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

Revision: final

Prepared by Hydrologic Solutions Limited for the U.S. National Weather Service under Subcontract Agreement 2012-04 with Len Technologies Incorporated (in fulfillment of Deliverable No. 2)

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

Thursday, May 16, 2013

Abstract

Retrospective forecasts of precipitation, temperature 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) using forcing inputs from the frozen version of the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS).The streamflow hindcasts cover a ~20 year period from 1979-1999 with a forecast horizon of 14 days. The hindcasts are verified conditionally upon forecast lead time, magnitude of the observed and forecast variables, season, and aggregation period. 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). In order to distinguish between the contributions of the MEFP and the EnsPost to the quality of the streamflow forecasts, verification is performed against simulated streamflow (removing hydrologic biases) and against observed streamflow. Interpretation of the verification results leads to guidance on the expected performance and limitations of the HEFS for medium-range forecasting with the GFS, together with recommendations on future enhancements.

Document history

Action	Person edited	Date
Complete first draft	James Brown	12th February 2013
Added Appendix C (suggested by Haksu Lee)	James Brown	15th February 2013
Incorporated feedback from Limin Wu	James Brown	20th February 2013
Incorporated feedback from Satish Regonda	James Brown	26th February 2013
Added section “How to read this report”	James Brown	27th February 2013
Incorporated feedback from Kevin He	James Brown	28th February 2013
Incorporated feedback from Haksu Lee	James Brown	28th February 2013
Added a glossary of terms	James Brown	1st March 2013
Completed final revisions to first draft	James Brown	3rd March 2013
Second draft to include reviewer comments	James Brown	27th March 2013
Incorporated feedback from Geoff Bonnin	James Brown	27th March 2013
Incorporated feedback from Mark Fresch	James Brown	28th March 2013
Incorporated feedback from Andy Wood	James Brown	4th April 2013
Incorporated feedback from Ernie Wells	James Brown	8th April 2013
Final version	James Brown	12th April 2013

Acknowledgements

This report was prepared by Hydrologic Solutions Limited under Subcontract Agreement 2012-04 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. The report was reviewed by Haksu Lee, Limin Wu, Satish Regonda, Kevin He, Geoff Bonnin, Mark Fresch, Ernie Wells, Andy Wood and D-J Seo.

CONTENTS

1. How to read this report

The aims of this report are twofold, namely to: a) provide a detailed scientific evaluation (verification) of the medium-range temperature, precipitation and streamflow forecasts from the HEFS; and b) to communicate the strengths and weaknesses of the HEFS to operational forecasters. In keeping with a), one or more scientific papers will be developed from this report. This section aims to guide readers with limited time or experience of ensemble forecasting or verification to the main results and conclusions. For these readers, the following sections are particularly important:

I. Executive summary and recommendations. This describes the structure of the study 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 (and understanding the risks of extrapolation);

III. Appendix C. This shows a selection of the paired streamflow forecasts and observations from which the verification results are derived. The relative scatter of the observations within the ensemble forecast distribution provides some insight into the quality of the streamflow forecasts. In general, the 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. The results are shown for different forcing inputs and before and after streamflow post-processing;

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). Key attributes of forecast quality are briefly described in Section

4.4, while Appendix B summarizes the key measures of forecast quality used throughout this report. In addition, a Glossary of terms and acronyms is provided towards the end of the report; and

V. Section 5.3. The verification results are presented separately for the meteorological forecasts, the "raw" streamflow forecasts and the bias-corrected streamflow forecasts. Products developed for operational use will generally comprise the bias-corrected streamflow forecasts for which verification results are presented in Section 5.3.

Some plots are simpler to understand than others. In general, plots that show specific attributes of forecast quality are more informative, but require more technical knowledge. 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 over another (0→1, although negative values are possible). For example, Figure 5 shows the fractional improvement of the MEFP precipitation forecasts with GFS forcing versus "raw" climatology and with an enhanced or "resampled" climatology. Figure 25 shows the overall skill of the post-processed streamflow forecasts with GFS forcing. The baseline comprises the streamflow forecasts with climatological forcing and without streamflow post-processing. In addition to the overall skill, the contributions from the MEFP with GFS forcing and the EnsPost are shown separately. Figure 25 is an important result, and a full explanation is provided on page 45.

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 diagnostic information about multiple calibration scenarios. Secondly, the report covers only a small fraction of the locations, conditions and forecasting scenarios under which the HEFS will be used operationally. In particular, it focuses on headwater basins and does not consider the quality of the forecasts in regulated rivers. Thirdly, the report does not provide any comparisons between the HEFS streamflow forecasts and the single-valued streamflow forecasts issued by the RFCs (for which archives are

limited). Finally, the report considers forcing inputs from NCEP's frozen GFS only, using climatology as a baseline. It does not consider any of the scheduled enhancements to the HEFS, notably to use ensemble forecasts from NCEP's operational GFS (also known as the Global Ensemble Forecast System, GEFS), which will be considered in a subsequent report.

2. Executive summary and recommendations

- Ensemble forecasts of precipitation, temperature and streamflow were generated with the NWS Hydrologic Ensemble Forecast Service (HEFS) for a ~20-year period between 1979 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 basins include a headwater and one immediately downstream basin. Precipitation and temperature forecasts were produced with the Meteorological Ensemble Forecast Processor (MEFP). Inputs to the MEFP comprised "raw" precipitation and temperature forecasts from the frozen (circa 1997) version of NCEP's Global Forecast System (GFS) and a climatological ensemble, which involved resampling historical observations in a moving window around the forecast valid date ("resampled climatology"). In both cases, the forecast horizon was 1-14 days. The streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS) and were bias-corrected with the Ensemble Post-processor (EnsPost).
- The HEFS is being evaluated in several phases. Subsequent phases will consider the long-range forecasts (out to ~1 year) with forcing inputs from the Climate Forecast System (CFSv2) and medium-range forecasts with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS). 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 MEFP with GFS forcing and the EnsPost 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 to illustrate how hindcasting and verification of the HEFS might be conducted in future.

- The precipitation, temperature and streamflow forecasts were verified with the Ensemble Verification System (EVS). Several datasets were verified, namely: 1) the “raw” forcing forecasts from the GFS; 2) the bias-corrected forcing forecasts from the MEFP with GFS inputs (MEFP-GFS); 3) the raw streamflow forecasts with forcing inputs from the MEFP-GFS; and 4) the bias-corrected streamflow forecasts from the EnsPost. In addition, resampled climatology was used to establish the skill of the MEFP-GFS forecasts. The forcing and streamflow forecasts were verified against corresponding meteorological and hydrologic observations. In addition, the raw streamflow forecasts were verified against simulated streamflows. This allowed the meteorological uncertainties to be separated from the total (meteorological and hydrologic) uncertainties, as the simulated streamflows do not contain any forecast uncertainties. The verification results are presented by forecast lead time, season, and magnitude of the observed and forecast variables. In an appendix to the main report, a selection of the paired streamflow forecasts and observations are shown for the downstream basin in each RFC. These allow for visual inspection of the strengths and weaknesses of the HEFS forecasts for specific hydrologic events.
- In general, the precipitation forecasts from the MEFP with GFS forcing are more skillful than resampled climatology during the first week, but comprise little or no skill during the second week. In contrast, the temperature forecasts improve upon resampled climatology at all forecast lead times. However, there are notable differences between RFCs and for different seasons, aggregation periods and magnitudes of the observed and forecast variables. A direct comparison between the raw inputs to the MEFP and the bias-corrected outputs was hampered by the different spatial scales of the inputs and outputs (i.e. a single grid node versus a basin average). Nevertheless, the MEFP-GFS forecasts generally preserve or improve upon the correlations in the raw GFS forecasts.
- The precipitation forecasts from the MEFP-GFS show the highest correlations and greatest skill in the CNRFC basins, particularly during the wet season (November-April). This is associated with the greater predictability of large storms

in the coastal mountain ranges of Northern California (the North Coast Ranges). However, the forecasts are much less skillful during the dry season, particularly at longer lead times (beyond ~5 days). In MARFC, the precipitation forecasts are skillful at early forecast lead times, with the greatest skill occurring during the wet season (December-March) and at moderate precipitation amounts. However, the forecast skill declines rapidly with increasing forecast lead time, particularly during the dry season. In the summer months, forecasts of high precipitation are constrained by the limited ability of the frozen GFS to model convection and by a range of biases in the precipitation forecasts. In AB- and CB-RFCs, the seasonal patterns are less pronounced, but the quality of the MEFP-GFS precipitation forecasts is also lower. 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 in the MEFP-GFS forecasts.

- The MEFP precipitation forecasts contain a range of "unconditional" and "conditional" biases; that is, overall biases in the forecasts and biases that occur under particular conditions. In particular, there is a tendency for the MEFP to underestimate the probability of precipitation (PoP), both with the GFS and resampled climatology as inputs. For medium-range forecasting, unbiased estimates of the PoP are generally less important for hydrologic applications than unbiased forecasts of non-zero precipitation amounts. Nevertheless, dry conditions are common, and any biases in PoP will, therefore, be conspicuous.

Recommendation 1: the systematic underestimation of PoP is indicative of a problem in the modeling, estimation, or implementation of the MEFP. This requires further investigation, as the forecasts of PoP with the MEFP-GFS are substantially worse than climatology in some basins.

- Alongside the underestimation of PoP, the precipitation forecasts are conditionally biased with increasing amounts of observed precipitation. Thus, non-zero precipitation amounts are systematically underestimated in all basins, and large amounts are increasingly underestimated (a "Type-II conditional bias"). The severity of this conditional bias varies with forecast lead time, RFC, season,

and aggregation period. For example, it increases with increasing forecast lead time and declines with increasing aggregation period. In CNRFC, the biases are sufficiently small, and the spread sufficiently large, that the highest precipitation amounts are generally forecast with some (non-zero) probability of occurrence, at least for early forecast lead times. In other basins, the largest precipitation totals frequently occur without warning and are routinely underestimated by as much as the observed precipitation amount.

Recommendation 2: the systematic underestimation of high precipitation amounts by the MEFP originates from the (in)ability of the frozen GFS to identify unusually high precipitation amounts, particularly at long forecast lead times. As a statistical technique, the MEFP cannot "see" conditions that are undetected by the GFS. The calibration of the MEFP to favor good performance under varied conditions (i.e. unconditional unbiasedness) is a secondary factor responsible for these conditional biases. Several approaches may be considered in future, including: 1) selective post-processing to favor unbiased estimates of high precipitation amounts for specific applications; and 2) improved sources of raw forcing, including the operational Global Ensemble Forecast System (GEFS), the Short-Range Ensemble Forecast System (SREF), and quantitative precipitation forecasts from the Weather Prediction Center (WPC), for which an archive of single-valued forecasts is available. A long and consistent record of historical forecasts is a prerequisite to considering additional sources of forcing in the MEFP. Thus, it requires a commitment by the operational forecasting agencies, such as the NCEP, to conduct hindcasting at strategic intervals and to continue forecasting with legacy models (providing a window for the HEFS to be re-calibrated).

- While the MEFP temperature forecasts are conditionally unbiased for most observed temperatures, the coldest temperatures are systematically overestimated. However, as with the highest precipitation totals, these are often associated with rare events for which the GFS was not adequately initialized or failed to predict the observed weather accurately, such as the extreme cold of January 1979 that affected much of the contiguous U.S. One exception concerns the systematic over-estimation of moderately cold temperatures in CBRFC, where average elevations in the headwater basin (DRRC2) exceed 2500m.

Recommendation 3: in CBRFC, the overestimation of temperatures by the MEFP requires further investigation. In particular, the MEFP forecasts with GFS forcing are increasingly biased as the forecast lead time increases. For example, after 14 days, the ensemble mean of the MEFP-GFS forecasts overestimates the coldest 10% of observed temperatures by ~5°C, on average, both in DRRC2, where the 10th percentile is -9.4°C, and in DOLC2, where the 10th percentile is -5.6°C. These variations in temperature originate from the mountaneous terrain surrounding the Dolores River. The GFS cannot be expected to resolve such variations. However, the inputs to the MEFP may be improved through careful interpolation of the GFS forecasts, conditionally upon the local terrain, or by using temperature and precipitation forecasts that better resolve the local terrain. As atmospheric models become more resolved, a gridded formulation of the MEFP may be preferred over a basin-averaged formulation.

- In all river basins, the streamflow forecasts are substantially more skillful when using the MEFP-GFS together with the EnsPost than using the MEFP with resampled climatology alone. In general, however, both the raw and bias-corrected streamflow forecasts have lower biases, stronger correlations and are more skillful in CB- and CN-RFCs than AB- and MA-RFCs. Aside from location, there are strong variations in forecast quality with streamflow amount, forecast lead time, season and aggregation period. The relative importance of the meteorological and hydrologic uncertainties also varies between basins and is modulated by the same controls on forecast quality.
- For moderate and high streamflow amounts, the MEFP-GFS accounts for the majority of skill in CNRFC, while the EnsPost contributes valuable skill under dry conditions. During the wet season, the GFS benefits from the increased predictability of large storms in the North Coast Ranges. The observed streamflow is consistently low in both CN- and CB-RFCs during the dry season, and the meteorological uncertainties are much less important. When factoring out the hydrologic initial conditions, the EnsPost contributes valuable skill at low and moderate streamflows, although for a narrower range of forecast lead times in

CNRFC than CBRFC. These improvements largely stem from an increase in reliability, or a reduction in bias, following streamflow post-processing.

- In CBRFC, unlike CNRFC, most of the skill during the wet season originates from the EnsPost, particularly at low to moderate streamflow amounts. During the winter months, much of the precipitation in CBRFC falls as snow, where light and moderate storms are integrated into the snowpack alongside heavy storms. As such, the forecasts of high streamflow, generally associated with snowmelt, are less dependent on unbiased forecasts of heavy precipitation. Also, by monitoring the snowpack, the initial conditions and parameters of the hydrologic models can be improved during snowmelt. Nevertheless, the contribution from the MEFP should not be completely ignored. In general, snow accumulates over several months, for which medium-range weather forecasts are inherently less useful. However, melting occurs in days or weeks, rather than months, and biases in the temperature forecasts, or inaccurate predictions of rain-on-snow, can lead to errors in the timing of the streamflow forecasts during snowmelt. The raw streamflow forecasts increasingly overestimate both the observed and simulated streamflows in the CBRFC basins as the forecast lead time increases. While these timing errors are partly (indirectly) removed by streamflow post-processing, the EnsPost models the total uncertainty in streamflow volume, not timing errors explicitly.

Recommendation 4: automated data assimilation (DA) is the preferred approach to adjusting model states, but is not currently implemented in the HEFS. Some RFCs manually modify ("mod") the inputs, parameters and states of the hydrologic models, either to account for additional sources or sinks or to offset other errors in the forecasting process. This must be reviewed in the context of the HEFS. The HEFS relies on a reasonable assessment of the individual sources of bias and uncertainty in the streamflow forecasts. This, in turn, depends on a reasonable assessment of the historical forecast errors, which are diagnosed from observations. Attempts to "improve predictions for the wrong reasons" will result in flawed estimates of the predictive uncertainties surrounding the forcing and streamflow forecasts. In contrast, automated DA is repeatable (e.g. for hindcasting purposes) and relies on objective measures for

reducing uncertainty. In general, the contribution of DA will increase with increasing persistence in the initial conditions and states of the hydrologic models and, thus, with increasing distance downstream. However, DA will also contribute valuable skill in headwater basins, particularly those dominated by snow accumulation and melting. A clear strategy is needed to implement (even rudimentary) DA in a future version of the HEFS.

- In AB- and MA-RFCs, the raw and bias-corrected streamflow forecasts generally have larger biases, weaker correlations and lower skill, while the relative contributions from the MEFP and the EnsPost are more variable. At early forecast lead times, most of the skill in the bias-corrected streamflow forecasts originates from the EnsPost, where the reliability of the forecasts is improved at low and moderate streamflow amounts. However, the forecast skill declines rapidly with increasing forecast lead time and amount of streamflow, particularly in ABRFC. In MARFC, the inflows to the Cannonsville Reservoir (CNNN6) were estimated rather than observed. Specifically, they were estimated by the New York City Department of Environmental Protection (NYCDEP) using the change in daily storage and outflow, without accounting for evaporation. Under dry conditions, differences in the quality of the streamflow forecasts between CNNN6 and the upstream basin, WALN6, suggest that the inflow estimates may contain errors.
- In all basins, the raw and post-processed streamflow forecasts contain biases at the highest observed streamflows, although the biases are generally smaller in CB- and CN-RFCs than AB- and MA-RFCs. This originates from similar biases in the MEFP precipitation forecasts. Unbiased predictions of high streamflow are important in operational forecasting (but not necessarily at the expense of good performance for low and moderate streamflow). For example, if the forecast streamflows severely underestimate the observed streamflows when flooding occurs, this could hamper any efforts to mitigate flood damage.
- 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 study covers only a small

fraction of the locations, conditions and forecasting scenarios under which the HEFS will be used operationally.

Recommendation 5: 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 across all RFCs, for a range of forcing inputs, and for a broader range of river basins, including regulated rivers and outlets. The baseline forecasting system should comprise forcing inputs from the GEFS, the Climate Forecast System (CFSv2) and climatology for periods longer than 9 months. Several hindcasting scenarios may be required, in order to guide the calibration and configuration of the HEFS. Forecasts from the HEFS should also be compared to the RFC single-valued forecasts. In addition, there is a need to evaluate models and decision support systems that rely on the HEFS. Such applications will show varying sensitivities to the HEFS forecasts, including their space-time properties, which are difficult to verify directly.

3. Introduction

Uncertainties about the inputs, parameters and structures of hydrologic models lead to uncertainties about the model predictions. In the presence of uncertainty, single-valued predictions are misleading, and can result in unfair decisions or persistent conflict and indecision (Handmer et al., 2001). As an important input to environmental decision making, hydrologic forecasts should properly account for the uncertainties inherent in hydrologic modeling (Beven, 2000; Brown et al., 2010a; Demeritt et al., 2013). In recent years, a plethora of quantitative and qualitative approaches has emerged for assessing uncertainty and for using uncertain information in decision making (for reviews, see Matott et al., 2009; Brown et al., 2010a; Ramos et al., 2012). In hydrologic modeling, a range of "bottom up" and "top down" approaches are used to quantify the uncertainties in model outputs. Bottom-up approaches quantify the total uncertainty as a combination of specific sources of uncertainty. For example, it is common to distinguish between the meteorological or "forcing uncertainties" and the "hydrologic uncertainties", which comprise uncertainties in the initial conditions, parameters and structures of the hydrologic models (Brown and Heuvelink, 2005; Kavetski et al., 2006a/b; Schaake et al., 2006; Seo et al., 2006; Liu and Gupta, 2008; Matott et al., 2009; Cloke et al., 2013). Using a combination of uncertainty propagation (forward modeling) and recursive estimation (inverse modeling), uncertainties in the model outputs can then be determined numerically, through Monte Carlo simulation (Heuvelink, 1998; Pappenberger et al., 2005; Helton et al., 2006). Top-down approaches directly model the total uncertainty with statistical techniques. For example, in one type of Model Output Statistics (MOS; Glahn and Lowery, 1972; Gneiting et al., 2005; Regonda et al., 2013), historical forecasts and observations are used to estimate a linear regression in which the observed outcome is the dependent variable and the single-valued forecast is the independent variable. Operationally, a linear adjustment is then applied to the real-time forecast and a measure of predictive uncertainty obtained from the historical residuals surrounding the linear regression.

In principle, source-based approaches to quantifying uncertainty are generally preferred over lumped treatments. They provide the flexibility to include prior information

or expert judgment on the key sources of uncertainty (Kadane and Wolfson, 1998; Rojas et al., 2009), which can lead to targeted improvements in modeling (Heuvelink, 1998; Oakley and O'Hagan, 2004; Saltelli et al., 2008). Nevertheless, top-down approaches, such as hydrologic MOS (Regonda et al, 2013), may be preferred where resources are limited or when the sources of uncertainty are difficult to quantify reliably. More generally, bottom-up approaches require several assumptions and approximations. This can lead to the omission of key sources of uncertainty or to inaccurate models of their joint probability distribution (Norton et al., 2006; Pappenberger and Beven, 2006). For example, the failure to consider uncertainties in the inputs to a hydrologic model can lead to biases in the model parameters (Kavetski et al., 2002; Chowdhury and Sharma, 2007). Also, when dealing with multiple sources of uncertainty, including inputs that vary in space or time or originate from different measurement scales (e.g. categorical and numerical), the problem of capturing their statistical dependencies is not trivial (Heuvelink, 2002). Thus, bottom-up approaches may require a combination of numerical uncertainty propagation and statistical modeling of the residual uncertainties and biases, also known as post-processing. Increasingly, Hydrologic Ensemble Prediction Systems (HEPS) use a combination of uncertainty propagation and statistical post-processing to account for the residual sources of uncertainty and bias (Pappenberger et al., 2005; Montanari and Grossi, 2008; van Andel et al., 2013). For example, the European Floods Awareness System (EFAS) uses ensemble forecasts of temperature and precipitation from the Global Atmospheric Model of the European Centre for Medium-Range Weather Forecasts (ECMWF). 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 original data in wavelet space (Bogner and Pappenberger, 2011).

The first operational HEPS was developed in the late 1970s by the U.S. National Weather Service (NWS; Day, 1985). Using historical observations of precipitation and temperature, the NWS River Forecast System was used to generate ensemble streamflow predictions (ESP) for up to 90 days into the future. In subsequent

developments, the forcing inputs to ESP were improved with seasonal “climate outlooks” from NOAAs Climate Prediction Center, and short-range quantitative precipitation forecast from the NWS River Forecast Centers (RFCs; Perica, 1998). More recently, a Hydrologic Ensemble Forecast Service (HEFS) has been developed for ensemble forecasting of temperature, precipitation and streamflow at lead times ranging from one hour to one year (Seo et al., 2010; Demargne et al., 2013). The HEFS quantifies the total uncertainty in streamflow as a combination of specific sources of uncertainty (Seo et al., 2010). The meteorological uncertainties are modeled with the Meteorological Ensemble Forecast Processor (MEFP). The MEFP generates ensemble forecasts of precipitation and temperature conditionally upon a raw, single-valued, forecast (Wu et al., 2011). Currently, the “frozen” version of NCEPs Global Forecast System (GFS) is used for medium range forecasting (Hamill et al, 2006), while the Climate Forecast System (CFSv2) is used for long-range forecasting. The hydrologic uncertainties 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; Seo et al., 2006). The EnsPost accounts for the hydrologic uncertainties and reduces any hydrologic biases. Specifically, it models the joint probability distribution of the observed and simulated streamflows in normal quantile space (NQT; Kelly and Krzysztofowicz, 1997; Seo et al., 2006).

The HEFS is being implemented in several phases, with the initial version (HEFSv1) scheduled for operational use in all RFCs by 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 forcing. Verification is being conducted with the Ensemble Verification System (Brown et al., 2010b), for which the first results are reported here.

Whether they explicitly account for uncertainty or not, hydrologic forecasts are subject to error. These errors may be correlated in space and time and may be

systematic. The skill of an ensemble forecasting system can depend largely on its systematic biases (Hashino et al., 2006; Wilczac et al., 2006; Brown and Seo, 2013). Forecast evaluation or 'verification' is necessary to identify these biases and to establish the skill of the forecasting system under a range of observed and forecast conditions (Jakeman et al., 2006). In ensemble forecasting, biases are manifest as differences between the forecast probabilities and the corresponding observed outcomes over a large sample of forecasts and verifying observations (Wilks, 2006; Jolliffe and Stephenson, 2011). By conditioning on the observed and forecast variables, these residuals can be factored into more detailed attributes of forecast quality. For example, a flood forecasting system is "reliable", on average, if flooding is observed twenty percent of the time when it is forecast with probability 0.2 (this can be repeated for all forecast probabilities). An ensemble forecasting system is discriminatory with respect to flooding if it consistently forecasts the occurrence of flooding with a probability higher than chance (i.e. the climatological probability of flooding) and consistently forecasts its non-occurrence with a probability lower than chance. Recent examples of hindcasting and verification with medium-range HEPS include Jaun and Ahrens (2009); Bartholmes et al. (2009); Renner et al. (2009); Thirel et al. (2010); Van den Bergh and Roulin (2010); Demargne et al. (2010); and Schellekens et al. (2011).

In this report, hindcasts of temperature, precipitation and streamflow are generated with the HEFSv1 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, magnitude of the observed and forecast variables, season, and aggregation period. Limited combinations of these attributes are also considered. Verification results are presented for the temperature and precipitation forecasts from the MEFP and for the streamflow hindcasts before and after streamflow post-processing with the EnsPost. In order to distinguish between the contributions of the MEFP and the EnsPost to the quality of the streamflow hindcasts, verification is performed against simulated streamflow (eliminating hydrologic biases) and against observed streamflow. 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 frozen GFS, together with recommendations on future enhancements.

4. Materials and methods

4.1 Study area

Four pairs of basins, each comprising one headwater and one immediately downstream basin, were used to evaluate the HEFSv1. Figure 1 and Table 1 show the location, the nearest GFS grid node, the drainage area and the mean elevation of each basin. Table 1 also shows the annual precipitation, the runoff coefficient (runoff/precipitation) and the "climate index" or ratio of precipitation to potential evaporation. 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. They all originate from RFCs for which testing of the HEFSv1 is currently underway, namely AB-, CB-, CN-, and MA-RFCs. As the uncertainties and biases propagate from upstream to downstream locations, it is important to understand the quality of the HEFSv1 in headwater basins. It is also important for operational forecasting in headwaters, where practical applications may include river regulations and diversions, flash-flooding, and recreation. Further downstream, the HEFS will be impacted by additional sources of bias and uncertainty, of which some may be poorly captured by the HEFSv1 (e.g. 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 river basins, together with the daily mean runoff for the headwater and

downstream basins separately. The averages are shown for each calendar month and were derived from gauged temperature, precipitation, and streamflow over a 20 year period between 1979 and 1999 (see Section 5.3). 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 between basins and between RFCs. The strongest patterns occur 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 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. They lie in “tornado alley”, where 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 during the summer months.

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 dramatic increase in runoff between April and July. 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, there is little or no precipitation and streamflow along the upper reaches of the Eel River. 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. The streamflows increase in ~November once the soil moisture is replenished (figure 2) and then continue increasing with precipitation until ~January. In DOSC1 and FTSC1, the predictability of heavy precipitation 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. 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 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. As with CBRFC, the daily precipitation amounts are relatively constant throughout the year. 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 rapidly 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, specifically the impacts of sedimentation on reservoir turbidity, 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 uncertainty is factored into two main sources of uncertainty, 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 the random variables for simplicity,

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

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" traces of precipitation and temperature through the hydrologic models. These traces include the forcing uncertainty, which is modeled explicitly

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

where $\mathbf{m}_f$ denotes the future (observed) forcing. The forcing uncertainties are quantified by the Meteorological Ensemble Forecast Processor (MEFP). The MEFP models the observed forcing conditionally upon a raw forecast, $\mathbf{r}_f$; that is, by estimating the joint distribution, $f_6(\mathbf{m}_f, \mathbf{r}_f)$, and factoring out $\mathbf{r}_f$ in real time

$$f_5(\mathbf{m}_f) = \int f_6(\mathbf{m}_f, \mathbf{r}_f)\ d\mathbf{r}_f . \quad (3)$$

The raw forcing may comprise the ensemble mean of NCEP's GFS or single-valued quantitative precipitation forecasts from the RFCs, among others (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 additional uncertainties are modeled in the adjusted streamflow forecast using the Ensemble Postprocessor (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. Here, the parameters of the HEFS were estimated from the same historical period (1979-1999) used for the streamflow hindcasting and verification. 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.

4.3 Datasets

Hindcasts of mean areal temperature (MAT) and mean areal precipitation (MAP) were generated with the MEFP for a ~20 year period from 1979-1999. 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 frozen (circa 1997) version of NCEP's GFS (Hamill et al., 2006) and by resampling the historical observations in a moving window of 30 days either side of the forecast valid date ("resampled climatology"). Raw streamflow hindcasts were generated with the HEP using the precipitation and temperature forecasts from the MEFP. The hydrologic modeling was conducted with the CHPS using the operational models implemented 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). 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 weighed 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). For each pair of forecast locations, the computational times were approximately 2 hours for the MEFP hindcasts (per variable, on one Virtual Machine, VM) and 20-90 hours for the streamflow hindcasts (distributed evenly across four VMs).

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) for the period 1979-1999. They 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 CBRFC, respectively. As the forecasts were verified at an aggregated support of one day or larger (see below), this timing error was considered unimportant. 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 ABRFC, where the forecasts are issued hourly, and in ABRFC, where the 6-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., 2010b). The forecasts were verified conditionally upon season, forecast lead time, magnitude of the observed and forecast variables, and aggregation period. 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 against observed streamflow at the outlet of each basin. In addition, the “raw” streamflow forecasts, i.e. before applying EnsPost, were verified against simulated streamflow. Verification against simulated streamflow allows the total uncertainty to be separated from the meteorological uncertainties, as the hydrologic simulations and forecasts both comprise hydrologic uncertainty. In short, 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).

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; 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 different 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 observations collected between 1979 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 were only interpreted for climatological exceedence probabilities greater than 0.005, corresponding to ~38 samples at a daily aggregation.

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, $f_{Y|X}(y \mid x)$, 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, $f_{X|Y}(x \mid y)$, 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. $f_Y(y) = \delta(y)$, 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 summarizes the key 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 are verified against observed MAP and MAT, respectively. The results are presented by forecast lead time, magnitude of the forcing variable, season and aggregation period.

5.1.1 Forecast lead time

Figure 3 shows the correlation of the ensemble mean forecast and observed precipitation by forecast lead time, while figure 4 shows the relative mean error (RME) of the ensemble mean forecast (Appendix B). The results are shown for the upstream and downstream basins in each RFC and, in CB- and CN-RFCs, comprise a weighed average over multiple sub-basins (see Section 4.1). Results from the MEFP with GFS forcing (MEFP-GFS) are shown together with "resampled climatology" (MEFP-CLIM), which measures the strength of the climate signal, or the inherent predictability of temperature and precipitation. For basins with a strong seasonal climatology (figure 2), MEFP-CLIM generally improves upon "raw" climatology (see below), as the resampling is performed within a 61-day moving window, which is centered on the forecast valid date over all historical years. Figure 3 also shows the correlations for the raw GFS forecasts used as input to the MEFP. A direct comparison between the MEFP inputs and outputs is complicated by the different support of the raw GFS forecasts (one grid node) versus the MEFP outputs and observations (one basin or sub-basin: see figure 1). However, by comparing the correlations of the inputs and outputs, any scale-related differences should be minimized. Specifically, the correlation coefficient is insensitive to the means and variances of the forecasts and observations. Nevertheless, the results should only be considered indicative; that is, an attempt to screen the MEFP for any substantial failure to preserve the correlations in the inputs.

As indicated in figure 3, the correlation declines smoothly with increasing forecast lead time in all basins, approaching the "background signal" of MEFP-CLIM after ~8-days in AB-, CB-, and MA-RFCs and ~10 days in CBRFC. The largest correlations are found in CN- and MA-RFCs, where they approach ~0.85 and ~0.75 for the first 24-hour period, respectively. At early lead times, the relative strength of MEFP-GFS over MEFP-CLIM is greatest in WALN6 and CNNN6 (MARFC), where resampled climatology performs similarly to raw climatology. The MEFP-GFS forecasts benefit from the regulating effects of the ocean on atmospheric predictability in the Mid-Atlantic region, which initially leads to greater forecast skill, particularly during the winter months (figure 3 and below). The weakest signal from the MEFP-GFS is observed in the high-altitude

basins of DRRC2 and DOLC2 in CBRFC, where a large fraction of the winter precipitation falls as snow. In general, the MEFP preserves or improves upon the correlations in the raw GFS forecasts (figure 3). However, the results are more variable in CBRFC, where the correlations are slightly improved in DRRC2 and reduced in DOLC2. The relative bias (RME) of the precipitation forecasts is also greatest in the CBRFC basins, where the ensemble mean of the MEFP-GFS consistently underestimates the observed precipitation amount by ~10% (figure 4). In MARFC, the MEFP-GFS underestimates the observed precipitation by only ~5% during the first week, but this increases to ~20% during the second week.

Figure 5 shows the Continuous Ranked Probability Skill Score (CRPSS) of the precipitation forecasts, which measures the overall skill of the MEFP with GFS forcing, as well as climatological forcing, relative to "raw" climatology. As an integral measure of skill across all precipitation amounts, the CRPSS is sensitive to a combination of relative bias and correlation of the ensemble mean forecast, as well as contributions from the individual ensemble members (Appendix B). The CRPSS is greatest in CNRFC, where the correlations are highest and the RME is lowest. The CRPSS is also high in MARFC, where the correlations are strong at early lead times. However, at longer lead times, the skill deteriorates rapidly as the RME increases and the correlations decline, leading to slightly negative skill after ~7 days. In CBRFC, the MEFP-GFS forecasts are relatively unskillful. Here, a combination of weak correlation (figure 3) and an unconditional bias in the ensemble mean (figure 4) leads to precipitation forecasts that are indistinguishable from raw climatology after only ~3-4 days. Elsewhere, the MEFP-GFS precipitation forecasts are more skillful than raw climatology for ~6 days in AB- and MA- RFCs and ~10 days in CNRFC.

The MEFP-GFS temperature forecasts are highly skillful across all basins and forecast lead times, with correlations above 0.9, and CRPSS above 0.6, increasing to ~0.8 for early lead times (results not shown). Figure 6 shows the mean error of the ensemble mean forecast (in °C) for the upstream and downstream basins in each RFC and for each source of forcing in the MEFP. As indicated in figure 6, the temperature forecasts are unbiased at all forecast lead times in AB- and MA-RFCs and comprise

only small biases in CNRFC and CBRFC. However, in CBRFC, there is an unconditional bias in the MEFP-GFS and MEFP-CLIM ensembles, which is constant for MEFP-CLIM, but increases with forecast lead time in the MEFP-GFS ensembles. This bias is small in absolute terms, approaching 0.15°C after 14 days for the MEFP-GFS forecasts. Nevertheless, the MEFP-GFS forecasts systematically overestimate the coldest observed temperatures. For example, after 14 days, the lowest 10% of observed temperatures are overestimated by ~5°C, on average, both in DRRC2, where the 10th percentile is -9.4°C, and in DOLC2, where the 10th percentile is -5.6°C (results not plotted). These variations in temperature are associated with the mountainous terrain surrounding the Dolores River. The GFS cannot be expected to resolve such variations. Indeed, a single grid-node (Figure 1) is used to calibrate the MEFP for both DRRC2 and DOLC2, without any interpolation or change of support.

5.1.2 Magnitude of the forcing variable

Figure 7 shows the correlation, relative bias and skill of the MEFP-GFS precipitation forecasts for increasing amounts of observed precipitation. The correlations are shown for the raw GFS inputs, as well as the MEFP-GFS outputs. Skill is measured with the Brier Skill Score (BSS) against sample climatology (Appendix B). The results are shown for the downstream basin in each RFC and at selected forecast lead times. Each lead time comprises the 24-hour period ending at the nominal lead time (e.g. 14 days denotes the period from 312 to 336 hours). The scores are plotted against climatological exceedence probability on a probit scale (similar to a log 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; this corresponds to 37mm in CBNK1 (table 1). As indicated in Section 4.4, the RME and correlation do not include dry conditions as they were derived from the subsample of pairs whose observed value exceeded the threshold. In contrast, the BSS includes the boundary between dry and wet conditions or the Probability of Precipitation (PoP). The origin of each curve in figure 7 corresponds to the climatological PoP in that basin.

The quality of the MEFP-GFS ensemble forecasts is sensitively dependent upon precipitation amount (figure 7). There is a systematic decline in correlation and an increase in conditional bias with increasing precipitation amount. The rate of decline in correlation is greatest during the first 24-hour period, where the forecasts are most skillful. The differences between basins are also greatest during the first 24-hour period with the strongest correlations and lowest biases in FTSC1, followed by CNNN6, BLKO2 and DOLC2. However, CNNN6 shows the weakest skill for PoP and small precipitation amounts, where the MEFP-GFS forecasts are significantly *worse*, in terms of BSS, than sample climatology. This stems from a Type-I conditional bias, or lack of reliability, in the forecast PoP (see below). In keeping with the overall correlations (figure 3), the MEFP-GFS forecasts generally preserve or improve upon the correlations in the raw GFS forecasts with increasing precipitation amount (figure 7).

Overall, the relative bias of the ensemble mean forecast is less sensitive to forecast lead time than the correlation coefficient. Nevertheless, it does increase with increasing forecast lead time (figure 7). At longer lead times, it also increases more rapidly with increasing precipitation amount; that is, the RME in figure 7 becomes more convex. This is not surprising as the GFS forecasts are unskillful after ~7-10 days and unconditional climatology is, by definition, conditionally biased. At light precipitation amounts, the ensemble mean forecast generally shows higher correlations and less conditional bias. However, as indicated above, the MEFP-GFS is much less skillful in forecasting PoP than the occurrence of moderate precipitation (figure 7). Theoretically, this is important, because precipitation intermittency is an important aspect of the joint distribution modeled by the MEFP (Wu et al., 2011). In practice, however, reliable forecasts of precipitation amount are generally more important for hydrologic applications than reliable forecasts of PoP, at least in the medium-range.

By computing the Brier Score (BS) over several categories of observed and forecast probability, the BS may be separated into multiple sources of error (Appendix B). Normalizing the BS by the climatological variance leads to the BSS and its associated contributions in terms of relative error (Appendix B). Figure 8 shows selected components of the BSS for the MEFP-GFS precipitation forecasts in each downstream

basin, namely the relative reliability or Type-I conditional bias (BSS-REL), the relative Type-II conditional bias (BSS-T2) and the relative sharpness (BSS-SHA). High relative sharpness is associated with forecasts that are substantially different than climatology or “go out on a limb”. For sharp forecasts to be skillful, they must be conditionally unbiased. The BSS-REL and BSS-T2 are conditional averages, grouped by forecast and observed probability, respectively (Appendix B). Larger values are associated with more conditional bias than climatology. Figure 9 shows the reliability diagram for selected events during the first 24-hour period in each downstream basin. The reliability diagram compares the average observed and forecast probabilities across several categories of forecast probability, together with the sample sizes in each forecast category (sharpness).

While the MEFP-GFS forecasts are broadly reliable for moderate to large precipitation thresholds across a range of forecast lead times and basins, they are unreliable for PoP and light precipitation amounts, particularly in CNNN6. This is apparent in the high values of BSS-REL (figure 8) and in the biased forecasts of PoP and light precipitation. Indeed, the MEFP-GFS forecasts consistently underestimate the observed PoP (figure 9), particularly at low and moderate forecast probabilities. During the first week, the sharpness and Type-II conditional biases vary more with forecast lead time, location, and precipitation threshold. For example, during the first 24-hour period, the sharpness is consistently higher and the Type-II biases consistently lower in FTSC1, followed by CNNN6, BLKO2 and DOLC2. In keeping with the decomposition of the BSS (Appendix B), any increase in sharpness must be consolidated by forecasts that are more discriminatory or less conditionally biased, otherwise the overall skill will decline (see Bradley et al., 2004). At longer lead times, the forecasts become less sharp and more conditionally biased when grouped by observed probability. Also, while the BSS-REL remains low at moderate and heavy precipitation amounts, the BSS is, by definition, driven by non-occurrences at high thresholds (Appendix B). Of the few occurrences at these high thresholds, even fewer are predicted with high probability, due to the consistent underestimation of heavy precipitation amounts. Consequently, at longer lead times, there are too few samples from which to develop reliability diagrams for moderate and high precipitation amounts.

As indicated in figure 8, there are large Type-II conditional biases at both low and high precipitation thresholds. This is further illustrated in figure 10, which shows box plots of errors in the MEFP-GFS precipitation forecasts. Each box represents one ensemble forecast from the first 24-hour period. 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. As indicated in figure 10, the MEFP-GFS forecasts consistently underestimate heavy precipitation. For the highest precipitation totals, there are strong conditional biases in the median (figure 10), as well as the mean (figure 7). In the absence of sufficient spread, the highest precipitation totals generally occur without warning in DOLC2. In FTSC1, the conditional biases are sufficiently small and the forecast spread sufficiently large to provide some warning in most cases. This is understandable because FTSC1 lies on the windward slopes of the North Coast Ranges, where the predictability of heavy precipitation is enhanced during the winter months by the onshore movement and orographic lifting of storms from the Pacific coast.

Box plots of errors in the MEFP-GFS temperature forecasts are shown in figure 11. The forecasts comprise daily mean temperatures for the first 24-hour period at the downstream locations. In contrast to precipitation, the temperature forecasts are conditionally unbiased at most observed temperatures. However, the lowest temperatures are consistently overestimated by the MEFP-GFS, with biases in the forecast median approaching the median value itself. Although significant in magnitude, these biases are associated with only a small number of events. In particular, they are associated with well-known events, such as the widespread cold and snow of January 1979. During the winter of 1978-79, temperatures were 12-16°C below average for much of the US and shifted rapidly with an unusually strong frontal system (Wagner, 1979). As a statistical technique, the MEFP cannot “see” conditions that are undetected by the GFS. Rather, it corrects for biases conditionally upon the raw forecasts and cannot add any information without additional predictors (see Brown and Seo, 2013). However, recent improvements in weather observations and data assimilation schemes, such as those employed in the operational GFS, can update the trajectory of a forecast when the state of the atmosphere changes unexpectedly (Park and Xu, 2009).

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, expressed as climatological probabilities. In order to compare the verification results between seasons, the climatological probabilities were derived from the overall observed sample, ensuring fixed thresholds of precipitation and temperature throughout the year. Figure 12 shows the unconditional bias of the MEFP-GFS forecasts for each season and for the overall sample. The results are plotted for selected forecast lead times in each downstream basin. Figure 13 shows the CRPSS of the MEFP-GFS forecasts relative to sample climatology.

For BLKO2 and DOLC2, the MEFP-GFS precipitation forecasts show similar RME in the wet and dry seasons. The relative bias depends largely on precipitation amount and forecast lead time (during the first week). Thus, it increases rapidly in both seasons with increasing precipitation amount and is greater across all thresholds at longer forecast lead times. While the RME also depends strongly on precipitation amount and forecast lead time in FTSC1 and CNNN6, the effects of seasonality are greater, particularly at early lead times. During the dry season, the RME is significantly larger in both FTSC1 and CNNN6 for the first 24-hour period (figure 12). In FTSC1, the increase in RME affects all precipitation amounts. During the wet season, the smaller relative bias in FTSC1, as well as the higher CRPSS at longer lead times (figure 13), may be explained by the relative predictability of large storms in the North Coast Ranges. In contrast to FTSC1, CNNN6 experiences very little seasonality in precipitation amount (figure 2). However, the relative biases are significantly lower and the CRPSS significantly higher during the wet season. This may be explained by the difficulty of forecasting in the Mid-Atlantic during the summer months where thunderstorms may evolve and dissipate rapidly, leading to reduced predictability.

While the RME of the ensemble mean forecast is relatively constant between seasons in BLKO2 and DOLC2, the precipitation forecasts are significantly more skillful during the dry season (also the warm season). Again, this reduction in skill of the MEFP-GFS forecasts may originate from the quality of the raw GFS forecasts in warm conditions, rather than a seasonal dependence on the skill added by the MEFP specifically. Indeed, a decomposition of the CRPSS into relative reliability, resolution and uncertainty (not shown), indicates that the loss of skill during the summer months is primarily associated with a loss of resolution in the precipitation forecasts. In both CBNK1 and DOLC2, the reliability of the forecast probabilities depends largely on precipitation amount, and not on season.

Figure 14 shows the mean CRPSS of the MEFP-GFS temperature forecasts against sample climatology. The results are shown at selected forecast lead times in the downstream basins. The seasonality of the forecast skill shown in figure 14 is controlled by three factors. First, the precipitation and temperature 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. Secondly, the CRPSS measures the *relative* quality of the forecasts. Under warm (summer) conditions, the MEFP-GFS forecasts are more skillful in the tails of the climatological distribution, where sample climatology is conditionally biased. They are less skillful at moderate temperatures, where sample climatology performs reasonable well. Under cool (winter) conditions, the MEFP-GFS forecasts are more skillful at relatively warmer temperatures, again because sample climatology does not predict conditionally upon the state of the atmosphere (and warm temperatures are unusual in cool months). Thirdly, as indicated in figure 11, the MEFP-GFS forecasts are conditionally biased at the coldest temperatures. As forecast skill partly depends on the conditional biases (Appendix B), the advantage of the MEFP-GFS forecasts is reduced at the lowest temperatures during the cold season. Nevertheless, the MEFP-GFS temperature forecasts are considerably more skillful than sample climatology at all forecast lead times and for all observed temperatures.

5.1.4 Aggregation period

The MEFP precipitation forecasts were verified for several aggregated periods, ranging from six hours to multi-day averages. In order to isolate the effects of aggregation period, the verification results were averaged over a constant forecast horizon. Specifically, the forecasts and observations were aggregated into sub-periods of 0.25, 1-, 2-, 3-, 4-, 6- and 12- days, before being paired and verified. The results were then averaged over the 12-day forecast horizon. For example, in forming the 6-day aggregations, the precipitation was accumulated from 1-6 days and 7-12 days. The two forecast accumulations were paired with corresponding observed accumulations, and the verification results were then averaged over the two sub-periods. The aggregation was performed separately for each ensemble trace, in order to preserve the statistical dependencies between forecast lead times. When aggregating the verification results, the sample data may be pooled and the statistics computed from the pooled sample or average performance computed from the statistics of the sub-periods. Unless the measures are additive, these two approaches will produce different results. In this case, averaging was performed, as the results will be smoother when the sub-periods comprise very different central tendencies (as might be expected for a broad range of forecast lead times). In order to compare the verification results conditionally upon observed value, separate climatologies were derived for each accumulation period. The verification thresholds were then fixed at common quantiles of the aggregated climatologies.

Figures 15 and 16 show, respectively, the RME of the ensemble mean forecast and the correlation of the ensemble mean forecast and observation for increasing amounts of observed precipitation. Figure 17 shows the CRPSS of the MEFP-GFS forecasts against raw climatology. In general, there is a systematic relationship between forecast skill and accumulation period in the MEFP-GFS precipitation forecasts. This is consistent with studies of other Numerical Weather Prediction (NWP) models, including raw precipitation forecasts from NCEP's Short Range Ensemble Forecast (SREF) system (e.g. Brown et al., 2012). It is also consistent with studies of HEPS, where forecast skill generally improves with increasing basin size (e.g. Hou et al., 2009;

Pappenberger et al., 2009), as well as accumulation volume. Indeed, the MEFP indirectly accounts for this relationship between forecast quality and aggregation period by predicting at one temporal scale with so-called “canonical events” derived at several other scales (Appendix A). When two cross-correlated variables are aggregated in time or space, the cross-correlations between them will increase, as random errors at finer scales are averaged out. Positive precipitation amounts are underestimated by the ensemble mean forecast in all RFCs, but the relative bias declines with increasing accumulation period.

In general, the correlation increases with increasing accumulation period in FTSC1 and DOLC2 while the CRPSS increases in BLKO2, FTSC1 and DOLC2. Given the separation in time between snow accumulation and melting, this sensitivity to aggregation period may be important in CBRFC, at least for long-range forecasting, where the forecast horizon covers a larger fraction of the accumulation period. In CNNN6, the dependence of forecast skill on aggregation period is less clear. An examination of the CRPS reveals a decline (improvement) in the CRPS with increasing aggregation period for both sources of forcing, but the GFS forecasts are less skillful than climatology at long forecast lead times (e.g. figure 5). Thus, in CNNN6, the precipitation forecasts with climatological forcing benefit more from the increase in aggregation period than those with GFS forcing.

5.2 Quality of the raw streamflow forecasts

The raw streamflow forecasts were verified against observed streamflow and simulated streamflow. The results are presented by forecast lead time, season and amount of streamflow.

5.2.1 Forecast lead time and season

Figure 18 shows the RME of the ensemble mean forecast with increasing forecast lead time. The results are shown for the upstream and downstream basins separately. During the wet season, the forecast ensemble mean is relatively unbiased in MA- and CN-RFCs, with a slight tendency to overforecast at early lead times and

underforecast at late lead times in MARFC. Verification against the simulated streamflows indicates that the unconditional biases contributed by the MEFP are small during the wet season (<10%), but increase during the dry season. When comparing the unconditional biases in the forcing ensembles to those in the streamflow forecasts, the streamflow forecasts are relatively unbiased; that is, the hydrologic models attenuate some of the biases in the meteorological ensembles. In CNRFC, the unconditional biases in precipitation increase during the dry season (not shown), with negative biases of 5-15%. These translate into small negative biases in the streamflow forecasts when compared to simulated streamflows. In MARFC, the MEFP-GFS forecasts consistently underestimate the observed precipitation by ~20% at lead times of ~8-12 days (figure 4 also). Again, this translates into a smaller negative bias when verifying against the simulated streamflows. However, during the dry season, the ensemble mean forecast significantly overestimates the observed streamflow in most basins, i.e. contains a wet bias. This is not surprising, as the SAC-SMA and continuous API models are calibrated to favor wet conditions and, particularly in DOSC1 and FTSC1, there is little streamflow during the summer months (figure 2). In CBRFC, the unconditional biases are also relatively small during the dry season when verifying against simulated streamflow. Here, the forecast precipitation underestimates the observed precipitation by ~5-10%, on average (not shown), but the streamflow biases are much smaller. However, the ensemble mean *overestimates* the observed streamflow by ~40% in DOLC2. As indicated above, the calibration of the hydrologic models will tend to favor wet conditions, and there is a strong seasonality in CBRFC, with relatively low flows throughout the dry season (figure 18). Nevertheless, this may reflect a more substantial source of hydrologic bias in DOLC2.

During the wet season in CBRFC, the ensemble mean forecast increasingly overestimates both the observed and simulated streamflows at longer lead times. This reflects a combination of meteorological and hydrologic biases. These biases are probably related to the onset of snowmelt, which is sensitively dependent upon air temperature and liquid precipitation (“rain-on-snow”). While the biases in precipitation are relatively constant (figure 4), the coldest observed temperatures are consistently

overestimated. Despite the high correlations between the observed and forecast temperatures (>0.9, not shown), the MEFP was unable to remove these biases.

In ABRFC, the ensemble mean forecast consistently underestimates the observed streamflow during the wet season in the upstream basin (CBNK1) and consistently overestimates the observed streamflow during the dry season. Thus, while the overall unconditional biases in CBNK1 are relatively small, the seasonal biases are quite large, particularly at long forecast lead times. Although the MEFP consistently underestimates high precipitation amounts (e.g. figure 10), and the wet season receives significantly more precipitation than the dry season (figure 2), these biases are hydrological in origin. Indeed, when verifying against simulated streamflow, the unconditional biases are relatively consistent between the dry and wet seasons (figure 18).

Figure 19 shows the correlations between the ensemble mean forecast and the observed and simulated streamflows. The results are shown for the MEFP with GFS forcing and, for the simulated streamflows only, the MEFP with climatological forcing; that is the potential skill with climatological forcing after factoring out any hydrologic uncertainties. Figure 20 shows the CRPSS of the streamflow forecasts with GFS forcing against those with climatological forcing. In principle, the forecasts with GFS forcing should not be skillful until some period after initialization, as the initial conditions are common to both the MEFP-GFS and resampled climatology forecasts. In practice, some skill is apparent, as the forecasts are verified at a daily aggregation and these basins respond quickly to precipitation. For this reason, the CRPSS increases during the first day, peaks with the residual contribution from the MEFP-GFS and declines thereafter.

In ABRFC, the correlations between the ensemble mean forecast and the observed streamflow are not significantly impacted by hydrologic uncertainty. In other words, the ensemble mean forecast is no more correlated with the hydrologic simulations than the observations. However, the correlations deteriorate rapidly with increasing forecast lead time, in keeping with the relatively poor quality of the precipitation forecasts (figure 3) and the lack of hydrologic persistence in these basins.

Also, at early lead times, there is a decline in CRPSS when verifying against observed streamflow (figure 20), which reflects an increase in hydrologic bias (e.g. figure 18).

Unlike ABRFC, the correlations are consistently high in CBRFC (figure 3). Here, the seasonal accumulation and melting of snow leads to hydrologic persistence over several months. As such, the initial conditions contribute substantial skill to the streamflow forecasts. Clearly, however, this skill cannot be attributed to the HEFS specifically. In terms of the HEFS, the GFS component of the MEFP contributes only modest skill to precipitation (figure 5). When combined with the skill of the temperature forecasts (e.g. figure 14) and the non-linearity of the hydrologic models, this does contribute *potential* skill to the streamflow forecasts. For example, when verifying against simulated streamflow, the GFS component of the MEFP improves upon climatological forcing by 20-30% at early lead times. In practice, however, this skill cannot be realized without streamflow post-processing, as the hydrologic uncertainties and biases eliminate any positive contribution from the MEFP (figure 20).

As in CBRFC, the hydrologic uncertainties reduce the quality of the streamflow forecasts in MARFC. Thus, while the ensemble mean of the streamflow forecasts is relatively unbiased (figure 18), the correlations and CRPSS are both impacted by hydrologic uncertainty (figures 19 and 20). As the EnsPost is calibrated with hydrologic simulations, the forecasts should benefit from streamflow post-processing (see below). During the dry season, however, the correlations are systematically lower in CNNN6 than WALN6 when verifying against streamflow observations (figure 19), despite the proximity of these two basins. As indicated above, the inflows to Cannonsville Reservoir were estimated from the (change in) daily storage and outflows, without accounting for evaporation. Unlike the uncertainties contributed by the hydrologic modeling, errors in the streamflow observations may lead to spurious adjustments by the EnsPost that cannot be identified through verification.

In keeping with the skill of the meteorological ensembles in CNRFC, (e.g. figure 18), the streamflow forecasts show good correlations (figure 19) and reasonable skill during the wet season (figure 20). Indeed, in the absence of hydrologic uncertainties,

the streamflow correlations are substantially higher when using the GFS forcing in the MEFP than climatological forcing (figure 19). While the correlations are higher during the dry season, this originates from hydrologic persistence, rather than the MEFP specifically. Thus, when factoring out the hydrologic uncertainties and biases, the potential skill is much lower under dry conditions than wet conditions (figure 20). When including the hydrologic uncertainties and biases, the actual skill is further diminished. However, the raw streamflow forecasts should be improved by streamflow post-processing, as the EnsPost benefits from hydrologic persistence under dry conditions (Appendix A).

5.2.2 Magnitude of streamflow

The raw streamflow forecasts were verified conditionally upon the magnitude of streamflow at each forecast lead time. Figure 21 shows the RME of the ensemble mean forecast for "low", "moderate" and "high" streamflows with climatological probabilities (C_p) of 0.1, 0.75 and 0.95, respectively (table 1). Figure 22 shows the correlations between the ensemble mean forecast and the observed streamflows, as well as the simulated streamflows. As indicated in Section 4.4, continuous measures, such as the RME, were derived from the subset of verification pairs whose observed (or simulated) value exceeded the threshold. In other words, "low flows" comprise the 90% of flows that exceed $C_p = 0.1$ and not the 10% of flows that fall below this threshold. For discrete measures, such as the Brier Score, the forecast event ($X > F_n^{-1}(C_p)$) and its complement ($X \leq F_n^{-1}(C_p)$) have the same error in absolute terms.

In AB-, CN- and MA-RFCs, the raw streamflow forecasts are conditionally biased in the ensemble mean with increasing streamflow amount (figure 21). At moderate and high streamflows in MA- and CN-RFCs, there is a smooth increase in conditional bias over the forecast horizon. However, in ABRFC, the conditional bias increases rapidly over the first week, reaching ~50% by the second week for "moderate" streamflows and ~75% for "high" streamflows. These conditional biases originate from the meteorological forcing rather than the hydrologic modeling. Indeed, the biases are similar when

verifying against observed and simulated streamflows (figure 21) and reflect a large conditional bias in the precipitation forecasts in all RFCs (figure 7). However, in CBRFC, the streamflow forecasts are relatively unbiased with increasing streamflow amount. This stems from the importance of snowmelt in generating large streamflows in DRRC2 and DOLC2. By definition of the skewed probability distribution of precipitation, snow accumulation will be dominated by small or moderate storms, implying a weaker dependence of high streamflows on large storms (and the conditional biases therein), except when snowmelt is accelerated by heavy rainfall (i.e. rain-on-snow).

Figure 23 shows the BSS of the raw streamflow forecasts with GFS forcing against those with climatological forcing. The skill from the initial conditions does not factor into the BSS, as the (warm) states used to initialize the hydrologic models are the same for both sources of forcing. Rather, when verifying against observed streamflows, the BSS reflects the hydrologic uncertainties without the skill contributed by the initial conditions. When verifying against simulated streamflows, the BSS shows the potential contribution of the MEFP with GFS forcing after factoring out *all* of the hydrologic uncertainties and biases, including any contribution from the initial conditions.

While the precipitation forecasts contain large conditional biases, the forcing ensembles have the potential to contribute meaningful skill in all basins and at all streamflow thresholds. In general, the potential skill declines more rapidly at high streamflow thresholds as the forecast lead time increases. This is not surprising, as high precipitation is more difficult to predict at longer lead times. However, at early lead times, there is good potential skill in all RFCs. Owing to the predictability of winter storms in the North Coast Ranges, and the calibration of the SAC-SMA model to favor wet conditions, the potential skill increases with increasing streamflow amount in CNRFC. Indeed, when using the GFS forecasts as input to the MEFP, the correlations are substantially greater for high streamflow amounts in CNRFC than those associated with climatological forcing (figure 22).

Despite the potential contribution from the MEFP with GFS forcing, the actual contribution is reduced by hydrologic uncertainty. This is evidenced by the decline in

BSS at all streamflow thresholds when verifying against observed streamflows, except in CNRFC (figure 23). This may be partly addressed through streamflow post-processing (see Section 5.3). Indeed, in CBRFC, the correlations between the ensemble mean forecast and simulations decline only gradually over time, and remain strong at all streamflow thresholds (figure 22). This reflects the high degree of persistence in the streamflow models, which should benefit the EnsPost. For the same reason, however, the practical impacts may be limited. The raw streamflow forecasts are already skillful and the EnsPost cannot directly improve the model states or account for timing errors associated with the onset of snowmelt. In ABRFC, the raw streamflow forecasts should also benefit from post-processing at early forecast lead times, particularly during the wet season in CBNK1, where the hydrologic biases are significant (figure 21) and the correlations are reasonable (figure 22). However, as the amount of streamflow increases, the correlations decline rapidly over the first week, which reduces the scope for streamflow post-processing (figure 22). This is mirrored in the temporal autocorrelation of the observed streamflows, as well as the cross-correlations between the observed and simulated streamflows (not shown) on which the EnsPost relies to make skillful adjustments (Appendix A).

5.3 Quality of the bias-corrected streamflow forecasts

The bias-corrected streamflow forecasts are verified against observed streamflows in each basin. The results are presented by forecast lead time and season, amount of streamflow, and aggregation period. The overall skill is determined relative to the “raw” streamflow forecasts with climatological forcing, and the contributions from the MEFP with GFS forcing 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.

5.3.1 Forecast lead time and season

Figure 24 shows the RME of the bias-corrected streamflow forecasts by season and forecast lead time. For ease of comparison, the RME of the raw streamflow forecasts is also shown. The ability of the EnsPost to eliminate bias will depends on several factors, including any residual meteorological biases from the MEFP (not addressed by the EnsPost) and the strength of the predictors, among other things. Also, there are various measures of unconditional and conditional bias. These measures have different sensitivities and practical implications. For example, the ensemble mean forecast is sensitive to outliers in a skewed probability distribution, but is nevertheless important for operational forecasting.

In general, the unconditional biases in the raw streamflow forecasts are reduced or eliminated by the EnsPost. However, the improvements are not uniform. For example, in CBRFC, the biases are effectively eliminated during the wet season. During the dry season, the tendency of the raw streamflow forecasts (GFS) to overestimate the observed streamflow is largely corrected, particularly in the lower basin (DOLC2), but the ensemble mean forecast is somewhat "over-adjusted." Thus, there is a negative bias of 5-10% in DOLC2 that increases with increasing forecast lead time. In the downstream basins of MARFC, and ABRFC, the EnsPost introduces a larger negative bias of ~15% during the wet and dry seasons, respectively. When estimating the parameters of the EnsPost, the estimates are not guaranteed to produce ensembles that are unconditionally unbiased (Seo et al., 2006). Rather, the predictions aim to minimize the mean CRPS of the post-processed ensembles. The unconditional bias is only one component of the CRPS, specifically an element of the reliability term (Hersbach, 2000). In practice, second-order effects may introduce unconditional biases into the post-processed ensembles. During the wet season in CNRFC, the raw streamflow forecasts are relatively unbiased (figure 24), and remain unbiased after streamflow post-processing. However, during the dry season, the raw streamflow forecasts overestimate the observed streamflows by ~5-10% in FTSC1 and ~15% in DOSC1, which is largely corrected by the EnsPost.

Figure 25 shows the overall skill of the post-processed streamflow forecasts with GFS forcing. The reference forecast comprises the raw streamflow with climatological forcing. In addition to the overall skill, the CRPSS is factored into contributions from the MEFP with GFS forcing and the EnsPost

$$\underbrace{\frac{CRPS_{CLIM} - CRPS_{GFSPOST}}{CRPS_{CLIM}}}_{\text{Total skill}} = \underbrace{\frac{CRPS_{CLIM} - CRPS_{GFS}}{CRPS_{CLIM}}}_{\text{MEFP - GFS skill}} + \underbrace{\frac{CRPS_{GFS} - CRPS_{GFSPOST}}{CRPS_{CLIM}}}_{\text{EnsPost skill}}, \quad (4)$$

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 (figure 25). The skill is greatest in CBRFC and CNRFC, particularly during the dry season, where the EnsPost benefits from hydrologic persistence. However, the relative contributions from the MEFP and the EnsPost are very different in CNRFC than CBRFC. During the wet season in CNRFC, the majority of skill originates from the GFS component of the MEFP, with little skill contributed by the EnsPost. Not surprisingly, this is reversed during the dry season, when the basins in CNRFC experience very low flows (figure 2). Here, the EnsPost accounts for the majority of skill in DOSC1 (upper) and a large fraction of the skill in FTSC1 (lower). In contrast, most of the skill in CBRFC originates from the EnsPost, with weak (< 10%) or negative contributions from the MEFP at a daily aggregation. This stems from the relatively poor quality of the GFS precipitation forecasts in CBRFC. Crucially, however, it also reflects the sensitive dependence of streamflow on snow accumulation and melting. Figure 26 shows the paired streamflow forecasts and observations in DOLC2, before and after hydrologic post-processing. The results are shown for selected calendar years and for a forecast lead time of 307-331 hours, where the biases in the MEFP-GFS forecasts are greatest (figure 6). As indicated in figure 26, the raw streamflow forecasts generally predict the onset of snowmelt too early, whether using MEFP-GFS or resampled climatology. However, this timing error is, to some degree, corrected by the EnsPost.

The EnsPost benefits from the strong hydrologic persistence in DOLC2 and applies an amplitude correction using the recent observed streamflow, which indirectly accounts for the timing error. Thus, while the hydrologic uncertainties are relatively more important than the meteorological uncertainties in CBRFC, they also operate on different temporal scales. Snow accumulates on a monthly basis, but melts in days. For this reason, some caution is needed when interpreting the relative contributions of the meteorological and hydrologic uncertainties in DOLC2 and in other mixed regimes.

In AB- and MA- RFCS, the overall CRPSS is generally lower and the relative contributions from the MEFP-GFS and the EnsPost are more variable (figure 25). Nevertheless, in WALN6, the bias-corrected forecasts with GFS forcing improve upon the raw forecasts with climatological forcing by up to ~40% during the wet season and ~30% during the dry season. At the earliest forecast lead times (~1-2 days), the majority of skill in AB- and MA-RFCs originates from the EnsPost, while the MEFP-GFS contributes a significant fraction of the overall skill after ~2 days.

5.3.2 Magnitude of streamflow

Figure 27 shows the Brier Skill Score (BSS) for increasing amounts of observed streamflow at a forecast lead time of ~18-42 hours. The scores are plotted against climatological exceedence probability on a probit scale, but are labeled with actual probability. For example, 0.1 denotes the daily mean streamflow that is exceeded, on average, only once in every 10 days. The BSS measures the gain in skill (or reduction in BS) of the streamflow ensembles with GFS forcing, and with the combination of GFS forcing and EnsPost, relative to those with climatological 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 climatological forcing, which removes any contribution from the hydrologic initial conditions.

In MA-, AB- and CNRFCs, the streamflow forecasts with GFS forcing are relatively unskillful at all streamflow thresholds (when measured in terms of the BSS). However, in CNRFC, the GFS contributes significant, and increasing, skill for observed streamflows above the median. Indeed, for streamflow amounts that are exceeded, on average, only once every 10 days, all of the skill in DOSC1 and FTSC1 originates from the GFS component of the MEFP.

As indicated in figure 27, the “raw” streamflow forecasts become increasingly reliable as the observed streamflow increases while, under dry conditions, the EnsPost also improves the forecast skill and reliability. However, the BSS is a relative measure, and climatological forcing is inherently less reliable at high precipitation amounts. Figure 28 shows the reliability diagrams for the bias-corrected streamflow forecasts with GFS forcing. The results are shown for selected thresholds and for the downstream basins only. 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. Notwithstanding sampling uncertainties, the streamflow forecasts in FTSC1 are reliable at a broad range of streamflow thresholds (figure 28). Thus, in CNRFC, at least for early forecast lead times, the HEFS benefits from a combination of the EnsPost (under dry conditions) and the GFS component of the MEFP (under wet conditions). However, in other RFCs, most of the skill from the HEFS originates from improvements in reliability accrued by streamflow post-processing, particularly at low to moderate streamflow thresholds. 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. In MARFC, following streamflow post-processing, the probabilities of exceeding low or moderate streamflow thresholds are

much less skillful in Cannonsville (CNNN6) than Walton (WALN6). This stems from a combination of lower resolution and higher Type-I conditional bias (figure 27). This may originate from a truncation of the estimated inflows to Cannonsville Reservoir under dry conditions (see above).

Figure 29 shows the likelihood-base-rate decomposition of the BSS, comprising the "relative Type-II conditional bias", the "relative discrimination" and the "relative sharpness". The forecast probabilities are relatively sharp if they have greater variance than the reference forecast (Appendix B). The variance is greatest when an equal fraction of the forecasts predict with certainty that an event will occur (probability 1) or not occur (probability 0). The forecasts are relatively discriminatory if, when grouped by observed outcome, the forecast probabilities are more dispersed around their unconditional average than the reference forecasts are around their own unconditional average. When grouped by observed outcome, the Type-II conditional bias comprises the mean square difference between the conditional averages of the observed and forecast probabilities. As indicated in Appendix B, the Type-II conditional bias cannot be zero unless the forecasts are perfectly sharp. Figure 30 shows the Relative Operating Characteristic (ROC) for the bias-corrected streamflow forecasts with GFS forcing. The results are shown for selected thresholds and for the downstream basins only. 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).

When conditioning on observed outcome (figure 29), increases in relative sharpness are generally offset by similar increases in relative discrimination. Although sharpness is sometimes viewed as a desirable property of an ensemble forecast (Gneiting et al., 2005), it is only desirable in a conditional sense, i.e. relative to other attributes, and contributes negatively to the overall skill score (Appendix B). As indicated in figure 30, the streamflow forecasts are most discriminatory in DOLC2 and FTSC1. In BLKO2, the forecasts are less discriminatory at low and moderate streamflow thresholds and, in CNNN6, they are less discriminatory at all streamflow

thresholds. Nevertheless, the bias-corrected forecasts are substantially more discriminatory than sample climatology in all basins and at all streamflow thresholds. Since the relative sharpness and relative discrimination are competing factors, and show similar patterns with observed threshold (figure 29), most of the differences between basins are driven by the Type-II conditional bias. While the EnsPost consistently improves the relative reliability of the forecasts (figure 29), the reductions in Type-II conditional bias are less consistent. For example, the conditional bias is reduced for low and moderate streamflow thresholds in AB-, CB- and CN-RFCs, but is similar or higher for streamflows that are exceeded only 10% of the time in AB-, CN- and MA-RFCs. In other words, in AB-, CN- and MA-RFCs, the tendency of the streamflow forecasts to systematically underestimate the highest observed streamflows is not significantly improved by the EnsPost. This is understandable, because the conditional bias originates from the precipitation forcing (e.g. figure 10).

In order to assess the practical impacts of this Type-II conditional bias on forecast error and the capacity to warn about unusually high streamflows, box plots of forecasting error were computed from the post-processed streamflow forecasts. The results are shown in figure 31 for each downstream basin and for a forecast lead time of ~18-42 hours. The box plots are organized by increasing amount of observed streamflow. For high streamflow amounts, they clearly show where the ensemble range has failed to capture the observed streamflow (zero error line). As indicated in figure 31, the Type II conditional biases are relatively constant with increasing streamflow amount in DOLC2. Thus, while the ensemble range does not always capture the observed streamflow, the tendency to systematically underestimate these streamflows is much lower in DOLC2 than other basins. In contrast, the highest observed inflows to CNNN6 are frequently missed, or systematically underestimated, due to a combination of large conditional bias and insufficient spread, at least with ~50 ensemble members. In BLKO2, the conditional bias increases with the increasing streamflow amount. However, this is partially offset by increased spread, i.e. heteroscedacticity, in the streamflow forecasts. In FTSC1, the Type II conditional bias is smaller and the forecast spread generally captures the observed streamflow. At longer forecast lead times (not shown), the correlations decline, and the forecasts increasingly resemble raw climatology

(although much more slowly in CBRFC). By definition, raw climatology has constant spread and is conditionally biased at high streamflow amounts.

5.3.3 Aggregation period

The streamflow forecasts were verified for several aggregated periods, ranging from 1 to 12 days, and then averaged over a constant forecast horizon of 12 days. For example, 3-day aggregations were formed from the streamflow forecasts with lead times of 1-3 days, 4-6 days, 7-9 days and 10-12 days. The aggregate forecasts were derived separately for each ensemble trace. The verification results were then averaged over these four sub-periods. Verification was conducted at increasing thresholds of the observed variable. Separate climatological distributions were derived for each accumulation period and the verification was performed at matching quantiles in the aggregated climatologies. When computing the aggregate statistics, averaging was preferred over pooling of the sample data, as this resulted in smoother estimates of the aggregated quantities (due to a strong dependence on forecast lead time). Results are shown for the post-processed streamflow forecasts with meteorological forcing from the GFS. Figure 32 shows the RME of the ensemble mean forecast for the downstream basin in each RFC. Figure 33 shows the correlation of the ensemble mean forecast, while figure 34 shows the CRPSS; the reference forecast comprises the raw streamflow forecasts with climatological forcing.

In keeping with the sensitivity of the precipitation forecasts to temporal aggregation (see Section 5.1.4), the streamflow forecasts also improve with increasing aggregation period. In all basins, there is a conditional bias in the ensemble mean forecast with increasing observed streamflow. This conditional bias declines as the aggregation period increases. For example, in MARFC, streamflows that are exceeded only 10% of the time are, on average, underestimated by ~60% at a 1-day aggregation versus ~40% at a 12-day aggregation. The lowest conditional biases and, hence, the smallest impacts of aggregation are seen in DOLC2, where hydrologic persistence leads to good performance over the full forecast horizon. Similarly, the correlation of the ensemble mean forecast and observed streamflow increases systematically with

increasing aggregation period. For example, in CNNN6, the top 10% of observed streamflows show an increase in the correlation from ~0.2 at an aggregation period of 1 day to ~0.55 at an aggregation period of 12 days (figure 33).

In some basins, the reductions in conditional bias and increases in correlation are accompanied by similar increases in CRPSS. For example, in FTSC1, the top 5% of observed streamflows show a CRPSS of ~0.275 at a 1-day aggregation and ~0.45 at a 12-day aggregation. However, the patterns in CRPSS are generally less clear (see Section 5.1.4 also). For example, there is a progressive decline in CRPSS with increasing aggregation period in CNNN6 (figure 34), reflecting a similar decline in the precipitation forecasts (figure 5). An examination of the mean CRPS (not shown) reveals an improvement in mean CRPS with increasing aggregation period for both sources of forcing, but the GFS forecasts are less skillful than climatology at long forecast lead times (e.g. figure 5), even at high precipitation and streamflow amounts (e.g. figure 7). Thus, in CNNN6, the streamflow forecasts with climatological forcing benefit more from the increase in aggregation period than those with GFS forcing, which is reflected in the CRPSS.

6. Discussion and conclusions

Ensemble forecasts of precipitation, temperature and streamflow were generated with the NWS HEFS for a ~20 year period between 1979 and 1999. The hindcasts were produced for two basins in each of four RFCs, namely AB-, CB-, CN- and MA-RFCs. The basins comprised a headwater and one immediately downstream basin, of which those in CB- and CN-RFCs were separated into various sub-basins to accommodate the varied elevations there. Precipitation and temperature forecasts were produced with the MEFP using “raw” precipitation and temperature forecasts from NCEP’s frozen GFS (frozen circa 1997) and by resampling the climatology in a moving window of 61 days (“resampled climatology”). The streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS). In AB-, CB- and CN-RFCs, the hydrologic models comprised 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 precipitation, temperature and streamflow forecasts were verified with the Ensemble Verification System (Brown et al., 2010b). The results are presented by forecast lead time, season, and magnitude of the observed variable. In order to separate the meteorological uncertainties from the total (meteorological and hydrologic) uncertainties, the raw streamflow forecasts were verified against simulated streamflows, as well as observed streamflows. Also, when verifying the bias-corrected streamflow forecasts, the total skill was decomposed into contributions from the MEFP with GFS forcing versus the EnsPost.

In general, the precipitation forecasts from the GFS component of the MEFP are more skillful than resampled climatology during the first week, but comprise little or no skill during the second week. In contrast, the temperature forecasts improve upon resampled climatology at all forecast lead times. However, there are notable differences between RFCs, between seasons, and for different magnitudes of the observed and forecast variables. A direct comparison between the raw inputs to the MEFP and the bias-corrected outputs was hampered by the different spatial scales of the inputs and outputs (i.e. a single grid node versus a basin average). Nevertheless, the MEFP-GFS forecasts generally preserve or improve upon the correlations in the raw GFS forecasts.

The GFS component of the MEFP contributes the highest correlations and greatest skill in the CNRFC basins, particularly during the wet season. This is associated with the greater predictability of large storms in the North Coast Ranges during the winter months. Indeed, the skill was much lower during the dry season, particularly at longer lead times. In MARFC, the GFS contributes significant skill at early forecast lead times, with the greatest skill occurring during the wet season and at moderate precipitation amounts. However, the forecast skill declines rapidly with increasing forecast lead time, particularly during the dry season. In the summer months, the MEFP is constrained by the limited ability of the frozen GFS to model convection and by a range of unconditional and conditional biases in the precipitation forecasts. In AB- and CB-RFCs, the seasonal patterns are less pronounced, but the quality of the MEFP-GFS precipitation forecasts is also lower. This originates from a combination of

reduced predictability in the southern plains and in the intermountain region of the western US, together with residual biases that were not removed by the MEFP. For example, in CBRFC, the MEFP-GFS forecasts are only moderately correlated with the observed precipitation, but improve upon climatology until ~7 days. However, in terms of the CRPSS, the forecasts are indistinguishable from climatology after only 3-4 days.

The MEFP precipitation forecasts comprise a range of unconditional and conditional biases. In particular, there is a tendency for the MEFP to underestimate the probability of precipitation (PoP), both with the GFS and resampled climatology. This requires further investigation, as the forecasts of PoP are substantially worse than raw climatology in some basins. For both sources of raw forcing, the MEFP employs a smooth approximation of the climatological distribution, and these biases may originate from the handling of precipitation intermittency in the fitted distribution (among other things). Alongside the underestimation of PoP, the precipitation forecasts are conditionally biased with increasing amounts of observed precipitation. Thus, non-zero precipitation amounts are systematically underestimated in all basins, and large amounts are increasingly underestimated. This originates from a Type-II conditional bias in the MEFP precipitation forecasts. The severity of this conditional bias varies with forecast lead time, RFC, season, and aggregation period. For example, it increases with increasing forecast lead time and declines with increasing aggregation period. In CNRFC, the biases are sufficiently small, and the spread sufficiently large, that the highest precipitation totals are generally forecast with some (non-zero) probability of occurrence, at least for early forecast lead times. However, in other basins, the largest precipitation totals frequently occur without warning and are routinely underestimated by as much as the observed precipitation amount. While the temperature forecasts are conditionally unbiased for most observed temperatures, the coldest temperatures are systematically overestimated. However, as with the highest precipitation totals, these are often associated with rare events, for which the GFS was not adequately initialized or failed to evolve the observed weather, such as the extreme cold of January 1979.

Ultimately, the MEFP is constrained by the ability of the raw forecasts to provide suitable conditioning. The MEFP estimates the observed variable conditionally upon the

raw forecasts. If the raw forecasts do not adequately capture the synoptic conditions, the MEFP has no additional information from which to estimate those conditions or to apply a bias correction that is appropriate to the conditions. For example, many of the highest precipitation totals underestimated by the MEFP were “seen” as light or moderate precipitation amounts and adjusted accordingly. Thus, improving the raw forecasts available to the MEFP is critical. However, this requires a long period of hindcasting, in order to estimate the parameters of the MEFP with reasonably small sampling uncertainty. It also requires a commitment by the weather forecasting agencies, such as the NCEP, to conduct hindcasting at strategic intervals and to continue forecasting with legacy models (providing a window for the HEFS to be re-calibrated), and by the RFCs, to re-calibrate the HEFS and to repeat hindcasting and verification at selected locations (e.g. a test bed).

As with the types and degrees of bias, their practical implications will vary with location and application. For medium-range forecasting, unbiased estimates of the PoP are generally less important for hydrologic applications than unbiased forecasts of moderate and heavy precipitation. Nevertheless, dry conditions are common and any biases in PoP will, therefore, be conspicuous. In contrast, many hydrologic applications rely on unbiased forecasts of heavy precipitation, such as flood warning and water quality forecasting (e.g. of turbidity in the NYCDEP water supply reservoirs). However, sensitivities will vary with basin hydrology. For example, where snow accumulation is important, light and moderate storms are integrated into the snowpack alongside heavy storms, which should reduce the sensitivity to conditional bias. In short, some of the meteorological uncertainty is transferred to the initial conditions of the hydrologic models, which may be improved by monitoring the snowpack. Nevertheless, as the onset of snowmelt is sensitive to air temperature and liquid precipitation (“rain-on-snow”), biases in the meteorological forecasts may be important. In CBRFC, the MEFP systematically overestimates moderately cold temperatures. For example, after 14 days, the ensemble mean of the MEFP-GFS forecasts overestimates the coldest 10% of observed temperatures by ~5°C, on average, both in DRRC2, where the 10^{th} percentile is -9.4°C, and in DOLC2, where the 10^{th} percentile is -5.6°C. These variations in

temperature originate from the mountaneous terrain surrounding the Dolores River. The frozen GFS cannot be expected to resolve such variations. However, the inputs to the MEFP may be improved through careful interpolation of the GFS forecasts, conditionally upon the local terrain (e.g. Hengl et al., 2004), by selecting the neighboring grid cell that is most skillful (e.g. consistent with a leeward or windward slope), or by using forecasts that better resolve the local terrain. In practice, some RFCs manually modify ("mod") the observed inputs to the hydrologic models, in order to account for biases introduced elsewhere in the forecasting process. This practice must be discouraged in the context of the HEFS. The HEFS relies on a reasonable assessment of the individual sources of bias and uncertainty in the streamflow forecasts. This, in turn, depends on a reasonable assessment of the historical forecast errors, which are diagnosed from observations.

In general, both the raw and bias-corrected streamflow forecasts have lower unconditional and conditional biases, stronger correlations and more skill in CB- and CN-RFCs than in AB- and MA-RFCs. However, there are strong variations in forecast quality with the amount of streamflow, forecast lead time, season and aggregation period. The relative importance of the meteorological and hydrologic uncertainties also varies between basins and is modulated by the same controls on forecast quality. Thus, while the meteorological forecasts are an important source of skill and uncertainty in the streamflow forecasts, they propagate and combine with other sources of uncertainty in the hydrologic modeling. For the same reason, the MEFP and the EnsPost contribute varying amounts of skill to the streamflow forecasts, both within and between basins. Further separating between the quality of the statistical modeling (the value added by the HEFS) and the underlying challenges of meteorological and hydrologic modeling is not always straightforward. Nevertheless, by verifying the raw streamflow forecasts against the simulated streamflows, the hydrologic biases were factored out of the total uncertainty. The effects of the hydrologic initial conditions were isolated by using a reference forecast that comprised the same initial conditions. In verifying the post-processed streamflow forecasts, the relative contributions from the MEFP and the EnsPost were isolated through a simple algebraic decomposition of the CRPSS.

While the overall quality of the raw and bias-corrected streamflow forecasts is generally higher in CB- and CN-RFCs, there are important differences in the relative contributions from the MEFP and the EnsPost, particularly during the wet season. Under wet conditions, the GFS component of the MEFP accounts for the majority of skill in CNRFC, while the EnsPost contributes valuable skill under dry conditions. During the wet season, the GFS benefits from the increased predictability of large storms in the North Coast Ranges. The observed streamflow is consistently low in both CN- and CB-RFCs during the dry season, and the meteorological uncertainties are much less important. When factoring out the initial conditions, the EnsPost contributes valuable skill at low and moderate streamflow amounts, although for a narrower range of forecast lead times in CNRFC than CBRFC. These improvements largely stem from an increase in reliability or a reduction in "Type-I conditional bias" following streamflow post-processing. In CBRFC, unlike CNRFC, most of the residual skill during the wet season originates from the EnsPost, particularly at low to moderate streamflow amounts. Here, "residual" refers to the contribution from the EnsPost after factoring out the hydrologic initial conditions. Clearly, this contribution should not be overstated in CBRFC because the raw streamflow forecasts are also exceptionally reliable and skillful *precisely because* of the hydrologic initial conditions. Likewise, the contribution from the MEFP should not be completely ignored. In general, snow accumulates over several months, for which medium-range weather forecasts are inherently less useful. However, melting occurs in days, rather than months, and biases in the temperature forecasts can lead to errors in the timing, and hence amplitude, of the streamflow forecasts during snowmelt. Indeed, the raw streamflow forecasts increasingly overestimate both the observed and simulated streamflows in DRRC2 and DOLC2 as the forecast lead time increases. While these timing errors are partly (indirectly) removed by streamflow post-processing, the EnsPost models the total uncertainty in streamflow volume, not timing errors explicitly. Data assimilation is the preferred approach to adjusting model states (Liu et al., 2012), but is not currently supported by the HEFS.

In AB- and MA-RFCs, the raw and bias-corrected streamflow forecasts generally have larger unconditional and conditional biases, weaker correlations and lower skill, while the relative contributions from the MEFP and the EnsPost are more variable. At

early forecast lead times (~1-2 days), most of the skill in the bias-corrected streamflow forecasts originates from the EnsPost, where the reliability of the forecasts is improved at low and moderate streamflows. However, the forecast skill declines rapidly with increasing forecast lead time and amount of streamflow, particularly in ABRFC. Likewise, as the temporal autocorrelations decline, the EnsPost becomes less skillful and, in WALN6, the relative contribution from the MEFP increases slightly, at least for low and moderate streamflows. At higher streamflows, both the raw and bias-corrected forecasts contain large Type-II conditional biases. In short, as the observed streamflows increase, the forecasts systematically, and increasingly, underestimate those observed streamflows. This is not surprising, because the MEFP-GFS precipitation forecasts contain large conditional biases, particularly in MA- and AB-RFCs (see above). As the EnsPost is calibrated with hydrologic simulations, rather than forecasts, it cannot "see" these conditional biases (and would, in any case, struggle to remove them; Brown and Seo, 2013). Unbiased predictions of high streamflow are important in operational forecasting, although not necessarily at the expense of good performance for low and moderate streamflow. For example, if the forecast streamflows severely underestimate the true streamflows when flooding occurs, this could hamper any efforts to mitigate flood damage. These biases could be reduced through selective post-processing of the GFS forecasts to favor unbiased estimates of heavy precipitation (e.g. Brown and Seo, 2013). However, improving the quality of the raw meteorological forecasts would be preferred, as selective optimization inevitably incurs come loss of unconditional skill. In this context, the MEFP is currently being augmented with precipitation and temperature forecasts from NCEPs operational GFS (GEFS). Corresponding streamflow hindcasts will be generated with the HEFS, in order to establish the benefits of recent improvements in the GFS for operational streamflow forecasting.

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)

CFS/v2 – 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

MAP – Mean Areal Precipitation over a basin/watershed

MAT – Mean Areal Temperature over a basin/watershed

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.

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. This generally has a "smoothing" effect or reduces the sampling uncertainty associated with using a single date across all historical years

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.

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

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, US National Weather Service.

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

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

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.

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.

Bröcker, J., and Smith, L.A. 2007. Increasing the reliability of reliability diagrams. *Weather and forecasting* **22**(3), 651-661.

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 Softwa*re **25**, 854-872.

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. 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. 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.

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.

Chowdhury, S. and Sharma, A. 2007. Mitigating parameter bias in hydrological modelling due to uncertainty in covariates. *Journal of Hydrology* **340**, 197-204.

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

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS, *Journal of Water Resources Planning and Management*, **111**(2), 157–170.

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., Brown, J., Lee, H., He, M., Seo, D-J., Hartman, R., Fresch, M. and Zhu, Y. 2013. The science of NOAA's operational Hydrologic Ensemble Forecast Service. Submitted to *Bulletin of the American Meteorological Society*, 6th November 2012.

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

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.

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.

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

Hashino T., Bradley, A.A., and Schwartz, S.S. 2006. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences Discussions* **3**, 561-594.

Hengl, T., Heuvelink, G.B.M, and Stein, A. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **120** (1-2): 75–93. doi:10.1016/j.geoderma.2003.08.018

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.

Heuvelink, G.B.M. 1998: *Error Propagation in Environmental Modelling with GIS*. Taylor and Francis: London.

Heuvelink, G.B.M. 2002: Analysing uncertainty propagation in GIS: Why is it not that simple? In: Foody, G.M. and Atkinson, P.M. (eds) *Uncertainty in Remote Sensing and GIS*. Chichester: John Wiley and Sons, 155-165.

Hou, D., Mitchell, K., Toth, Z., Lohmann, D. and Wei, H. 2009. The effect of large scale atmospheric uncertainty on Streamflow predictability. *Journal of Hydrometeorology* **10**, 717-733.

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

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.

Jaun, S., and Ahrens, B. 2009. Evaluation of a probabilistic hydrometeorological forecast system. *Hydrology and Earth System Sciences* **13**, 1031-1043.

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.

Kadane, J.B. and Wolfson. L.J. 1998. Experiences in elicitation. *The Statistician* **47**, 3–19.

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.

Kavetski, D., Franks, S.W. and Kuczera, G. 2002. Confronting input uncertainty in environmental modeling. In: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A.N. and Turcotte, R. (eds) *Calibration of Watershed Models*. AGU Books: Washington DC, 49-68.

Kavetski, D., Kuczera, G. and Franks, S.W. 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research* **42**, W03407, doi:10.1029/2005WR004368.

Kavetski, D., Kuczera, G. and Franks, S.W. 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research* **42**, W03408, doi:10.1029/2005WR004376.

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., and Gupta, H. V. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research* **43**, W07401, doi:10.1029/2006WR005756.

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 Discussions* **9**, 3415–3472.

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.

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.

Norton, J.P., Brown, J.D. and Mysiak, J. 2006. To what extent, and how, might uncertainty be defined? Comments engendered by “Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support” Walker *et al*. Integrated Assessment 4: 1 (2003). *Integrated Assessment* **6**(1), 83-88.

Oakley, J. and O'Hagan, A. 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society* ***B*66**, 751–769.

Pappenberger, F. and Beven, K.J. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research* **42**, W05302, doi:10.1029/2005WR004820.

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.

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 Science*s **9**, 381-393.

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

Perica, S. 1998. Integration of Meteorological Forecasts/Climate Outlooks into an Ensemble Streamflow Prediction System. *14th Conference on Probability and Statistics in the Atmospheric Sciences*, American Meteorological Society, Phoenix, Arizona.

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, Bureau of Reclamation, and National Oceanic and Atmospheric Administration, 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, 2012

Regonda, S., 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. Accepted for publication in *Journal of Hydrology*.

Renner, M., Werner, M.G.F., Rademacher, S. and Sprokkereef, E. 2009. Verification of ensemble flow forecasts for the River Rhine. *Journal of Hydrology* **376**(3-4), 463-475.

Rojas, R., Feyen, L. and Dassargues, A. 2009. Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling. *Hydrological Processes* **23**(8), 1131-114.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D. Saisana, M., and Tarantola, S. 2008. *Global sensitivity analysis, The Primer*. John Wiley and Sons: Chichester.

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.

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.

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.

Thirel, G., Regimbeau, F., Martin, E., Noilhan, J. and Habets, F. 2010. Short- and medium-range hydrological ensemble forecasts over France. *Atmospheric Science Letters* **11**(2), 72-77.

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

van den Bergh, J., and Roulin, E. 2010. Hydrological ensemble prediction and verification for the Meuse and Scheldt basins. *Atmospheric Science Letters*, **11**(2), 64-71.

Verbunt, M., Zappa, M., Gurtz, J. and Kaufmann, P. 2006. Verification of a coupled hydrometeorological modelling approach for Alpine tributaries in the Rhine basin. *Journal of Hydrology* **324**, 224-238.

Wagner, J.A. 1979. Weather and Circulation of January 1979. *Monthly Weather Review* **107**, 499–506.

Wilczak, J., McKeen, S., Djalalova, I., Grell, G., Peckham, S., Gong, W., Bouchet, V., Moffet, R., McHenry, J., McQueen, J., Lee, P., Tang, Y. and Carmichael, G. R. 2006. Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004. *Journal of Geophysical Research* **111**, D23S28, doi:10.1029/2006JD007598.

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

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
Latitude (outlet)	37.1292	36.8086	37.6389	37.4725	39.71	40.22	42.1661	42.0628
Longitude (outlet)	-97.6017	-97.2775	-108.06	-108.497	-123.32	-123.63	-75.1403	-75.3747
Latitude (GFS node)	37.5	37.5	37.5	37.5	40.0	40.0	42.0	42.0
Latitude (GFS node)	-97.5	-97.5	-107.5	-107.5	-122.5	-122.5	-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

10. Figures

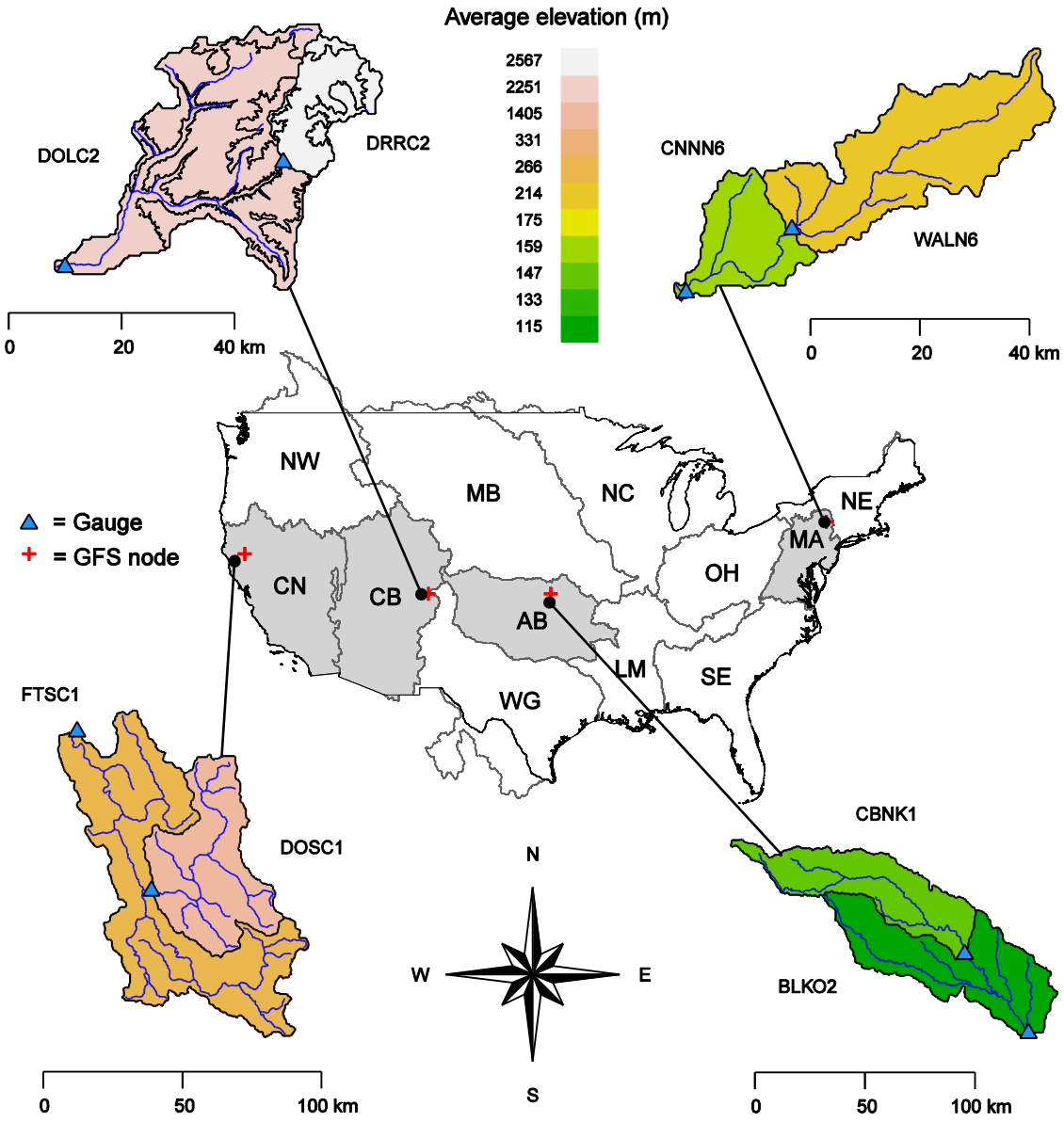

Figure 1: The eight study basins, comprising one upstream and one downstream basin in each of AB-, CB-, CN- and MA-RFCs.

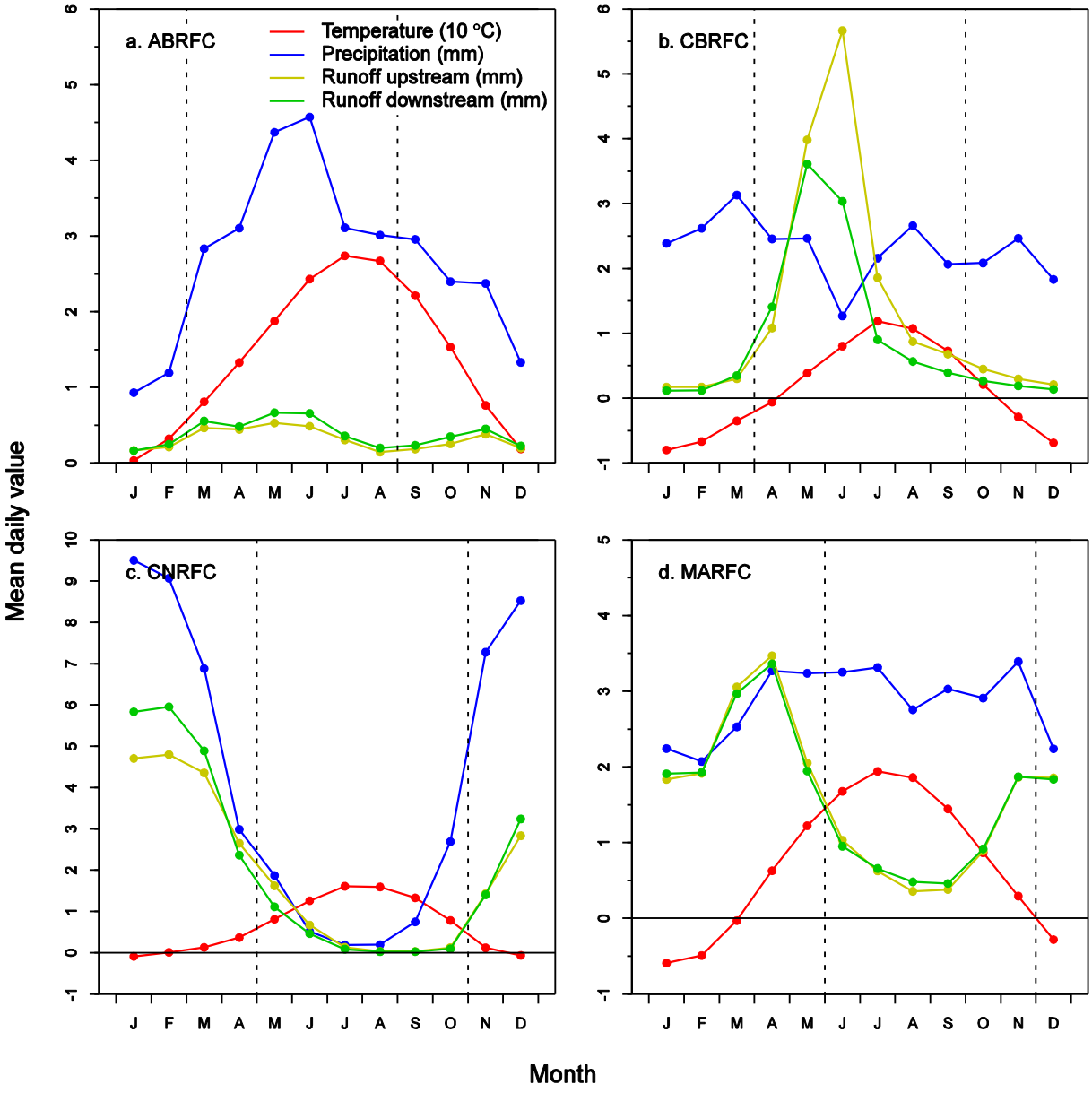

Figure 2: Daily averages of temperature, precipitation and runoff by calendar month for each study basin. The meteorological variables are averaged over the upstream and downstream basins (weighed by basin area).

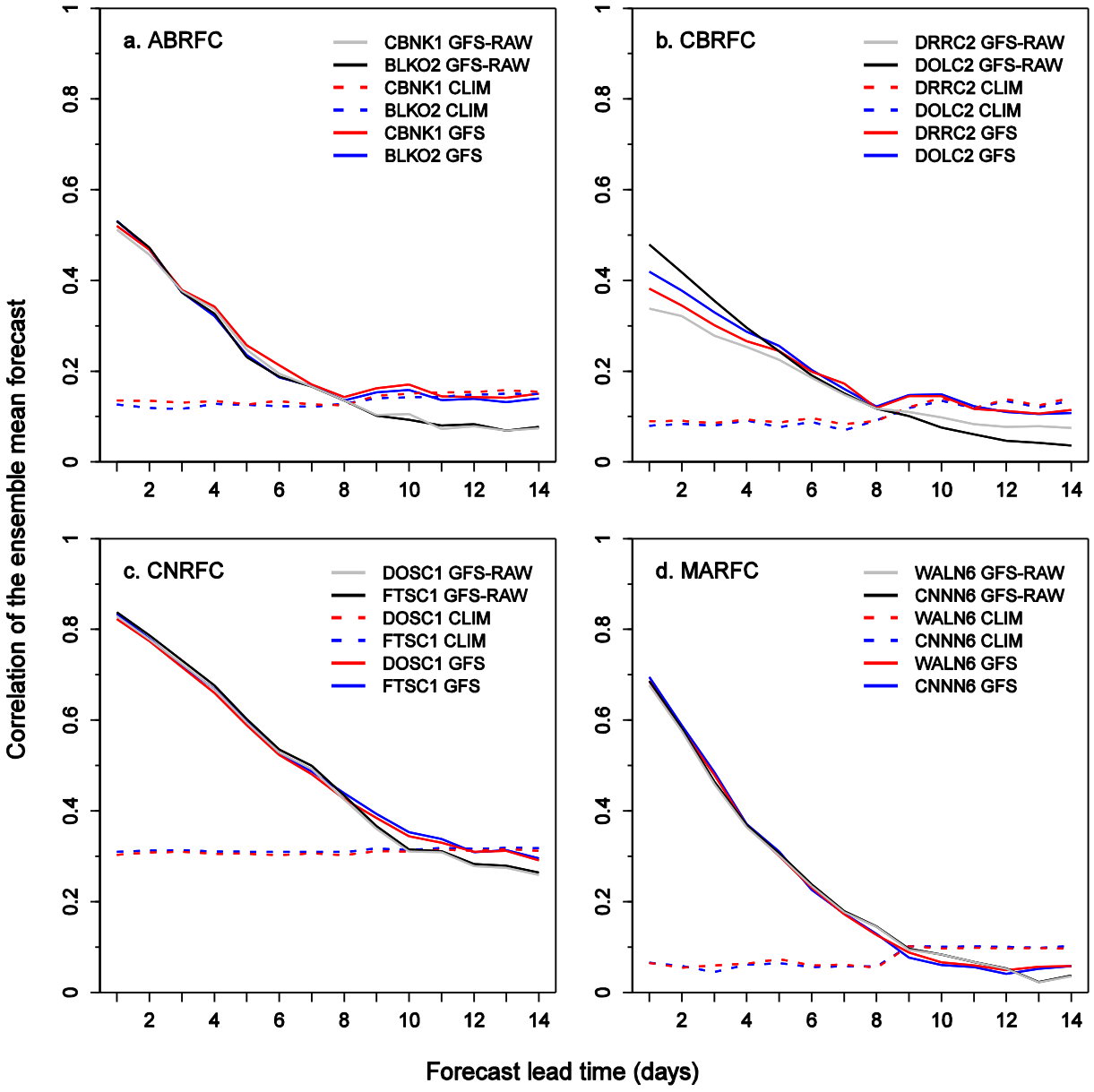

Figure 3: Correlation of the ensemble mean forecast and observed precipitation amounts by forecast lead time for each source of forcing from the MEFP, namely resampled climatology (CLIM) and GFS. Results are also shown for the raw GFS precipitation forecasts (GFS-RAW) used as input to the MEFP.

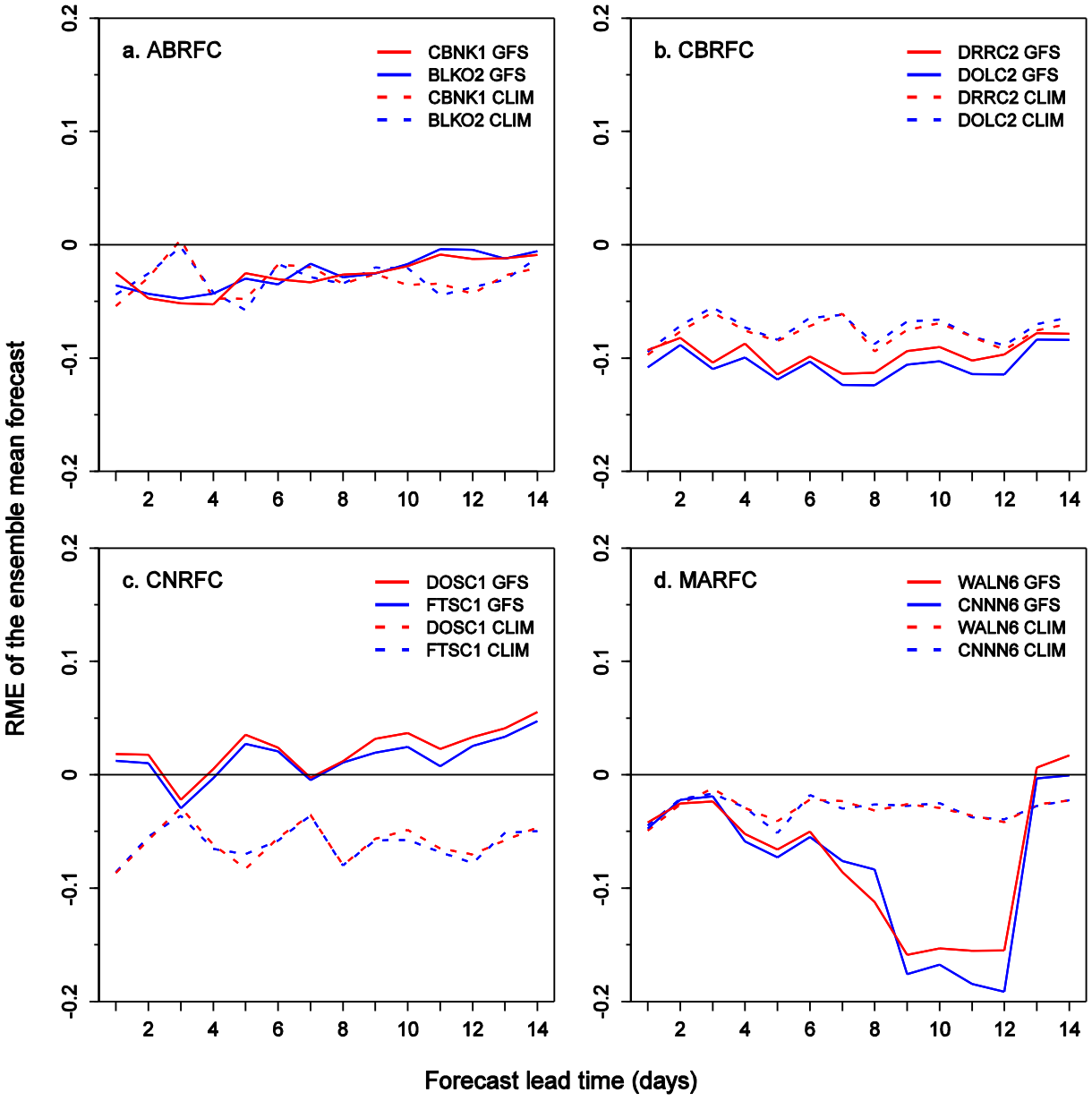

Figure 4: Relative mean error of the ensemble mean forecasts of precipitation by forecast lead time for each source of forcing from the MEFP, namely resampled climatology (CLIM) and GFS.

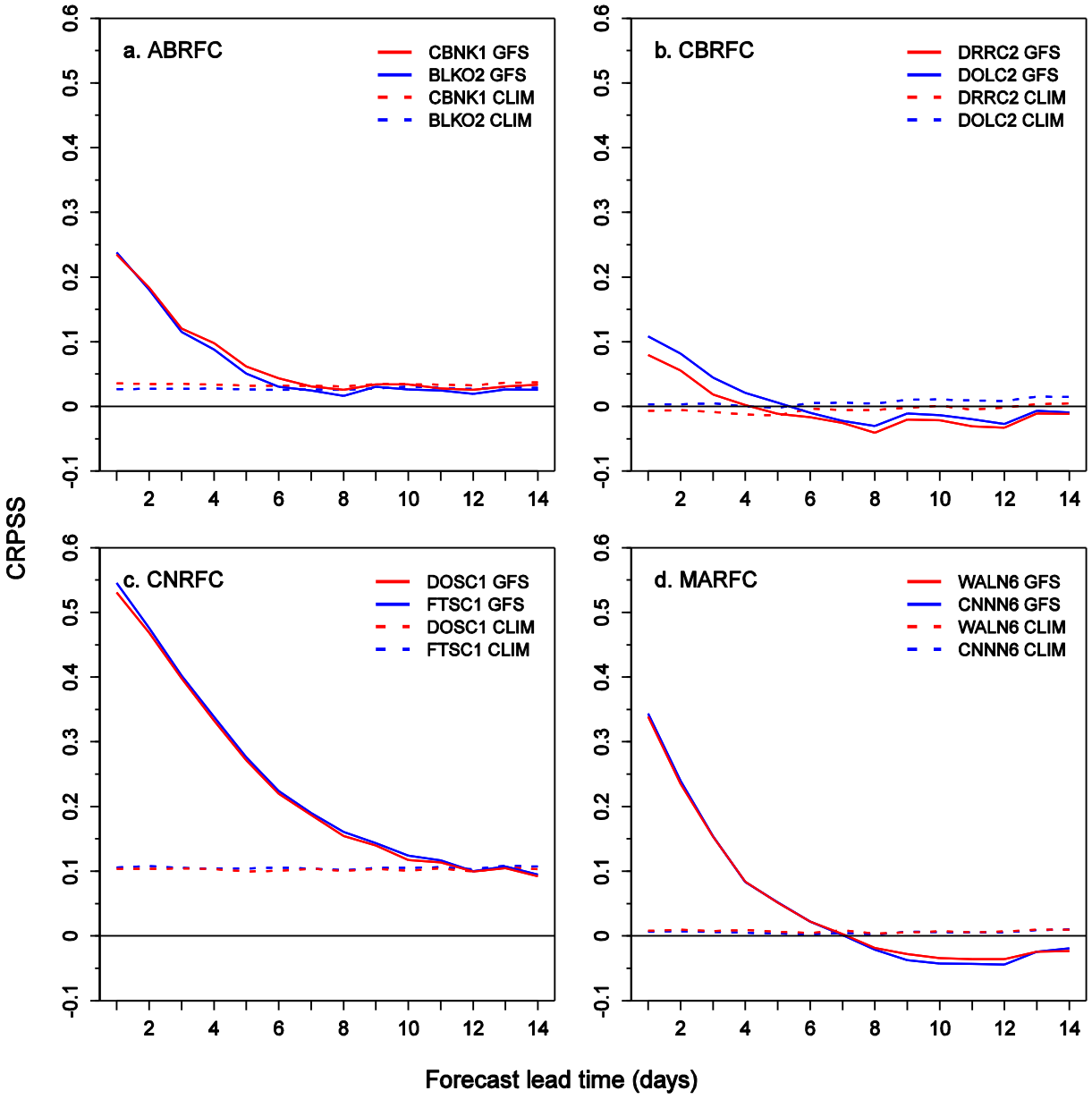

Figure 5: An in figure 4, but for the CRPSS with sample climatology as the reference forecast.

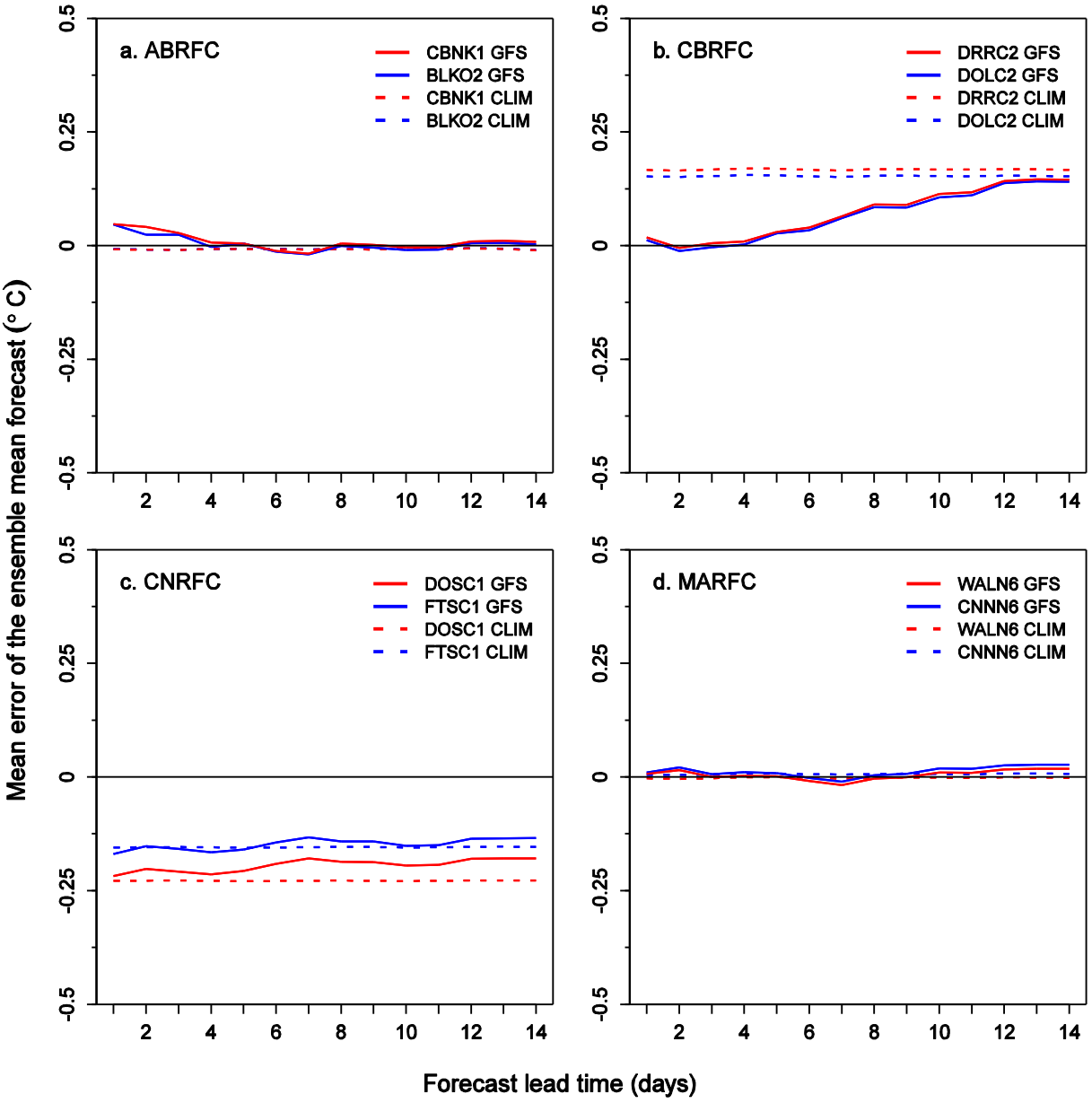

Figure 6: Mean error of the ensemble mean forecasts of temperature (in °C) by forecast lead time for each source of forcing from the MEFP, namely resampled climatology (CLIM) and GFS.

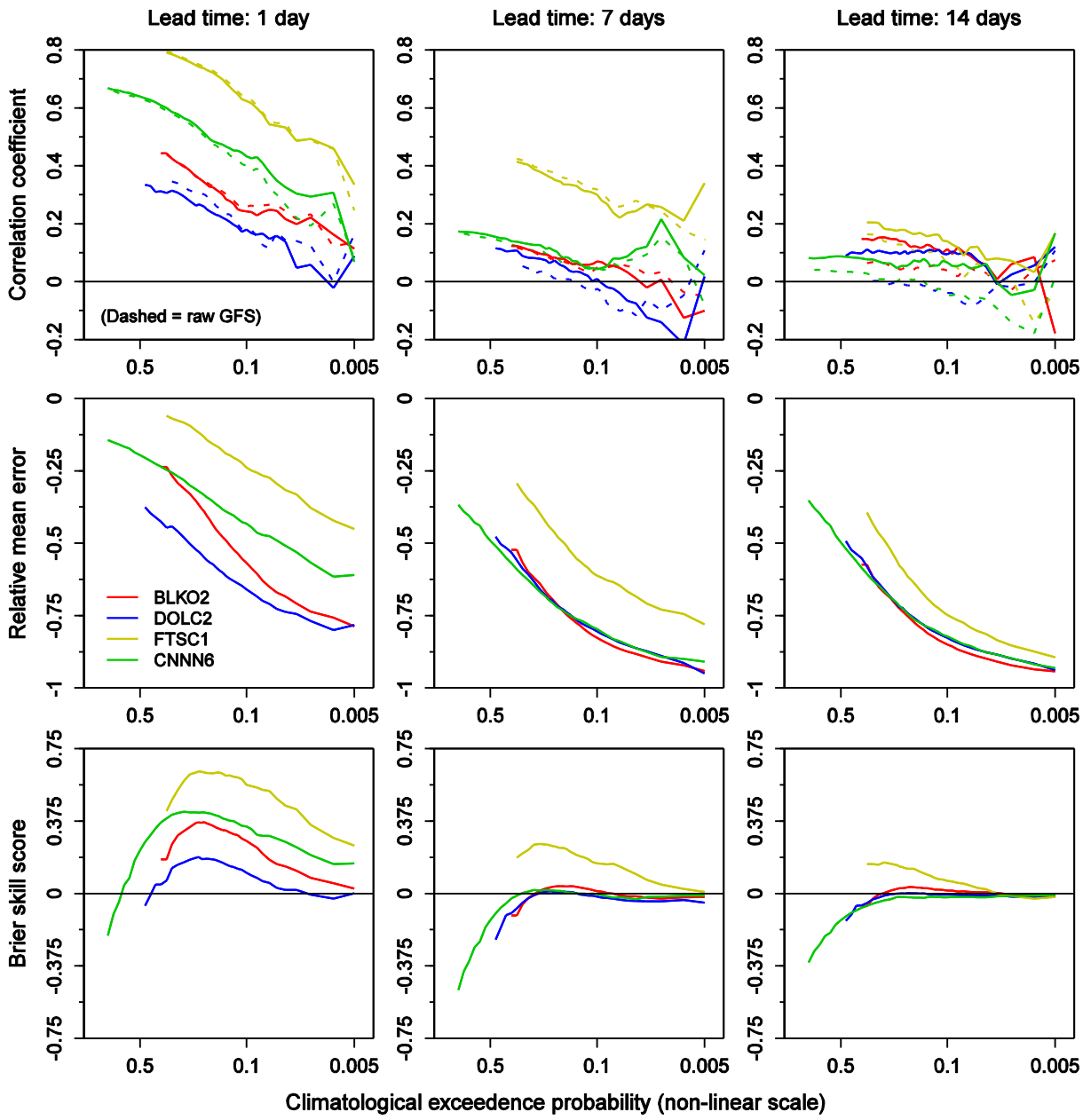

Figure 7: Selected verification metrics for the MEFP precipitation forecasts with input forcing from the GFS at 1, 7 and 14 days. The correlations are also shown for the raw GFS forcing used as input to the MEFP (dashed lines). The results are shown for increasing amounts of observed precipitation, expressed as climatological exceedence probabilities and plotted on a probit scale (but labeled with actual probability).

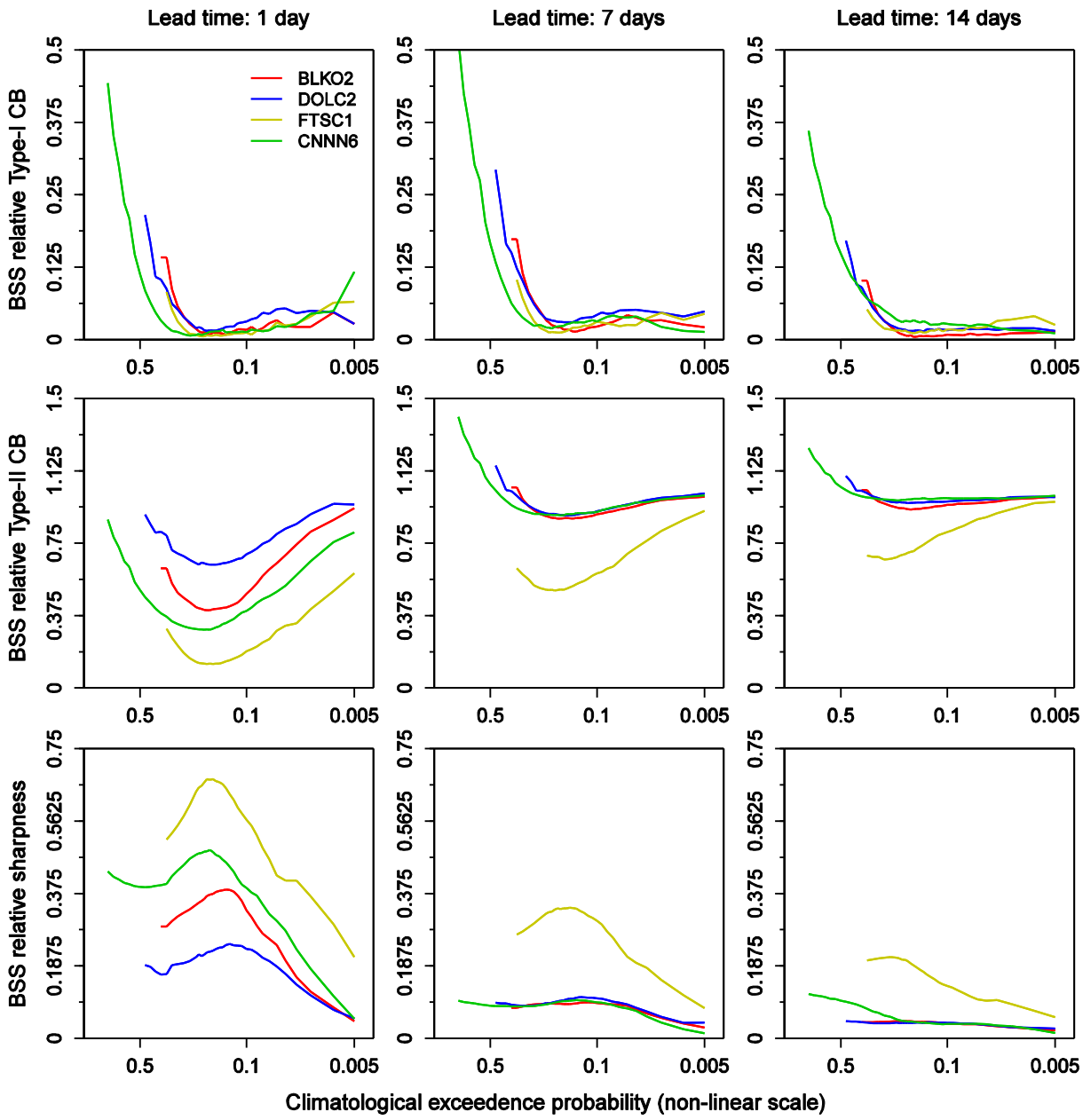

Figure 8: As in figure 7, but for selected components of the BSS.

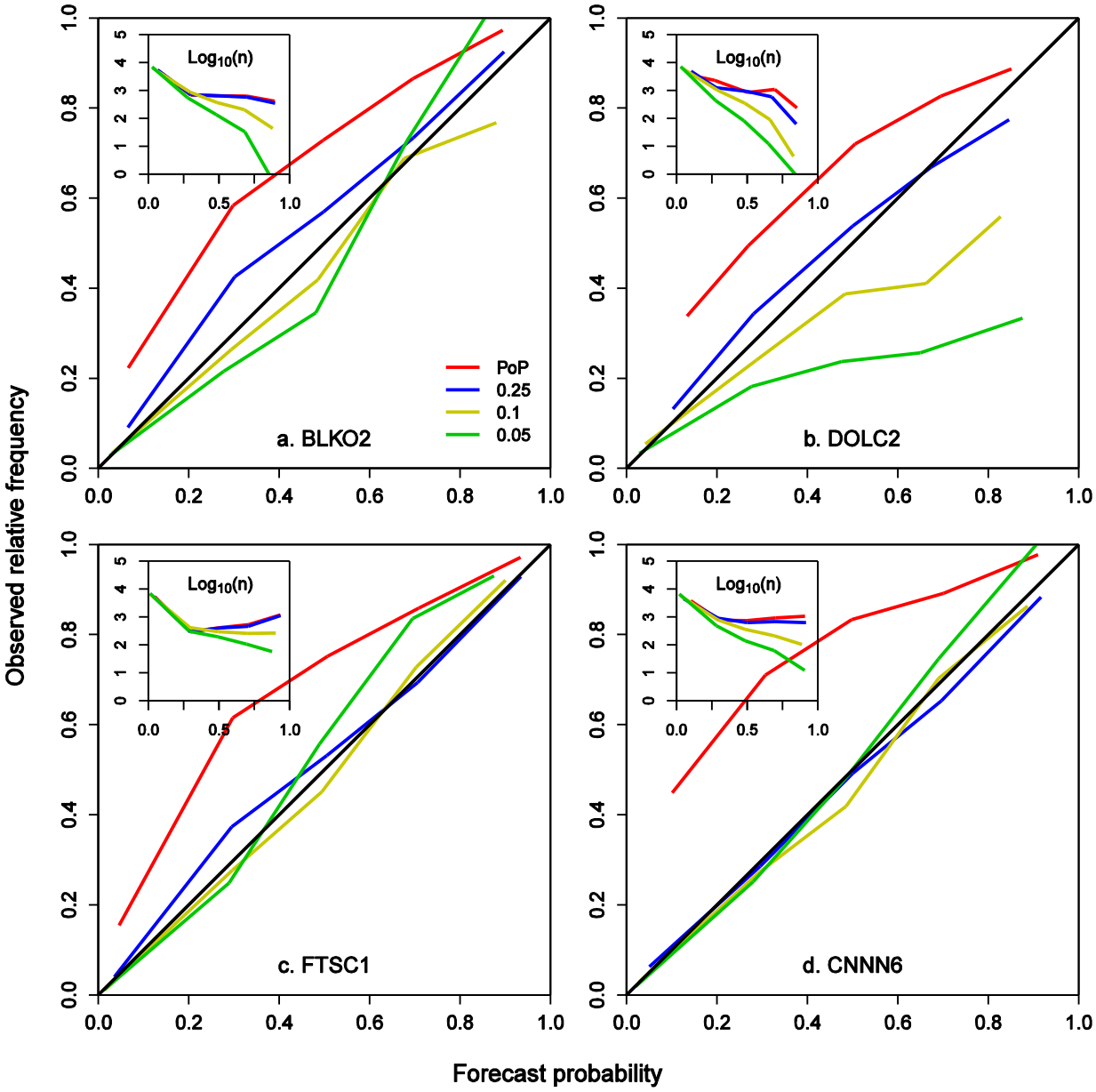

Figure 9: Reliability diagrams for the MEFP-GFS precipitation forecasts in each downstream basin. The results are shown for PoP and selected other thresholds, expressed as climatological exceedence probabilities.

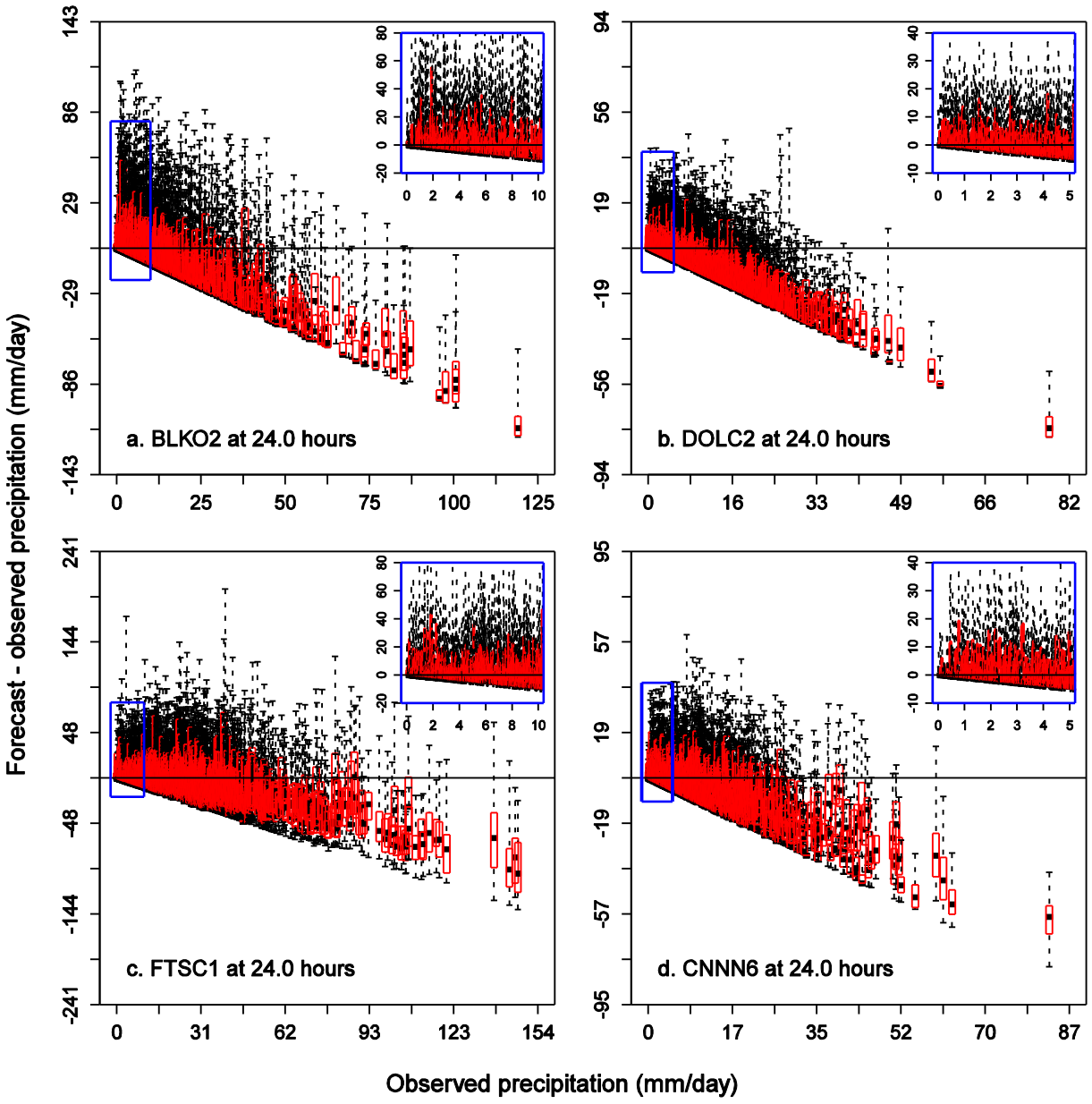

Figure 10: Box plots of errors in the MEFP precipitation forecasts with GFS forcing. The results are shown for the downstream basin in each RFC and for a forecast lead time of 0-24 hours. The boxes are ordered by increasing amounts of observed precipitation, with an inset plot for the smallest observed amounts.

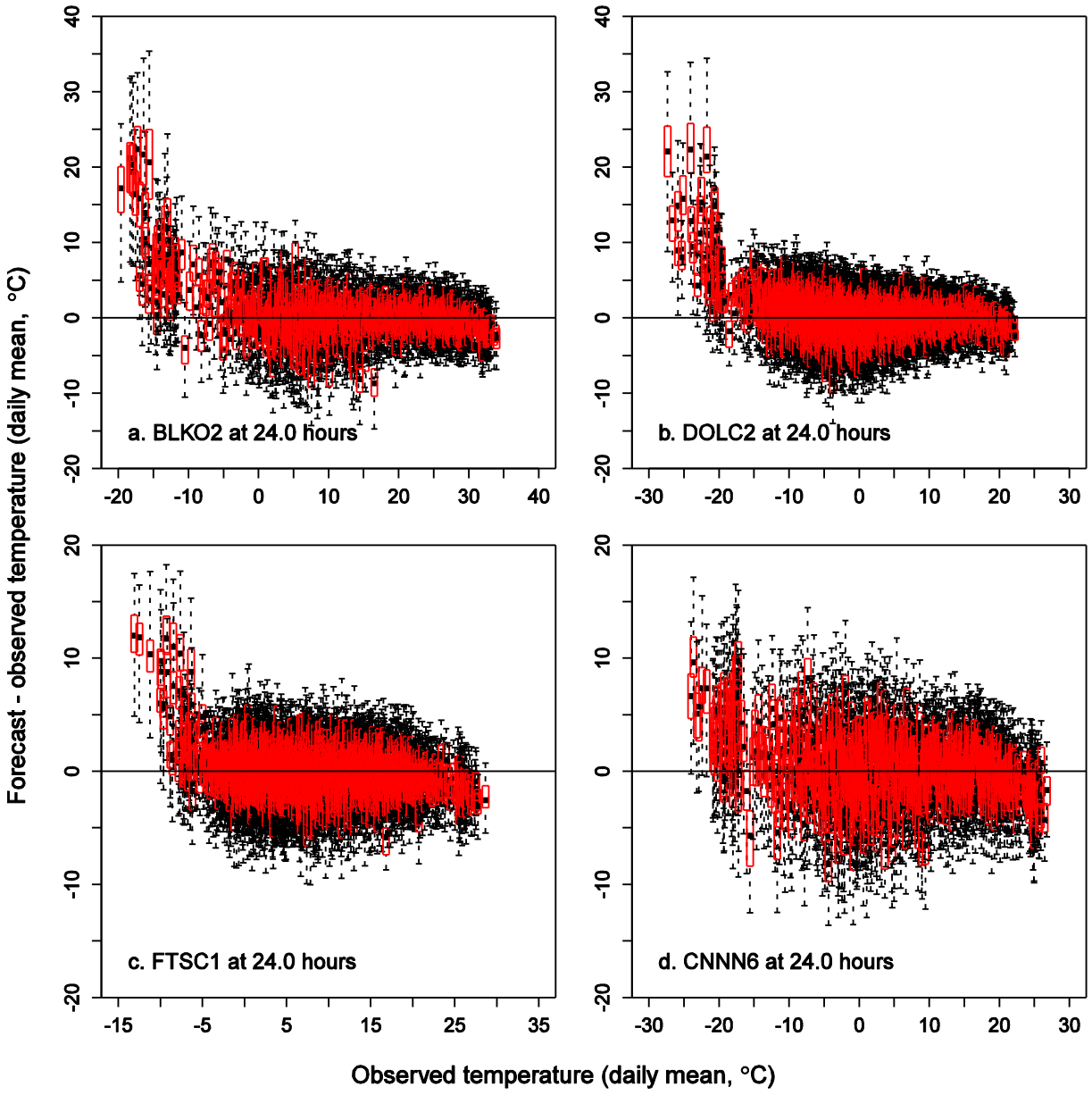

Figure 11: As in figure 9, but for temperature (with no inset).

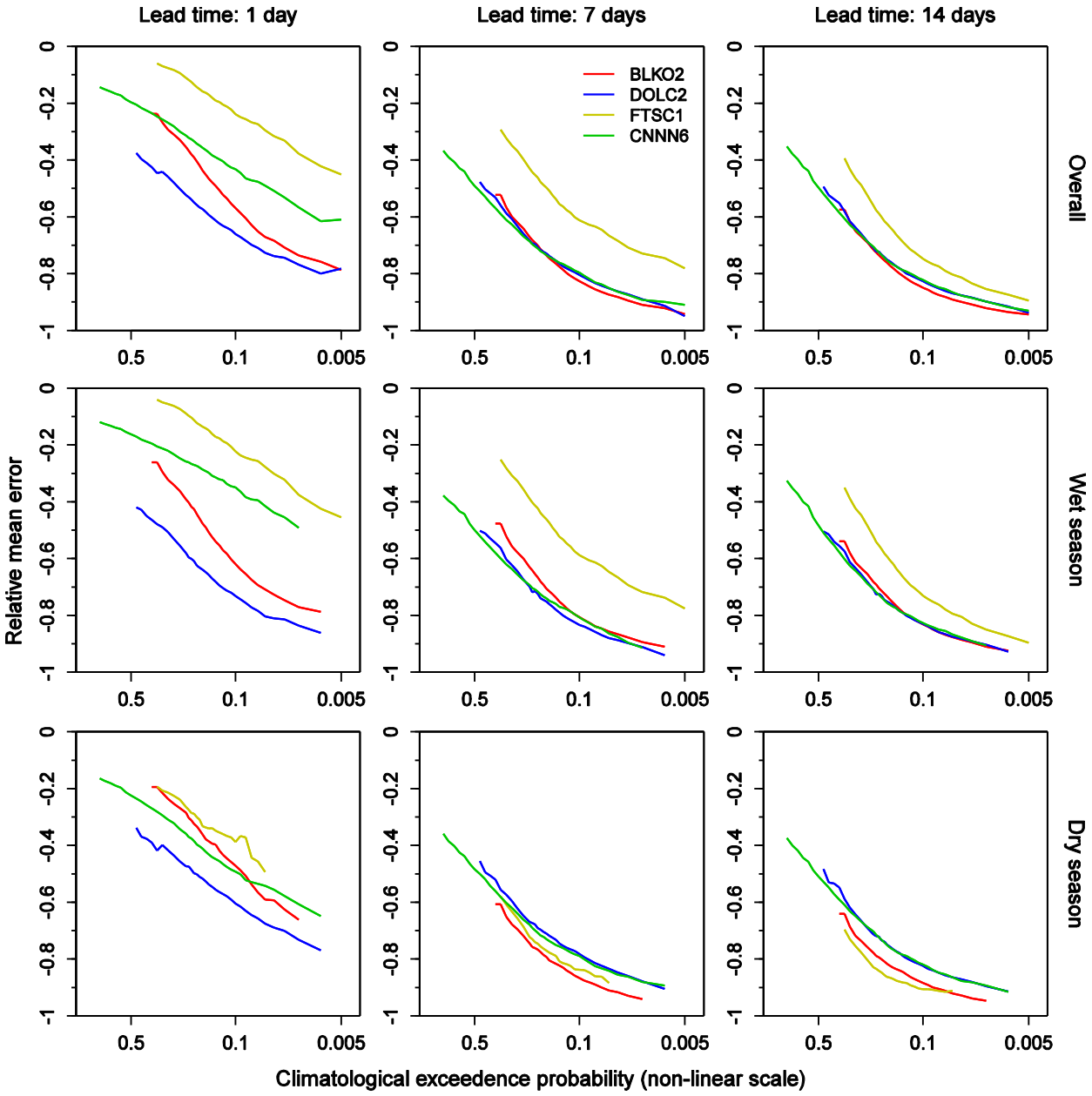

Figure 12: Relative mean error of the MEFP-GFS precipitation forecasts (ensemble mean) with input forcing from resampled climatology (CLIM) and the GFS at 1, 7 and 14 days. The results are shown for the overall period and for the wet and dry seasons separately. The score values are ordered by increasing amounts of observed precipitation, which are expressed as climatological exceedence probabilities. The thresholds are plotted on a probit scale, but labeled with actual probability.

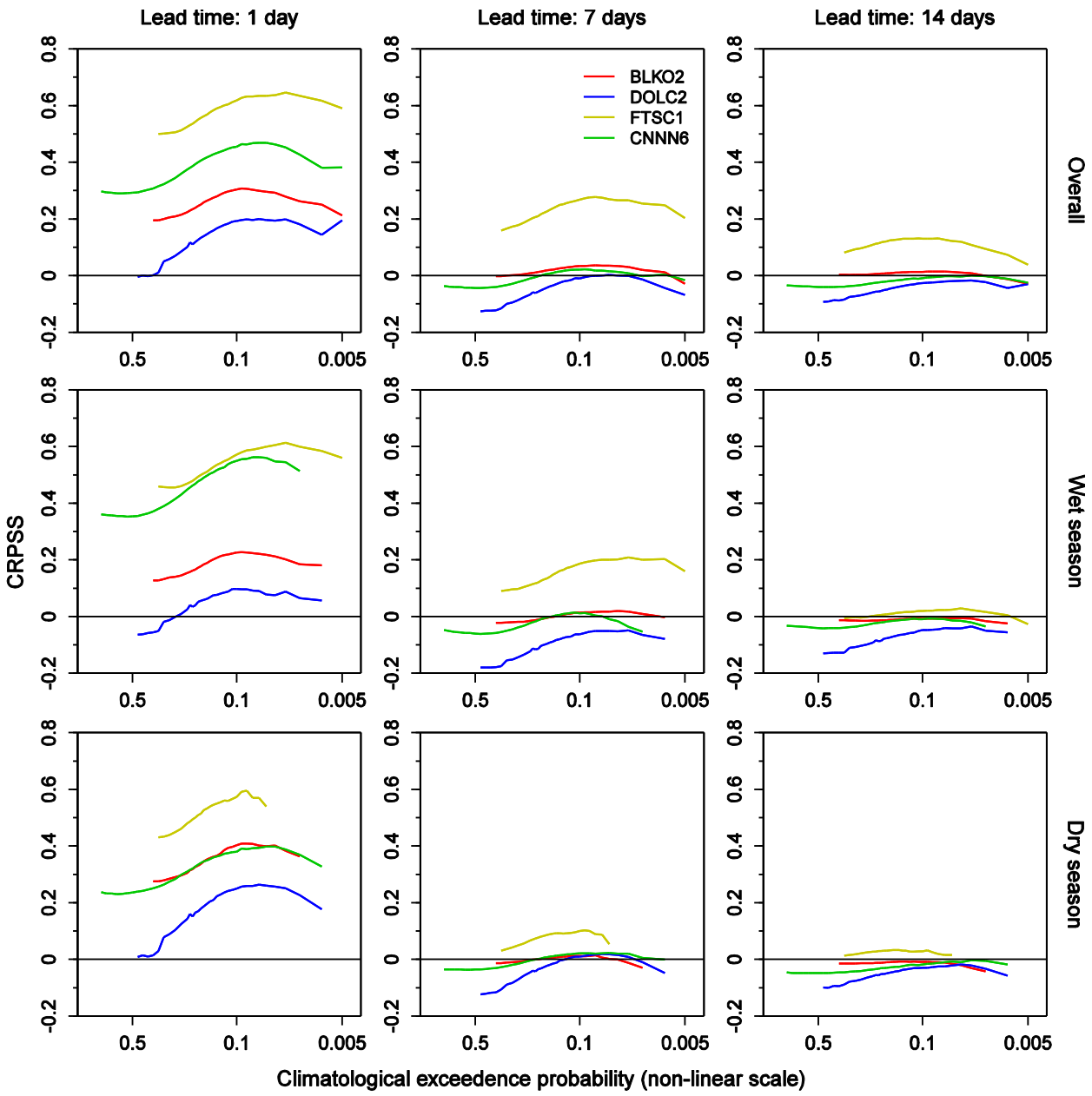

Figure 13: As in figure 12, but for the CRPSS, with sample climatology as the reference forecast.

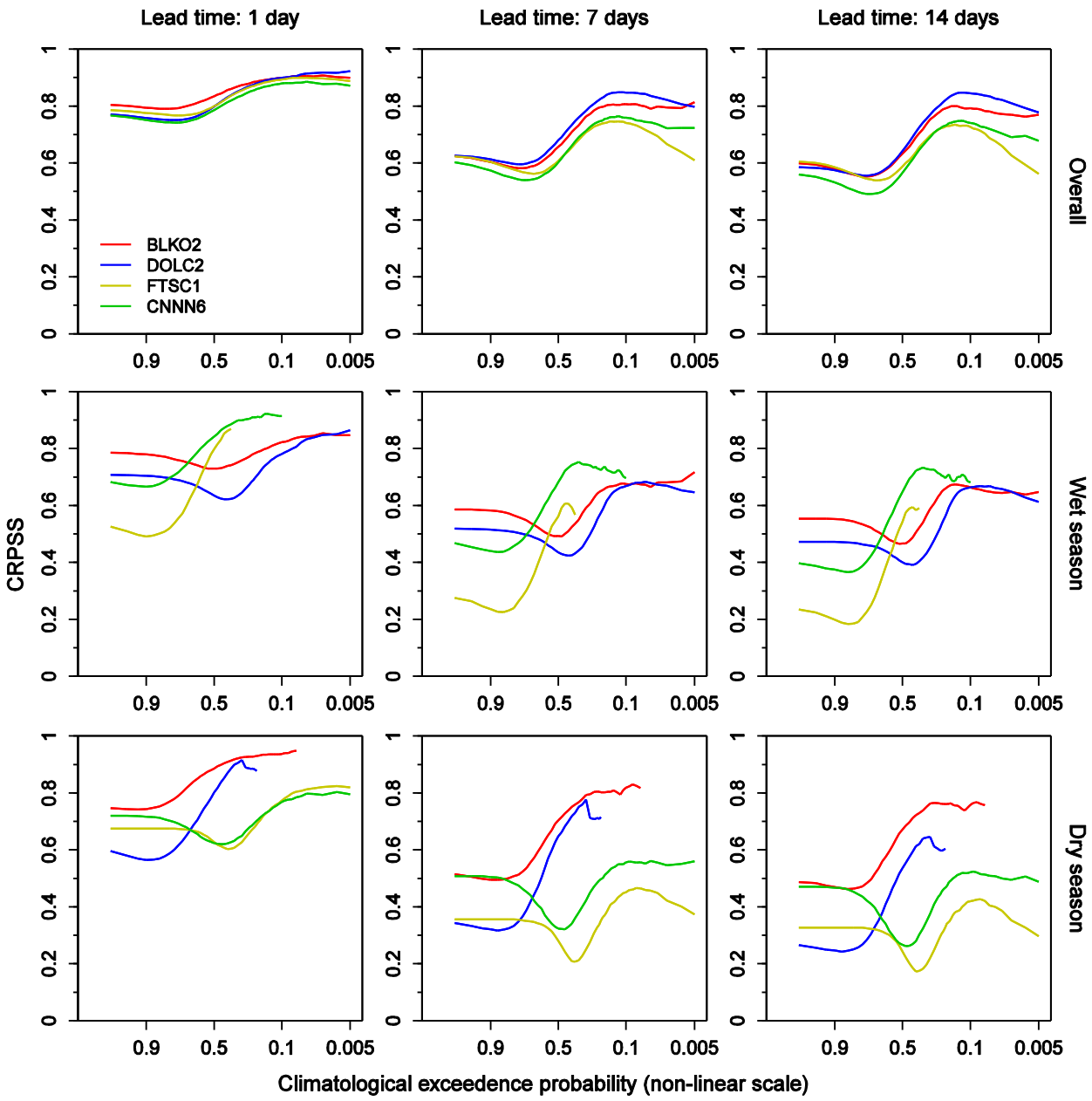

Figure 14: As in figure 13, but for temperature.

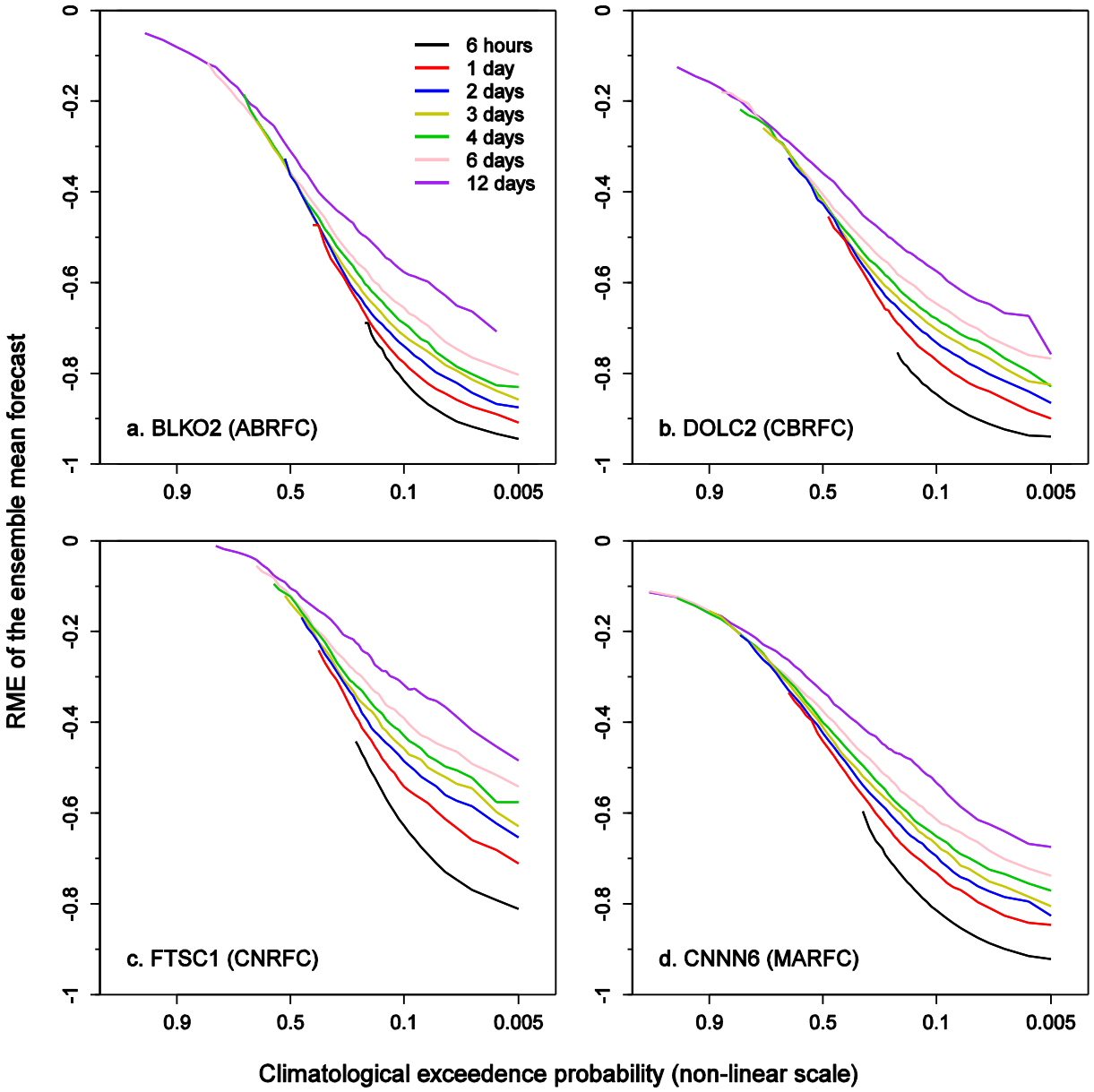

Figure 15: Relative mean error of the MEFP-GFS precipitation forecasts (ensemble mean) for increasing aggregation periods within a 1-12 day forecast horizon. The results are shown for the downstream basin in each RFC and are plotted for increasing amounts of observed precipitation.

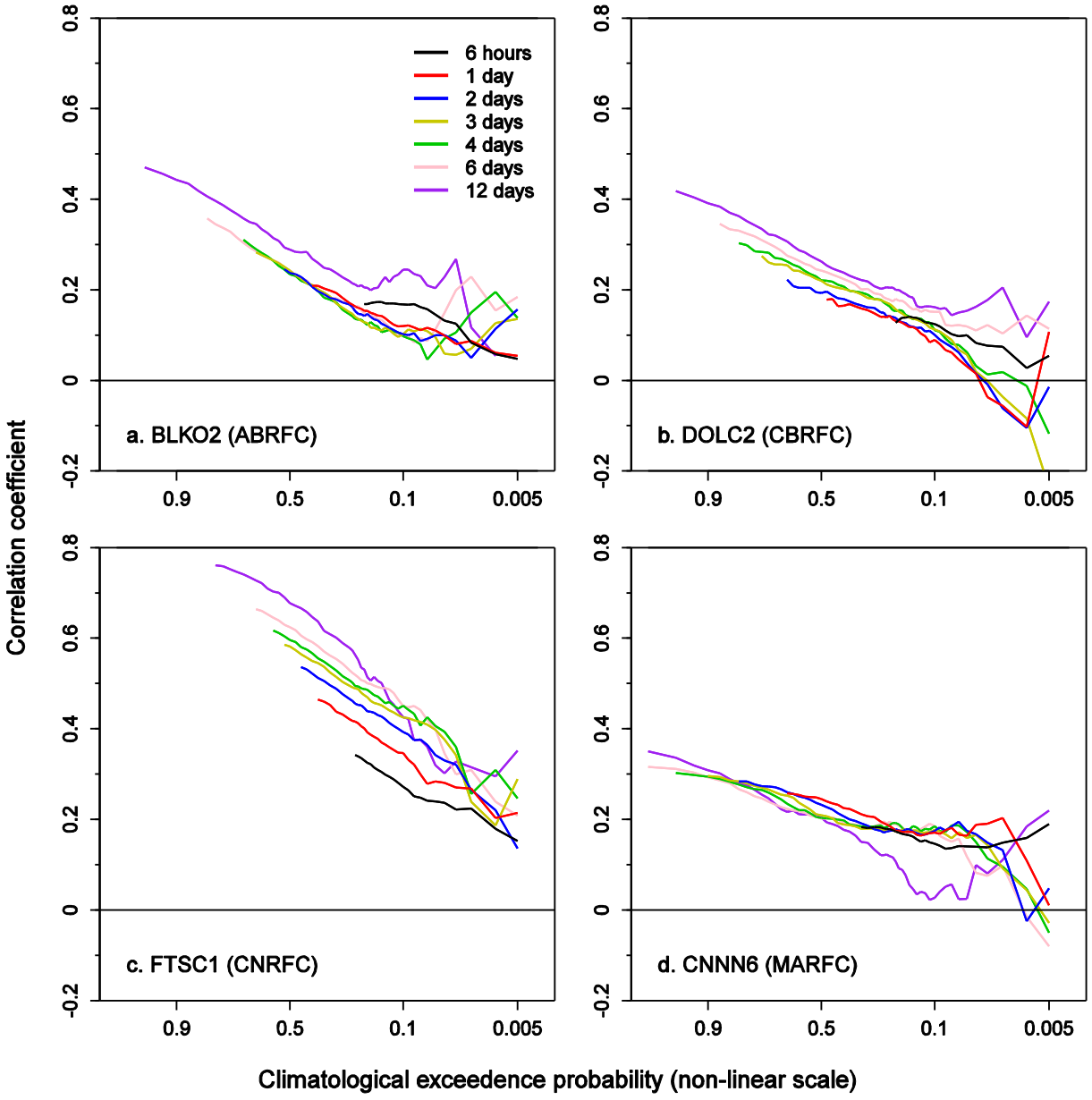

Figure 16: As in figure 15, but for the correlation coefficient.

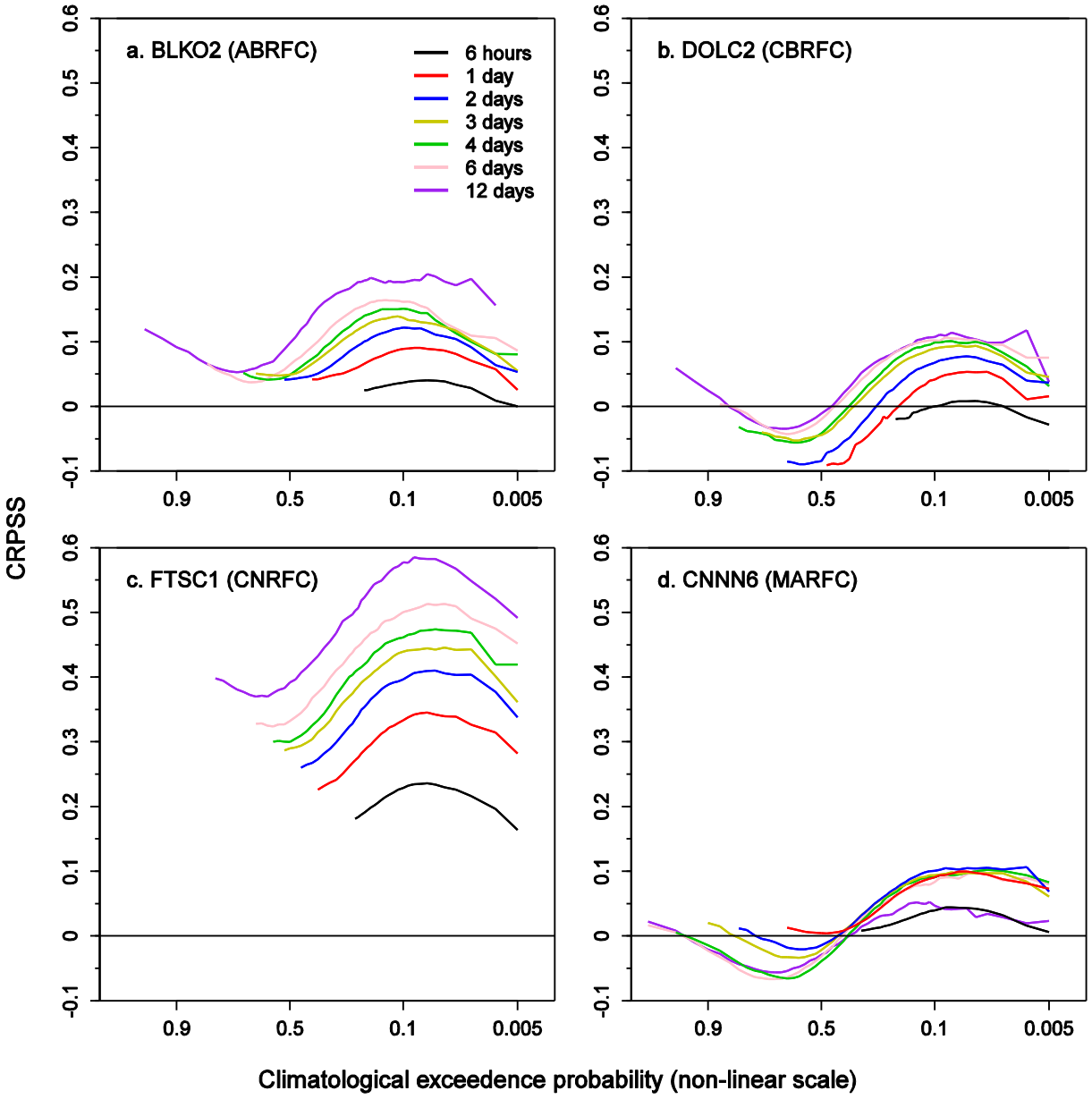

Figure 17: As in figure 16, but for the CRPSS, with sample climatology as the reference forecast.

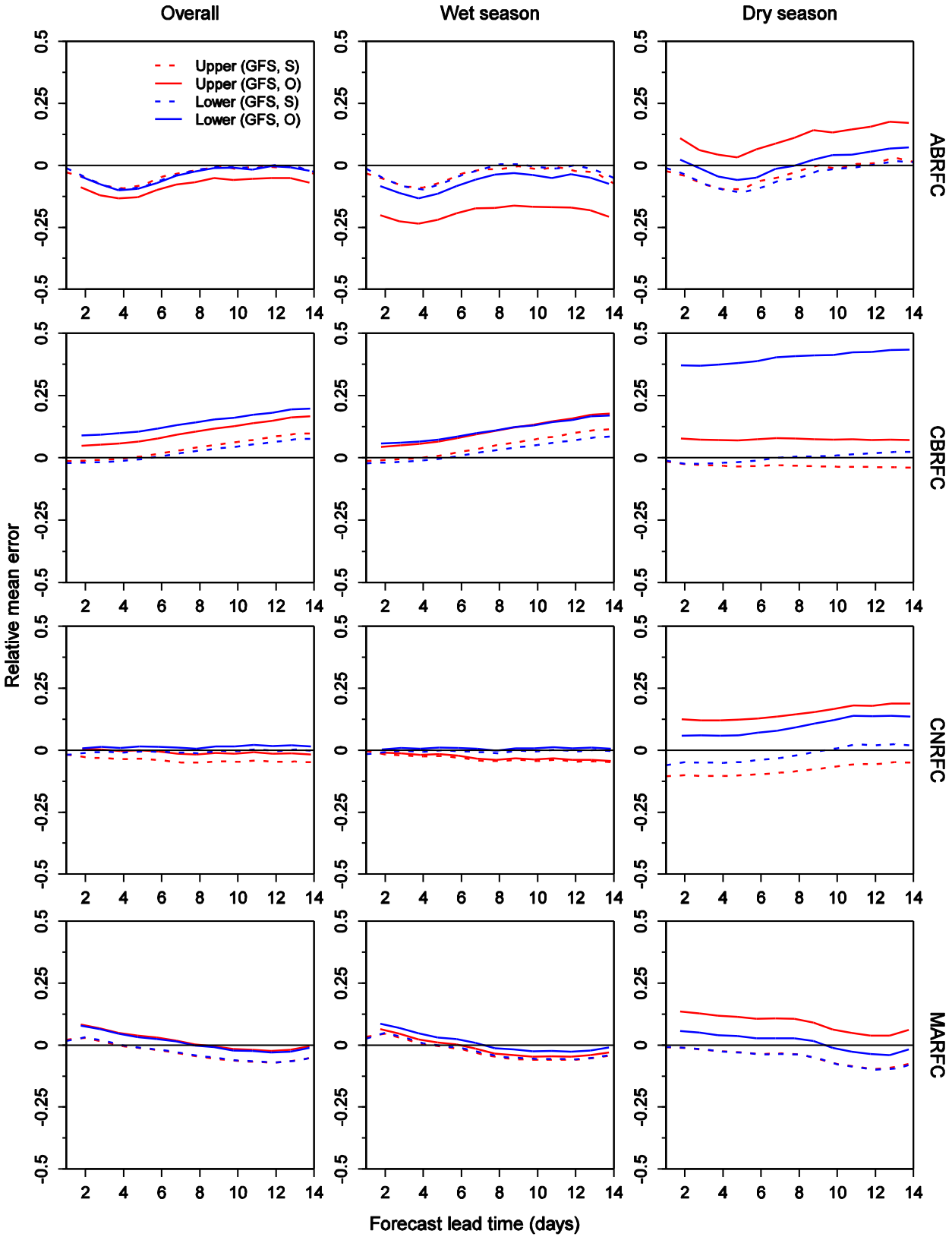

Figure 18: Relative mean error of the raw streamflow forecasts (ensemble mean) with MEFP-GFS forcing against observed (O) and simulated (S) streamflows. The results are shown for the upstream and downstream basins in each RFC and for the wet and dry seasons separately.

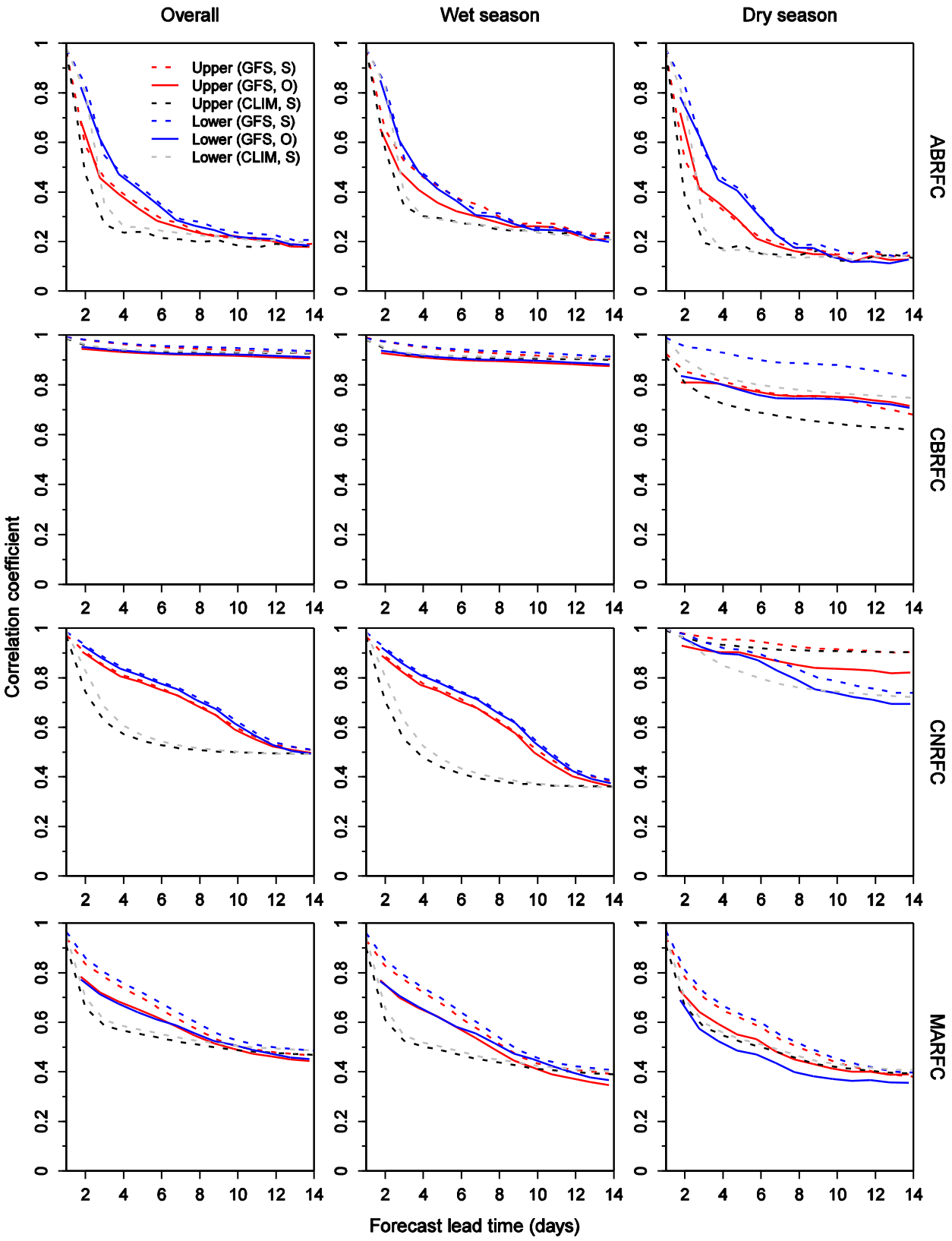

Figure 19: As in figure 18, but for the correlation coefficient.

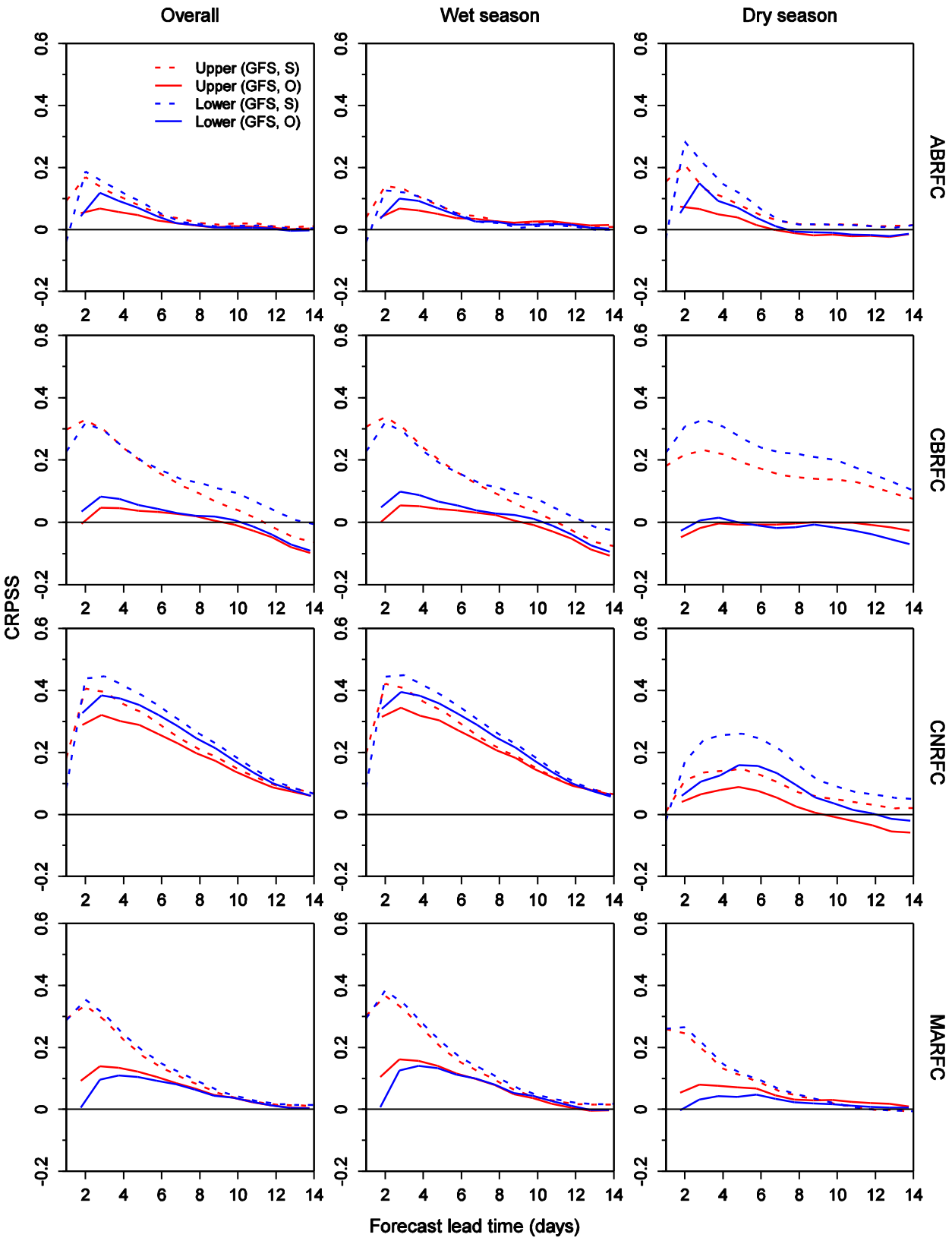

Figure 20: As in figure 18, but for the CRPSS. The reference streamflow forecasts comprise forcing from the MEFP with resampled climatology as input.

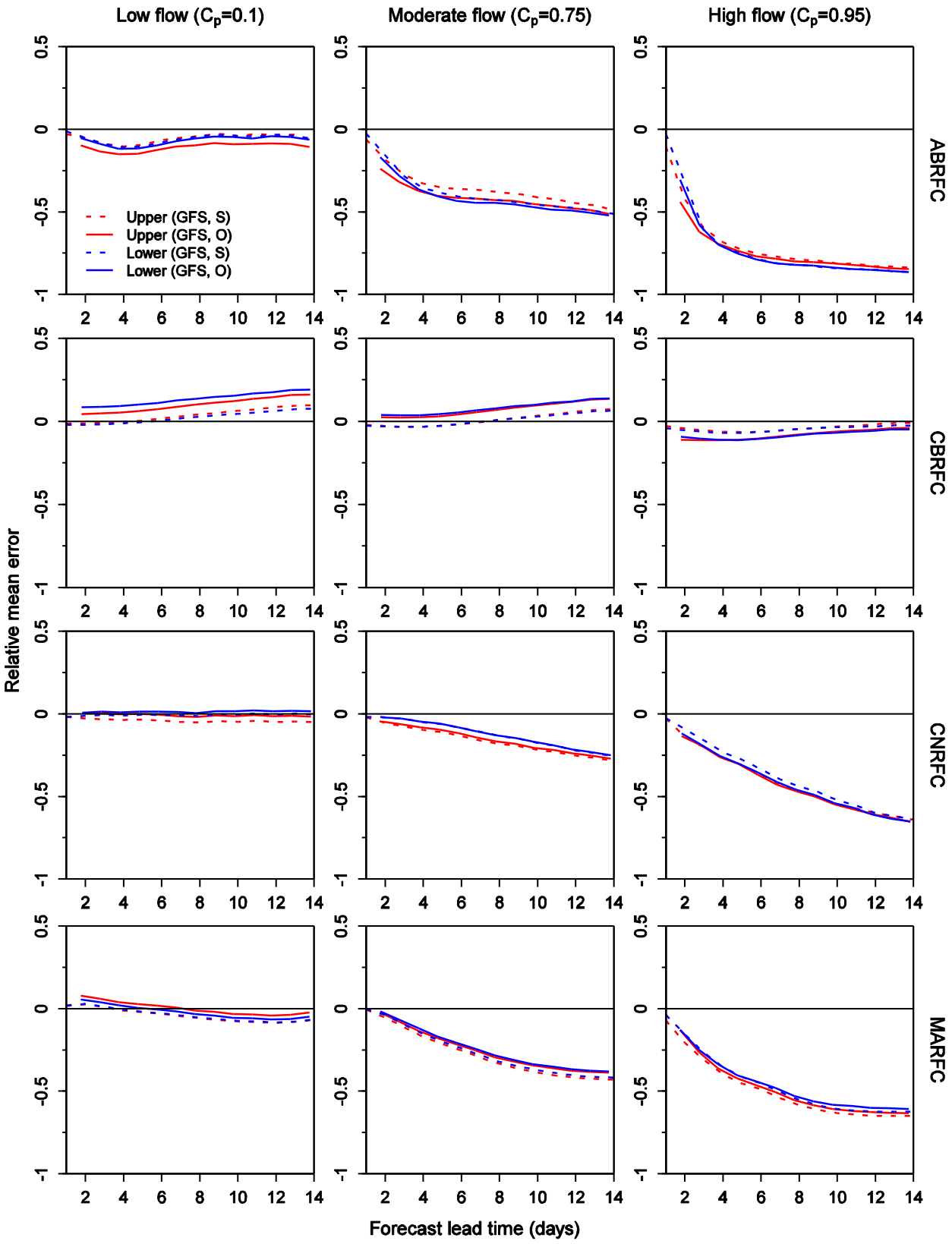

Figure 21: Relative mean error of the raw streamflow forecasts (ensemble mean) with MEFP-GFS forcing against observed (O) and simulated (S) streamflows. The results are shown for the upstream and downstream basins in each RFC and for daily streamflow amounts that are exceeded, on average, 90%, 25% and 5% of the time.

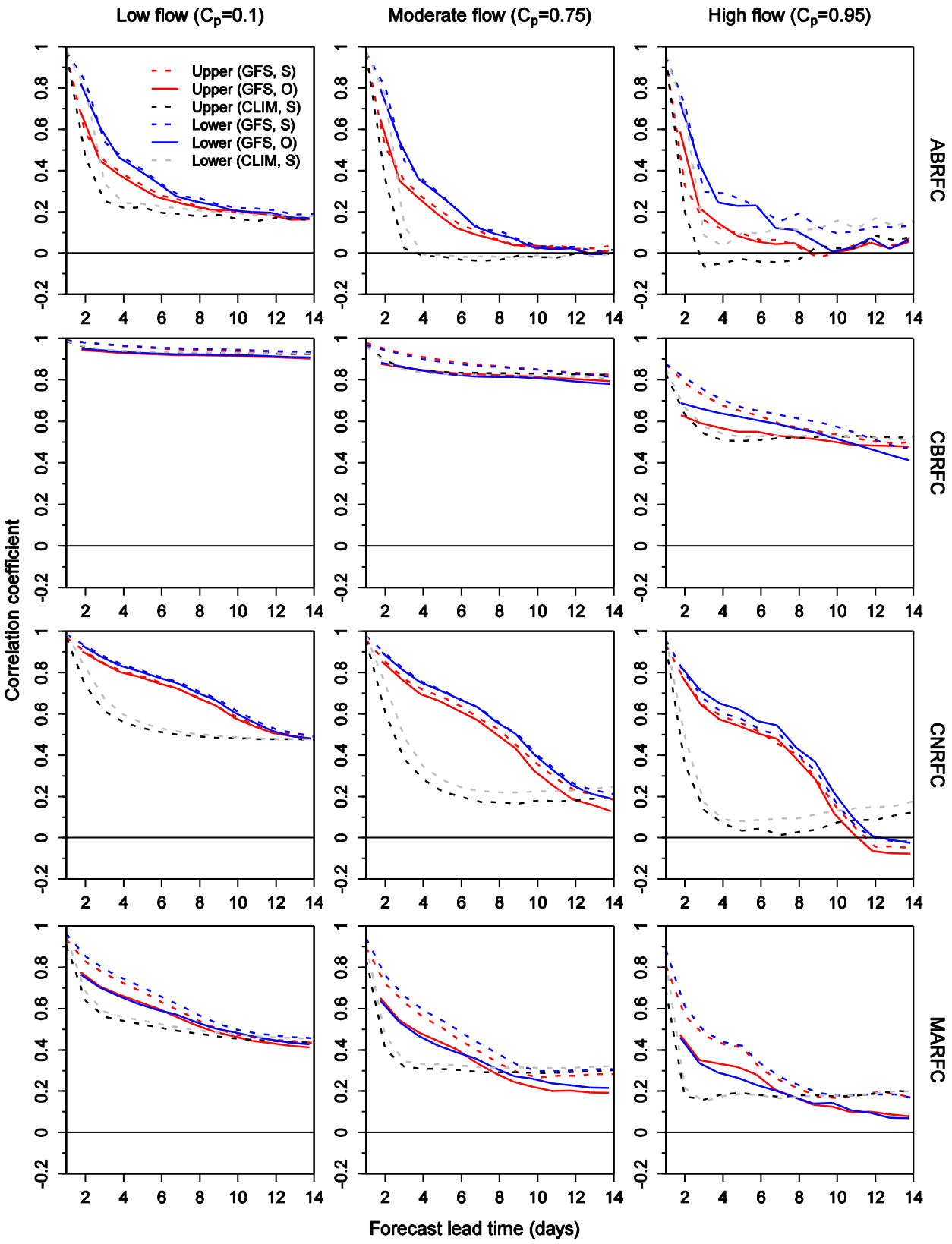

Figure 22: As in figure 20, but for the correlation coefficient. Also, the results are shown for the streamflow forecasts with climatological forcing (against simulated streamflows), as well as forcing from the MEFP-GFS.

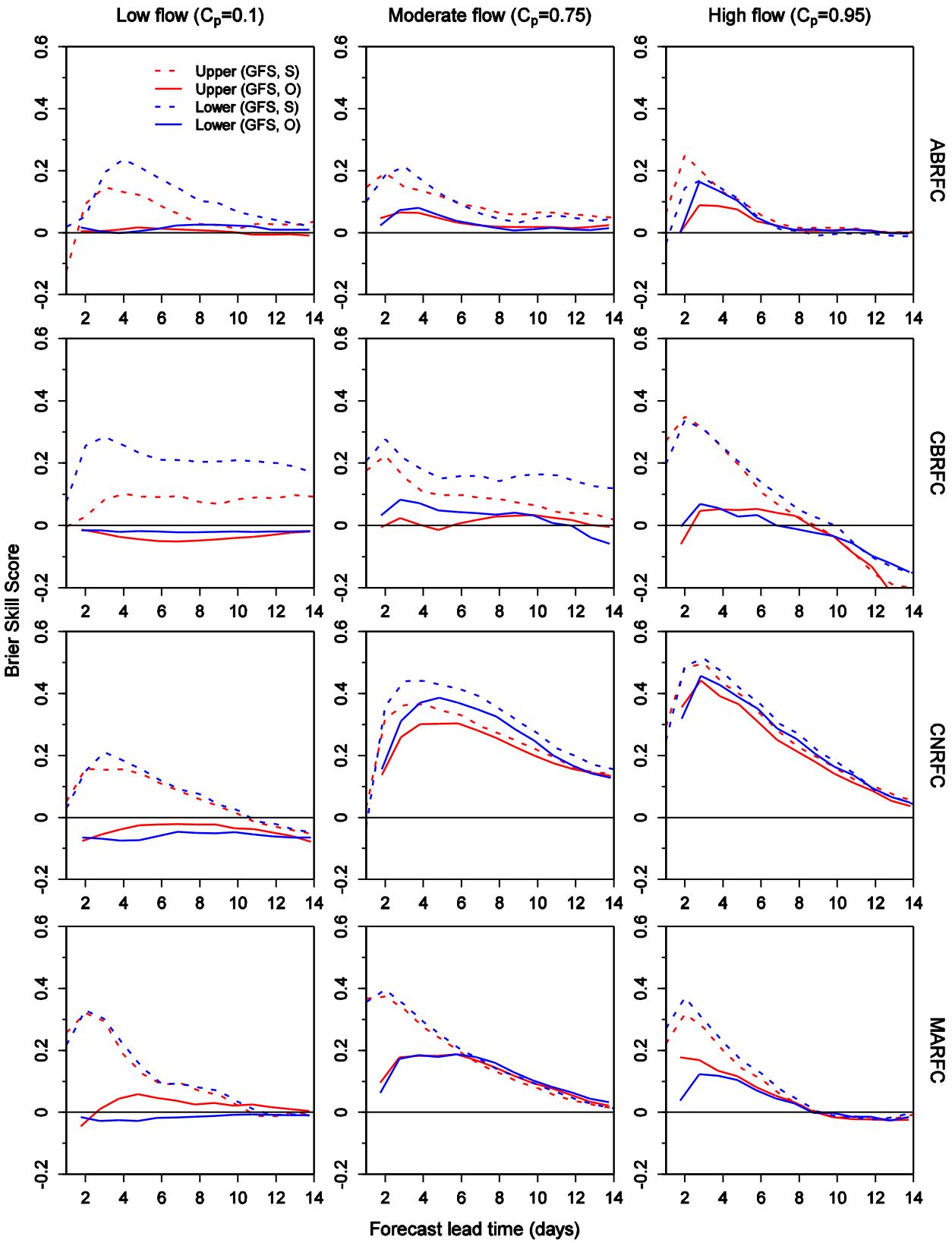

Figure 23: As in figure 21, but for the BSS. The reference streamflow forecasts comprise forcing from the MEFP with resampled climatology as input.

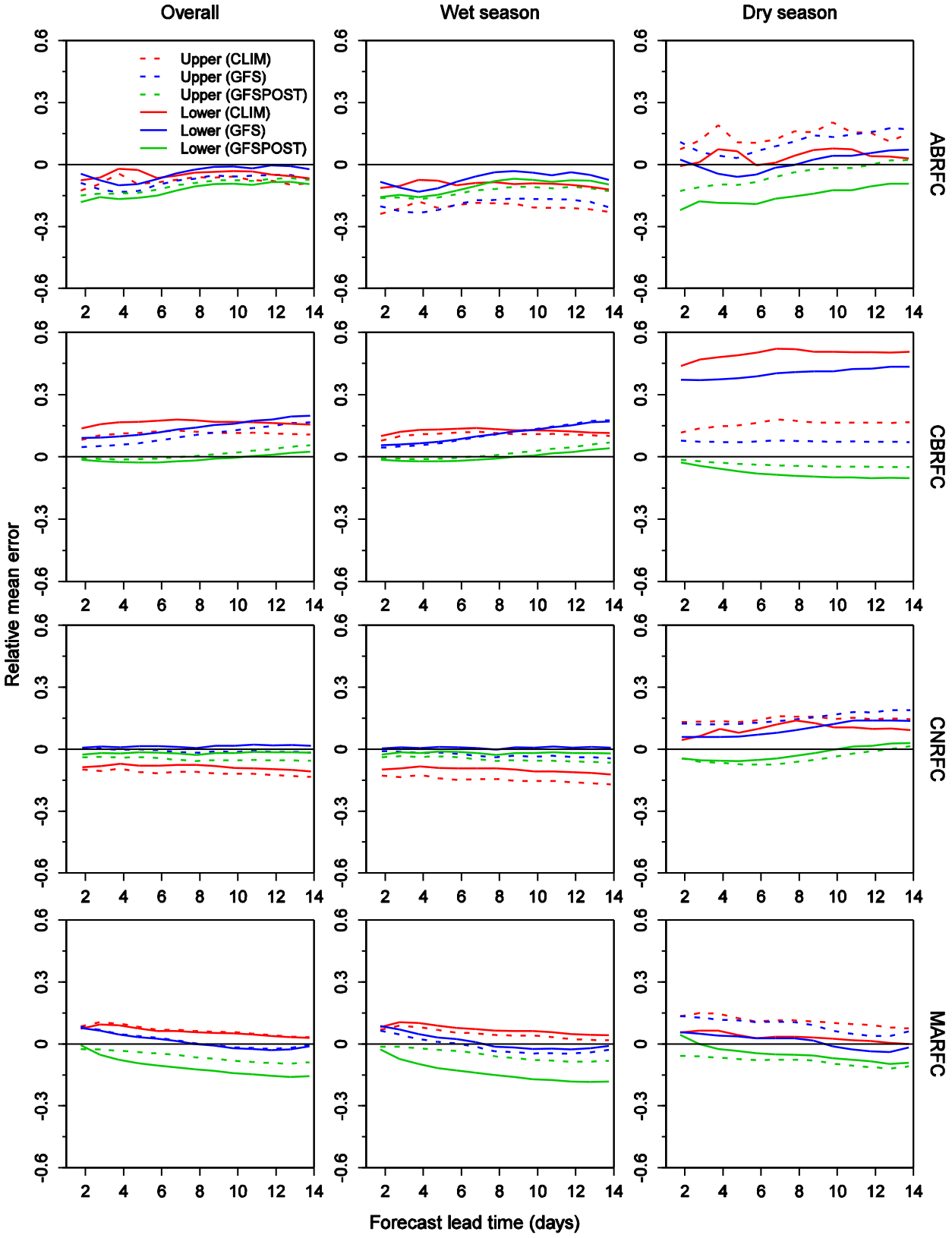

Figure 24: Relative mean error of the raw (GFS) and bias-corrected (GFSPOST) streamflow forecasts (ensemble mean) with forcing from the MEFP using GFS as input. The results are also shown for the raw streamflow forecasts with climatological forcing (CLIM). Results from upstream and downstream basins are shown together with the dry season, wet season and overall period.

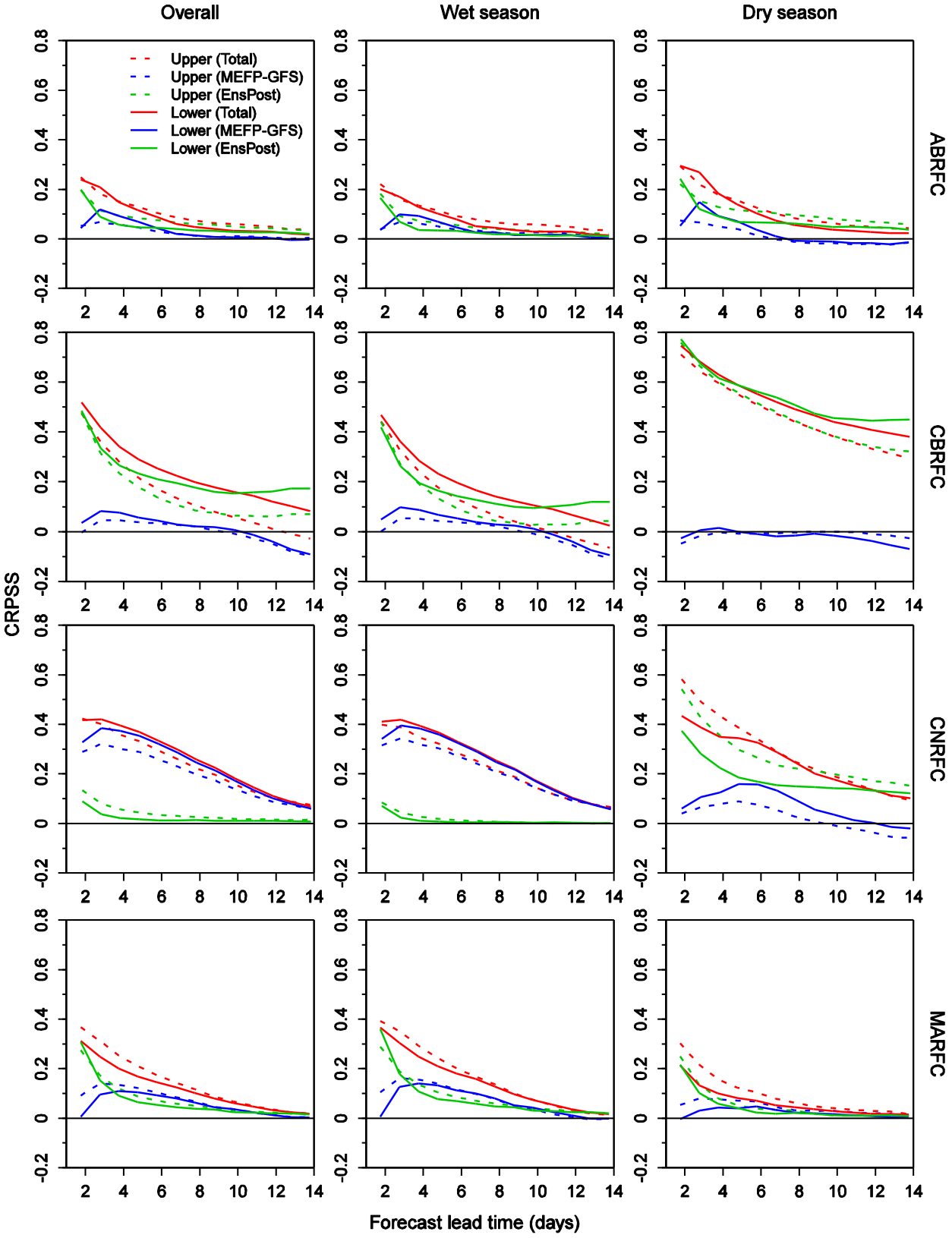

Figure 25: CRPSS of the bias-corrected streamflow forecasts against the raw streamflow forecasts with climatological forcing. The CRPSS is further decomposed into contributions from the MEFP with GFS forcing and the EnsPost. The results are shown for the upstream and downstream basins in each RFC and for the dry season, wet season and overall period.

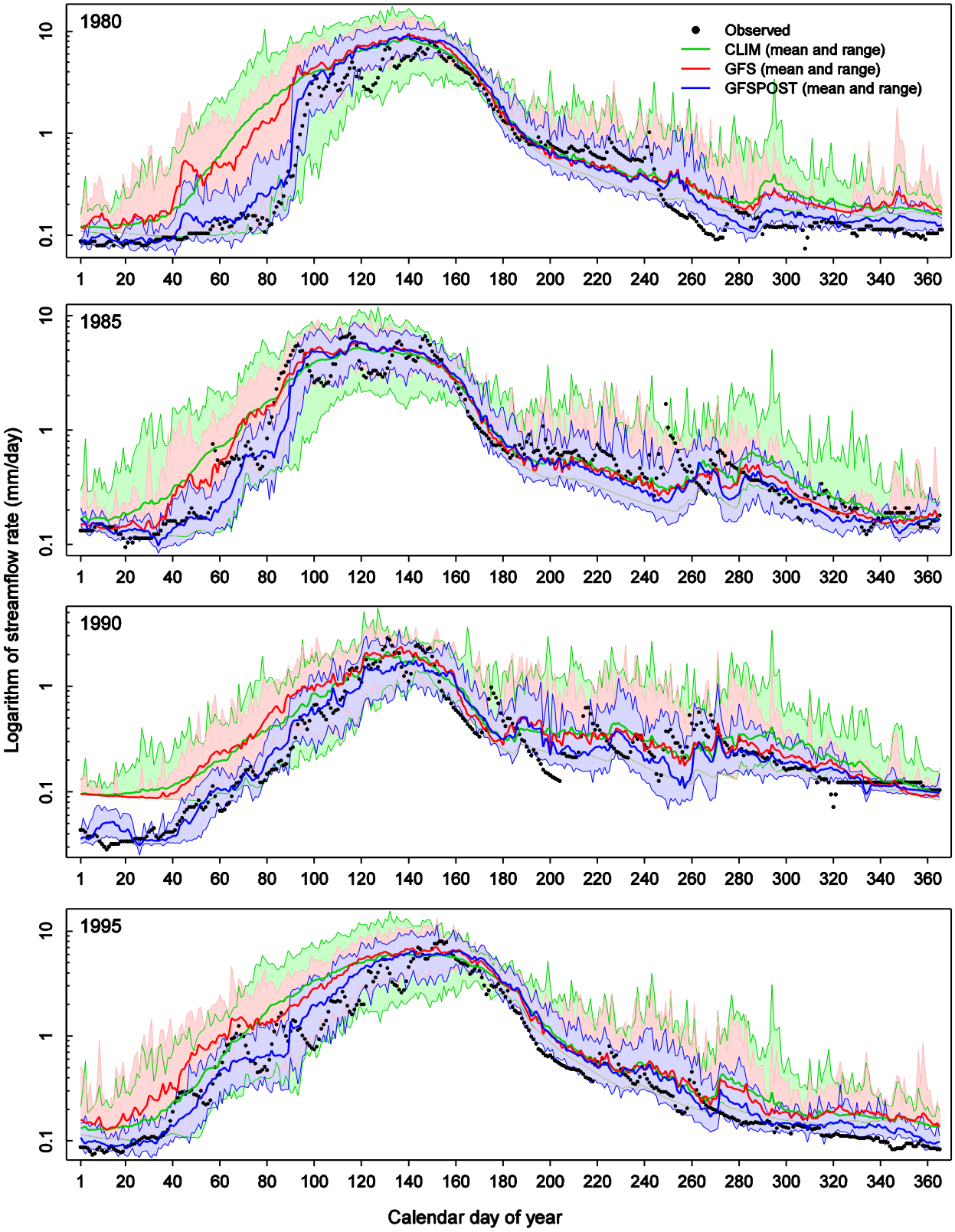

Figure 26: The ensemble mean and range of the raw and biased-corrected streamflow forecasts for selected calendar years in DOLC2. The results are shown for a forecast lead time of 307-331 hours and comprise forcing inputs from the MEFP with resampled climatology (CLIM) and GFS, together with the post-processed streamflow forecasts (GFSPOST).

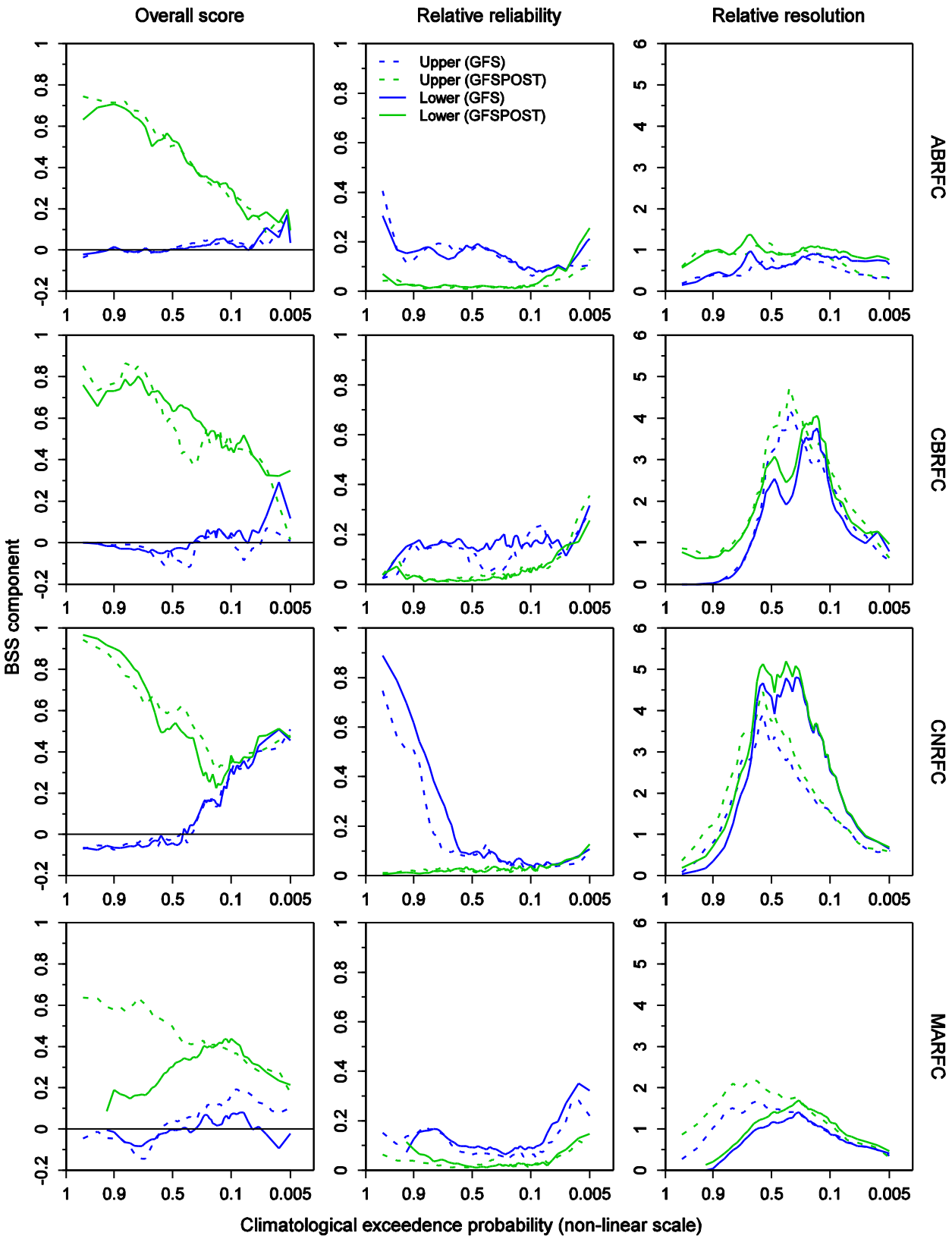

Figure 27: BSS of the raw and bias-corrected streamflow forecasts with MEFP-GFS forcing against the raw streamflow forecasts with climatological forcing. The results are shown for the upstream and downstream basins in each RFC. The BSS is further decomposed into the calibration-refinement factors of "relative reliability" and "relative resolution."

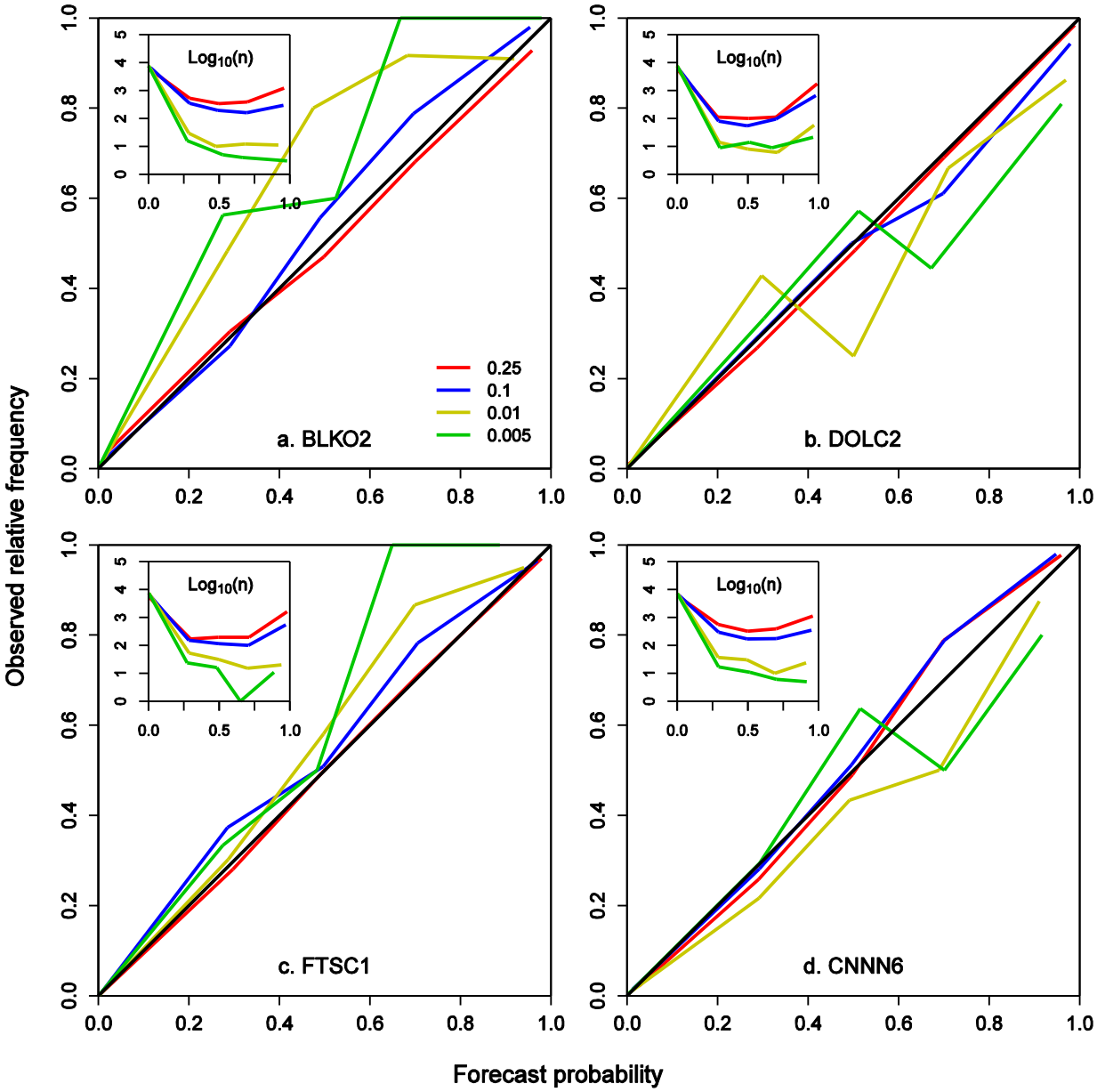

Figure 28: Reliability diagrams for the bias-corrected streamflow forecasts with forcing from the MEFP-GFS. The results are shown for each downstream basin and for selected streamflow thresholds, expressed as climatological exceedence probabilities.

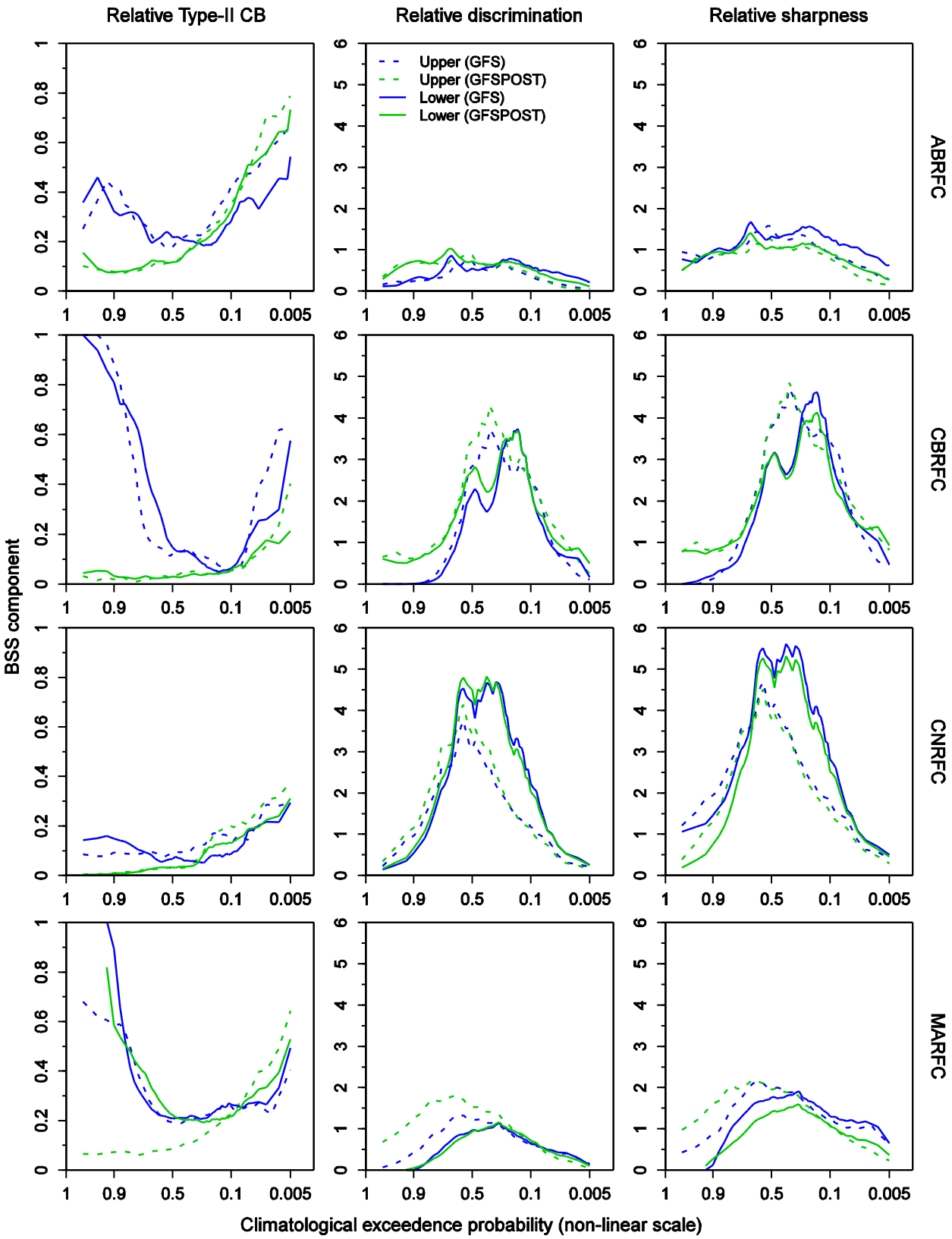

Figure 29: As in figure 27, but for the likelihood-base-rate factorization of the BSS, comprising the "relative Type-II conditional bias", "relative discrimination" and "relative sharpness."

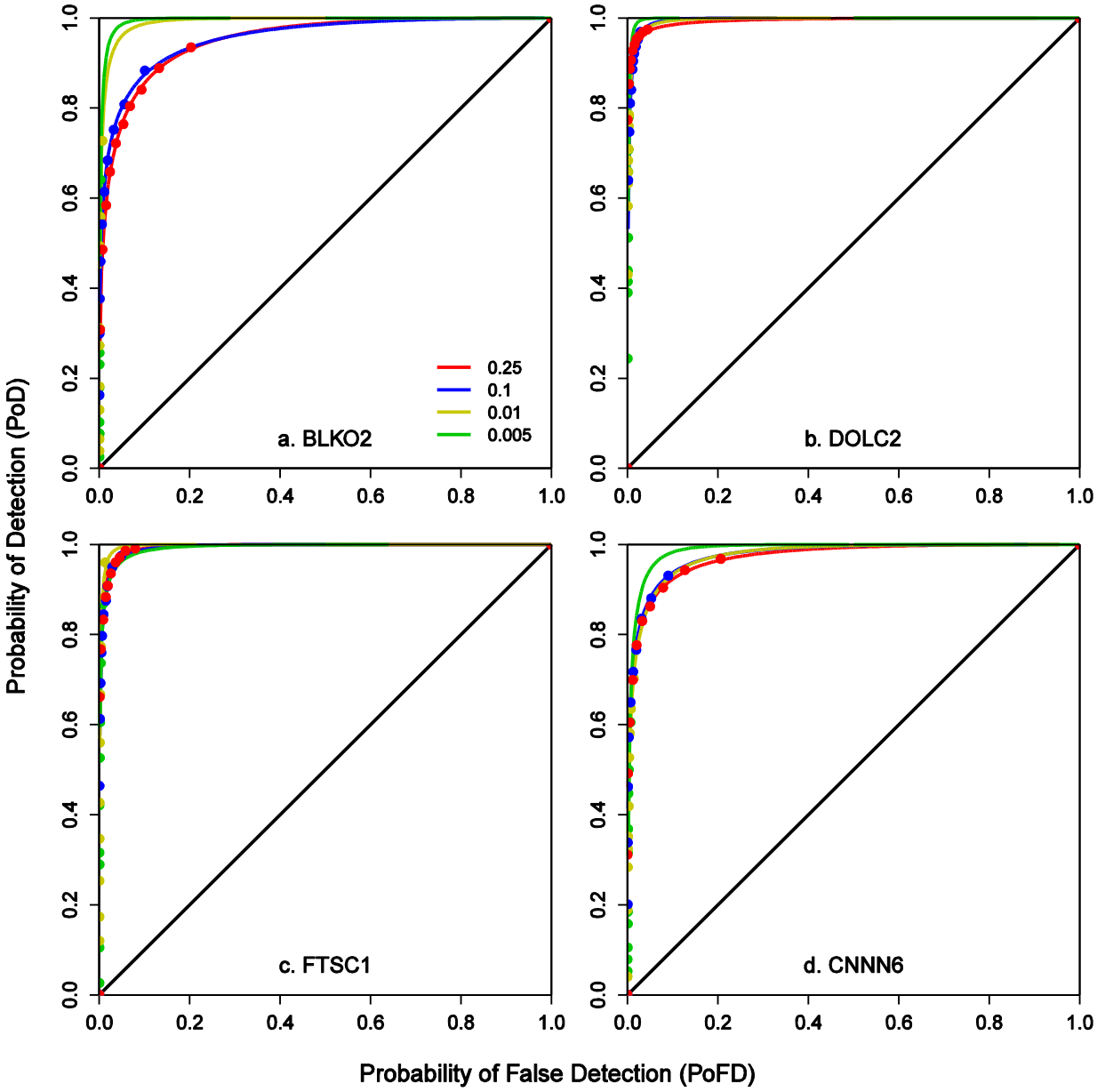

Figure 30: As in figure 28, but for the Relative Operating Characteristic (ROC). The lines comprise the fitted values of the ROC (under an assumption of bivariate normality), with the empirical values overlain as points.

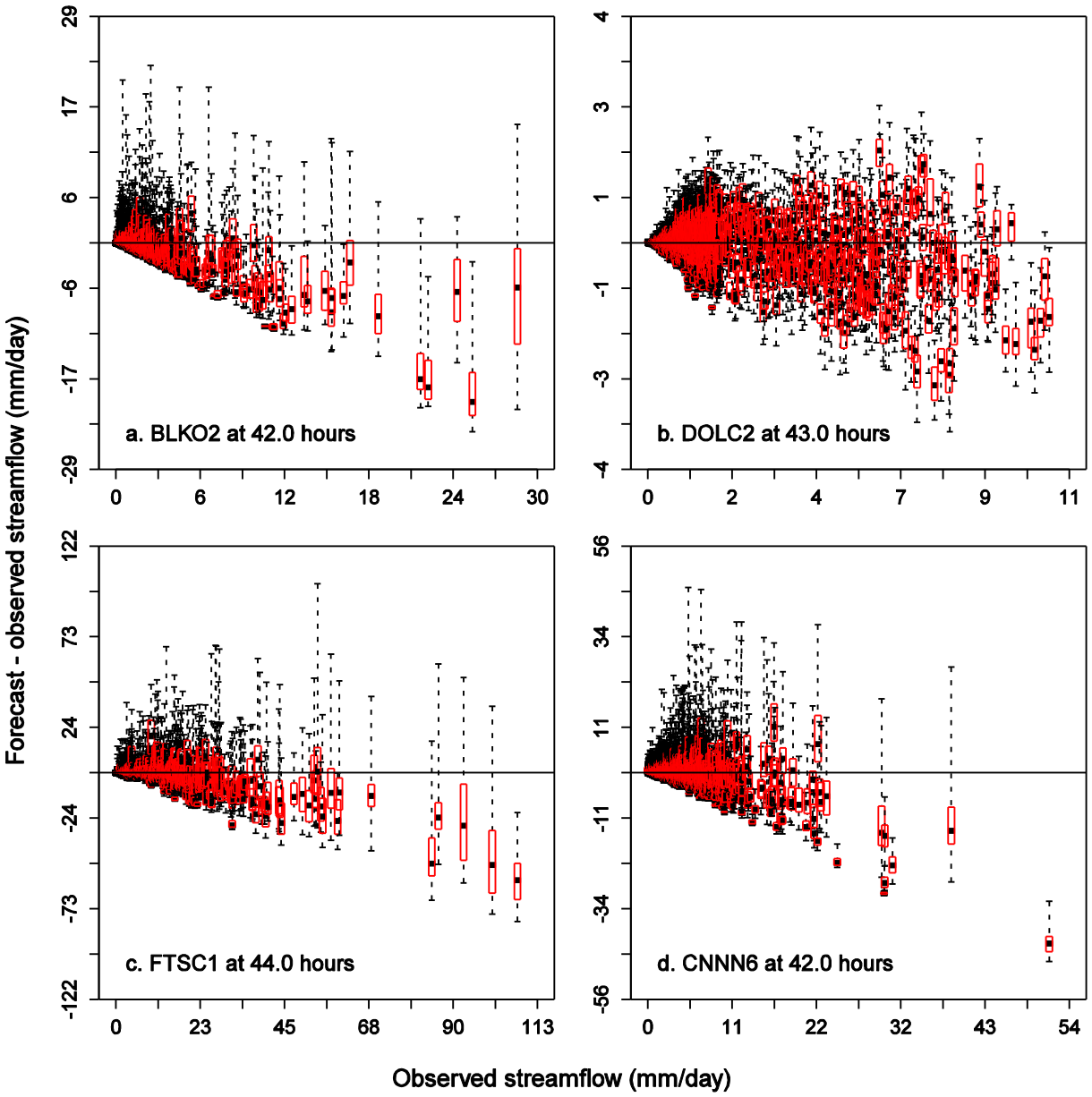

Figure 31: Box plots of errors in the bias-corrected streamflow forecasts with MEFP-GFS forcing. The results are shown for the downstream basin in each RFC and for a forecast lead time of ~18-42 hours. The boxes are ordered by increasing amounts of observed streamflow.

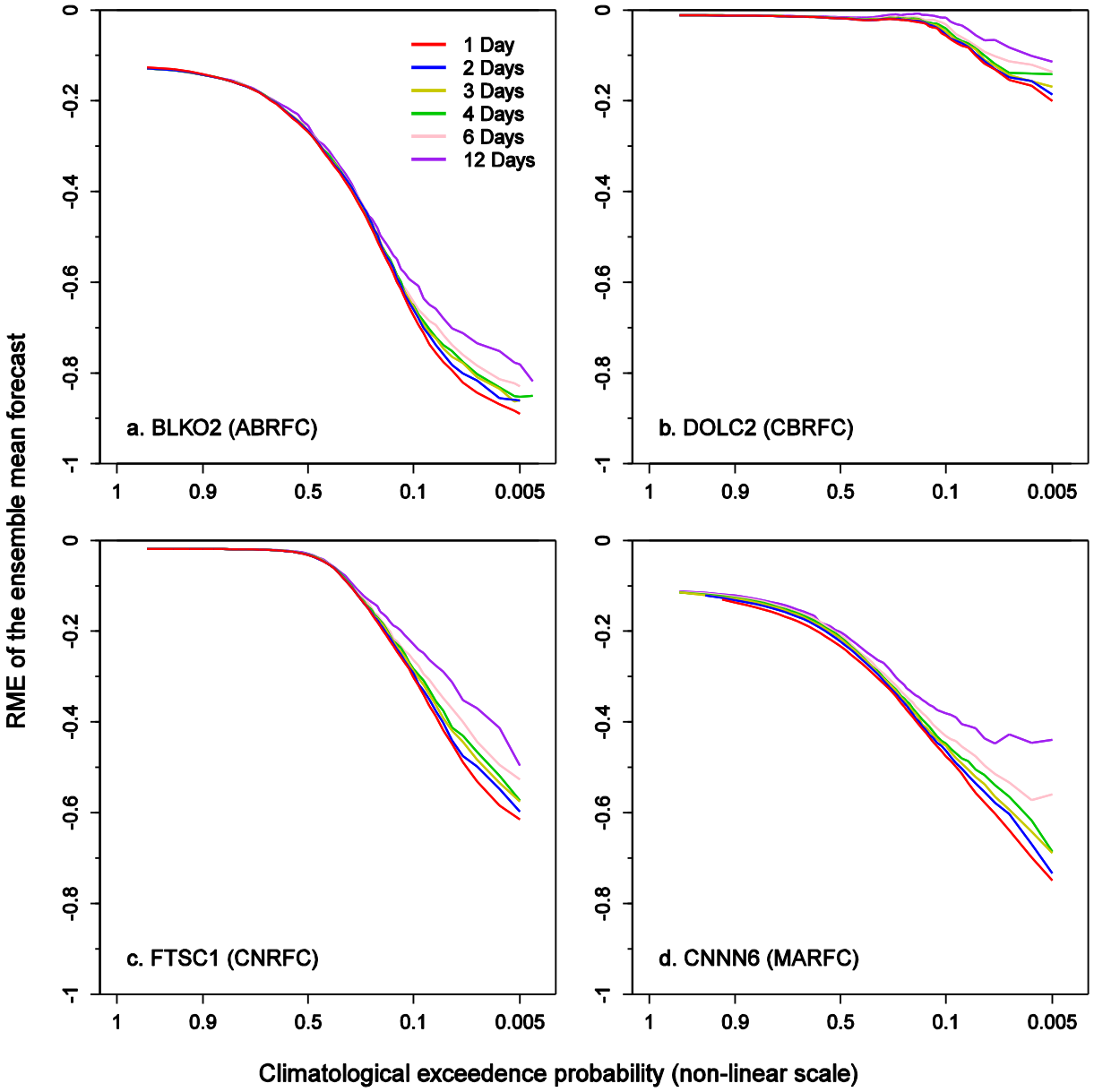

Figure 32: Relative mean error of the bias-corrected streamflow forecasts (ensemble mean) with MEFP-GFS forcing for increasing aggregation periods within a 1-12 day forecast horizon. Results are shown for the downstream basin in each RFC and are plotted for increasing amounts of observed streamflow.

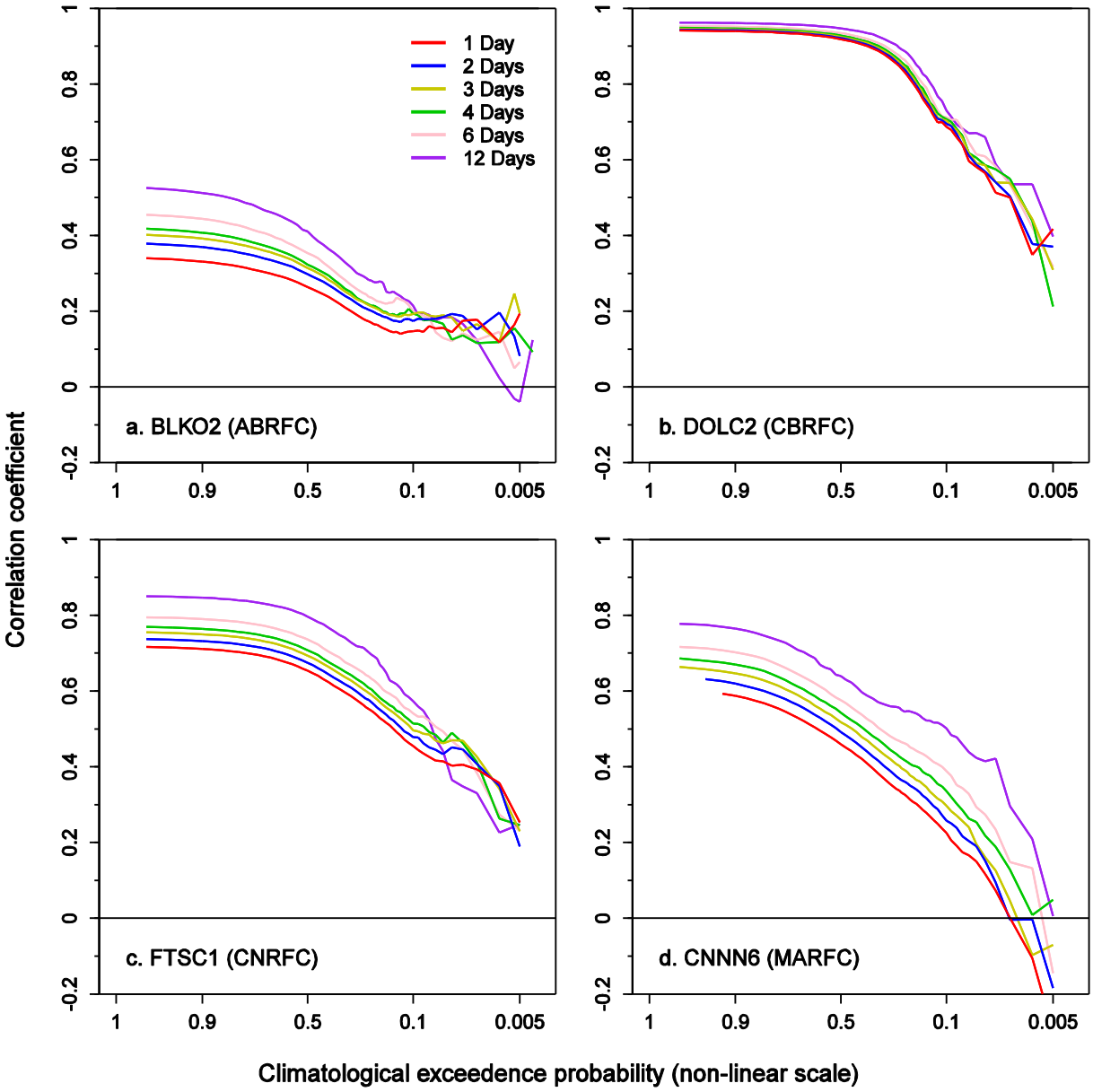

Figure 33: As in figure 32, but for the correlation coefficient.

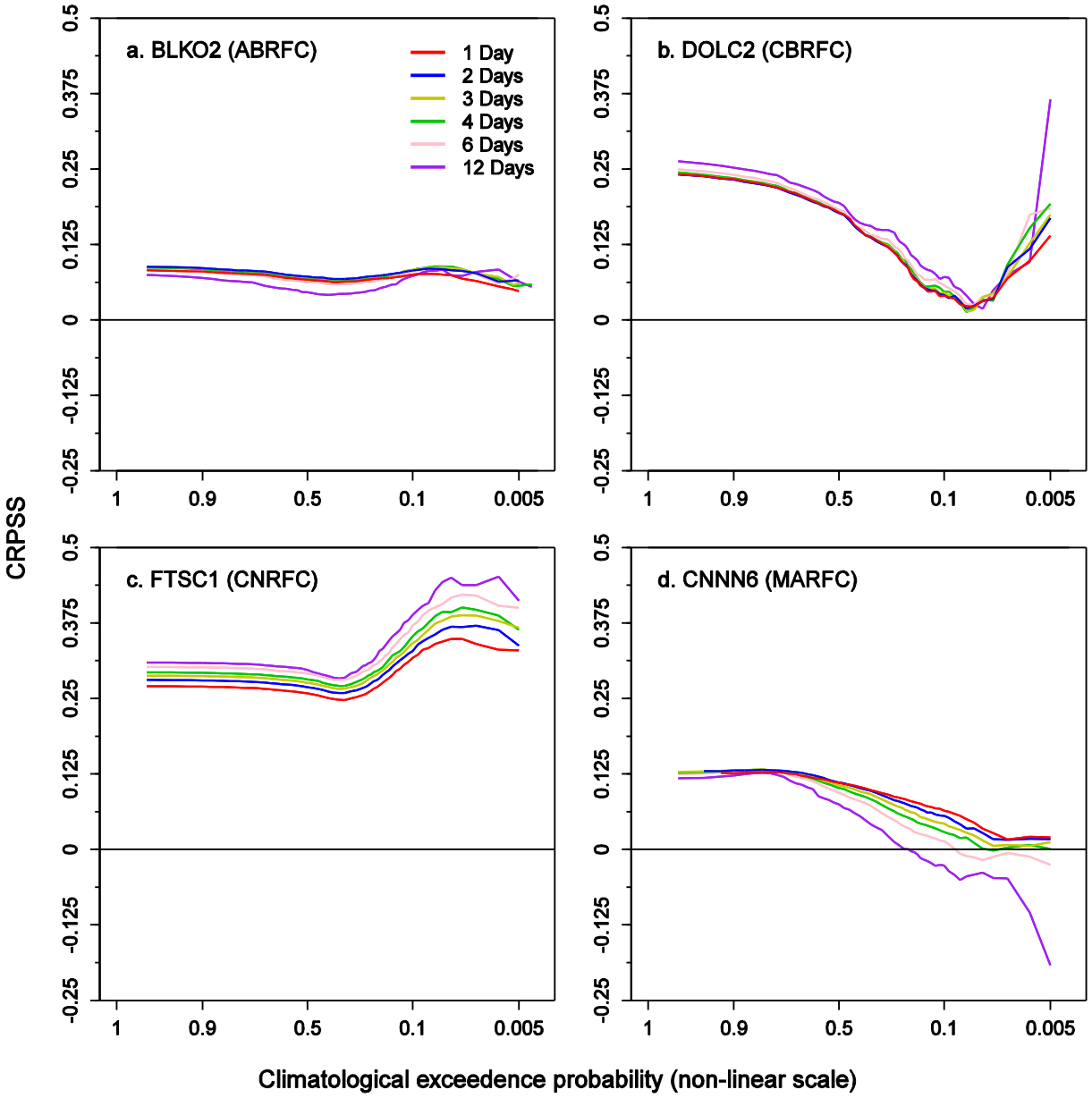

Figure 34: As in figure 32, but for the CRPSS. The reference streamflow forecasts comprise forcing from the MEFP with resampled climatology as input.

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. (2013), 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, $f_1(\mathbf{q}_f \mid \mathbf{q}_c)$, may be factored into a "raw" streamflow forecast, $f_3(\mathbf{q}_r \mid \mathbf{q}_c)$, and an "adjusted" streamflow forecast, given the raw forecast, $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$

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

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, $f_3(\mathbf{q}_r \mid \mathbf{q}_c)$, may be further separated into specific sources of uncertainty in the hydrologic modeling,

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

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, $f_4(\mathbf{q}_r \mid \mathbf{m}_f, \mathbf{i}, \mathbf{p}, \mathbf{q}_c)$, 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. 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. 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,

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

Thus, while the hydrologic uncertainties are not factored into specific contributions, their aggregate effects on $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$ 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 $f_2(\mathbf{q}_f \mid \mathbf{q}_c, \mathbf{q}_r)$ 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. (3). 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. (1) 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: Key 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

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

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)

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

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

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

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),

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

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 average probability that Y is observed to exceed q. By definition of B1, $\bar{F}_{Y_k}(q)$ is either zero or one, and the Type-II CB can only be zero if the forecasts are perfectly sharp. The BSS is then given by,

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

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

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

The mean CRPS comprises the CRPS averaged across n pairs of forecasts and observations. While less accessible than eqn. B2, and 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. Reliability diagram

The reliability diagram plots the average probability with which an event is observed to occur, conditionally upon the forecast probability, against its forecast probability of occurrence (Hsu and Murphy, 1986; Bröcker and Smith, 2007). For example, over a large number of cases where flooding is forecast to occur with a probability of 0.95, it should be observed to occur ~95% of the time. In practice, the forecasts are binned into discrete probability intervals and the observed relative frequencies are plotted against the average forecast probability in each bin. For a forecast event defined by the exceedence of some threshold, *q*, the average probability of the forecasts that fall in the *k*th forecast bin, B_k, is given by

$$\bar{F}_{X_k}(q) = \frac{1}{|I_k|} \sum_{I_k} F_{X_i}(q), \ \textit{where } I_k = \{i : i \in B_k\}. \quad \text{(B9)}$$

The corresponding fraction of observations is

$$\bar{F}_{Y_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{Y_i}(q), where\ F_{Y_i}(q) = \begin{cases} 1, \mathrm{Y_i} \le q; \\ 0, otherwise \end{cases}. \quad \text{(B10)}$$

The reliability diagram comprises a plot of $\bar{F}_{X_k}(q)$ against $\bar{F}_{Y_k}(q)$ for each B_k, together with the number of forecasts, $|I_k|$, in each bin or the "sharpness."

e. 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

$$PoD = \sum_{i=0}^{n} I_{X_i}\left(F_{X_i}(q) > d \,|\, Y_i > q\right) \Big/ \sum_{i=0}^{n} I_{Y_i}(Y_i > q). \quad \text{(B11)}$$

where I denotes the indicator function. The empirical PoFD is

$$PoFD = \sum_{i=0}^{n} I_{X_i}\left(F_{X_i}(q) > d \,|\, Y_i \le q\right) \Big/ \sum_{i=0}^{n} I_{Y_i}(Y_i \le q). \quad \text{(B12)}$$

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

$$PoD = \Phi\left\{a + b\Phi^{-1}\left(PoFD\right)\right\}\ where\ a = \frac{\mu_{PoD} - \mu_{PoFD}}{\sigma_{PoD}}\ and\ b = \frac{\sigma_{PoFD}}{\sigma_{PoD}}, \quad \text{(B13)}$$

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 raw and bias-corrected streamflow forecasts with forcing inputs from the GFS component of the MEFP. The results are also shown for the raw streamflow forecasts with climatological forcing. The plots include the single-valued streamflow observations, together with the ensemble range (maximum – minimum value) of the corresponding streamflow forecast on each 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 1980, 1985, 1990, and 1995. 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.

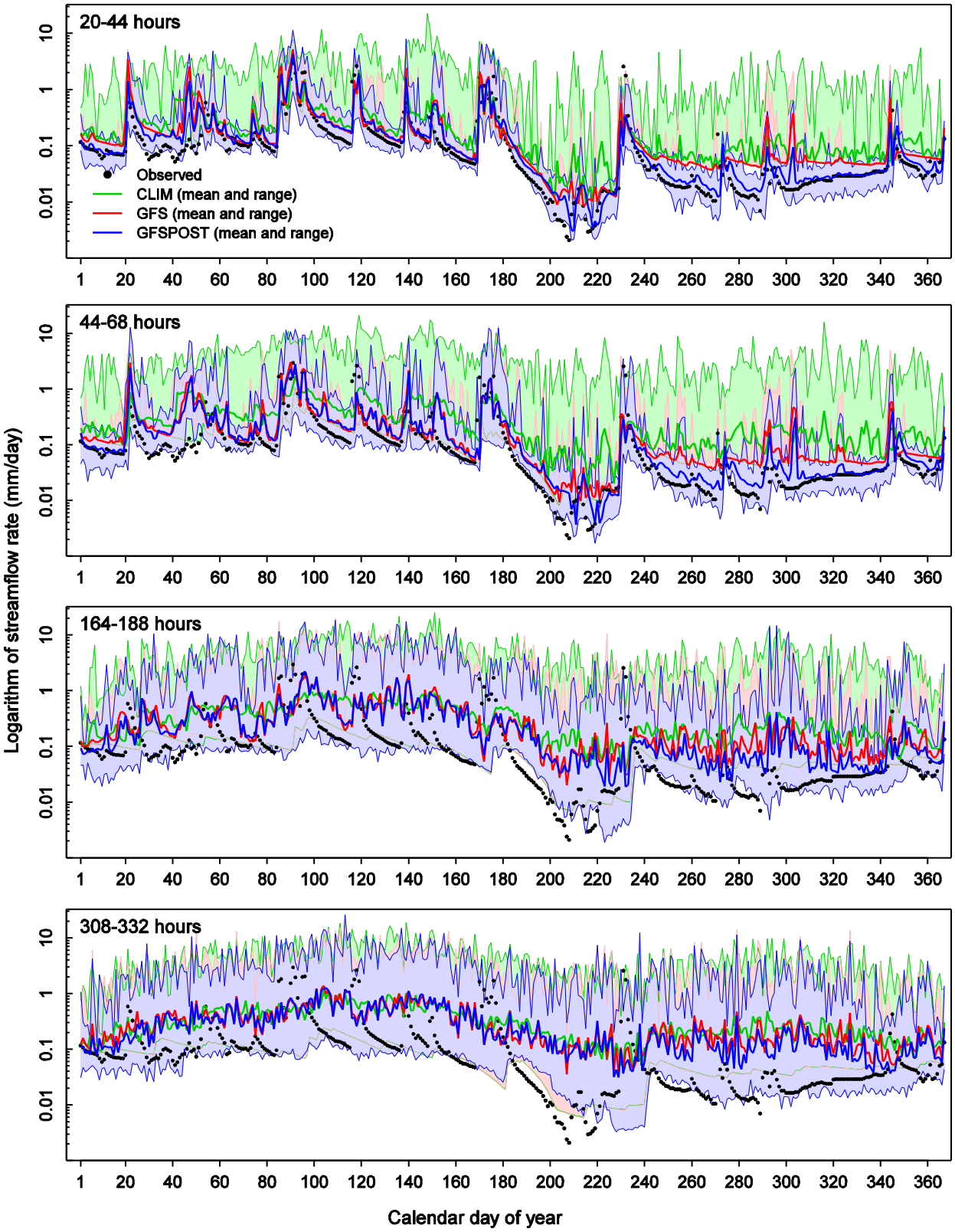

Figure C01: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1980 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

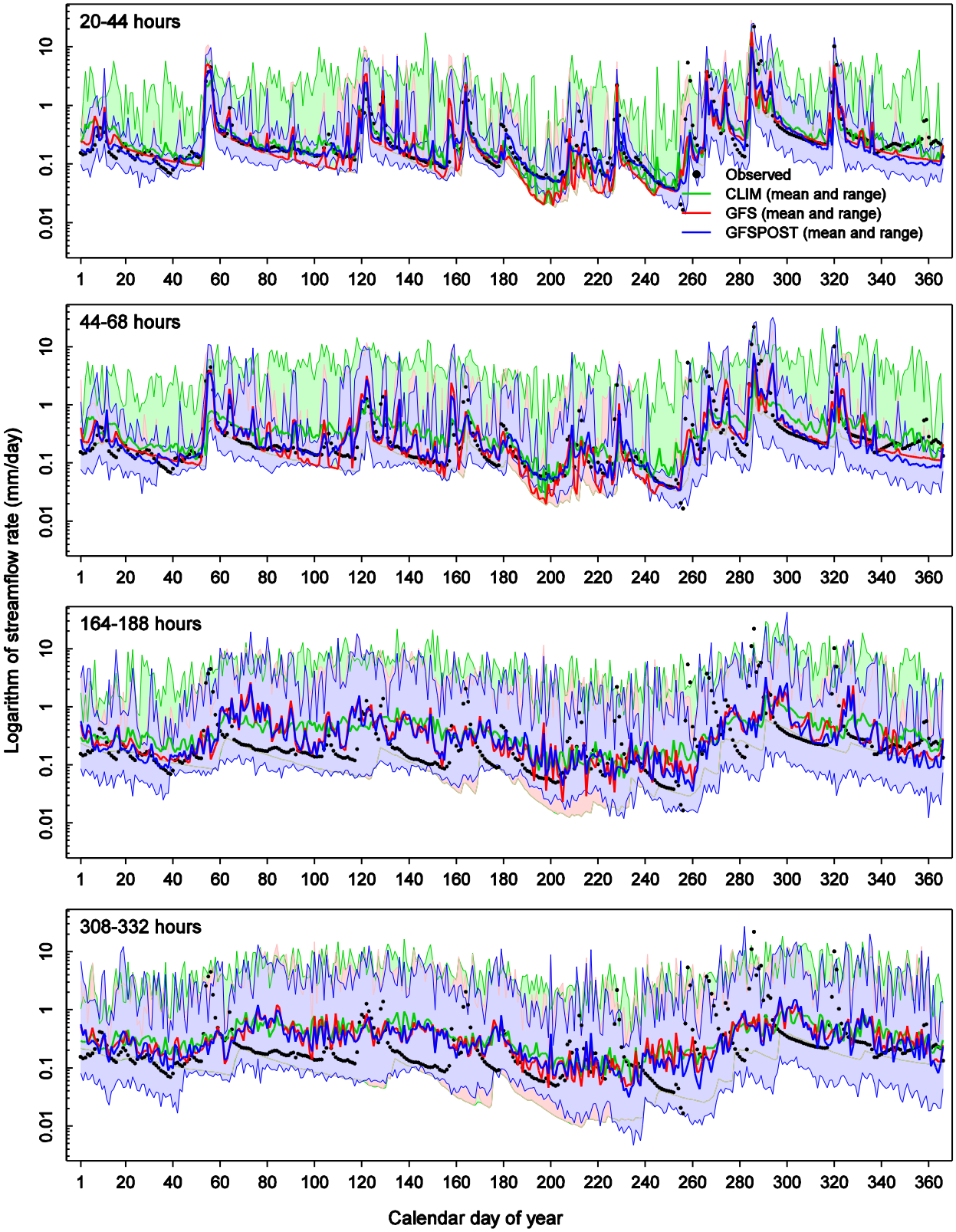

Figure C02: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1985 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

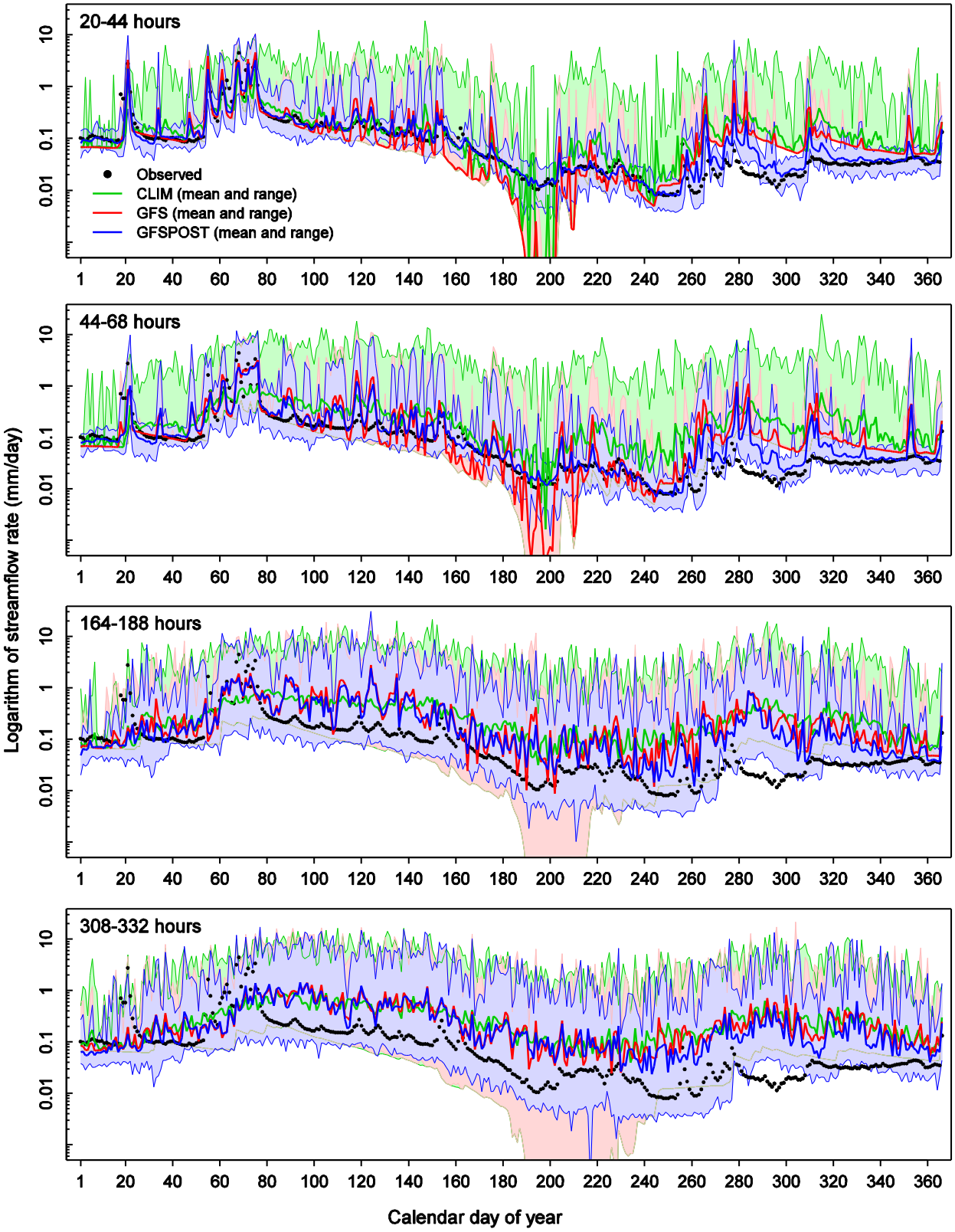

Figure C03: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

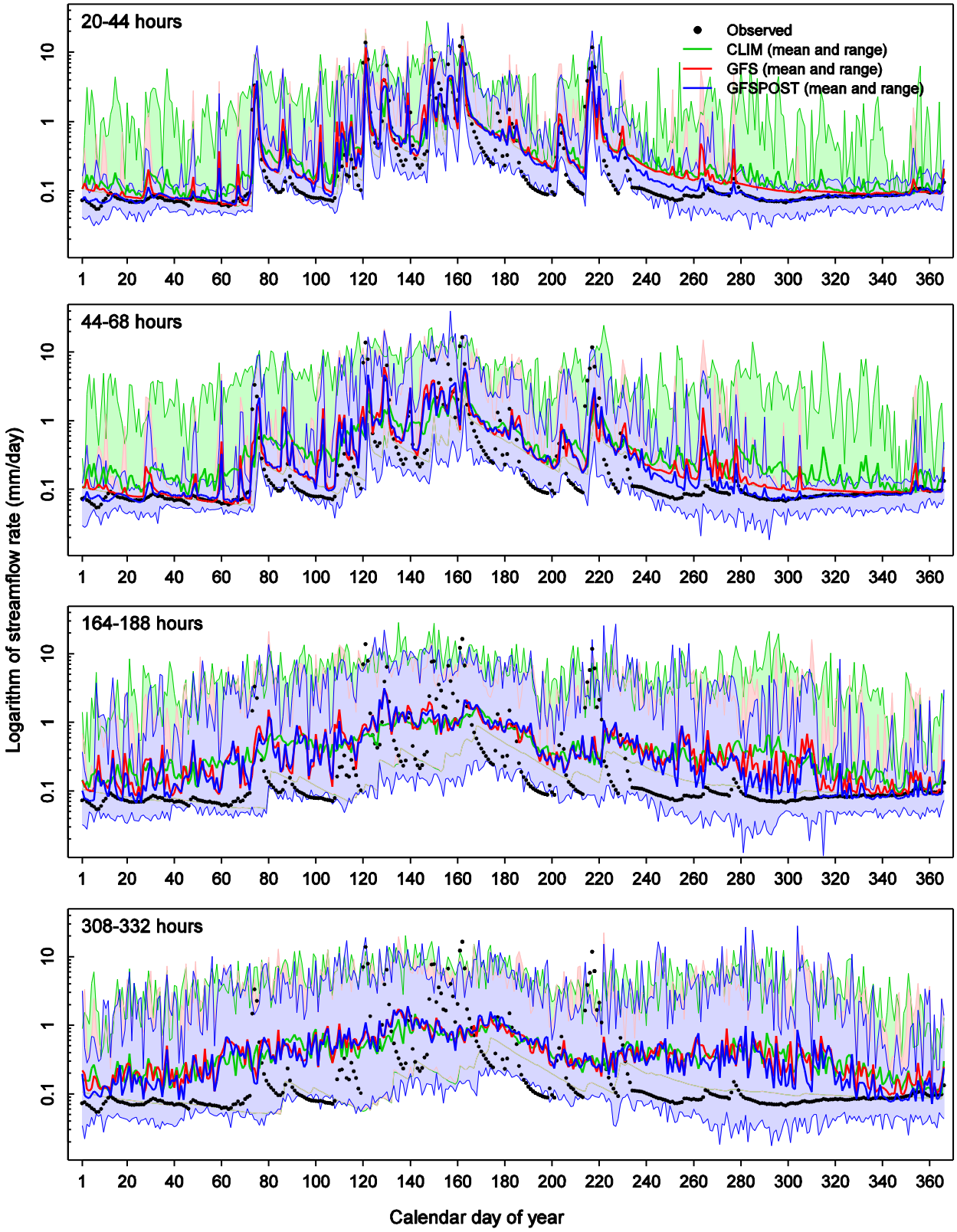

Figure C04: Mean and range of the streamflow forecasts in BLKO2. The results are shown by forecast valid date in 1995 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

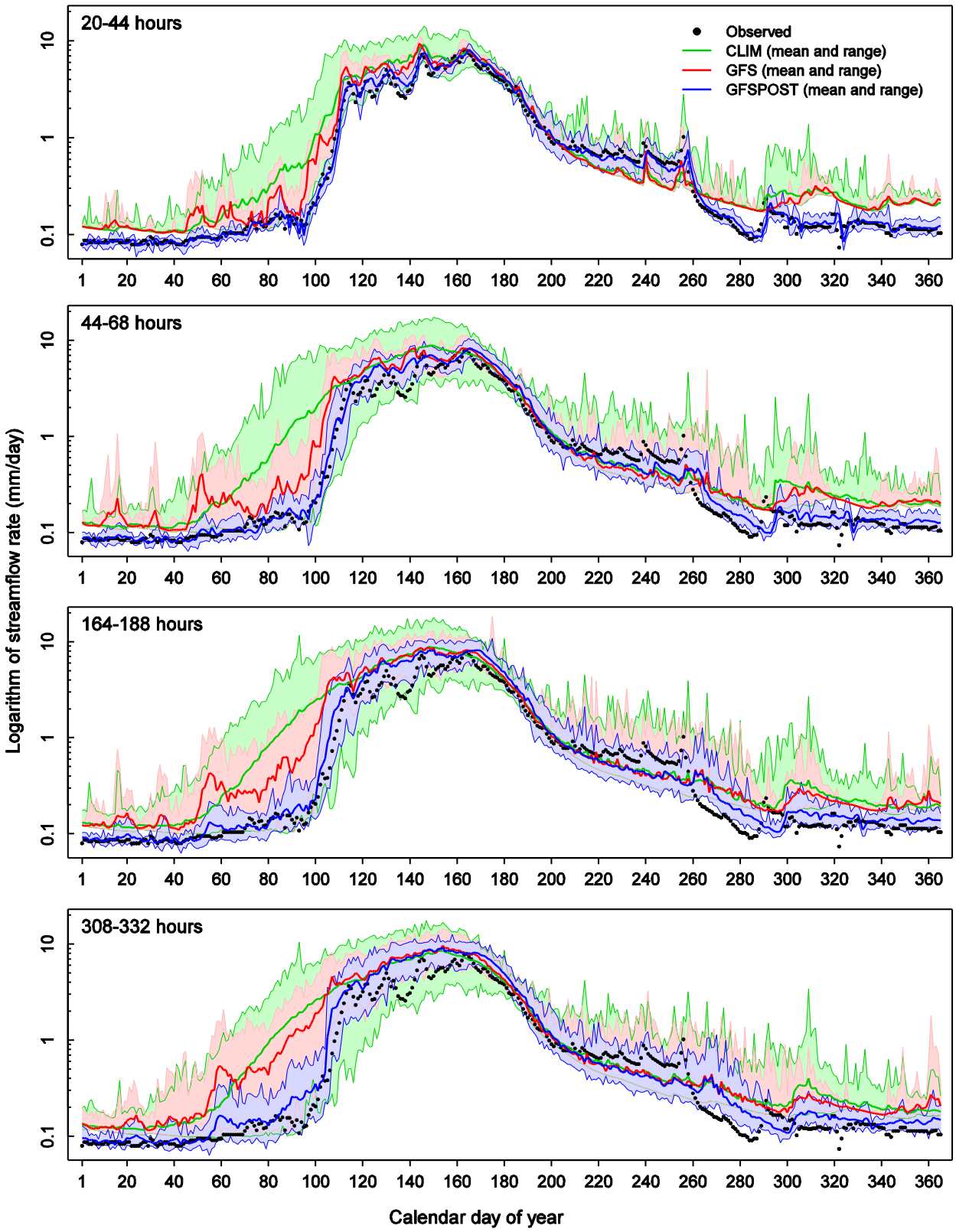

Figure C05: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1980 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

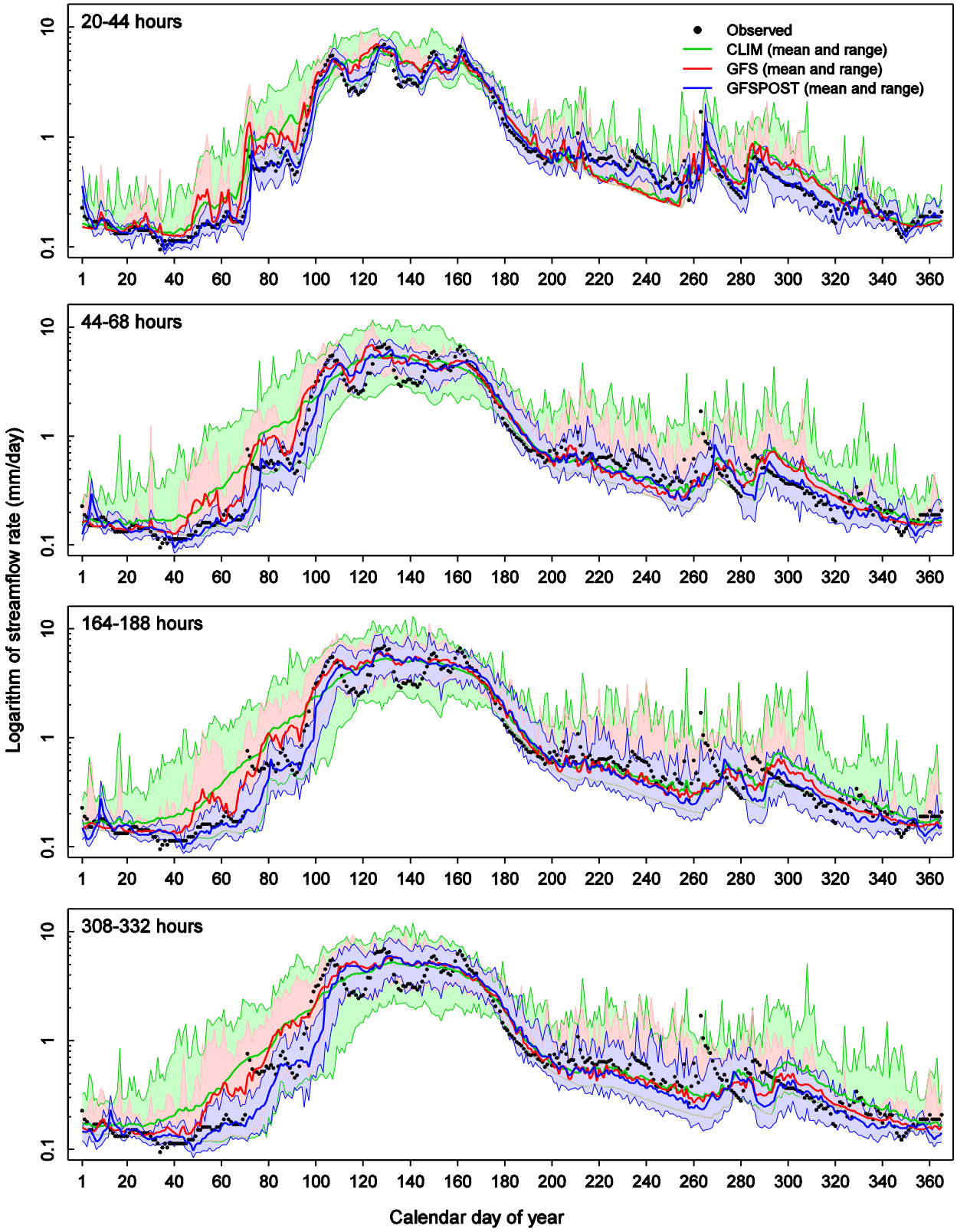

Figure C06: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1985 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

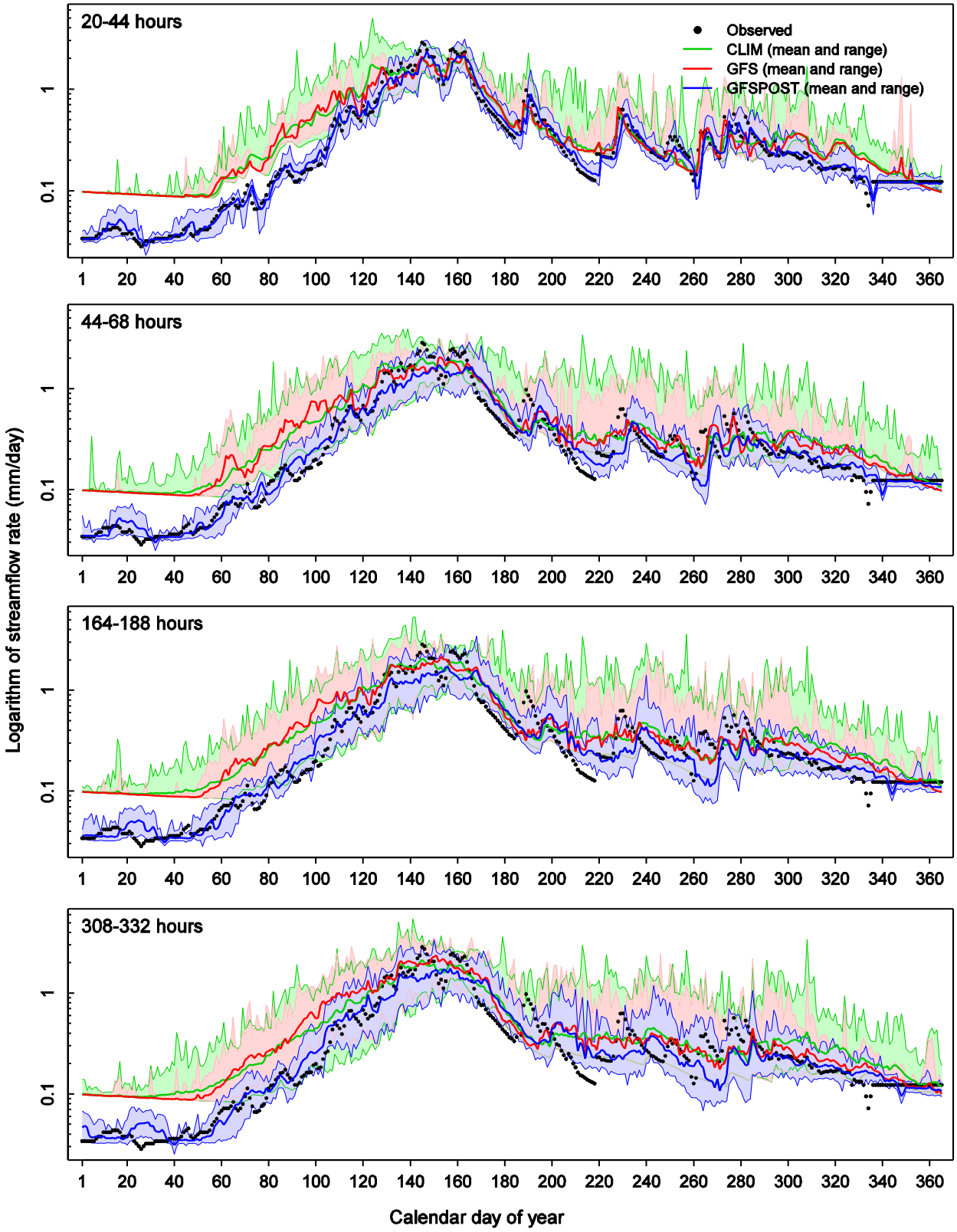

Figure C07: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

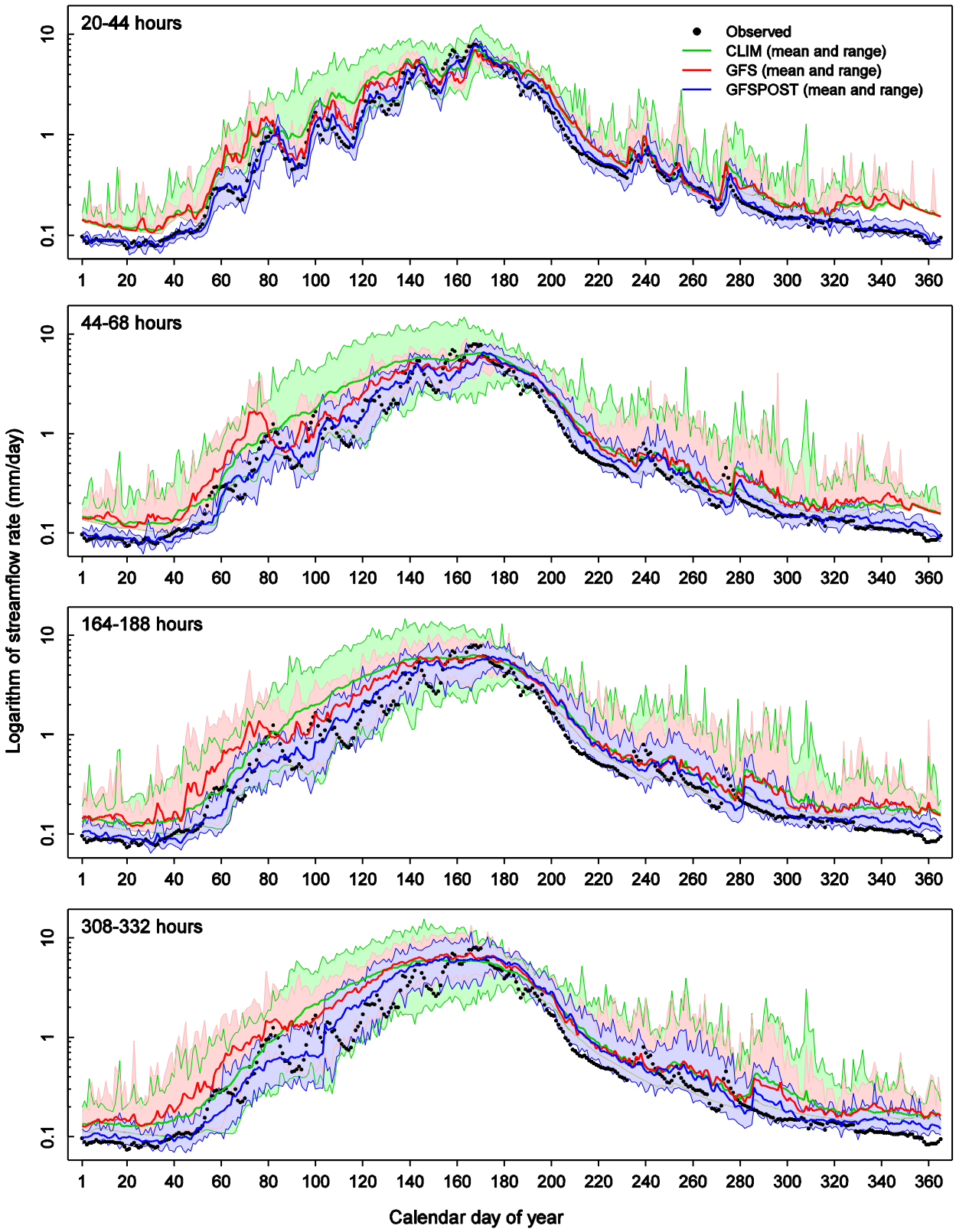

Figure C08: Mean and range of the streamflow forecasts in DOLC2. The results are shown by forecast valid date in 1995 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

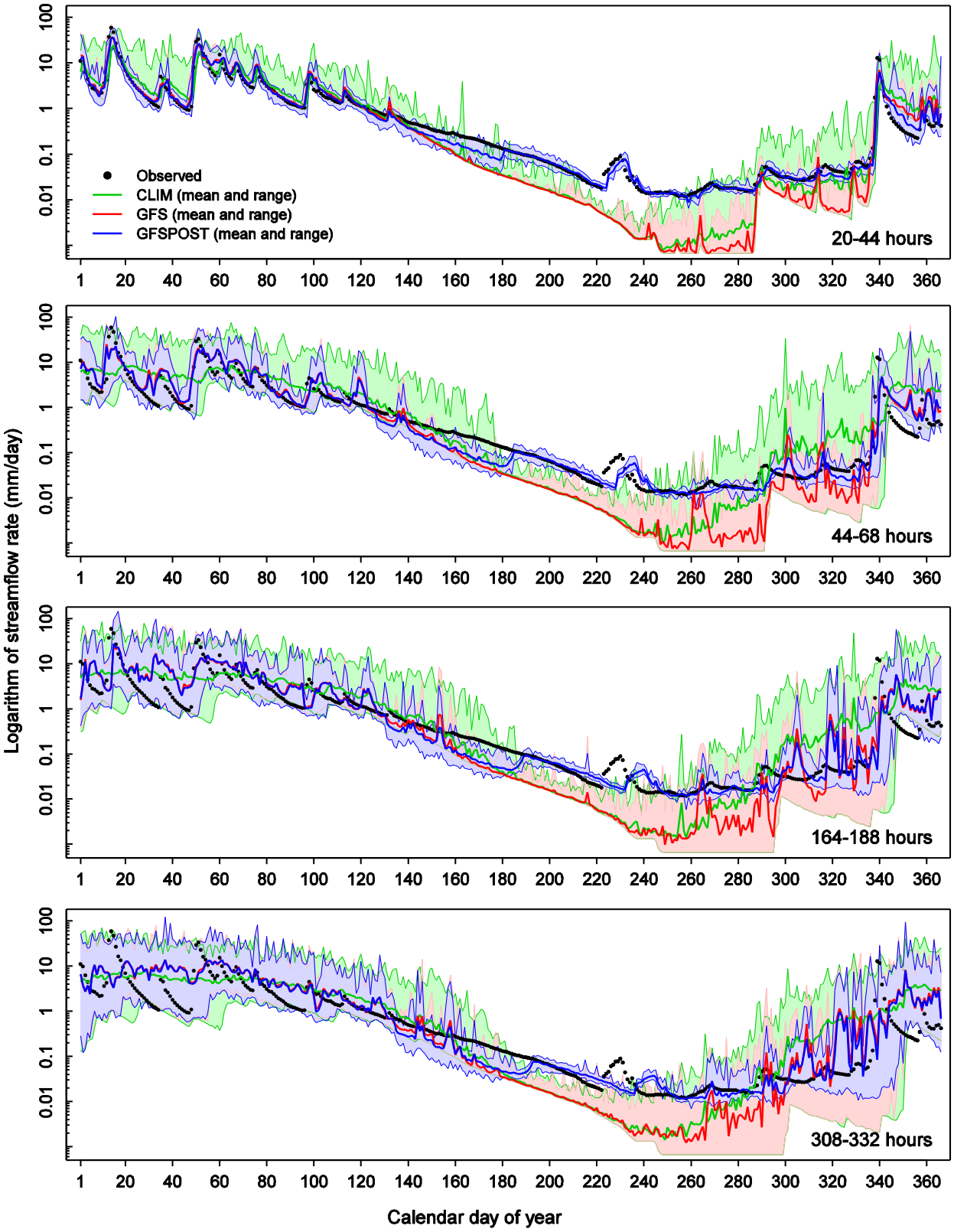

Figure C09: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1980 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

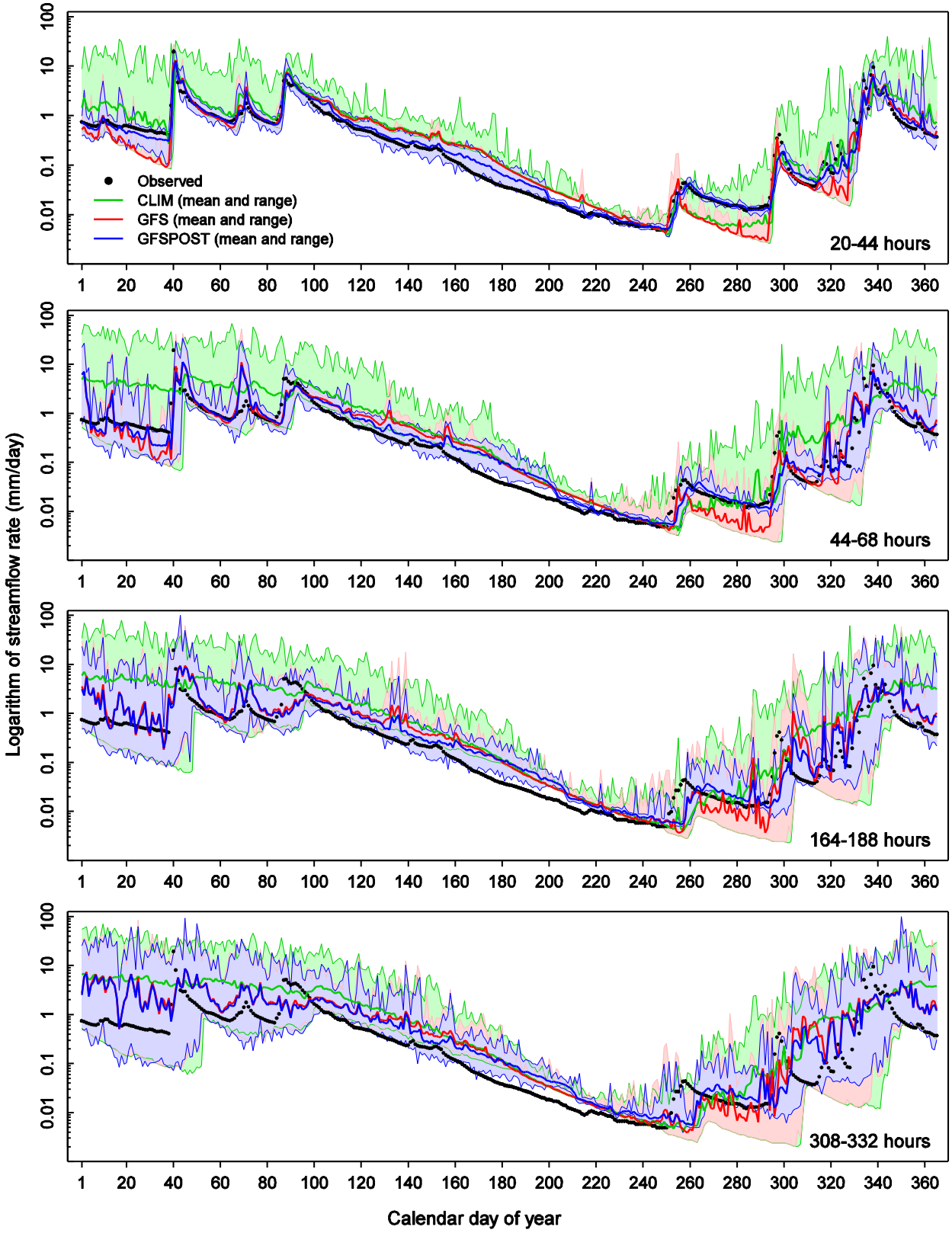

Figure C10: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1985 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

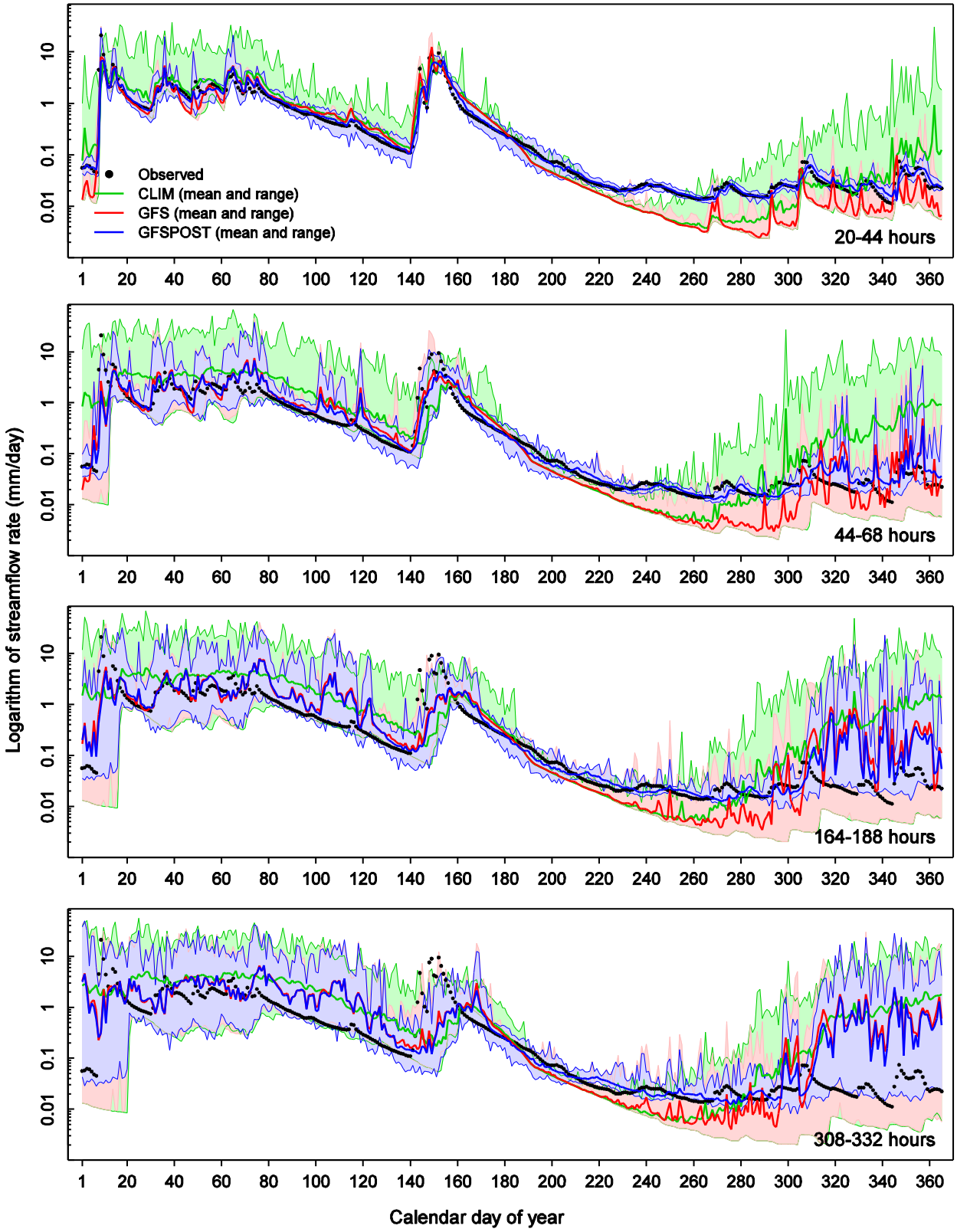

Figure C11: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

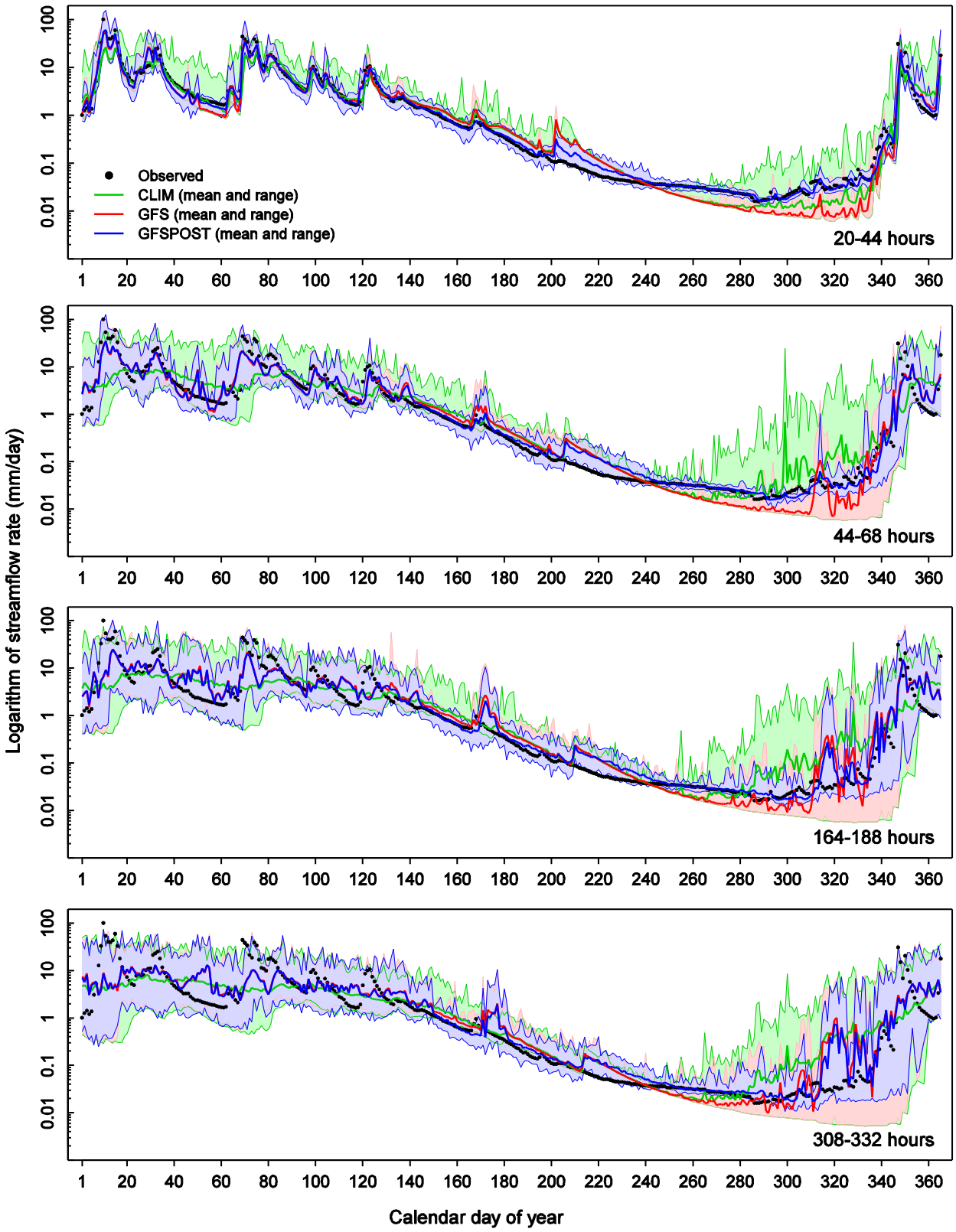

Figure C12: Mean and range of the streamflow forecasts in FTSC1. The results are shown by forecast valid date in 1995 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

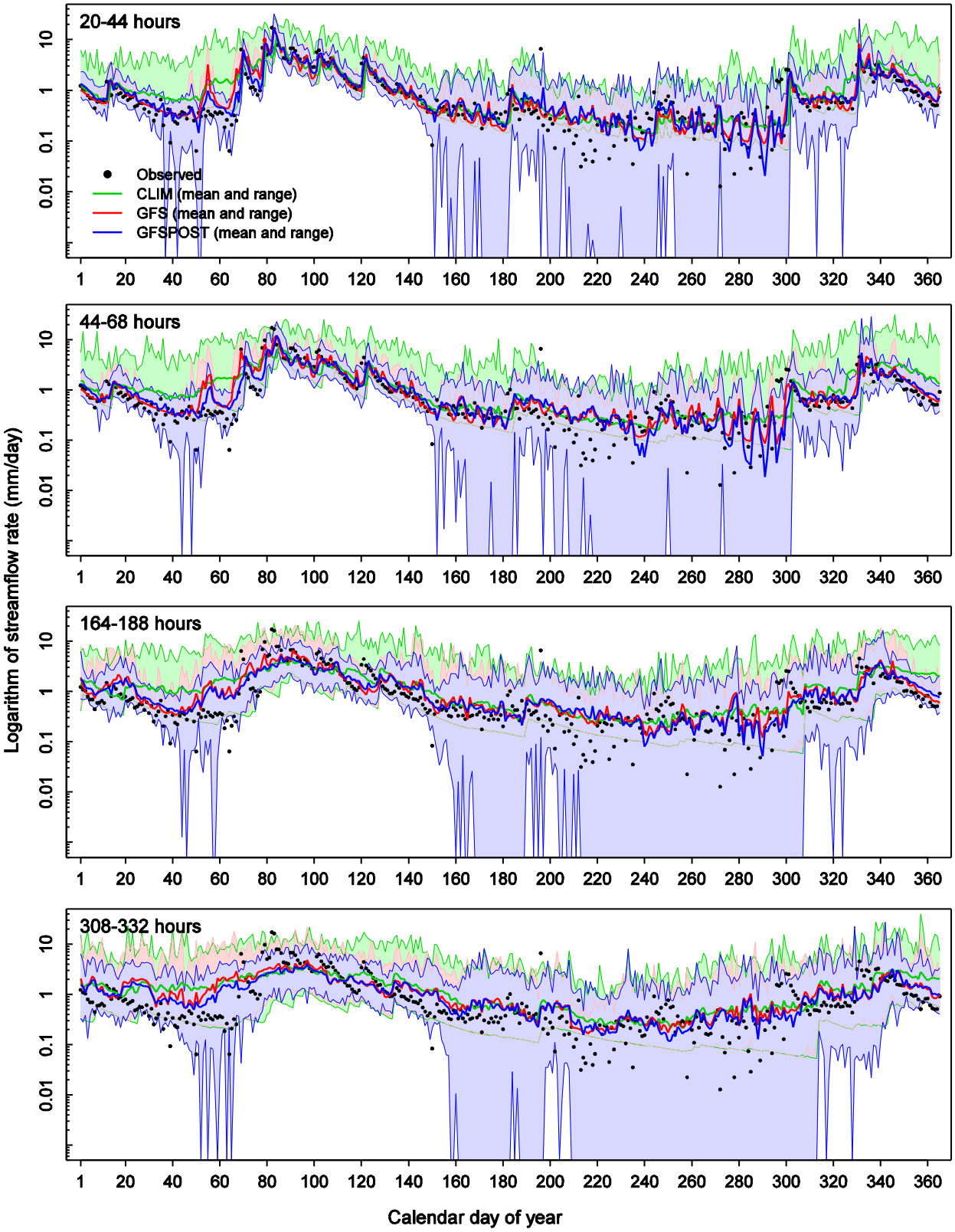

Figure C13: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1980 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

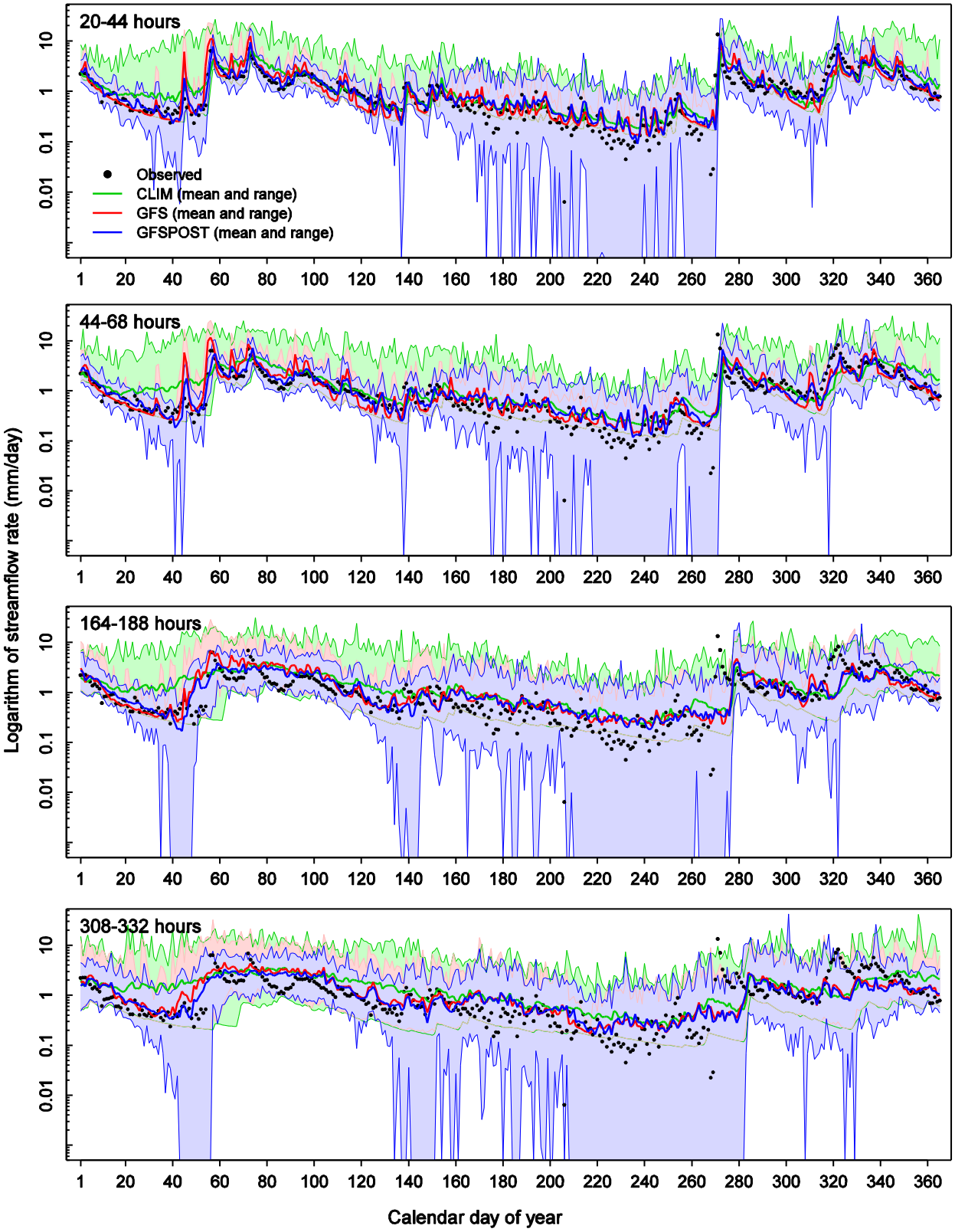

Figure C14: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1985 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

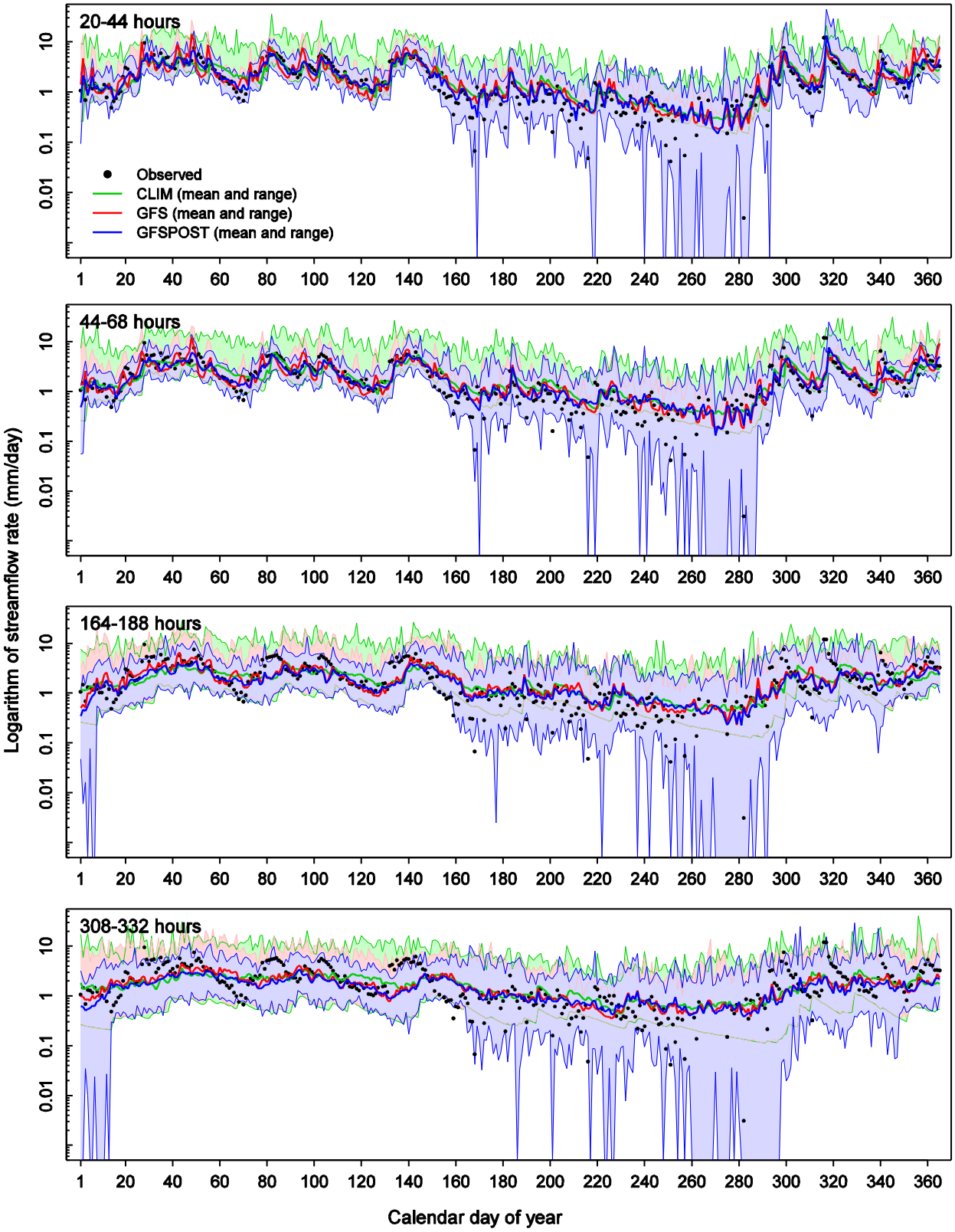

Figure C15: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1990 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).

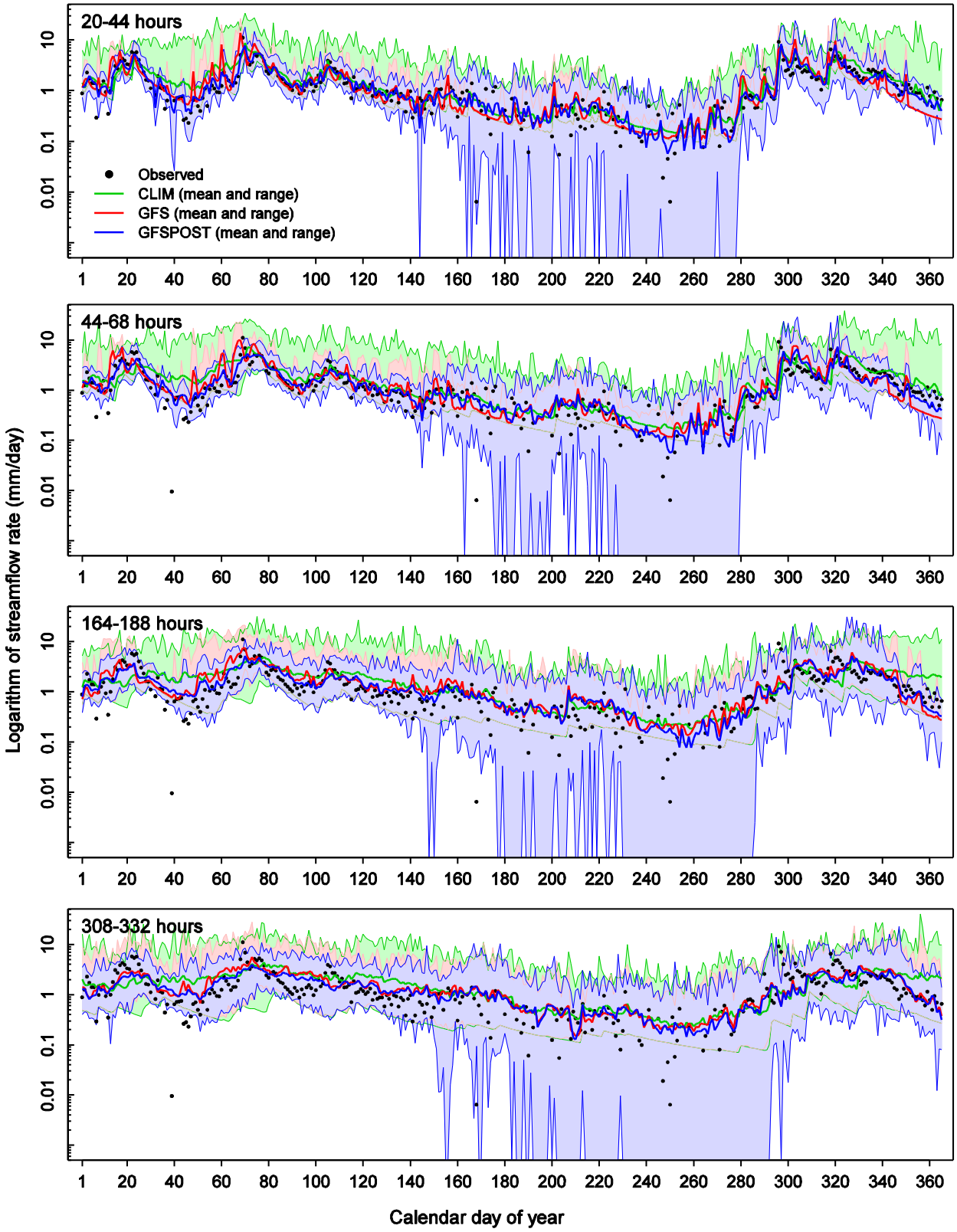

Figure C16: Mean and range of the streamflow forecasts in CNNN6. The results are shown by forecast valid date in 1995 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology (CLIM) and GFS as input (GFS). The post-processed streamflow forecasts comprise forcing from the MEFP with GFS as input (GFSPOST).