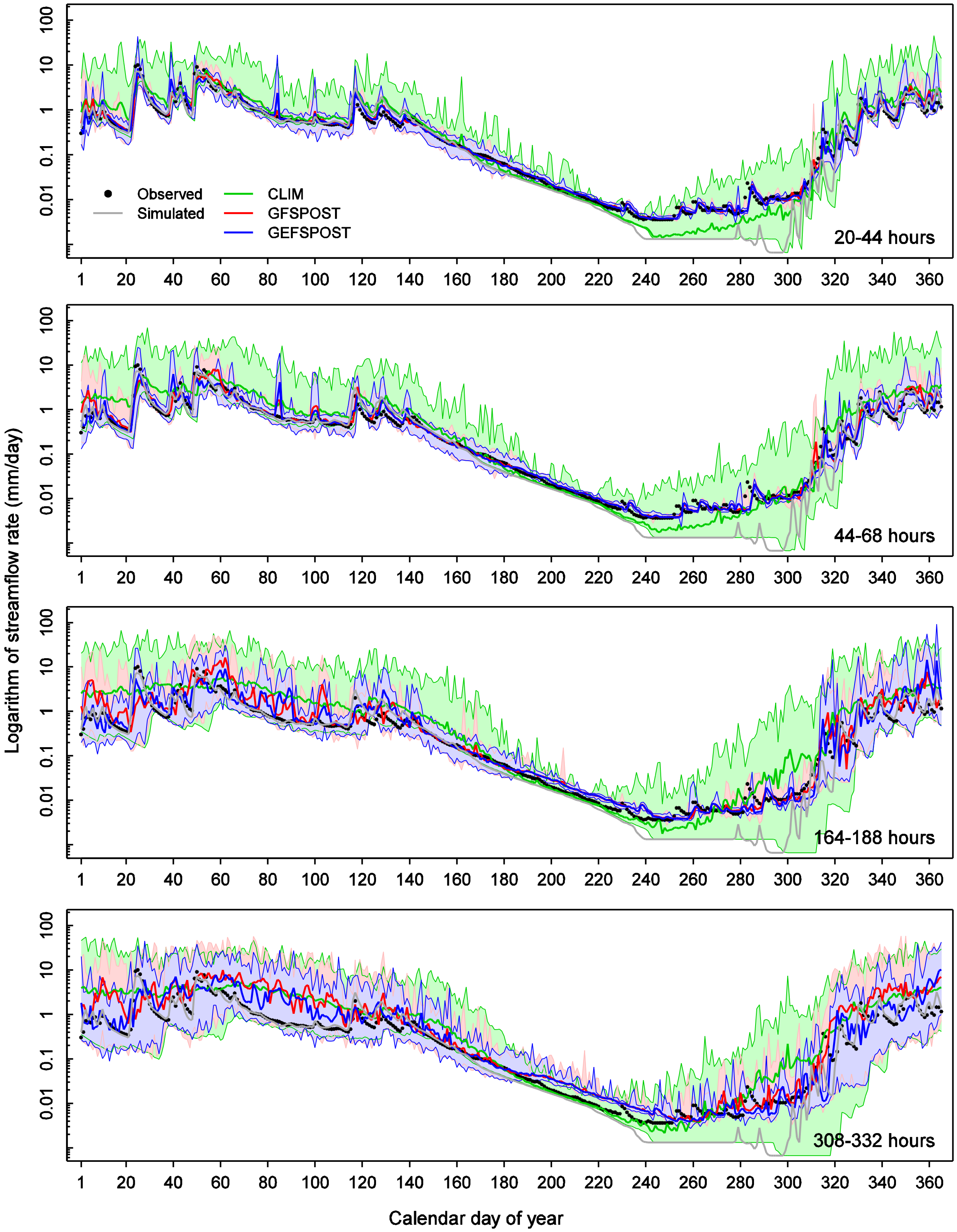

Figure C11: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1994 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

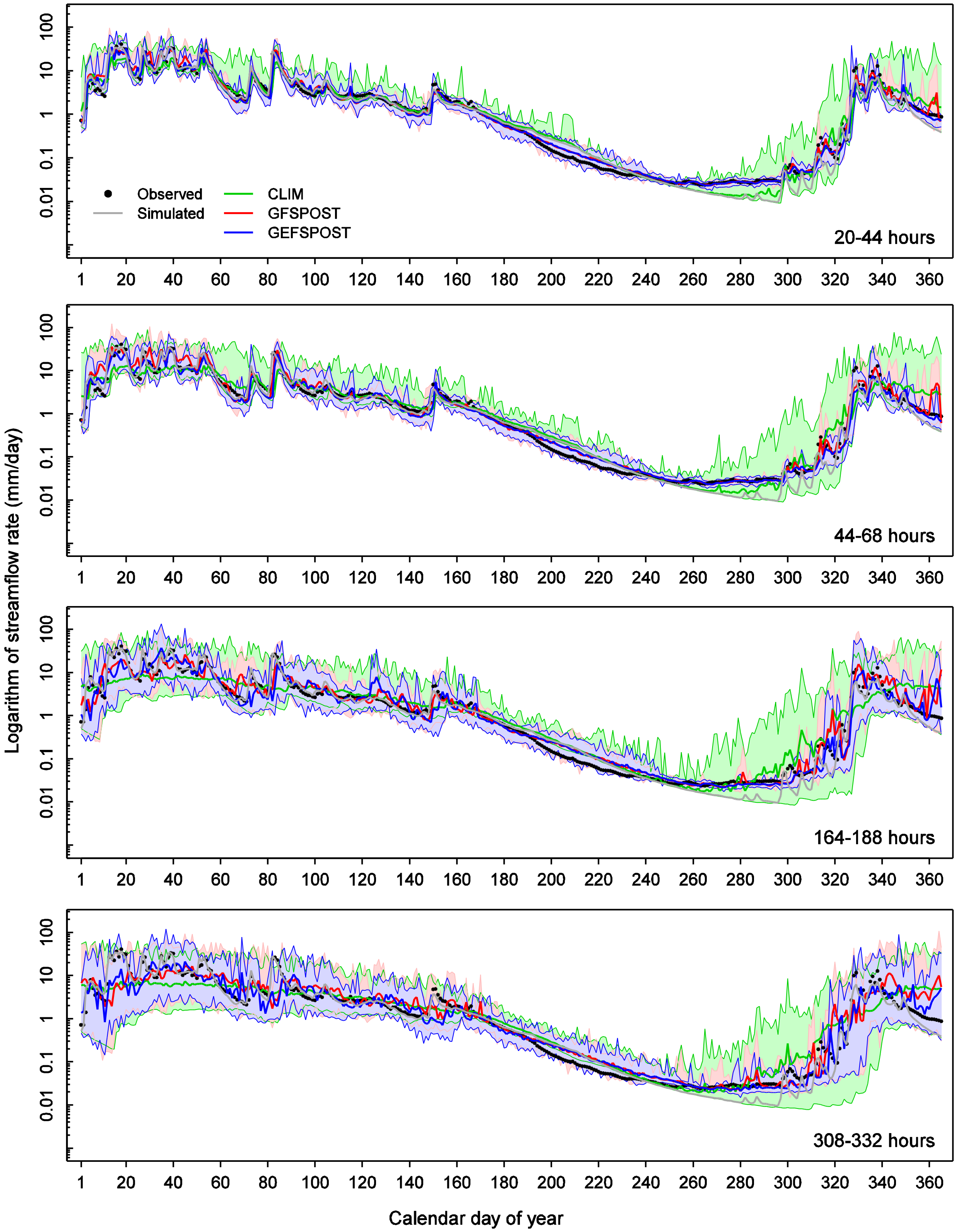

Figure C12: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1998 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).).

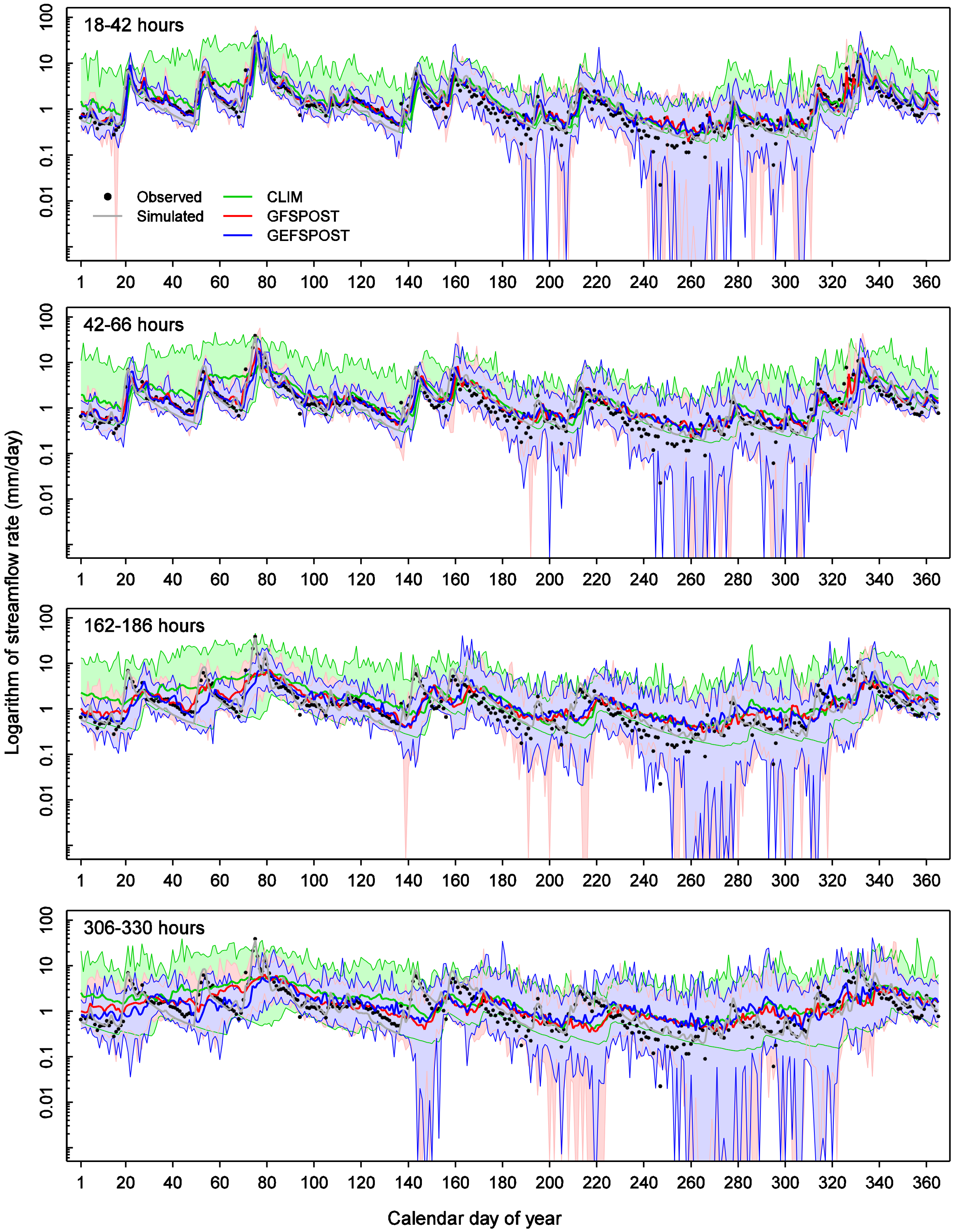

Figure C13: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

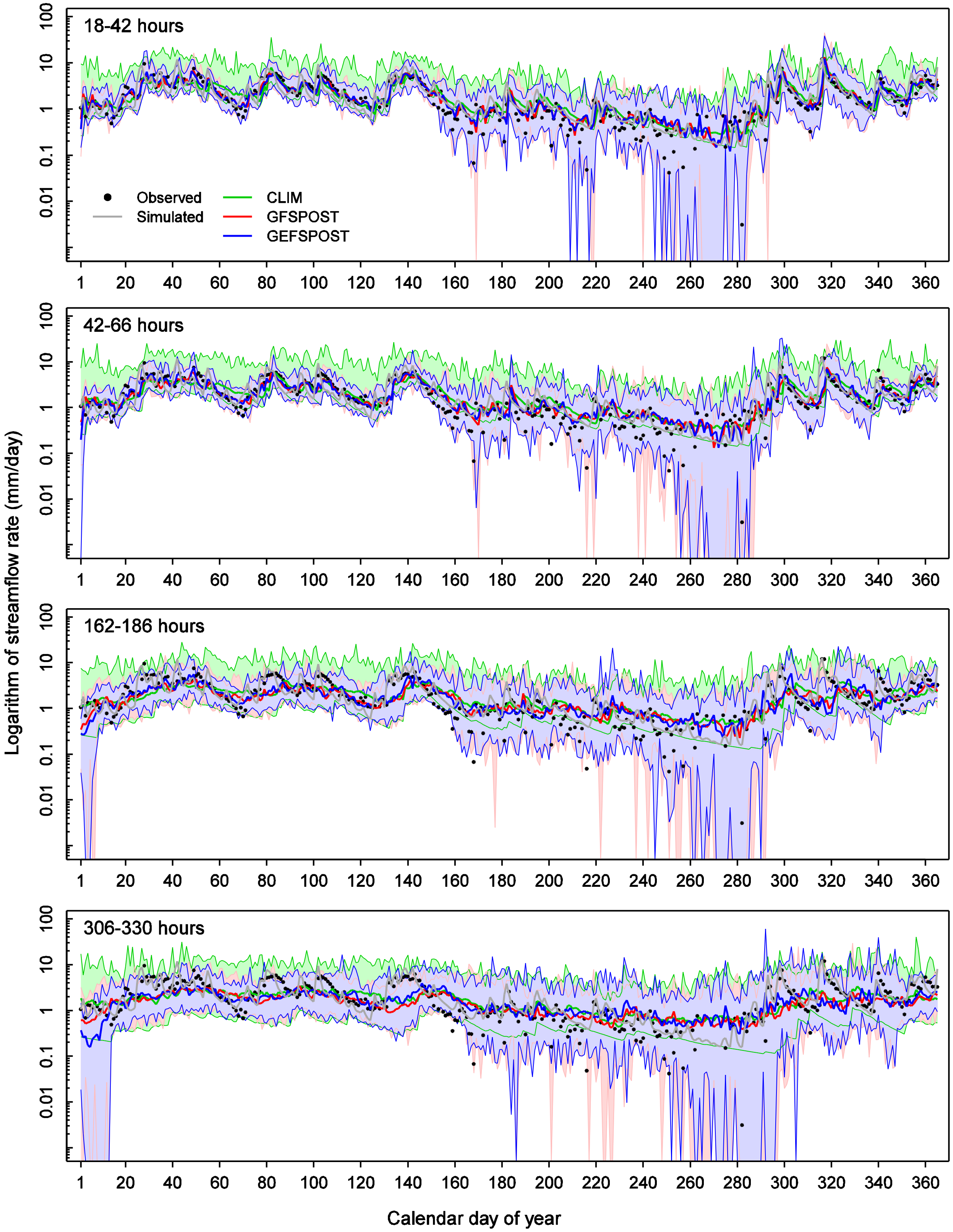

Figure C14: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

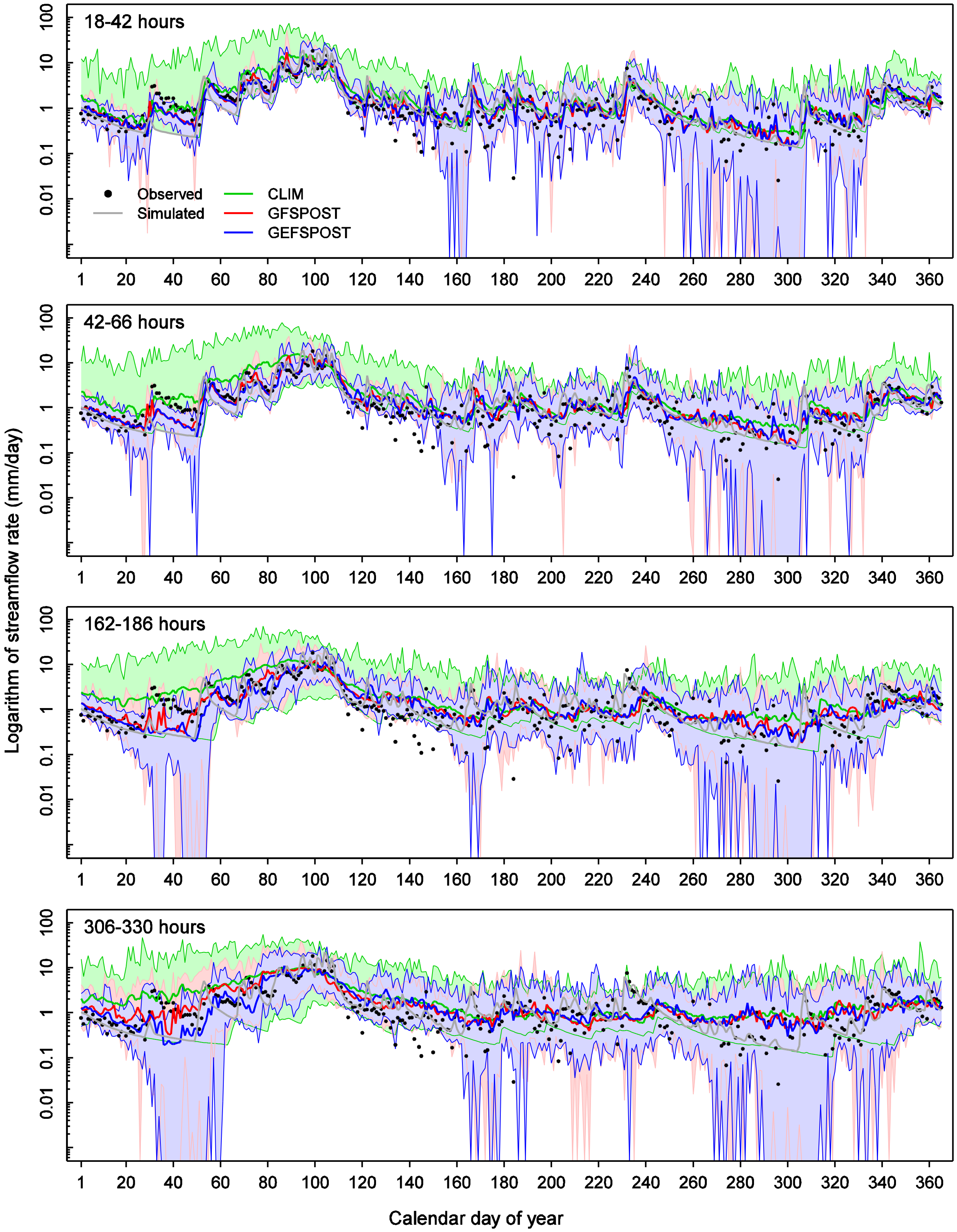

Figure C15: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1994 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).

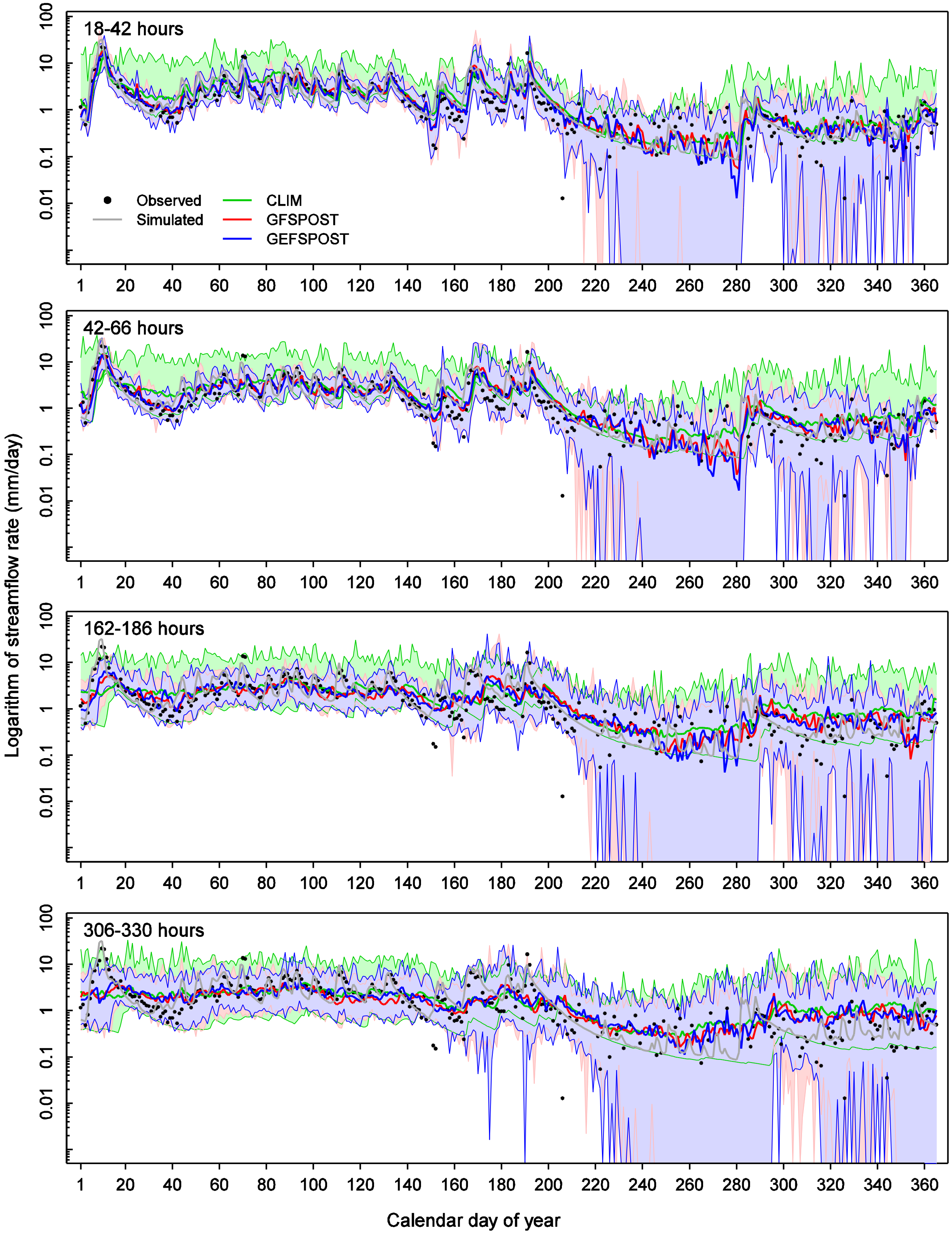

Figure C16: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1998 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM). The post-processed streamflow forecasts comprise forcing from the MEFP-GFS (GFSPOST) and the MEFP-GEFS (GEFSPOST).