Verification of long-range temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service

Revision number: 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 Nos. 3 and 5)

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

Wednesday, November 20, 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 the Mid-Atlantic River Forecast Center (MARFC) and the North East RFC (NERFC). The meteorological hindcasts were produced with the HEFS Meteorological Ensemble Forecast Processor (MEFP). The MEFP was calibrated with raw forcing from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) from 1-15 days and the NCEP Climate Forecast System Version 2.0 (CFSv2) for 16-270 days, together with climatological forcing from 271-330 days. The streamflow hindcasts cover a 15 year period between 1985 and 1999. The hindcasts were verified conditionally upon forecast lead time, magnitude of the observed and forecast variables, season, and aggregation period. Verification results are presented for the temperature, precipitation and streamflow forecasts. In order to distinguish between the contributions of the meteorological and hydrologic uncertainties to the quality of the streamflow forecasts, verification is performed against simulated streamflow (effectively 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 long-range forecasting, together with recommendations on future enhancements.

Document history

Action	Version	Person	Date
Complete first draft	1.0	James Brown	09/16/2013
Minor updates	1.1	James Brown	09/23/2013
Review from Satish Regonda	1.2	James Brown	09/27/2013
Review from Limin Wu	1.3	James Brown	09/30/2013
Review from Haksu Lee	1.4	James Brown	09/30/2013
Review from Kevin He	1.5	James Brown	09/30/2013
Final edits to draft	1.6	James Brown	09/30/2013
Review from Hank Herr	1.7	James Brown	11/18/2013
Review from Ernie Wells	1.8	James Brown	11/18/2013
Complete final version	Final	James Brown	11/20/2013

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), and Limin Wu (MEFP calibration and hindcasting). The streamflow hindcasting was commissioned by the New York City Department of Environmental Protection (NYCDEP). The Middle Atlantic RFC (MARFC) and the North East RFC (NERFC) provided guidance on the CHPS configurations and some of the data required for hindcasting and verification. Additional data and support was provided by Michael Thiemann and Jay Day of Riverside, Luke Wang of Hayzen and Sawyer, and John Schaake, an independent consultant. The report was reviewed by Haksu Lee, Limin Wu, Satish Regonda, Kevin He, Ernie Wells and Hank Herr.

CONTENTS

1. How to read this document

This document aims to: 1) provide a comprehensive scientific evaluation (verification) of the temperature, precipitation and streamflow forecasts from the HEFSv1 with forcing inputs from the Global Ensemble Forecast System (GEFS), the Climate Forecast System (CFSv2) and climatology (collectively referred to as GCC); and 2) communicate the strengths and weaknesses of the HEFSv1 for operational forecasting over the long-range. 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 report and the strengths and weakness of the forecasts in non-technical terms;

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

III. Appendix C. This shows a selection of the paired streamflow forecasts and observations from which the verification results are derived. The relative scatter of the observations within the ensemble forecast distribution provides some insight into the quality of the forecasts when using different forcing inputs. In general, the streamflow observations should fall "randomly" within the ensemble range. They should not fall consistently in one part of the ensemble forecast distribution or outside of the ensemble range;

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

V. Section 5.2. The verification results are presented separately for the meteorological forecasts and the "raw" streamflow forecasts (which do not include streamflow post-processing). The raw streamflow forecasts are verified against simulated flows, as well as observed flows. By verifying against simulated flows, the hydrologic biases and uncertainties are effectively removed.

Some plots are simpler to understand than others. Skill scores are generally simpler to understand and to compare between basins, partly because they are dimensionless. A skill score measures the fractional improvement of one forecasting system over another (0→1, although negative values are possible). For example, Figure 7 shows the fractional improvement of the MEFP precipitation forecasts with GCC forcing versus the unconditional observations (raw climatology) and with an enhanced or "resampled" climatology. Figure 18 shows the skill of the raw streamflow forecasts with GCC forcing when verified against the observed streamflows and the simulated streamflows. The baseline comprises the raw streamflow forecasts with climatological forcing.

It is also important to understand the limitations of this study. First, it does not provide any guidance on the calibration or configuration of the HEFS. Such guidance would require hindcasting and verification for multiple calibration and configuration scenarios. Second, the report covers only a small fraction of the locations and scenarios under which the HEFS will be used operationally. It focuses on headwater basins and downstream basins that are effectively treated as headwaters. All of the downstream basins are subject to river regulations, including flow diversions that are applied in real-time. Estimates of the local streamflows were provided by the New York City Department of Environmental Protection (NYCDEP), after accounting for reservoir releases and flow diversions. Ideally, the archived diversions and other regulations would be incorporated into the streamflow hindcasting, as the operational forecasts comprise residual uncertainties and biases from upstream locations, including those

from reservoir modeling. However, only the estimated local flows were available from NYCDEP and, hence, only the local contributions were verified at downstream locations. Thirdly, forecast products will generally comprise the bias-corrected streamflow forecasts. However, the Ensemble Postprocessor (EnsPost) was undergoing improvements and could not be considered here. In the absence of a formal bias-correction, the raw streamflow forecasts were verified against simulated flows, as well as observed flows, in order to factor out the hydrologic uncertainties and biases. Finally, the report does not explicitly benchmark the HEFSv1 against archived operational forecasts, notably from Ensemble Streamflow Prediction (ESP). However, the GCC streamflow forecasts are compared to streamflow forecast with climatological forcing, which are similar to those from ESP.

2. Executive summary and recommendations

- Ensemble forecasts of precipitation, temperature and streamflow were generated with the NWS Hydrologic Ensemble Forecast Service (HEFS) for a 15 year period between 1985 and 1999. The hindcasts were produced for eight river basins, comprising four basins in each of two RFCs, namely the Middle Atlantic River Forecast Center (MARFC) and the North-East River Forecast Center (NERFC). The basins include a range of headwater and downstream locations within the Delaware and Catskill systems of New York State. They are subject to extensive river regulations, including diversions to the New York City (NYC) municipal water supply. The four basins in MARFC comprise three locations on the Delaware River, namely Walton (WALN6), Callicoon (CCRN6) and Montague (MTGN4), and one location on the Neversink River, namely the Neversink Reservoir (NVXN6). The four basins in NERFC comprise two locations on the Esopus Creek, namely Mount Trempor (MTRN6) and Mount Marion (MRNN6), and two locations on the Schoharie Creek, namely Prattsville (PTVN6) and the Gilboa Dam (GILN6). The hindcasts were commissioned by the NYC Department of Environmental Protection (NYCDEP), in order to support the initial implementation of the HEFS at MARFC and NERFC and to improve the management of risks to water quantity and quality objectives in the NYC area.
- The HEFS is being evaluated in several phases. In the initial phase, verification was conducted for temperature, precipitation and streamflow hindcasts with forcing inputs from the “frozen” version of NCEP’s Global Forecast System (GFS). A subsequent phase will evaluate the HEFS with forcing inputs from NCEP’s operational Global Ensemble Forecast System (GEFS) and compare to those from the frozen GFS. This report focuses on the quality of the long-range forecasts from ~15 days to ~1 year. Specifically, it focuses on the temperature, precipitation and streamflow forecasts with forcing inputs from the GEFS, the Climate Forecast System Version 2.0 (CFSv2) and resampled climatology (defined below). While the focus is on the long-range forecasts, the HEFSv1 aims to provide “seamless” forecasts across multiple temporal scales. Depending

on basin characteristics, skillful forcing from the GEFS may persist for several weeks in the streamflow forecasts. Thus, the 1-15 day forecasts are also verified at an appropriate temporal scale. Collectively, the phased evaluation aims to: establish the expected performance and limitations of the HEFS; demonstrate that the outputs from the HEFS are reasonably unbiased and skillful; identify the key factors responsible for forecast error and skill in different situations; isolate the contribution of the meteorological and hydrologic components of the HEFS 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.

- Precipitation and temperature hindcasts were produced with the Meteorological Ensemble Forecast Processor (MEFP) using “raw” precipitation and temperature forecasts from multiple sources. Ensemble forecasts from NCEP’s GEFS were used for the period 1-15 days. Single-valued forecasts from the CFSv2 were used for the period 16-270 days. For the period 271-330 days, and as a reference forecast for the period 1-330 days, climatological ensembles were derived from historical observations of mean areal precipitation (MAP) and mean areal temperature (MAT). This involved resampling the MAP and MAT in a moving window of, respectively, 30 days and 15 days around the forecast valid date (“resampled climatology”). The GEFS, CFSv2 and resampled climatology are collectively denoted GCC, while resampled climatology is denoted CLIM. The streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS) using the operational hydrologic models and configurations provided by MARFC and NERFC.
- The precipitation, temperature and streamflow forecasts were verified with the Ensemble Verification System (EVS). The forecasts were verified conditionally upon season, forecast lead time, magnitude of the observed and forecast variables, and aggregation period. The raw streamflow forecasts were verified against simulated streamflows, as well as observed streamflows, in order to

separate the meteorological uncertainties from the total (meteorological and hydrologic) uncertainties. In practice, however, the simulated streamflows also include errors in the meteorological observations, while the hydrologic observations include errors from stage measurements, flow ratings and accounting for upstream regulations (see below).

- In general, the MEFP-GCC precipitation forecasts are both reliable and skillful during the short-range (1-5 days). This largely originates from the skill in the raw GEFS precipitation forecasts. Likewise, the MEFP-GCC temperature forecasts are reliable and skillful during the short-range. Beyond the first 1-5 days, the skill of the MEFP-GCC precipitation forecasts declines rapidly, while the temperature forecasts remain skillful throughout the medium-range. However, neither the precipitation nor the temperature forecasts are skillful beyond ~2 weeks. This originates from a lack of skill in the raw CFSv2 forecasts and, beyond 270 days, from resampled climatology, which is inherently unskillful. Indeed, for the period from ~15-330 days, the MEFP-GCC forecasts show similar conditional biases to the MEFP-CLIM forecasts. This includes a substantial underestimation of the largest precipitation totals and a smaller conditional bias in the temperature forecasts, whereby the lowest and highest observed temperatures are over- and under-estimated, respectively. While the MEFP precipitation forecasts are generally no worse than sample climatology, the forecasts of Probability of Precipitation (PoP) are consistently worse than climatology. This originates from a lack of reliability in the MEFP forecasts for PoP. Similar biases were observed when calibrating the MEFP with the frozen version of NCEP's GFS. Again, this suggests a problem in the modeling, estimation, or implementation of the MEFP for PoP and light precipitation amounts.
- Except for PoP and light precipitation, the MEFP-GCC forecasts are no worse than resampled climatology. This is an important attribute of any bias-correction technique whose unconditional distribution is climatology. Also, the MEFP maintains or improves upon the correlations between the raw forcing from the GEFS and CFSv2 and the corresponding observed precipitation amounts. For

example, the MEFP-CLIM shows background correlations of ~0.2 during the long-range, whereas the raw CFSv2 forecasts show correlations of ~0.0. Without skillful predictors, the MEFP cannot improve upon climatology; it can only issue forecasts that are unconditionally unbiased. In practice, the ensemble mean of the MEFP-CLIM forecasts contain small unconditional biases (<5%), which may be related to the modeling of PoP and light precipitation amounts.

- Seasonal water supply and other long-range applications are known to benefit from climatological forecasts (so-called Ensemble Streamflow Prediction, ESP). Enhancements to the MEFP may consider additional predictors that improve upon climatology. For example, given the strong autocorrelations in temperature, the MEFP may benefit from an autoregression of the future MAT on the most recently observed MAT, as well as the raw forecast. While precipitation generally shows much weaker autocorrelations, auxiliary variables, such as relative humidity, may improve forecast quality over the short- to medium-range. For seasonal and long-range prediction, the MEFP may benefit from auxiliary climate information, such as climate outlooks from NOAA's Climate Prediction Center (CPC), or indices of teleconnection patterns, such as the El-Niño Southern Oscillation (ENSO), the Pacific-North American teleconnection (PNA) and the Pacific Diurnal Oscillation (PDO).

Recommendation 1: In order to enhance the limited skill of the CFSv2 for long-range forecasting of precipitation and temperature, additional predictors may be included in the MEFP alongside the raw CFSv2 forecasts. In some regions and time periods, the MEFP may benefit from additional sources of climate information, such as the 90-day climate outlooks from NOAA's CPC. Other useful predictors may include indices of the ENSO, PDO and PNA or other regional climate patterns. In order to support the HEFS as a unified platform for ensemble forecasting, any local implementations of ESP that demonstrably improve upon the HEFS should be integrated into the HEFS and ESP should then be retired as a legacy platform. For short- to medium-range forecasting, the MEFP may benefit from an autoregression of the future MAT on the most recently observed MAT, while the precipitation forecasts may benefit from carefully-selected auxiliary variables (although precipitation is

inherently difficult to forecast beyond the medium-range).

- As well as producing reliable forecasts at discrete times and locations, the MEFP should maintain realistic patterns in space and time and between variables. These statistical dependencies and multi-scale properties are important for decision making. In general, decisions about water resources are based on products derived from hydrologic forecasts, such as aggregated quantities, or on additional modeling studies or rules embedded into decision support systems. As with hydrologic modeling, these calculations involve uncertainty propagation, for which the space-time and cross variable relationships are important. In this context, important attributes of the MEFP include the ability to: 1) preserve space-time and cross-variable relationships via the Schaake Shuffle; 2) derive skillful predictors at multiple space-time scales using “canonical events” (aggregated predictors); and 3) provide seamless predictions across multiple forecast horizons, depending on the raw forcing available. In practice, some discontinuities were observed in the verification statistics between 270-271 days, where the CFSv2 transitions to resampled climatology. This may originate from sampling uncertainty, including uncertainty resulting from the parameterization of the MEFP with too many canonical events (see below). Anecdotally, the reliability and skill of the MEFP forecasts is the same or better at aggregated scales. This is consistent with the temporal autocorrelations being modeled adequately. Nevertheless, further investigation is warranted into the limitations of the Schaake Shuffle, particularly for extreme events, and whether, at the basin-scale, other empirical structures, such as high-resolution forecasts or conditional climatologies, can better reproduce the space-time covariability.

Recommendation 2: Further investigation is warranted into the limitations of the Schaake Shuffle and the conditions under which other empirical structures (empirical copulas) may improve the modeling of space-time and cross-variable relationships. In practice, these relationships are conditional upon the state of the atmosphere at the forecast valid time, yet the Schaake Shuffle relies on unconditional structures only (i.e. conditional upon forecast valid time, but not on the state of the atmosphere). Among other things, there is a need to explore

the consequences of this assumption during high-impact events, for which suitable analogs are unlikely to appear on the same date in historical years. More generally, there is a need to evaluate the MEFP for high-impact events by deconstructing and evaluating specific forecasts. Here, there is a trade-off between the ability of the MEFP to learn from historical experience, in order to reduce conditional biases, and the need to preserve any novel information in the raw model forecasts (i.e. for which historical experience is limited).

- In operational forecasting, there is always a trade-off between model complexity, or the need to capture salient features of the observations, and practicality, or the need for a model whose parameters can be estimated reliably. It is questionable whether the current implementation (or parameterization) of the MEFP manages this trade-off effectively. The MEFP uses canonical events to sequentially adjust the climatological probability distribution. Each canonical event comprises a separate model of the joint probability distribution of the forecasts and observations. A canonical event defines a window centered on the forecast valid date (into which data are pooled from all historical years), together with the period of aggregation for which the joint distribution is estimated. The sample size used to calibrate the MEFP was not particularly small (15 years) and is consistent, or more favorable, than the expected operational practice. However, artificial periodicities were clearly visible in some of the verification statistics. They were also observed in the raw ensemble traces, particularly for temperature, which should otherwise vary smoothly. By experimenting with the parameterization of the MEFP, these discontinuities were found to originate from the use of canonical events. The addition or removal of particular canonical events led to discontinuities in the ensemble traces at corresponding timescales during the forecast horizon.

Recommendation 3: Further investigation is warranted into the use of canonical events and, more generally, the need to parameterize the multiscale properties of temperature and precipitation in the MEFP. Where explicit modeling is justified, it should be parsimonious, smooth and allow for reasonably small sampling uncertainty, whether using canonical events or other techniques. Outcomes of this investigation may include simplified

modeling approaches or further guidance on calibrating the MEFP with canonical events (e.g. based on geography and climatology).

- In regulated rivers, the hydrologic uncertainties reflect a combination of natural and engineering influences. Some regulations may be obscured from operational forecasters because they are operated with limited warning or specificity (e.g. because they involve rapidly changing conditions, multiple actors or agencies or commercially sensitive information). When information about diversions and other regulations is available in real-time, statistical post-processors, such as the Ensemble Post-processor (EnsPost), should ideally model the natural (local) flows, as upstream regulations are difficult to model statistically. However, the total flows are preferred for hindcasting and verification, as they include the residual uncertainties from upstream basins. In practice, only the estimated local flows were available from NYCDEP and, hence, only the local contributions were verified at downstream locations. In future, river regulations should be archived by the RFCs, in order to allow for hindcasting and verification of the total flows at downstream locations.

Recommendation 4: There are a number of challenges for the successful application of the HEFS in regulated rivers. These include inadequate archiving of real-time adjustments to operational forecasts (runtime modifications), which are also required for hindcasting and verification, and difficulties in adjusting regulated flows with statistical techniques. Where possible, the EnsPost should be calibrated on natural flows and any known regulations incorporated in real-time. In other cases (e.g. when the regulations are poorly defined), regulations may leave signatures in the streamflow observations that can be modeled indirectly, whether using deterministic or stochastic techniques. In order to guide practical applications of the HEFS in regulated rivers, further evaluation is needed, including evaluation of the total flows at downstream locations. In this context, collaborations between the NWS and other agencies, such as the NYCDEP, should be encouraged. Ultimately, improvements in the modeling of upstream regulations will reduce the need for indirect accounting and increase the operational readiness of the HEFS. Aside from these collaborations, the RFCs should archive *all* adjustments to their operational forecasts, in order to support routine hindcasting and verification in regulated rivers.

- In keeping with the MEFP-GCC precipitation forecasts, the GCC streamflow forecasts are substantially more skillful than the climatological forecasts during the first week. Beyond the short-range, they are as skillful as the climatological forecasts. In general, the hydrologic biases are greater under low flow conditions, where the streamflow forecasts systematically over-estimate the observed flows, particularly in CCRN6 and MTGN4. While the hydrologic models do not target specific applications or flow conditions, the high flows receive particular scrutiny during model calibration, and the "Continuous API" model used by MARFC (for CCRN6 and MTGN4) is less well-suited to dry conditions. For long-range forecasting, the meteorological biases are more important than the hydrologic biases, as the MEFP-GCC precipitation forecasts resemble climatology after ~5 days. In particular, they underestimate the heaviest precipitation amounts by up to ~80% at longer forecast lead times.
- Further work is needed to compare the long-range streamflow forecasts from the HEFS against the RFC operational forecasts, which include ESP and statistical modeling on monthly and seasonal timescales. Given the lack of skill in the CFSv2, the opportunities to improve on climatological forcing and, thus, on ESP may appear limited. However, this does not imply similar performance in other regions or time periods, where long-range prediction is more straightforward (e.g. the coastal mountain ranges of California and the Pacific Northwest). Alongside the resampling procedure adopted by the MEFP, there are other notable differences between the HEFS and ESP. For example, some RFCs incorporate runtime modifications into the hydrologic model states from which the ESP forecasts are produced operationally. Whether the HEFS can improve on ESP will also depend on basin characteristics. For headwater basins with longer memory (e.g. due to snow accumulation or soil characteristics), and for downstream basins in general, the skill from the GEFS will persist for longer in the streamflow forecasts. Also, the EnsPost should eliminate any unconditional biases, which may originate from weaknesses in the structure or calibration of the hydrologic models, and may reduce conditional biases where the streamflow correlations are strong. For example, at MRNN6 and GILN6, the observed

streamflows were consistently underestimated during the late spring and early summer. At CCRN6 and MTGN4, the observed streamflows were consistently overestimated during the summer months.

- Scientific evaluation of the HEFS is an ongoing activity; it requires a sustained effort and a dedicated infrastructure for archiving data and for hindcasting and verification, as well as communicating verification concepts and results. This study covers only a small fraction of the locations, conditions and scenarios under which the HEFS will be used operationally. In order to guide a broader range of applications 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. Furthermore, there is a need to evaluate decision support systems and other applications that rely on the HEFS (e.g. water quality models, water regulation/supply decisions). Such applications will show varying sensitivities to the HEFS forecasts and may lead to targeted improvements in the HEFS, as well as new ensemble products.

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. Further work is needed to compare the long-range streamflow forecasts from the HEFS against the RFC operational forecasts, which include ESP and statistical modeling on monthly and seasonal timescales. While such comparisons are not straightforward (e.g. because the raw forcing data used by the HEFS is not used for operational forecasting), they are necessary to benchmark the HEFS and to show that, overall, the forecasts improve on existing products. In addition, there is a need to evaluate decision support systems and other applications that rely on the HEFS, such as water quality, ecology, river navigation, water supply, and civil engineering design. Such applications will show varying sensitivities to the HEFS forecasts and are necessary to demonstrate the wider, societal and economic, benefits of the HEFS and ensemble forecasting more generally.

3. Introduction

Uncertainties are manifest in all aspects of environmental modeling (Brown, 2010a) and they contribute to risks in environmental decision making (Handmer et al., 2001; Beven, 2000; Ramos et al., 2012; Demeritt et al., 2013). In order to evaluate, communicate and manage these risks effectively, operational forecasting agencies, such as the U.S. National Weather Service (NWS), must properly account for, and quantify, the uncertainties associated with model predictions. Whether using physically-based models, statistical models or some combination of the two, the inputs, structure, and parameters of these models are all uncertain (Matott et al., 2009). Uncertainties propagate through the modeling system and lead to uncertainties about the model outputs (Brown and Heuvelink, 2005). Broadly, there are two approaches to quantifying and propagating uncertainty, namely source-based modeling ("bottom up"), where specific sources of uncertainty are combined and integrated numerically (Gneiting and Raftery, 2005; Helton et al., 2006; Cloke et al., 2013), and statistical modeling ("top down"), where the total uncertainty is modeled empirically (Glahn and Lowery, 1972). A hybrid of these approaches involves statistical post-processing of ensemble forecasts (Gneiting et al., 2007; Montanari and Grossi, 2008; van Andel et al., 2013). The latter uses historical observations to correct for biases in the forecast probabilities.

The NWS Hydrologic Ensemble Forecast Service (HEFS) provides ensemble forecasts of temperature, precipitation and streamflow at lead times ranging from one hour to one year (Seo et al., 2010; Demargne et al., 2014). 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). The raw forcing may comprise operational quantitative precipitation forecasts (QPF) and quantitative temperature forecasts (QTF) from the NWS River Forecast Centers (RFCs) or the ensemble mean of NCEP's Global Ensemble Forecast System (GEFS), among others. For the period from 16 days to 9 months, the MEFP uses raw forcing from the Climate Forecast System Version 2.0 (CFSv2) and, beyond 9 months

or as a baseline for evaluating other forecasts, various types of conditional climatology. The total uncertainty in the streamflow forecasts is modeled in two stages (see Kelly and Krzysztofowicz, 1997 also). 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. Second, the raw streamflow forecasts are post-processed with the Ensemble Postprocessor (EnsPost). The EnsPost accounts for the hydrologic uncertainties and reduces any systematic biases in the streamflow forecasts (Seo et al., 2006).

The HEFS is being implemented in several phases, with the initial version (HEFSv1) scheduled for operational use at all RFCs by the end of 2014. In order to establish a baseline for future enhancements, and to guide the operational use of the HEFSv1, several phases of hindcasting and verification are also underway. This involves retrospective forecasting of temperature, precipitation, and streamflow at selected RFCs and for selected sources of meteorological forcing. In an earlier phase of evaluation (see Brown, 2013), temperature, precipitation and streamflow hindcasts were generated with the HEFSv1 using forcing inputs from the “frozen” version of NCEP’s Global Forecast System (GFS; Hamill et al., 2006). In a subsequent phase of evaluation, temperature, precipitation, and streamflow hindcasts will be generated with the HEFSv1 using forcing hindcasts from NCEP’s Global Ensemble Forecast System (GEFS; Hamill et al., 2013). This report focuses on the quality of the long-range forecasts from ~15 days to ~1 year. Specifically, it focuses on the temperature, precipitation and streamflow forecasts with forcing inputs from the GEFS, CFSv2 and climatology. While the focus is on the long-range forecasts, the HEFSv1 aims to provide “seamless” forecasts across multiple temporal scales and, depending on basin characteristics, skillful forcing from the GEFS may persist for several weeks in the streamflow forecasts.

Approaches to long-range forecasting vary between RFCs, but most use statistical modeling, physically-based modeling or a subjective combination of the two. Ensemble Streamflow Prediction (ESP) was developed in the late 1970s (Day, 1985) and is used operationally by many RFCs. For example, it is used in the western U.S. to

forecast seasonal water supply (Wood and Lettenmaier, 2006), while the North Central RFC uses ESP to evaluate the probability of flooding from snowmelt in the following spring. By initializing the NWS River Forecast System (NWSRFS) and looping through historical time-series of observed temperature and precipitation, ESP provides ensemble forecasts of streamflow that are consistent with the historical climatology. Enhancements to ESP include sampling of the raw climatology with conditioning variables, such as the 90-day climate outlooks from NOAA's Climate Prediction Center (Perica, 1998) or large-scale climate indices (Najafi et al., 2012). Statistical models for long-range forecasting generally employ multiple linear regression (Garen, 1992). Common predictors include snow water equivalent (SWE), precipitation, large-scale climate indices, and antecedent streamflow, among others (e.g. Robertson and Wang, 2013). In order to avoid collinearity, the original covariates may be aggregated or translated into fewer principal components (Regonda, 2006; Garen and Pagano, 2007). Combinations of ESP and statistical modeling are also common, and may involve post-processing ESP forecasts (e.g. Wood and Schaake, 2008) or a subjective blending of ESP and regression models. For example, until recently, the NWS coordinated with the Natural Resources Conservation Service (NRCS) to provide "consensus" forecasts of water supply for 700 basins in the western U.S. In negotiating a best estimate and spread from the individual forecasts, the consensus was necessarily subjective. However, it avoided confusion among end-users and provided a single forecast for critical decisions about water supply (Pagano et al., 2013). Currently, the operational practice varies between RFCs, with some using unconditional ESP (e.g. NWRFC), and others using a subjective combination of ESP and statistical modeling (e.g. CBRFC).

In parts of the U.S., there are strong climate anomalies or teleconnection patterns that significantly impact temperature, precipitation and streamflow on seasonal to decadal timescales (Regonda, 2006). These include the El-Niño Southern Oscillation (ENSO), the Pacific-North American teleconnection (PNA) and the Pacific Diurnal Oscillation (PDO). By capturing the phase and strength of these teleconnections in climate indices, statistical models may be augmented with additional predictors (Garen and Pagano, 2007, Robertson and Wang, 2012), probability distributions sampled conditionally upon the auxiliary information (Perica, 1998) or statistical and dynamical

forecasts weighed and combined (e.g. Schepen et al., 2012). For example, Hamlet and Lettenmaier (1999) use climate indices of the ENSO and PDO to improve long-range streamflow forecasting in the Columbia River Basin. Elsewhere, Grantz et al. (2005) use large scale climate indices to improve streamflow forecasting on the Truckee and Carson Rivers in Nevada for two seasons ahead. However, the ENSO, PNA and PDO are not uniformly strong, and impacts on streamflow are generally weaker and more variable in the interior west (Cayan, 1996; Regonda, 2006). Alongside climate indices, and in areas where a significant fraction of the annual precipitation falls as snow, measures of SWE are also used in ensemble forecasting and statistical models of water supply. For example, Clark et al. (2001) use a combination of large-scale climate indices and information on SWE to improve long-range streamflow forecasts in the Columbia and Colorado River Basins. In alpine regions, snow accumulation and melting is driven by air temperature, as well as precipitation, and both are important in long-range forecasting. For example, in a study of streamflow responses to climate change in the Colorado Basin, Nash and Gleick (1991) found that increases in temperature of 2-4°C would reduce the mean annual runoff by 4-20%, while changes in precipitation of 10-20% would alter the mean annual runoff by 10-20%.

In practice, both ESP and statistical models are imperfect tools for long-range forecasting. For example, regression models use observations of SWE that do not include the entire period of snow accumulation. Also, they may rely on teleconnection patterns that have limited explanatory power in some regions or invoke assumptions of stationarity that are complicated by intra- and inter-annual climate variability. Similarly, ESP relies on historical forcing and does not incorporate the latest information from global climate models, such as the CFSv2. In this context, Yuan et al. (2013) found that seasonal hydroclimate forecasts with the CFSv2 significantly improved upon ESP for many locations in the CONUS, but these improvements generally only materialized after streamflow post-processing and were strongly dependent on the variables, seasons and regions considered. In the absence of streamflow post-processing, ESP relies on hydrologic models that are well-calibrated or climatologically unbiased (Shi et al., 2008). In practice, ESP forecasts may comprise a range of unconditional and conditional

biases that could be addressed through statistical post-processing (Wood and Schaake, 2008; Shi et al., 2008). Some of these weaknesses are addressed by the HEFSv1, while others may be addressed in future. First, the HEFS uses an objective combination of ensemble forecasting and statistical modelling. Second, the HEFS uses raw forcing information from the GEFS and CFSv2, among others. While the GEFS forecasts are limited to 1-15 days, the skill from these forecasts may persist in streamflow for longer periods, depending on basin characteristics. Finally, the forcing and streamflow forecasts are corrected for biases, including biases in the meteorological forcing (MEFP) and in the hydrologic modelling (EnsPost). However, the MEFP does not include auxiliary information from the CPC's climate outlooks or large-scale climate indices, which may be useful in some basins.

Whether using statistical models, physically-based models or some combination of the two, hydrometeorological and 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 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; Demargne et al., 2010). Examples of hindcasting and verification for the long-range include Franz et al. (2003), Pagano et al. (2004), Bradley et al. (2004), Regonda (2006), Schepen et al. (2012) and Robertson and Wang (2013). In ensemble forecasting, biases produce systematic differences between the forecast probabilities of particular events and the corresponding observed outcomes [0,1] over a large sample of historical data (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 (repeated for all 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 and consistently forecasts its non-occurrence with a probability lower than chance.

In this report, hindcasts of temperature, precipitation and streamflow are generated with the HEFSv1 for selected river basins in the North East RFC (NERFC) and the Middle-Atlantic RFC (RFC). 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 raw streamflow forecasts, which do not include statistical post-processing. In order to distinguish between the meteorological and hydrologic uncertainties, the raw streamflow forecasts are verified against simulated streamflows, as well as observed streamflows. 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) and the raw streamflow forecasts (Section 5.2). Finally, the Discussion and Conclusions (Section 6) lead to guidance on the expected performance and limitations of the HEFSv1 for long-range forecasting, together with recommendations on future enhancements.

4. Materials and methods

4.1 Study basins

Eight river basins were considered in this study, of which four are located in MARFC and four in NERFC. Figure 1 and Table 1 show the latitude and longitude, drainage area and mean elevation of each basin, together with the nearest GEFS and CFSv2 grid nodes. Table 1 also shows the annual precipitation, the runoff coefficient (runoff/precipitation) and the ratio of precipitation to potential evaporation. The drainage areas range from 240 square kilometers (NVXN6) to 9013 square kilometers (MTGN4) and the runoff coefficients vary from 0.35 (PTVN6) to 0.88 (MTRN6). Figure 2a and Figure 2b show the daily means of temperature, precipitation and runoff for each basin in MARFC and NERFC, respectively. The averages are shown for each calendar month and were derived from gauged temperature, precipitation, and streamflow over a 15 year period between 1985 and 1999 (see Section 4.3). Nominally, two seasons are

identified for each RFC, namely a “wet” season and a “dry” season (Figure 2a/b). The forcing and streamflow hindcasts are verified separately for each of these seasons, as well as for the overall period (Section 5).

The eight river basins have similar climate and runoff characteristics (Figure 2a/b), with slightly higher precipitation and much higher temperatures during the summer months. The runoff peaks in April, when snowmelt from the Catskill Mountains contributes to higher streamflow in the Catskill and Delaware basins. Runoff is much lower between June and October, as flows are diverted for irrigation and water supply. Both forecast groups are subject to extensive river regulations and diversions. The Catskill and Delaware systems account for ~90% of the municipal water supply to New York City (NYC), with approximately 1.75×10^9 m^3 of water stored in six reservoirs (Figure 1). The Catskill Basin comprises the Schoharie and Ashoken Reservoirs, which drain the eastern portion of the Catskill Mountains. The Delaware Basin comprises the Cannonsville, Pepacton (Downsville), Neversink and Rondout Reservoirs, which drain the western portion of the Catskill Mountains. Water from the Catskill Basin is stored in the Ashoken Reservoir and diverted via the Catskill Aqueduct to NYC. Water from the Delaware Basin is stored in the Rondout Reservoir and distributed via the Delaware Aqueduct to NYC.

Figure 3 shows the topology of the eight river basins, together with the surrounding basins for which streamflow hindcasting was conducted. The four basins in MARFC comprise three locations on the Delaware River, namely Walton (WALN6), Callicoon (CCRN6) and Montague (MTGN4), and one location on the Neversink River, namely the Neversink Reservoir (NVXN6). The four basins in NERFC comprise two locations on the Esopus Creek, namely Mount Trempor (MTRN6) and Mount Marion (MRNN6), and two locations on the Schoharie Creek, namely Prattsville (PTVN6) and the Gilboa Dam (GILN6). Mount Trempor and Mount Marion are separated by the Ashoken Reservoir (ASEN6) and the Schoharie Reservoir lies between Prattsville and the Gilboa Dam.

In MARFC, flows are diverted from the Cannonsville Reservoir (CNNN6) and the Downsville Reservoir (DWNN6) to the NYC municipal water supply. The remaining flows, except for conservation releases and spills, are impounded for subsequent release in the lower Delaware Basin under dry conditions. Information about these diversions and releases is provided to MARFC in near real-time for operational forecasting. However, records of the individual diversions and releases were not available for hindcasting. As the diversions remove a significant fraction of the total flows at downstream locations, the hindcast (total) flows could not be compared with USGS gaged flows at basins downstream of CNNN6 and DWNN6 in MARFC. Instead, estimated natural flows were provided by NYCDEP, which adjust the gaged flows to account for the overall effects of diversions and other regulations. The estimated flows generally correspond to the local contributions at downstream locations. For example, at CCRN6, the estimated flows provided by NYCDEP correspond to the local contribution between HLEN6 and FSHN6 upstream and CCRN6 downstream. However, MTGN4 is modeled differently by MARFC than NYCDEP. Specifically, the local areas of MTGN4 and BRGN6 in MARFC are equivalent to those of Montague, Oakland Valley and Woodbourne in NYCDEP. Thus, in order to verify the streamflow forecasts at MTGN4, the observed flows from Oakland Valley and Woodbourne were routed to MTGN4 and added to the local contribution from Montague. In summary, the (total) hindcast flows were verified against USGS gaged flows at WALN6 and NVXN6 (see Section 4.3 for data sources), while estimated local flows were used to verify the (local) hindcast flows for CCRN6 and MTGN4.

Figure 4 provides a schematic of the flow pathways and regulations associated with the Ashoken Reservoir. Flows are diverted from the Ashokan Reservoir to NYC and from several upstream locations for local irrigation and water supply. These diversions remove a significant fraction of the total flows at Mount Marion (MRNN6). Estimated flows were provided by NYCDEP for the NERFC river basins. At MRNN6, the estimated flows comprise the local contribution to MRNN6 only, without any spillage or waste channel flows from the Ashoken Reservoir (Figure 4). As indicated in Figure 4, flows are routed through the Schandaken Tunnel (STUN6) from the Schoharie Reservoir (GILN6) to the Esopus Creek upstream of Mount Trempor (MTRN6) and then

to the Ashoken Reservoir. The Ashoken Reservoir comprises two storage basins, namely the East Ashoken (ASEN6) and the West Ashoken (ASWN6), which are separated by a concrete dividing weir and roadway.

At MTRN6, the total streamflow comprises the local contribution at MTRN6 plus the diverted flows from the Schoharie Reservoir. An archive of these diverted flows was provided by the USGS for the entrance to the Schandaken Tunnel. This was subtracted from the USGS gaged flow at MTRN6 for comparison with the streamflow hindcasts, which only comprise the local contribution at MTRN6. The diverted flow was *not* routed to MTRN6 before being subtracted from the USGS gaged flow. Thus, a small timing error should be expected in the estimated streamflows at MTRN6. In practice, this timing error is unlikely to be significant as the verification focuses on aggregated time-scales of 5 days or more (see Section 4.4). Finally, the hydrologic models were calibrated against the observed flows at MTRN6 without accounting for diversions. Thus, the model parameters and associated forecasts at MTRN6 may show reduced skill.

At GILN6, the hindcasts comprise inflows to the Schoharie Reservoir, which are not impacted by the diversions to MTRN6. The USGS gage at the Schoharie Reservoir is located in the reservoir pool, rather than the inflow. Thus, the NYCDEP estimated inflows were compared to the corresponding forecast inflows. 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.

In summary, the (total) hindcast flows were verified against the USGS gaged flows at PTVN6, whereas the hindcast flows at MTRN6, MRNN6 and GILN6 were verified against estimated flows provided by NYCDEP.

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,

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

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

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

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

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

The raw forcing may comprise the ensemble mean of NCEP's GEFS 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 Post-processor (EnsPost; Seo et al., 2006). In all cases, the parameters of future quantities are estimated from subsets of the historical data, for which a degree of stationarity is assumed. Here, the parameters of

the HEFS were estimated from the same historical period (1985-1999) 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 15 year period between 1985 and 1999. The hindcasts of MAP and MAT were produced at 12Z every 5 days. Each forecast comprised ~50 ensemble members, with lead times varying from 6 to 7,920 hours in six-hourly increments. Inputs to the MEFP comprised “raw” precipitation and temperature hindcasts from NCEP’s Global Ensemble Forecast System (GEFS; Hamill et al., 2013) and the Climate Forecast System Version 2.0 (CFSv2). For the period 1-15 days, the MEFP was calibrated with the ensemble mean of the GEFS hindcasts. For the period 16-270 days, the MEFP was calibrated with the single-valued forecasts from the CFSv2. As the CFSv2 forecasts were initialized only once every 5 days, the HEFS forcing and streamflow hindcasts were also produced at this frequency (i.e. 6-hourly forecasts with a T0 every 5 days). For the period 271-330 days, a “resampled climatology” was derived from the historical observations of MAP and MAT. Specifically, the MAP and MAT were resampled in a moving window of, respectively, 30 days and 15 days either side of the forecast valid date. A smooth probability distribution was then fitted to the resampled observations and ensemble members were derived from the fitted distribution. The MEFP forecasts with combined inputs from the GEFS, CFSv2 and resampled climatology are denoted MEFP-GCC. Resampled climatology was also generated for the period 1-330 days, in order to evaluate the skill of the MEFP-GCC forecasts. The resampled climatology forecasts are denoted MEFP-CLIM.

Raw streamflow hindcasts were generated with the HEFS 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 NERFC 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 SNOW17 model is used together with an empirical hydrologic model, based on the Antecedent Precipitation Index (API), but adapted for continuous simulations (the so-called "Continuous API" model). Where applicable, routing 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 several RFCs, an ADJUST-Q operation is used to blend the most recent observed streamflow into the operational forecast, although hydrologic persistence is generally limited for long-range forecasting. In the HEFS, ADJUST-Q is (largely) replicated by the EnsPost, which corrects for biases in the raw streamflow forecasts conditionally upon the prior observed flow, as well as the contemporary simulated flow (see Seo et al., 2006). However, following a preliminary application of the EnsPost for long-range forecasting, some enhancements were deemed necessary. Thus, only the raw streamflow forecasts are considered in this study. Nevertheless, in order to separate the meteorological uncertainties and biases (addressed by the MEFP) from the hydrologic uncertainties and biases (otherwise addressed by the EnsPost), the streamflow forecasts were verified against simulated streamflow as well as observed streamflow (see Section 4.4).

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 1949 and 1999. Streamflow observations were obtained from the United States Geological Survey (USGS) for the period 1985-1999. 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. However, the USGS gaged flows were only used to verify the streamflow forecasts at

three headwater locations, namely WALN6 and NVXN6 in MARFC and PTVN6 in NERFC. All other locations (including MTRN6, which receives diverted flows from the Schoharie Reservoir) were effectively treated as headwaters. Specifically, the forecast local flows were verified against estimated local flows provided by NYCDEP.

In practice, the estimated flows provided by NYCDEP are known to be imperfect. For example, reservoir inflows are estimated from gaged 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 flows, which are assigned zero if the inflow estimates are negative. In other cases, the contributing areas defined by NYCDEP differ from those used by the RFC and observed flows are estimated by routing and summing contributions from multiple sub-basins (e.g. MTGN4).

There are several challenges for applying the HEFS consistently in regulated rivers; that is, to maintain consistency between calibration and operational use and between hindcasting and operational use. Consistency between hindcasting and operational use is necessary to evaluate the HEFS and provide measures of forecast quality that can guide operational applications. Consistency between calibration and operational use is necessary to train the EnsPost on hydrologic biases and uncertainties that represent the operational reality. These issues are currently being explored, and recommendations developed, as part of a Concept Of Operations (CONOPS) for the HEFS. Elsewhere, Georgakakos et al. (2010) describe a methodology for accommodating river regulations in operational ESP.

4.4 Verification strategy

Verification was conducted with the 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. While limited combinations of these attributes were also considered, they were often constrained by the sampling uncertainties of the verification metrics. The sampling uncertainties were not explicitly quantified here (see Brown and Seo, 2013 for an example). However, the

verification results were only computed for samples of 30 or more verification pairs (the smaller of the number of occurrences and non-occurrences for discrete metrics).

In pairing the meteorological forecasts and observations, 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 hour in both MARFC and NERFC (UTC-5). As the forecasts were verified at an aggregated support of five days or larger (see below), this timing error was considered unimportant. Pairing of the observed and forecast streamflows was complicated by the daily frequency of the verifying observations and estimates. Specifically, the observations comprise daily mean flows from 5Z-5Z. Thus, in pairing the streamflow forecasts and observations, it was assumed that the observed streamflows adequately represent the period 6Z-6Z. The first three forecasts, which comprise valid times of 18Z, 0Z and 6Z (representing the period 12Z-6Z), were then ignored. As such, the first verification pair comprises the observed streamflow from 5Z-5Z and the average of the 6-hourly forecasts from 12Z, 18Z, 0Z, and 6Z with forecast lead times of 24, 30, 36 and 42 hours, respectively (nominally labelled 42 hours). For consistency, the first three forecasts were also dropped when pairing against the simulated flows.

While pairing was conducted for daily averages of temperature and runoff, and for accumulated precipitation, the verification was conducted for aggregated periods of five days or more, as: 1) this study focuses on the long-range forecasts, for which most practical applications benefit from aggregated quantities (i.e. daily averages have little skill for the long-range); and 2) the hindcasts were initialized only once every five days, which introduced an artificial cyclicity into the paired sample when verifying at a daily scale. The latter is illustrated in Table 2, where the indices of verifying observations are shown for a selection of forecast initialization times (T0) and lead times. As evidenced by the shading in Table 2, the composition of the observed sample varies systematically with forecast lead time (i.e. every five days) when verifying at a daily timescale. By aggregating the forecasts and observations into periods of five or more days, this sampling artifact was avoided.

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 are anticipated (with varying sensitivities to streamflow amount). However, "average conditions" generally imply 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 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 and biases (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 n observations collected between 1985 and 1999. Real-valued thresholds were then determined for $k \approx 100$ climatological exceedence probabilities, c_p, $\hat{F}^{-1}_{n,v,a,b}(c_p)$, where $c_p \in [0,1]$ and $p = 1,\ldots,k$. Verification measures that depend continuously on the data, 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.

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 5 shows the correlations of the ensemble mean forecast and observed precipitation amounts by forecast lead time. The results are shown for the raw forcing from the GEFS and CFSv2 for the period 1-270 days, together with the bias-corrected forcing from the MEFP-CLIM and MEFP-GCC for the period 1-330 days. The correlations between the ensemble mean of the MEFP-GCC precipitation forecasts and the corresponding observed precipitation generally exceeds 0.6 when averaged over the first 5-day period, but decline rapidly thereafter. Beyond 5-10 days, the correlations approach the background signal of 0.1-0.2 associated with resampled climatology. At some locations, such as GILN6 and PTVN6 in NERFC, the MEFP-GCC forecasts show lower correlations than the MEFP-CLIM forecasts between 10 and 15 days, but recover after 15 days. During the period of CFSv2 forcing (16-270) days, the ensemble mean of the MEFP precipitation forecasts is no more correlated with the observed precipitation amount than resampled climatology. However, the MEFP maintains or improves upon the correlations between the raw forcing from the GEFS and CFSv2 and the corresponding observed precipitation amounts.

Both the MEFP-GCC and MEFP-CLIM forecasts show cyclic variations in the correlation coefficient, with a cycle of ~30 days. This originates from the use of so-called “canonical events” in the MEFP, whereby predictors are formed from different aggregation periods for each forecast valid time (Appendix A). These canonical events are grouped into 30-day periods within the forecast horizon (or multiples thereof), with separate events applying to days 1-30, 31-60 etc. The observed cyclicity in some verification statistics may be an artifact of calibrating the MEFP with limited sample data, as the precipitation climatology should otherwise vary smoothly during the forecast horizon.

Figure 6 shows the relative mean error (RME) of the MEFP-CLIM and MEFP-GCC precipitation forecasts by increasing forecast lead time. On average, the ensemble mean of the MEFP underestimates the observed precipitation amount by ~5% for both sources of raw forcing, in all basins, and at most forecast lead times. However, in

absolute terms, these biases amount to less than 1mm accumulation over 5 days. As indicated in Figure 6, the MEFP-GCC precipitation forecasts show a slight discontinuity in the RME at 271 days, where the raw forecasts transition from CFSv2 to resampled climatology and the underforecasting bias increases slightly. This artifact is not visible in the mean error of the ensemble mean forecast (results not shown) and reflects the greater sensitivity of the RME to small changes in mean error under typical (i.e. dry) conditions.

Figure 7 shows the mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-GCC and MEFP-CLIM precipitation forecasts against sample climatology. Sample climatology comprises the unconditional probability distribution of precipitation between 1985 and 1999. During the first ~5 days of the forecast horizon, the MEFP-GCC precipitation forecasts are 20-25% more skillful than sample climatology. This originates from the skill of the raw GEFS forecasts during the first week. In contrast, the forecasts are only marginally more skillful than sample climatology after ~5 days. This originates from the (lack of) skill in the raw forcing beyond the first week. As climatological forcing is currently used by the RFCs for ESP, it is unlikely that the MEFP will improve upon the existing long-range temperature and precipitation forecasts at these locations.

Figure 8 shows the mean CRPSS of the MEFP-GCC and MEFP-CLIM temperature forecasts against sample climatology. On average, the MEFP-GCC temperature forecasts are ~80-90% more skillful than sample climatology at a forecast lead time of 1-5 days across all basins and ~65-70% more skillful than sample climatology after ~15 days. Unlike the precipitation forecasts, the raw GEFS forecasts and hence the MEFP-GEFS forecasts remain skillful after 5 days when compared to the MEFP-CLIM forecasts. This stems from the relative predictability and temporal autocorrelation of temperature versus precipitation. However, during the period of CFSv2 forcing, the MEFP-GCC forecasts are only slightly (<5%) more skillful than the MEFP-CLIM forecasts. In practice, the similarity between the MEFP-GCC and MEFP-CLIM forecasts is more informative than the absolute skill, as the latter depends on the inherent predictability of the forecast variable. Sample climatology, as defined here,

does not account for seasonal variations in temperature, so the absolute skill of the MEFP-GCC and MEFP-CLIM forecasts is high. However, the MEFP-GCC temperature forecasts hardly improve on the MEFP-CLIM forecasts.

5.1.2 Magnitude of the forcing variable

Figure 9a shows the Brier Skill Score (BSS) for the MEFP-CLIM and MEFP-GCC precipitation forecasts in MARFC and NERFC with sample climatology as the baseline. The results are shown for a 5-day precipitation total with a forecast lead time of 95-100 days. The precise lead time is not important, as the precipitation forecasts are similar to climatology after ~1 week. The BSS is plotted for multiple precipitation thresholds, which are expressed in terms of their climatological probability of exceedence. The threshold values are plotted on a non-linear (probit) scale, but are labelled with actual probability. For example, 0.1 denotes the 5-day precipitation total that is exceeded, on average, only once in every 50 days (i.e. 10 periods of 5 days). The origin of each curve denotes the BSS for the Probability of Precipitation (PoP) forecast.

As indicated in Figure 9a, the MEFP-GCC and MEFP-CLIM forecasts are equally unskillful. Indeed, the MEFP forecasts of PoP and light precipitation are generally worse than sample climatology, particularly at WALN6 in MARFC. Figure 9b shows the "calibration-refinement" factorization of the BSS (see Appendix B), which comprises the relative reliability (Type-I conditional bias) and relative resolution of the MEFP forecasts. As indicated in Figure 9b, the lack of skill in the MEFP forecasts of PoP and light precipitation originates from a conditional bias in the forecast probabilities. The largest bias occurs in WALN6 where the forecast PoP systematically underestimates the observed PoP. This is consistent with earlier studies of the MEFP that focused on the medium-range forecasts, where an underforecasting bias was identified for PoP (Brown, 2013). For moderate and large precipitation thresholds, the MEFP forecasts are slightly more skillful than sample climatology (Figure 9a).

Figure 10 shows the BSS for the MEFP-CLIM and MEFP-GCC temperature forecasts in MARFC and NERFC with sample climatology as the baseline. The results are shown for a 5-day mean temperature with a forecast lead time of 95-100 days. As in

Figure 8, the BSS is plotted for multiple thresholds, which are expressed in terms of their climatological probabilities of exceedence. In keeping with the precipitation forecasts, the MEFP temperature forecasts are no more skillful when using GCC forcing than resampled climatology. However, unlike the precipitation forecasts, they are substantially more skillful than sample climatology. The BSS is highest around the median temperature and declines for lower and higher temperatures. Again, this originates from the use of resampled climatology (rather than sample climatology) as the unconditional distribution in the MEFP. As indicated in Figure 2a/b, there are much stronger seasonal variations in temperature than precipitation in MARFC and NERFC. This leads to a more pronounced increase in BSS from using a conditional climatology for temperature than precipitation.

Figure 11 shows box plots of errors in the MEFP-GCC precipitation forecasts for both MARFC and NERFC. Each box represents one ensemble forecast of the 5-day precipitation total at a forecast lead time of 95-100 days. 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 11, the MEFP-GCC forecasts consistently underestimate the highest precipitation totals, for which there is a strong conditional bias in the ensemble median. The conditional biases are similar at all forecast locations and reflect the lack of skill in the CFSv2 forecasts at 95-100 days. Indeed, the MEFP-GCC forecasts closely resemble the MEFP-CLIM forecasts during the period of CFSv2 forcing. By definition, resampled climatology is conditionally biased with respect to precipitation amount, as the sampling is conditional upon forecast valid time only, not on precipitation amount. In most cases, there is sufficient spread in the MEFP-GCC forecasts to predict some probability of the highest precipitation totals that subsequently occur. In principle, when issuing a climatological probability forecast, the verifying observation should fall on the climatological quantile with the same relative frequency as the corresponding probability implies. However, for the most extreme precipitation amounts, these climatological quantiles will be sensitive to the period of record and the resampling window used by the MEFP-CLIM. Since the MEFP-GCC and MEFP-CLIM

forecasts are based on dependent validation, the ensemble spread should be treated as optimistic for the most extreme observed precipitation amounts. In practice, the ensemble spread may not capture novel conditions in operational forecasting, as this novelty is, by definition, absent from the historical record from which the MEFP is calibrated.

Figure 12 shows box plots of errors in the MEFP-GCC temperature forecasts for each basin in MARFC and NERFC. Each box represents one ensemble forecast of the 5-day average temperature at a forecast lead time of 95-100 days. The boxes are arranged by increasing observed temperature.

According to Figure 12, the median of the MEFP-GCC forecasts is unbiased for most observed temperatures, but the lowest and highest observed temperatures are generally over- and under-forecast, respectively. Again, this is understandable because the MEFP-GCC forecasts are no more skillful than the MEFP-CLIM forecasts and the MEFP-CLIM forecasts are conditionally biased, by definition. However, in keeping with the relative predictability of temperature, the conditional biases are much smaller for temperature than precipitation and the ensemble spread is generally sufficient to capture the highest and lowest observed temperatures that subsequently occur (except for the most extreme events). In general, the MEFP-GCC forecasts show the greatest conditional biases for the coldest observed temperatures, where the ensemble median is frequently 50% higher than the observed temperature. For hydrologic forecasting, the transition between freezing and above-freezing temperatures is important in determining the fraction of precipitation that falls as snow versus rain and for predicting snowmelt. Reliability diagrams (not shown) indicate that the MEFP-GCC temperature forecasts are highly reliable at predicting the probability of exceeding 0^{o}C.

5.1.3 Season

Figure 13 shows the monthly climatologies of the observed and forecast precipitation amounts at a 5-day accumulation volume. The ensemble mean was computed for each forecast whose valid time occurred within a given calendar month. Figure 13 shows the overall mean of those values. The conditional mean was then

determined separately for each forecast lead time. The range of these conditional mean values is also plotted in Figure 13; it shows the variability of the average forecast amount within a given month across all forecast lead times. The average observed precipitation is also shown for paired forecasts whose valid time occurred within a particular month. As indicated in Figure 13, the observed values are generally captured by the range of forecast values.

Figure 14a and Figure 14b show the mean CRPSS of the MEFP-GCC and MEFP-CLIM precipitation forecasts against sample climatology for MARFC and NERFC, respectively. The results are shown for a 5-day precipitation total with a forecast lead time of 95-100 days and comprise the overall results, together with the "wet" and "dry" seasons separately (see Figure 2a/b for definitions of the seasons). During the first week, the MEFP forecasts are slightly more skillful in the wet season than the dry season, both in MARFC and NERFC. However, the precipitation climatology is relatively constant in all basins (Figure 2a/b) and the absolute skill is low in both seasons. As indicated in Figure 14a/b, the cyclic variations in CRPSS are slightly more pronounced during the wet season than the dry season. This may reflect the reduced scope for parameter variability in dry conditions.

5.1.4 Aggregation period

Figure 15a and Figure 15b show the RME, correlation coefficient and CRPSS (versus sample climatology) of the MEFP-GCC precipitation forecasts for basins in MARFC and NERFC, respectively. The results are plotted by forecast lead time for three aggregation periods, namely 5-, 10- and 30- days.

In general, the relative bias of the ensemble mean forecast is insensitive to aggregation period, with a slight underestimation of the observed precipitation amount at all aggregation periods. The transition in forcing between the CFSv2 (15-270 days) and resampled climatology (271-330 days) is also apparent in the RME at all aggregation periods.

Unlike the RME, the correlation coefficient and the CRPSS are somewhat sensitive to aggregation period, although the differences are small in absolute terms (note the range axis dimensions). In general, the correlation coefficient increases within increasing aggregation period. This is understandable, as random errors are effectively smoothed by averaging. However, the precise sensitivities of the correlation coefficient to aggregation period vary with forecast location. The greatest sensitivities are observed in WALN6, PTVN6 and GILN6 and the smallest sensitivities are found in MTRN6 and MTGN4. These variations are not easily explained by basin size or elevation. In general, the CRPSS shows smaller sensitivities to aggregation period than the correlation, although the greatest sensitivities are found in WALN6, PTVN6 and GILN6 where the correlations are also more sensitive to aggregation period. This is understandable, as the CRPSS reflects a combination of first-order bias and correlation, as well as higher-order effects that contribute to the integral error measured by the CRPS.

The MEFP-CLIM forecasts show similar sensitivities to aggregation period as the MEFP-GCC forecasts (not shown). Indeed, the MEFP-GCC precipitation forecasts are no more skillful than the MEFP-CLIM forecasts at any of the aggregated scales considered here.

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 magnitude of streamflow.

5.2.1 Forecast lead time

Figure 16 shows the RME of the streamflow forecasts with forcing inputs from the MEFP. The results are shown by forecast lead time for each basin in both RFCs and for both sources of forcing (MEFP-CLIM and MEFP-GCC). The streamflow forecasts are verified against simulated flows (S) as well as observed flows (O).

When verified against simulated flows, errors in the streamflow forecasts originate from errors in the MEFP forcing and in the observed forcing from which the

simulations are produced (these are assumed to be negligible). When verified against observed streamflow, the verification results also include errors from the hydrologic modeling. Clearly, however, errors in the hydrologic modeling may offset, as well as accentuate, those in the meteorological forcing. Differences between the observed and simulated flows originate from errors in the hydrologic modeling or from the observed flows, which include biases in the stage-discharge relation or in the assumptions surrounding inflow estimates. At most forecast locations, the streamflows were estimated rather than observed. Estimated flows involve assumptions about unknown quantities, such as reservoir inflows, conditionally upon gaged quantities, such as reservoir outflows and storage levels. For example, evaporation is not considered in the mass balance relation from which inflows to the NYCDEP reservoirs are derived. While these errors may be small relative to the forecasting errors, some caution is needed in attributing errors and biases to any one source.

As indicated in Figure 16, the biases in the streamflow forecasts vary with basin, forecast lead time, forcing source, and with the source of verifying observations. For example, when verifying against observed streamflows, the relative biases are consistently small in WALN6, but are relatively large in GILN6. In general the relative biases are smaller and more variable in the MARFC basins. In the NERFC basins, the forecasts consistently underestimate the verifying observations and simulations, except at the earliest forecast lead times.

In Figure 6, the precipitation forecasts show a small underforecasting bias, which generally increases with increasing forecast lead time. The streamflow forecasts show a similar dependence on forecast lead time (Figure 16), with smaller biases at earlier lead times (for those basins with an underforecasting bias). Also, as indicated above, the precipitation forecasts, particularly the MEFP-CLIM forecasts, show a cyclicity in the RME and other verification statistics. This originates from the parameterization of the MEFP with canonical events, which attempt to capture the skill in the raw forcing forecasts at multiple temporal scales. In Figure 16, the MEFP-CLIM streamflow forecasts also show a cyclicity in the RME (Figure 16). While these variations in RME are small in absolute terms, they do *not* reflect any fundamental cyclicity in precipitation.

Rather, they originate from the parameterization of the MEFP and the estimation of those parameters with limited sample data.

In general, the RME is similar whether verifying against observed or simulated flows, indicating that the hydrologic biases are relatively small. However, there are larger differences in some basins. For example, there is an over-forecasting bias of ~10% at CCRN6, which originates from the hydrologic modeling rather than the forcing (Figure 16). In contrast, there is an underforecasting bias of ~20% at GILN6, which originates from a combination of the MEFP forcing and the hydrologic modeling; specifically, during the late spring and summer, where the streamflow forecasts consistently underestimate the observations (see Section 5.2.3).

Figure 17 shows the correlations between the streamflow forecasts with MEFP-CLIM and MEFP-GCC forcing against the corresponding observed and simulated streamflows. For the streamflow forecasts with MEFP-GCC forcing, the correlations are greatest during the first five days, particularly when verifying against simulated flows. This originates from the skill of the raw GEFS forecasts during the first week. The correlations then decline gradually over the long-range in most basins, as the initial conditions are progressively diluted and the forecasts approach the baseline skill of the long-range forcing (i.e. climatology). When using a common source of verifying observations, the forecasts with MEFP-GCC forcing generally show higher correlations than those with MEFP-CLIM forcing. Some persistence of skill from the GEFS should be expected beyond the medium-range, depending on basin characteristics, but these differences are small and are also impacted by sampling uncertainties.

Notwithstanding errors in the streamflow observations and simulations, the impacts of the hydrologic uncertainties are greatest in MTRN6, MRNN6, GILN6 and, at early forecast lead times, in MTGN4. This is evidenced by lower correlations when verifying against observed streamflows than simulated streamflows (noting that correlation is insensitive to bias). In MRNN6 and GILN6, these differences are substantial and originate from a climatological bias during the late spring and early

summer, where the spring snowmelt is forecast to decline much more rapidly than observed (see Section 5.2.3).

Figure 18 shows the mean CRPSS of the streamflow forecasts with MEFP-GCC forcing relative to those with MEFP-CLIM forcing. When verifying against simulated streamflows, the CRPSS reflects the contribution of the MEFP-GCC forcing *without* any hydrologic uncertainties and biases. The streamflow forecasts show the greatest skill during the first week, where the MEFP benefits from the GEFS forcing. The skill then declines progressively with increasing forecast lead time as the quality of the raw forcing from the GEFS and the CFSv2 declines. In keeping with the slightly higher correlations of the GCC streamflow forecasts (Figure 17), the CRPSS is generally positive between ~10-100 days. However, given the sampling uncertainties and the small magnitude of the CRPSS, this should not be over-emphasized. Moreover, it does *not* originate from the CFSv2 component of the MEFP, as the precipitation forecasts were shown to be unskillful beyond ~1 week (Figure 14a/b). Rather, it stems from hydrologic persistence; that is, persistence of the skill from the GEFS forecasts beyond one week, depending on basin characteristics.

In principle, the hydrologic uncertainties and biases are independent of the meteorological uncertainties and biases and do not, therefore, depend on forecast lead time (assuming a stationary streamflow climatology). However, the relative contributions of the meteorological and hydrologic uncertainties do vary with forecast lead time (Figure 18). At early forecast lead times (but depending on basin conditions), much of the skill in the HEFS originates from the initial conditions in the hydrologic models. As the meteorological forecasts propagate through the hydrologic models, and the initial states are updated, the meteorological uncertainties become important. Thus, the potential for streamflow post-processing varies with forecast lead time. During the first 5 days, the CRPSS is substantially higher when verifying the GCC streamflow forecasts against the simulated flows than the observed flows. As such, the EnsPost may contribute valuable skill in the first ~5 days (see Brown, 2013 also). During the long-range, the scope for streamflow post-processing is reduced. Here, the background skill is inherently low and hydrologic persistence is also diminished. Under these conditions,

the benefits of post-processing will depend largely on the climatological biases in the streamflow forecasts. Since the hydrologic models are reasonably well-calibrated, at least for moderate and high flows, there is less scope for post-processing. However, the EnsPost may contribute valuable skill at low flows, where hydrologic persistence is generally stronger and the hydrologic biases are greater (see Section 5.2.3). In this context, the CRPS is a measure of integral error and, therefore, dominated by moderate and high flows. Also, Figure 18 does not show the marginal skill from streamflow post-processing or approximate the specific contribution from the EnsPost (which also benefits from prior observed flows); it shows the expected contribution from the MEFP-GCC in the absence of any hydrologic uncertainties and biases. Thus, further work is needed to establish the benefits of the EnsPost for long-range forecasting under varied conditions.

The impacts of the hydrologic uncertainties and biases are greatest in MTGN4, where the CRPSS is much stronger at early forecast lead times when verifying against simulated flows. As indicated above, the local flows at Montague are modeled differently by MARFC than NYCDEP. Comparable flows were derived by routing the observed flows from Woodbourne and Oakland Valley to Montague and adding this upstream contribution to the local flows at Montague. Alongside the routing of observed flows, the hydrologic models were not explicitly calibrated with the streamflows from Oakland Valley or Woodbourne, so some residual hydrologic uncertainties and biases may be expected.

5.2.2 Magnitude of streamflow

Figure 19 shows the RME of the streamflow forecasts with forcing inputs from the MEFP-GCC and MEFP-CLIM when verified against the observed (O) and simulated (S) streamflows. The results are shown for a 5-day streamflow rate with a forecast lead time of 95-100 days. The precise lead time is not important, as the streamflow forecasts are similar to climatology beyond the medium-range. The RME is plotted for multiple streamflow thresholds and each threshold is expressed in terms of its climatological probability of exceedence. The threshold values are plotted on a non-linear (probit)

scale, but are labelled with actual probability. For example, 0.1 denotes the 5-day streamflow rate that is exceeded, on average, only once in every 50 days (i.e. 10 periods of 5 days). The origin of each curve denotes the unconditional bias in the streamflow forecasts; that is, the RME of all verification pairs where the observed streamflow exceeds the lowest historical observation.

In MARFC, the streamflow forecasts contain relatively small unconditional biases during the period 95-100 days, with a slight underforecasting bias at NVXN6 and a slight overforecasting bias at WALN6, CCRN6 and MTGN4 (versus the observed streamflow). In contrast, the streamflow forecasts in NERFC consistently underestimate the observed streamflow, both unconditionally and conditionally; that is, with increasing streamflow amount. While the unconditional biases are larger in NERFC than MARFC, the conditional biases are large in both RFCs. Indeed, the highest streamflow amounts are underestimated by up to 80%, on average. These conditional biases are further illustrated in Figure 20, which shows box plots of errors in the MEFP-GCC streamflow forecasts. Each box represents one ensemble forecast from the period 95-100 days. Selected quantiles of the forecasting errors are plotted together with the median error and range (extreme residuals) as whiskers. The boxes are arranged by increasing amounts of observed streamflow. As indicated in Figure 20, the MEFP-GCC streamflow forecasts consistently underestimate the moderate and high observed flows. This originates from a large conditional bias in the MEFP precipitation forecasts, particularly during the long-range (see Figure 11). Indeed, the MEFP-GCC precipitation forecasts are no more skillful than the MEFP-CLIM forecasts at 95-100 days and climatological forecasts are, by definition, conditionally biased.

As indicated above, any separation between the RME for the observed and simulated flows implies that the streamflow forecasts are impacted by hydrologic biases (i.e. the separation denotes the difference between the simulated and observed flows). For those basins with a strong separation (WALN6, CCRN6, MTGN4, PTVN6 and GILN6), the difference is greater at low flows than high flows. While the hydrologic models do not target specific applications or flow conditions, the high flows receive particular scrutiny during model calibration, and the “Continuous API” model used by

MARFC (for CCRN6 and MTGN4) is less well-suited to dry conditions (Michael Thiemann, pers. comm.). At high flows, the conditional biases largely originate from the MEFP precipitation forecasts (Figure 11), and the separation between the RME for the observed and simulated flows is much smaller, i.e. the impacts of the hydrologic biases are smaller.

Figure 21 shows the correlations between the ensemble mean of the streamflow forecasts with MEFP-GCC and MEFP-CLIM forcing against the corresponding observed and simulated streamflows. The correlations are shown by increasing streamflow threshold. As above, the thresholds are labelled by their climatological probabilities of exceedence. In general, the correlations decline with increasing observed streamflow. The greatest correlations occur under dry conditions, when the HEFS forecasts benefit from hydrologic persistence. At moderate to high streamflow rates, the GCC forecasts are slightly more correlated with the observed and simulated flows than the CLIM forecasts. However, as indicated above, this originates from the persistence of the GEFS forecasts, rather than any meaningful skill from the CFSv2. Moreover, the differences between the GCC and CLIM forecasts are probably not beyond the range of sampling uncertainty.

The correlations between the streamflow forecasts and observations generally decline with increasing streamflow rate, but increase at moderately high flows in several basins when verified against simulated flow. While the correlations are insensitive to hydrologic bias (in the mean sense), they are sensitive to hydrologic uncertainties and higher-order biases. As indicated in Figure 21, the hydrologic uncertainties have a significant impact at high flows, with substantial differences between the correlations for the observed and simulated flows in WALN6, CCRN6, MTGN4, MTRN6, MRNN6 and GILN6.

Figure 22 shows the mean CRPSS of the MEFP-GCC streamflow forecasts against those with MEFP-CLIM forcing. The streamflow forecasts are verified against both observed and simulated flows. Verification against simulated flows shows the expected skill of the streamflow forecasts in the absence of hydrologic uncertainties and

biases. As indicated in Figure 22, the streamflow forecasts with MEFP-GCC forcing show little or no skill when compared to those with MEFP-CLIM forcing for the period 95-100 days. At high streamflow rates, any benefits accrued from the calibration of the hydrologic models are offset by the large conditional biases in the MEFP precipitation forecasts. Indeed, the higher correlations between the streamflow forecasts and simulations (Figure 21) do not translate into higher CRPSS, as the forcing biases dominate at high streamflow thresholds.

While the streamflow forecasts with MEFP-GCC forcing are no more skillful at 95-100 days than those with MEFP-CLIM forcing, both are substantially more skillful than streamflow climatology. Figure 23 shows the Relative Operating Characteristic (ROC) of the streamflow forecasts with MEFP-GCC forcing. The results are shown for a headwater and an outlet in each RFC and for several streamflow thresholds (denoted by their climatological probabilities). The ROC curves were fitted under an assumption of bivariate normality between the Probability of Detection (PoD) and the Probability of False Detection (PoFD) (Appendix B). An unskillful forecast has an equal chance of correctly detecting an occurrence as incorrectly detecting a non-occurrence. Thus, a forecasting system is more discriminatory than climatology if the POD exceeds the PoFD. As indicated in Figure 23, the streamflow forecasts with MEFP-GCC forcing clearly improve upon climatology. However, the streamflow forecasts with MEFP-CLIM forcing are equally discriminatory (not shown). Thus, while the streamflow forecasts with MEFP-GCC forcing do not improve upon those with MEFP-CLIM forcing, they do improve upon sample climatology. This is understandable, as the MEFP-CLIM samples historical observations of temperature and precipitation conditionally upon forecast valid date. Also, at early forecast lead times, the initial conditions of the hydrologic models account for a substantial fraction of the overall skill in the streamflow forecasts.

5.2.3 Season

Figure 24 shows the monthly climatologies of the observed and forecast streamflow rates at a 5-day accumulation volume. The ensemble mean was computed for each forecast whose valid time occurred within a given calendar month. Figure 24

shows the overall mean of those values. The conditional mean was then determined separately for each forecast lead time and the range of those values is also plotted in Figure 24; it shows the variability of the average across all forecast lead times. The average observed streamflow is also shown for paired forecasts whose valid time occurred within a particular month. As indicated in Figure 24, the average forecast streamflows are generally too high in the summer and, to a lesser extent, too low around the spring peak in April. These patterns are likely to originate from a conditional bias that depends on streamflow rate, rather than season directly, as the hydrologic models are calibrated to provide reliable forecasts of high flows (spring) but may overestimate low flows (summer). However, in MRNN6 and GILN6, the spring snowmelt is forecast to decline much more rapidly than observed and the simulated flows also fail to capture the observed flows (Figure 24). The latter points to weaknesses in the calibration of the hydrologic models at MRNN6 and GILN6.

Figure 25a and Figure 25b show the mean CRPSS of the streamflow forecasts with MEFP-GCC forcing against those with MEFP-CLIM forcing for MARFC and NERFC, respectively. The CRPSS is shown for the overall period, together with the "wet" and "dry" seasons separately (see Figure 2a/b for definitions of the seasons). During the first week, the streamflow forecasts are generally more skillful in the wet season than the dry season, both in MARFC and NERFC. However, in keeping with the MEFP precipitation forecasts (Figure 14a/b), the streamflow forecasts show no appreciable skill in either season beyond one week. In all cases, the streamflow forecasts show significantly higher CRPSS during the first week when verified against simulated streamflow than observed streamflow. This is indicative of the hydrologic bias and uncertainty impacting the skill from the MEFP forcing (see above also).

5.2.4 Aggregation period

Figure 26a and Figure 26b show the RME, correlation and CRPSS of the streamflow forecasts with MEFP-GCC forcing in MARFC and NERFC, respectively. The results are shown for three aggregation periods, namely 5-, 10- and 30- days. In keeping with the MEFP precipitation forecasts (Figure 15a/b), the relative bias of the

streamflow forecasts is constant with increasing aggregation period. In the NERFC basins, there is an underforecasting bias of 5-15% at all aggregation periods, whereas the relative bias is smaller and more variable in the MARFC basins.

Unlike the RME, the correlations are highly sensitive to aggregation period, as random errors are effectively canceled out by averaging. Crucially, an increase in correlation will not translate into greater skill if the baseline forecast shows a similar increase in correlation (other factors being equal). Indeed, there is no appreciable skill in the aggregated forecasts at longer timescales, either when averaging over 10- or 30-day periods (Figure 26 a/b). The higher skill during the first 30-day period should not be confused as a gain in skill from the aggregation itself. Rather, it is a plotting artifact that stems from the averaging of skill from the GEFS period across the first 30 days. The results were similar when verifying against simulated streamflows (not shown).

6. Discussion and conclusions

Ensemble forecasts of precipitation, temperature and streamflow were generated with the NWS HEFS for a 15 year period between 1985 and 1999. The hindcasts were produced for 22 locations and verification was conducted at eight locations, comprising four basins in MARFC and four in NERFC. The basins include a range of headwater and downstream locations within the Delaware and Catskill systems. They are subject to extensive river regulations, including diversions to the NYC municipal water supply. The four basins in MARFC comprise three locations on the Delaware River, namely Walton (WALN6), Callicoon (CCRN6) and Montague (MTGN4), and one location on the Neversink River, namely the Neversink Reservoir (NVXN6). The four basins in NERFC comprise two locations on the Esopus Creek, namely Mount Trempor (MTRN6) and Mount Marion (MRNN6), and two locations on the Schoharie Creek, namely Prattsville (PTVN6) and the Gilboa Dam (GILN6). Mount Trempor and Mount Marion are separated by the Ashoken Reservoir (ASEN6), while the Schoharie Reservoir lies between Prattsville and the Gilboa Dam. The HEFS hindcasts were commissioned by the NYCDEP, in order to support the initial implementation of the HEFS at MARFC and

NERFC and to improve the management of risks to water quantity and quality objectives in the NYC area.

Precipitation and temperature hindcasts were produced with the MEFP using "raw" precipitation and temperature forecasts from multiple sources. Ensemble forecasts from NCEP's Global Ensemble Forecast System (GEFS) were used for the period 1-15 days. Single-valued forecasts from the Climate Forecast System Version 2.0 (CFSv2) were used for the period 16-270 days. For the period 271-330 days, and as a reference forecast for the period 1-330 days, climatological ensembles were derived by resampling the historical MAT and MAP. Specifically, the MAP and MAT were resampled in a moving window of, respectively, 30 days and 15 days either side of the forecast valid date. A smooth probability distribution was then fitted to the resampled observations and ensemble members were derived from the fitted distribution. The GEFS, CFSv2 and resampled climatology are collectively denoted GCC, while resampled climatology is denoted CLIM.

The streamflow forecasts were produced with the Community Hydrologic Prediction System (CHPS). In NERFC, the hydrologic models comprise the Sacramento Soil Moisture Accounting model (SAC-SMA) and the Snow Accumulation and Ablation Model (SNOW-17). In MARFC, the SNOW17 model is used together with an empirical hydrologic model, based on the Antecedent Precipitation Index (API), but adapted for continuous simulations (the so-called "Continuous API" model). The precipitation, temperature and streamflow forecasts were verified with the Ensemble Verification System (Brown et al., 2010b). The forecasts were verified conditionally upon season, forecast lead time, magnitude of the observed and forecast variables, and aggregation period. The raw streamflow forecasts were verified against simulated streamflows, as well as observed streamflows, in order to separate the meteorological uncertainties from the total (meteorological and hydrologic) uncertainties.

In general, the MEFP-GCC precipitation forecasts are both reliable and skillful during the short-range (1-5 days). This largely originates from the skill in the raw GEFS precipitation forecasts. However, the reliability of the MEFP-GCC forecasts implies that

the MEFP is pre-processing (downscaling and bias-correcting) the raw inputs adequately. Indeed, the MEFP maintains or improves upon the correlations between the raw forcing from the GEFS and CFSv2 and the corresponding observed precipitation amounts. Likewise, the MEFP-GCC temperature forecasts are reliable and skillful during the short-range. Beyond the first 1-5 days, the skill of the MEFP-GCC precipitation forecasts declines rapidly, while the temperature forecasts remain skillful throughout the medium-range. However, neither the precipitation nor the temperature forecasts are skillful beyond ~2 weeks. This originates from a lack of skill in the raw CFSv2 forecasts and, beyond 270 days, from resampled climatology, which is inherently unskillful. Indeed, for the period from ~15-330 days, the MEFP-GCC precipitation and temperature forecasts closely resemble the MEFP-CLIM forecasts. For example, the MEFP-GCC forecasts show similar conditional biases to the MEFP-CLIM forecasts. This includes a substantial underestimation of the largest precipitation totals and a smaller conditional bias in the temperature forecasts, whereby the lowest and highest observed temperatures are over- and under-estimated, respectively. While the MEFP precipitation forecasts are generally no worse than sample climatology, the forecasts of Probability of Precipitation (PoP) are consistently worse than climatology. This originates from a lack of reliability in the MEFP forecasts of PoP. Similar biases were observed when calibrating the MEFP with NCEP's Global Forecast System (Brown, 2013). Again, this suggests a problem in the modeling, estimation, or implementation of the MEFP for PoP and light precipitation amounts.

The lack of skill in the MEFP-GCC precipitation and temperature forecasts suggests that the MEFP may not improve upon the existing operational practice for long-range forecasting, which relies on climatological forcing (ESP). Nevertheless, except for PoP and light precipitation, the MEFP-GCC forecasts are no worse than climatology. This is an important attribute of any bias-correction technique whose unconditional distribution is climatology. Without skillful predictors, the MEFP cannot improve upon climatology; it can only issue forecasts that are unconditionally unbiased. Enhancements to the MEFP may consider additional predictors. For example, given the strong autocorrelations in temperature, the MEFP may benefit from an autoregression of the future MAT on the most recently observed MAT, as well as the raw forecast.

While precipitation generally shows much weaker autocorrelations, auxiliary variables, such as relative humidity, may improve forecast quality over the short- to medium-range (Applequist et al., 2002; Hamill and Whitaker, 2006). For seasonal and long-range prediction, the MEFP may benefit from auxiliary climate information, such as the CPC's climate outlooks used in "conditional ESP" (Perica, 1998), or indices of teleconnection patterns, such as the El-Niño Southern Oscillation (ENSO), the Pacific-North American teleconnection (PNA) or the Pacific Diurnal Oscillation (PDO). Again, in the absence of skillful predictors, the ability to provide unbiased forecasts is an important attribute of the MEFP. Indeed, seasonal water supply and other long-range applications are known to benefit from climatological ensemble forecasts (Wood et al., 2005; Georgakakos et al., 2010; Schepen et al., 2012; Robertson and Wang, 2013). Ultimately, the HEFS will replace ESP for operational streamflow forecasting beyond the medium-range. Thus, the HEFS and ESP should be compared through hindcasting and verification (see below). Finally, the lack of skill in these basins does not imply similar performance in other regions or time periods, where the CFSv2 may contribute valuable skill (e.g. Yuan et al., 2013), or in slow-responding basins more generally, where the skill from the GEFS may persist for longer periods in the streamflow forecasts.

As well as producing reliable forecasts of temperature and precipitation at discrete times and locations, the MEFP should maintain realistic patterns in space and time and between variables. These statistical dependencies and multi-scale properties are important for decision making. In general, decisions about water resources are based on products derived from hydrologic forecasts, such as aggregated quantities, or on additional modeling studies or rules embedded in decision support systems. As with hydrologic modeling, these applications involve uncertainty propagation, for which space-time covariability is important (e.g. Clark et al., 2004). In this context, important attributes of the MEFP include the ability to: 1) preserve space-time and cross-variable relationships via the Schaake Shuffle; 2) derive skillful predictors at multiple space-time scales using canonical events; and 3) provide seamless predictions across multiple forecast horizons, depending on the raw forcing available (see Appendix A). Currently, the MEFP does not smooth the transition between raw forcing sources, although the underlying climatological distribution should be preserved, both marginally and in terms

of the joint relationships (via the Schaake Shuffle). In practice, some discontinuities were observed in the verification statistics between 270-272 days, where the CFSv2 transitions to resampled climatology. This may originate from sampling and parameter uncertainty, including uncertainties introduced by canonical events (see below). Indeed, the calibration requirements of the MEFP warrant further investigation. In terms of accounting for space-time covariability, the strengths and weaknesses of the Schaake Shuffle are described elsewhere (e.g. Clark et al., 2004). Anecdotally, the reliability and skill of the MEFP forecasts is the same or better at aggregated scales. This is consistent with the temporal autocorrelations being modeled adequately. Nevertheless, further investigation is warranted into the limitations of the Schaake Shuffle, particularly for extreme events, and whether, at the basin-scale, other empirical structures, such as high-resolution forecasts or conditional climatologies, can better predict the multivariate relationships (Shefzik et al., 2013).

In operational forecasting, there is always a trade-off between model complexity, or the need to capture salient features of the observations, and practicality, or the need for a model whose parameters can be estimated reliably. It is questionable whether the current implementation of the MEFP, or the choice of calibration used in this study, manages this trade-off effectively. The MEFP uses canonical events to sequentially adjust the climatological probability distribution. Each canonical event comprises a separate model of the joint probability distribution of the forecasts and observations. A canonical event defines a window centered on the forecast valid date (into which data are pooled from all historical years), together with the period of aggregation for which the joint distribution is estimated. Given this complexity, including the scope for interactions between canonical events, parameter estimation is a significant concern. The sample size used to calibrate the MEFP was not particularly small (15 years) and is consistent, or more favorable, than the expected operational practice. However, artificial periodicities were clearly visible in some of the verification statistics (shown in Section 5). They were also observed in the raw ensemble traces, particularly for temperature, which should otherwise vary smoothly. By experimenting with the parameterization of the MEFP, these discontinuities were found to originate from the choice of canonical events. The addition or removal of particular canonical events led to discontinuities in

the ensemble traces at corresponding timescales during the forecast horizon. Further investigation should establish where and when explicit modeling of the multi-scale properties is warranted. In such cases, they should be modeled smoothly, parsimoniously and with reasonably small sampling uncertainty, whether using canonical events or other techniques.

The overall uncertainties and biases in the streamflow forecasts comprise a combination of meteorological uncertainties and biases (from the MEFP) and hydrologic uncertainties and biases. By verifying the streamflow forecasts against hydrologic simulations, the meteorological uncertainties and biases can be separated from the hydrologic uncertainties and biases. In regulated rivers, the hydrologic uncertainties and biases are complicated by a range of natural and engineering influences. Also, some regulations may be obscured from operational forecasters because they involve rapidly changing conditions, multiple actors or agencies or commercially sensitive information. When information about diversions and other regulations is available in real-time, statistical post-processors, such as the EnsPost, should ideally adjust the natural flows, as regulations are difficult to model statistically. Also, when available from a trusted source, such information may imply a deterministic adjustment to the natural flows. However, the total (regulated) flows are preferred for hindcasting and verification, as they include the residual uncertainties from upstream basins (e.g. from hydrologic routing, poorly defined regulations, and simplified reservoir modeling), which are important in operational forecasting. In practice, only the local contributions were verified here, as the historical regulations were not sufficiently resolved to include in hindcasting; they comprised an overall adjustment to the observed flow, rather than a separate contribution for each regulation. In future, the precise regulations should be archived by the RFCs, in order to allow for hindcasting and verification of the total flows in downstream basins.

In keeping with the MEFP-GCC precipitation forecasts, the GCC streamflow forecasts are substantially more skillful than the climatological forecasts during the first week. Beyond the short-range, they are no less skillful than the climatological forecasts. Also, the streamflow forecasts with climatological forcing, as well as those with GCC

forcing, are consistently more skillful than sample climatology (e.g. in terms of ROC area). In general, the hydrologic biases are more important under low flow conditions, where the streamflow forecasts systematically over-estimate the observed flows, particularly in CCRN6 and MTGN4. While the hydrologic models do not target specific applications or flow conditions, the high flows receive particular scrutiny during model calibration, and the Continuous API model used by MARFC (for CCRN6 and MTGN4) is less well-suited to dry conditions. Nevertheless, the hydrologic uncertainties are also important for moderate and high streamflows. Indeed, in WALN6, CCRN6, MTGN4, MTRN6, MRNN6 and GILN6, the correlations are substantially higher when verifying against simulated flows than observed flows. However, for long-range forecasting, the meteorological biases are more important than the hydrologic biases, as the MEFP-GCC precipitation forecasts resemble the MEFP-CLIM forecasts beyond the short-range. Climatology is, by definition, conditionally biased and the MEFP precipitation forecasts underestimate the heaviest precipitation amounts by up to ~80% at longer forecast lead times.

Further work is needed to compare the long-range streamflow forecasts from the HEFS against the operational streamflow forecasts from the RFCs, which include ESP and statistical modeling on monthly and seasonal timescales. Given the lack of skill in the CFSv2, the opportunities to improve on ESP may appear limited. However, as indicated above, this does not imply similar performance in other regions or time periods. Also, some RFCs may benefit from the use of auxiliary variables, whether from seasonal climate outlooks or other large-scale climate indices (for a list of indices, see http://www.esrl.noaa.gov/psd/data/climateindices/list/, accessed 10th September 2013). For example, in comparing ESP with streamflow forecasts driven by NCEP's Global Spectral Model (GSM), Wood et al. (2005) found practically no benefits of the GSM over ESP. However, when selecting the GSM forecasts conditionally upon ENSO strength, they found a significant improvement in forecast quality during the autumn and winter months, mainly in California but also in the Pacific Northwest and the Great Basin. Elsewhere, Yuan et al. (2013) found that the CFSv2 produced significantly more skilful hydrologic forecasts than ESP, both unconditionally and conditionally on ENSO strength, but these improvements generally only materialized after streamflow post-

processing and were strongly dependent on the variables, seasons and locations considered. Alongside the resampling procedure adopted by the MEFP (Appendix A), there are other notable differences between the HEFS and ESP. For example, some RFCs incorporate adjustments or “MODs” into the hydrologic model states from which the ESP forecast are produced operationally. The latter is problematic for the HEFS (and more generally) if the MODs are subjective or not otherwise reproducible. Whether the HEFS can improve on ESP will also depend on basin characteristics. For headwater basins with longer memory (e.g. due to snow accumulation or soil characteristics), and for downstream basins in general, the skill from the GEFS may persist for longer periods. Also, the EnsPost should eliminate any unconditional biases and possibly reduce the conditional biases where the streamflow correlations are strong. For example, at MRNN6 and GILN6, the observed streamflows were consistently underestimated during the late spring and early summer. In CCRN6 and MTGN4, the observed streamflows were consistently overestimated during the summer months.

Scientific evaluation of the HEFS is an ongoing activity; it requires a sustained effort and a dedicated 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. In order to guide a broader range of applications 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. Furthermore, there is a need to evaluate decision support systems and other models that rely on the HEFS (e.g. of water quality and ecology). Such applications will show varying sensitivities to the HEFS forecasts and may lead to targeted improvements in the HEFS and to new ensemble products.

7. Glossary of terms and acronyms

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Disaggregation – (see aggregation/disaggregation)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

MAP – Mean Areal Precipitation over a basin/watershed

MAT – Mean Areal Temperature over a basin/watershed

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

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

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

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

NWP – Numerical Weather Prediction

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

NYCDEP – New York City Department of Environmental Protection

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

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

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

Reforecast – See Hindcast. Commonly used in the atmospheric sciences.

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

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

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

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

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

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

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

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

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

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

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

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

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

WPC – Weather Prediction Center, formerly the Hydrometeorological Prediction Center

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

8. References

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

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

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

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

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

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

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

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

Brown, J.D. and Heuvelink, G. 2005. Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In: Anderson, M.

(Ed.) *The Encyclopedia of Hydrological Sciences*, Chichester: John Wiley and Sons, 1181–1195.

Brown, J.D., and Seo, D-J 2013. Evaluation of a nonparametric post-processor for bias-correction and uncertainty estimation of hydrologic predictions. *Hydrological Processes*, **27**(1), 83-105, doi: 10.1002/hyp.9263.

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.

Cayan, D. R. 1996. Climate variability and snow pack in the western United States. *Journal of Climate* **9**, 928-948.

Clark, M. P., Serreze, M. C. and McCabe, G. J. 2001. Historical effects of El Nino and La Nina events on the seasonal evolution of the montane snowpack in the Columbia and Colorado River Basins. *Water Resources Research* **37***,* 741-757.

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

Demeritt, D., Nobert, S., Cloke, H. L. and Pappenberger, F. 2013. The European Flood Alert System and the communication, perception, and use of ensemble

predictions for operational flood risk management. *Hydrological Processes* **27**, 147–157. doi: 10.1002/hyp.9419.

Franz, K.J., Hartmann, H.C., Sorooshian, S. and Bales, R. 2003. Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. *Journal of Hydrometeorology* **4**, 1105–1118.

Garen, D.C. 1992. Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management* **118**(6), 654–670.

Garen, D.C., and Pagano, T.C. 2007. *Statistical techniques used in the VIPER water supply forecasting software*. NRCS-USDA Engineering-Snow Survey and Water Supply Forecasting Technical Note 210-2, 18pp. [Available at: http://www.wcc.nrcs.usda.gov/ftpref/downloads/factpub/wsf/technotes/Tech_note_statistical_techniques_in_Viper.pdf, accessed 12th September, 2-13]

Georgakakos, A.P., Yao, H. and Georgakakos, K.P. 2010. *Upstream regulation adjustments to ensemble streamflow predictions*. HRC Technical Report No. 7. Hydrologic Research Center, San Diego, CA. 76pp.

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.

Grantz, K., Rajagopalan, B., Clark, M. and Zagona, E. 2005. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resources Research* **41**, W10410, doi:10.1029/2004WR003467.

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

Hamill, T.M., and Whitaker, J.S. 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review* **134**, 3209-3229.

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

Najafi, M.R., Moradkhani, H. and Piechota, T.C. 2012. Ensemble Streamflow Prediction: Climate signal weighting methods vs. Climate Forecast System Reanalysis. *Journal of Hydrology* **442**, 105-116.

Nash, L. N. and Gleick, P. H. 1991. Sensitivity of streamflow in the Colorado basin to climatic changes. *Journal of Hydrology* **125**, 221-241.

Pagano, T., Garen, D. and Sorooshian, S. 2004. Evaluation of official western US seasonal water supply outlooks. *Journal of Hydrometeorology* **5**, 896–909.

Pagano, T.C., Wood, A.W., Werner, K. and Tama-Tweet, R. 2013. Western US water supply forecasting: a tradition evolves. Submitted to *EoS*.

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].

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

Regonda, S.K. 2006. *Intra-annual to inter-decadal variability in the Upper Colorado hydroclimatology: diagnosis, forecasting and implications for water resources management*. Ph.D. dissertation, University of Colorado, boulder, CO. 151pp.

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

Robertson, D.E. and Wang, Q. J. 2012. A Bayesian approach to predictor selection for seasonal streamflow forecasting. *Journal of Hydrometeorology* **13**(1), 155-171.

Robertson, D.E. and Wang, Q. J. 2013. Seasonal forecasts of unregulated inflows into the Murray River, Australia. *Water Resources Management* **27**(8), 2747-2769.

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

Schepen, A., Wang, Q. J. and Robertson, D. E. 2012. Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian

seasonal rainfall. *Journal of Geophysical Research* **117***, D20107,* doi:10.1029/2012JD018011*.*

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.

Shi, X. Wood, A.W. and Lettenmaier, D.P. 2008. How essential is hydrologic model calibration to seasonal streamflow forecasting? *Journal of Hydrometeorology* **9**, 1350-1363.

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

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

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.

Wood, A.W., Kumar, A. and Lettenmaier, D.P. 2005. A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *Journal of Geophysical Research* **110***,* D04105 doi:10.1029/2004JD004508.

Wood, A. W. and Lettenmaier, D.P. 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bulletin of the American Meteorological Society* **87**, 1699–1712.

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

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

Yuan, X., Wood, E., Roundy, J. and Pan., M. 2013: CFSv2-based seasonal hydroclimate forecasts over conterminous United States. *Journal of Climate*, **26**, 4828-4847.

9. Tables

Table 1: characteristics of the study basins

Characteristic	MARFC				NERFC			
	NVXN6	WALN6	CCRN6	MTGN4	MTRN6	MRNN6	PTVN6	GILN6
Lat. of outlet	41.8269	42.1661	41.7567	41.3092	42.0142	42.0378	42.3194	42.3558
Long. of outlet	-74.6389	-75.1403	75.0578	74.7956	74.2708	73.9725	74.4369	74.4450
Lat. of GEFS node	41.8985	42.3667	41.8985	41.4304	42.3667	42.3667	42.3667	42.3667
Long. of GEFS node	-74.5312	-75.0000	-75.0000	-74.5312	-74.5312	-74.0625	-74.0625	-74.5312
Lat. of CFS node	42.0471	42.0471	42.0471	42.0471	42.0471	42.0471	42.0471	42.0471
Long. of CFS node	-75.0000	-75.0000	-75.0000	-75.0000	-74.0625	-74.0625	-74.0625	-74.0625
Area (total, km^2)	240	860	4714	9013	497	1085	614	816
Mean elev. (m)	209.8	180.1	232.4	100.5	169.8	138.4	197.4	172.8
Annual P (mm)	1308	1049	1117	1157	1463	1268	1101	956
Annual PE (mm)	692	692	692	782	622	701	633	643
P/PE	1.89	1.52	1.61	1.48	2.35	1.81	1.74	1.49
Annual runoff (mm)	932	588	489	527	1280	590	380	710
Runoff coefficient	0.71	0.56	0.44	0.46	0.88	0.47	0.35	0.74

P = precipitation
PE = potential evaporation

Table 2: indices of verifying observations (day of year) for different forecast lead times and multiple T0s

T0 (day of year)	Forecast lead time (days)											
	1	2	3	4	5	6	7	8	9	10	11	…
1	2	3	4	5	6	7	8	9	10	11	12	…
6	7	8	9	10	11	12	13	14	15	16	17	…
11	12	13	14	15	16	17	18	19	20	21	22	…
16	17	18	19	20	21	22	23	24	25	26	27	…
21	22	23	24	25	26	27	28	29	30	31	32	…
26	27	28	29	30	31	32	33	34	35	36	37	…
31	32	33	34	35	36	37	37	39	40	41	42	…
…	…	…	…	…	…	…	…	…	…	…	…	…

10. Figures

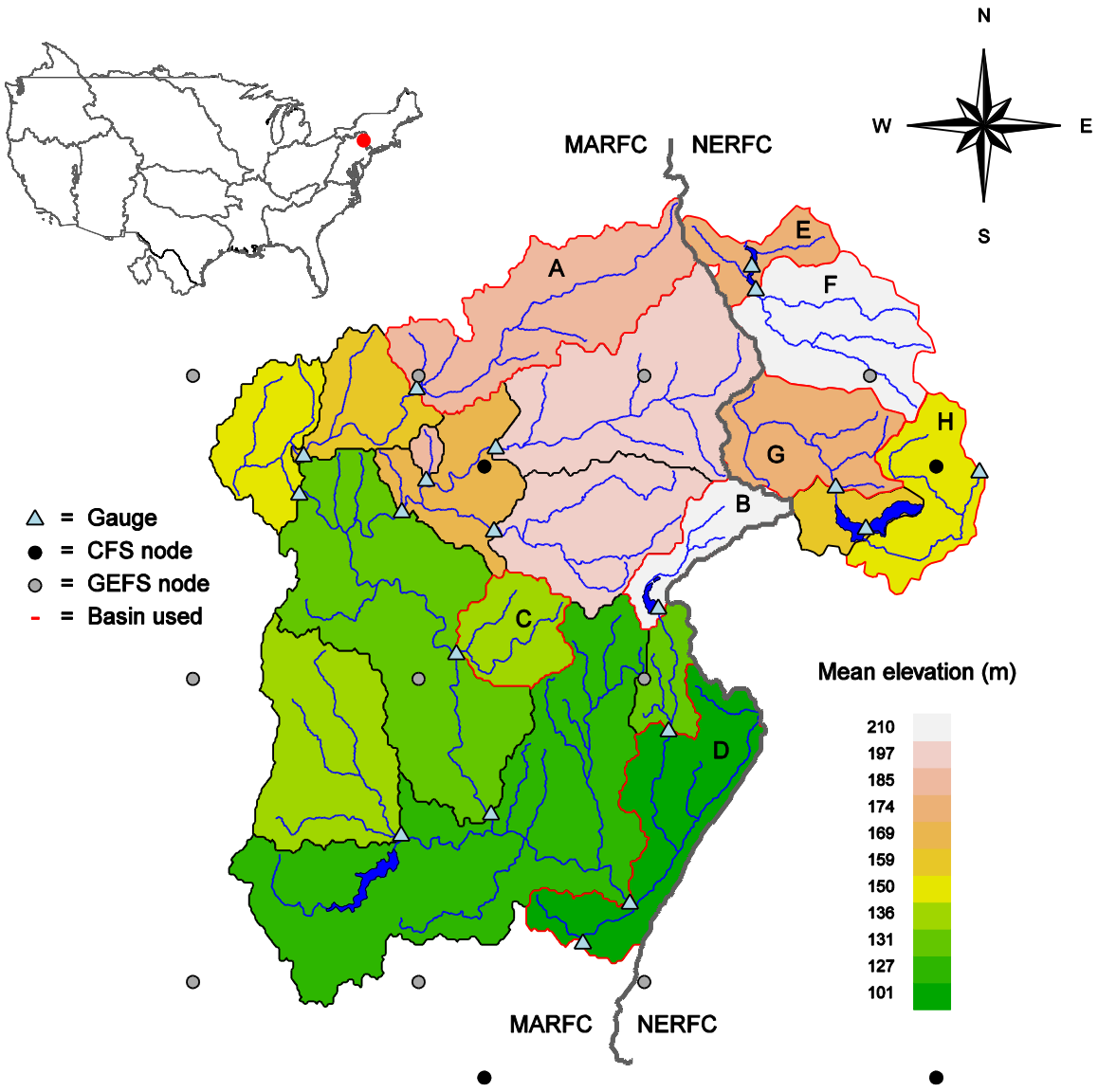

Figure 1: the study area, comprising four basins in MARFC and four basins in NERFC (highlighted), together with the surrounding basins. The MARFC basins include: WALN6 (A), NVXN6 (B), CCRN6 (C) and MTGN4 (D). The NERFC basins include: GILN6 (E), PTVN6 (F), MTRN6 (G) and MRNN6 (H).

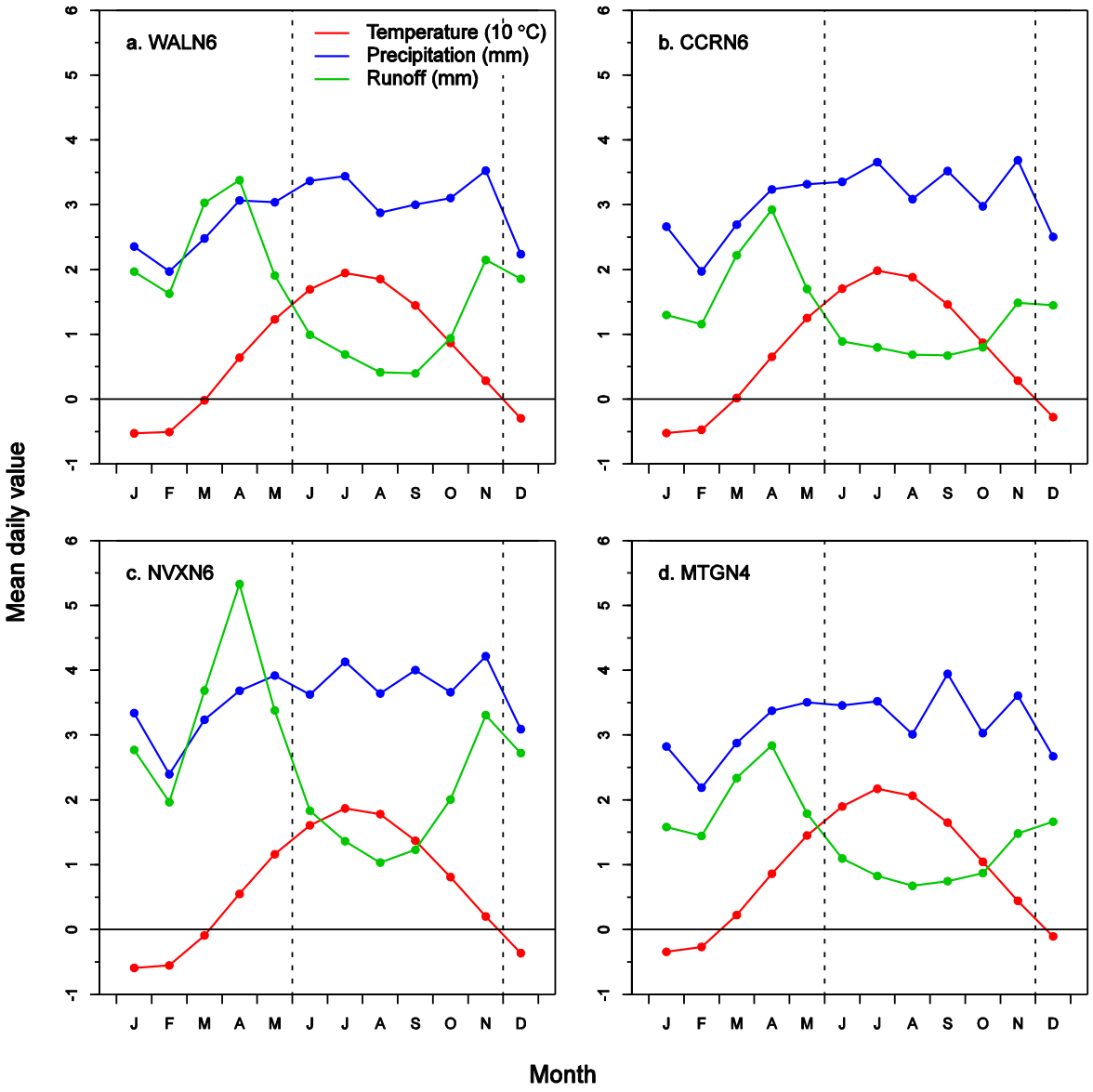

Figure 2a: Daily averages of temperature, precipitation and runoff by calendar month for each study basin in MARFC. Locations MTRN6 and PTVN6 in NERFC each comprise two sub-basins; the meteorological variables are averaged over these sub-basins, weighed by basin area.

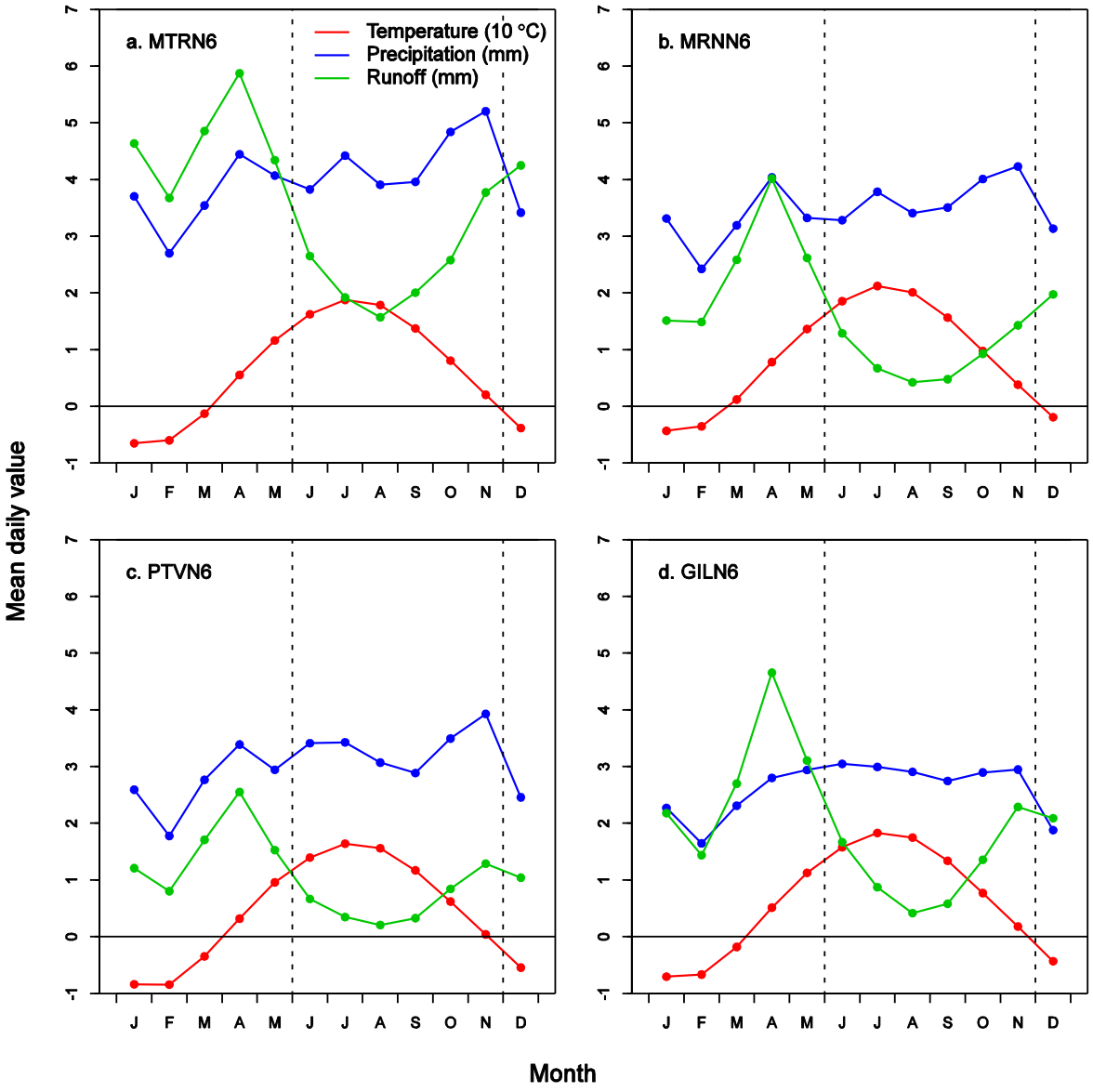

Figure 2b: Daily averages of temperature, precipitation and runoff by calendar month for each study basin in NERFC. Locations MTRN6 and PTVN6 each comprise two sub-basins; the meteorological variables are averaged over these sub-basins, weighed by basin area.

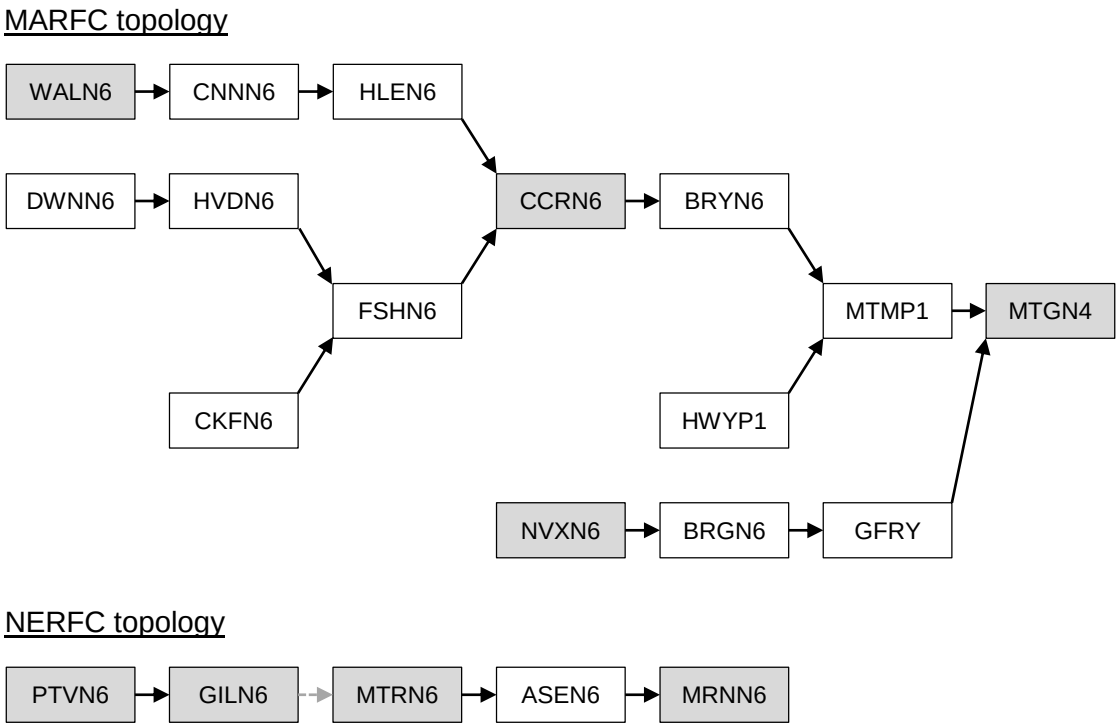

Figure 3: topology of the forecast locations for MARFC and NERFC. The shaded boxes denote the eight locations considered in this study. Mount Trempor (MTRN6) receives diverted flows from the Schoharie Reservoir (GILN6). In practice, GFRY is modeled as part of BRGN6.

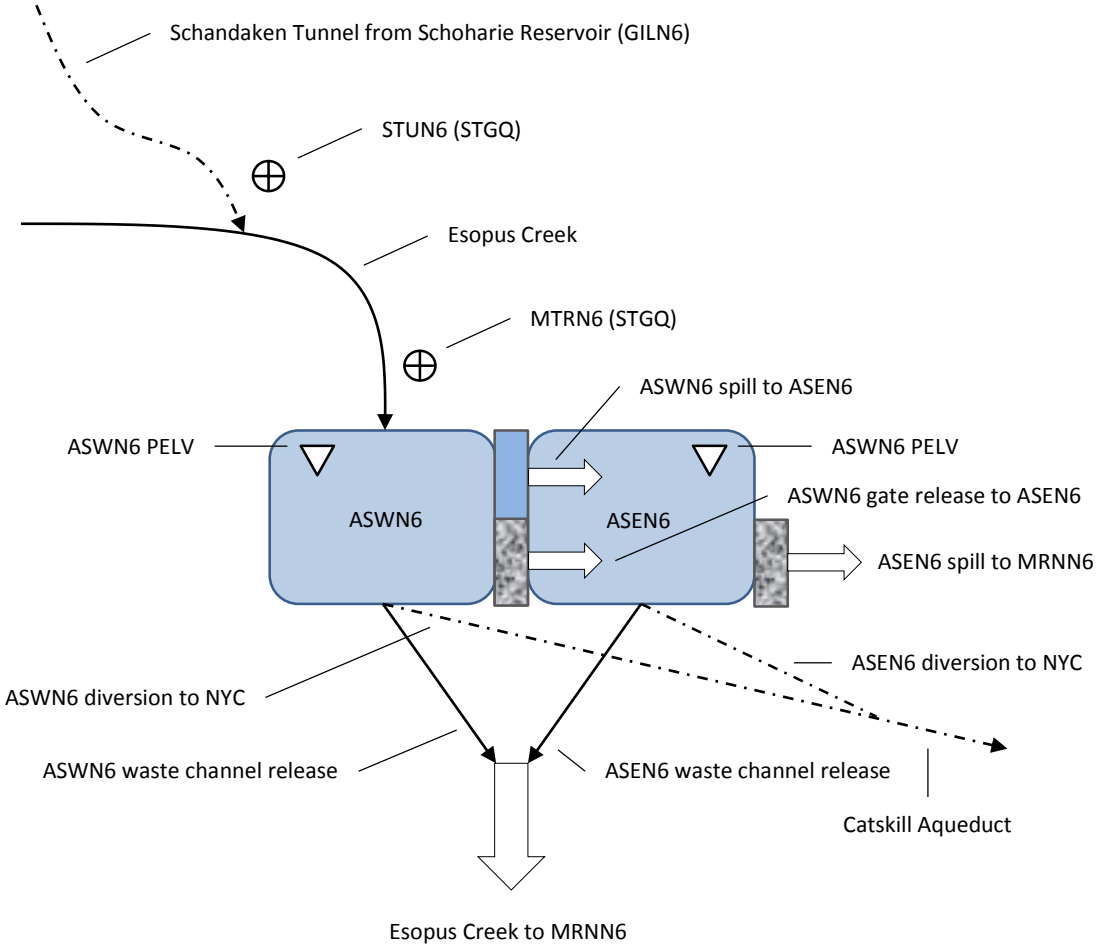

Figure 4: schematic of the flow pathways and regulations associated with the Ashoken Reservoir in NERFC.

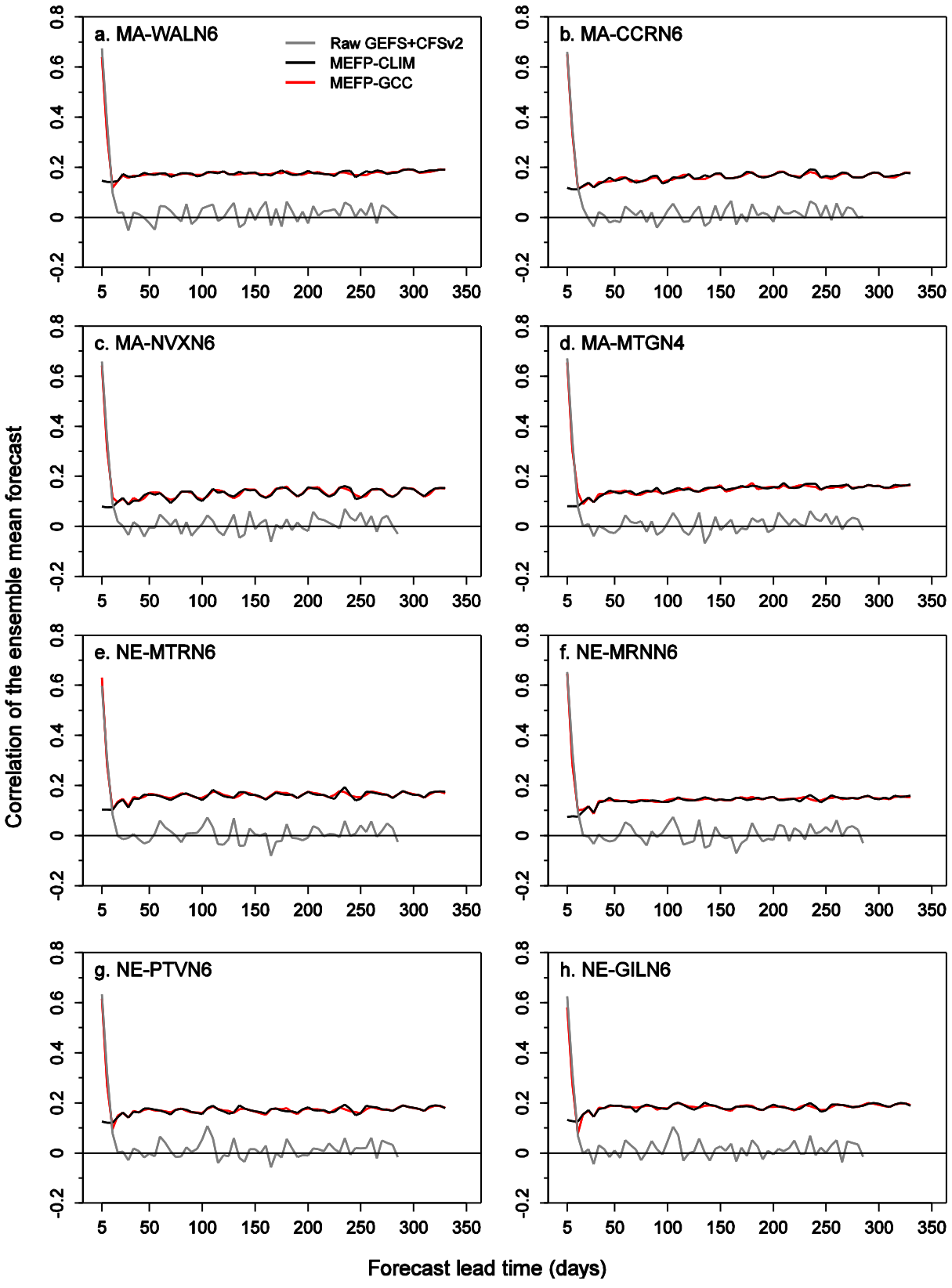

Figure 5: Correlation of the ensemble mean forecast and observed precipitation amounts by forecast lead time for each source of forcing from the MEFP, namely resampled climatology (CLIM) and GEFS+CFSv2+CLIM (GCC), together with the raw forcing from the GEFS and CFSv2 for the period 1-270 days.

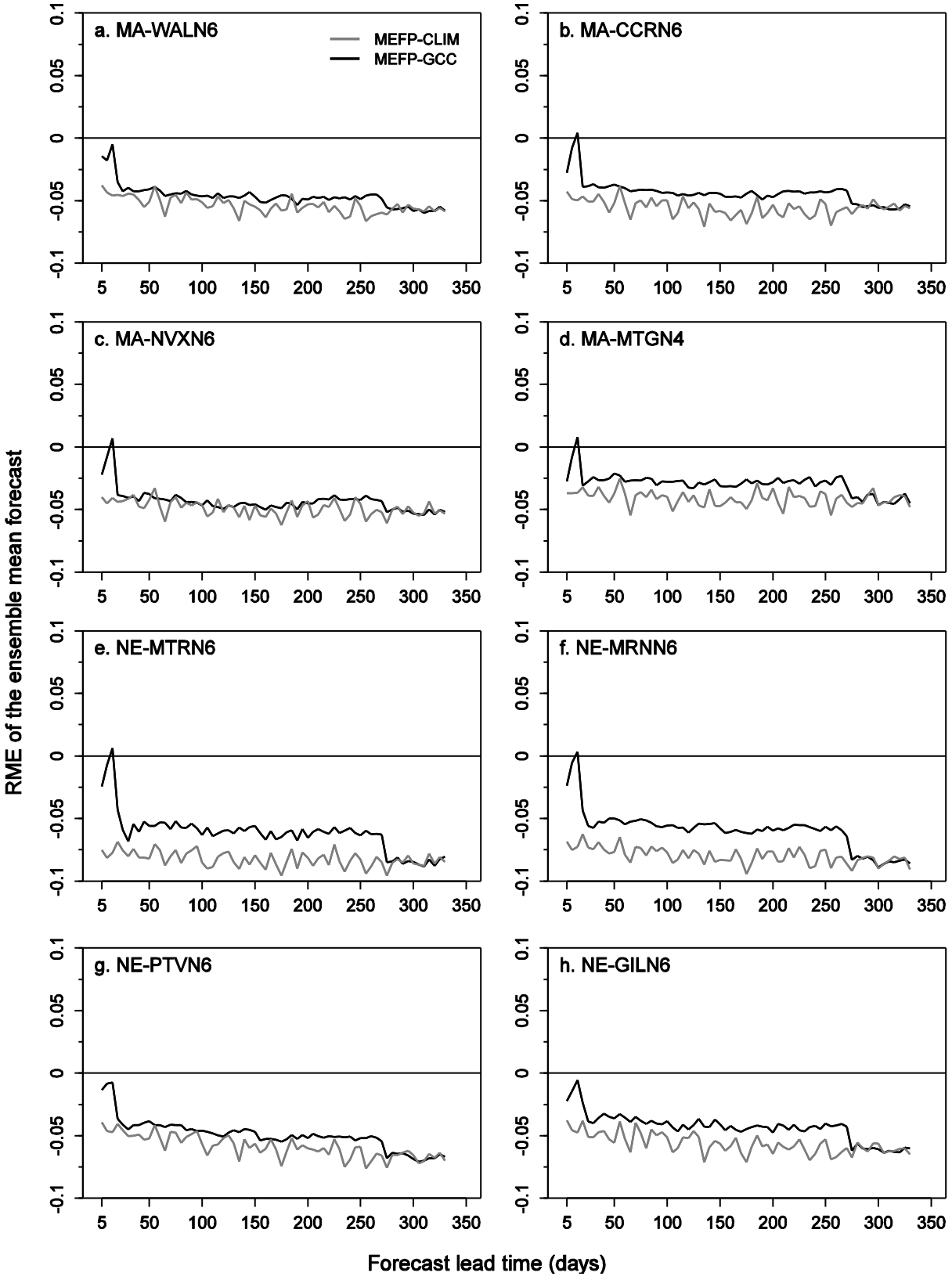

Figure 6: Relative mean error of the ensemble mean forecasts of precipitation by forecast lead time for each source of forcing from the MEFP, namely resampled climatology (CLIM) and GEFS+CFSv2+CLIM (GCC).

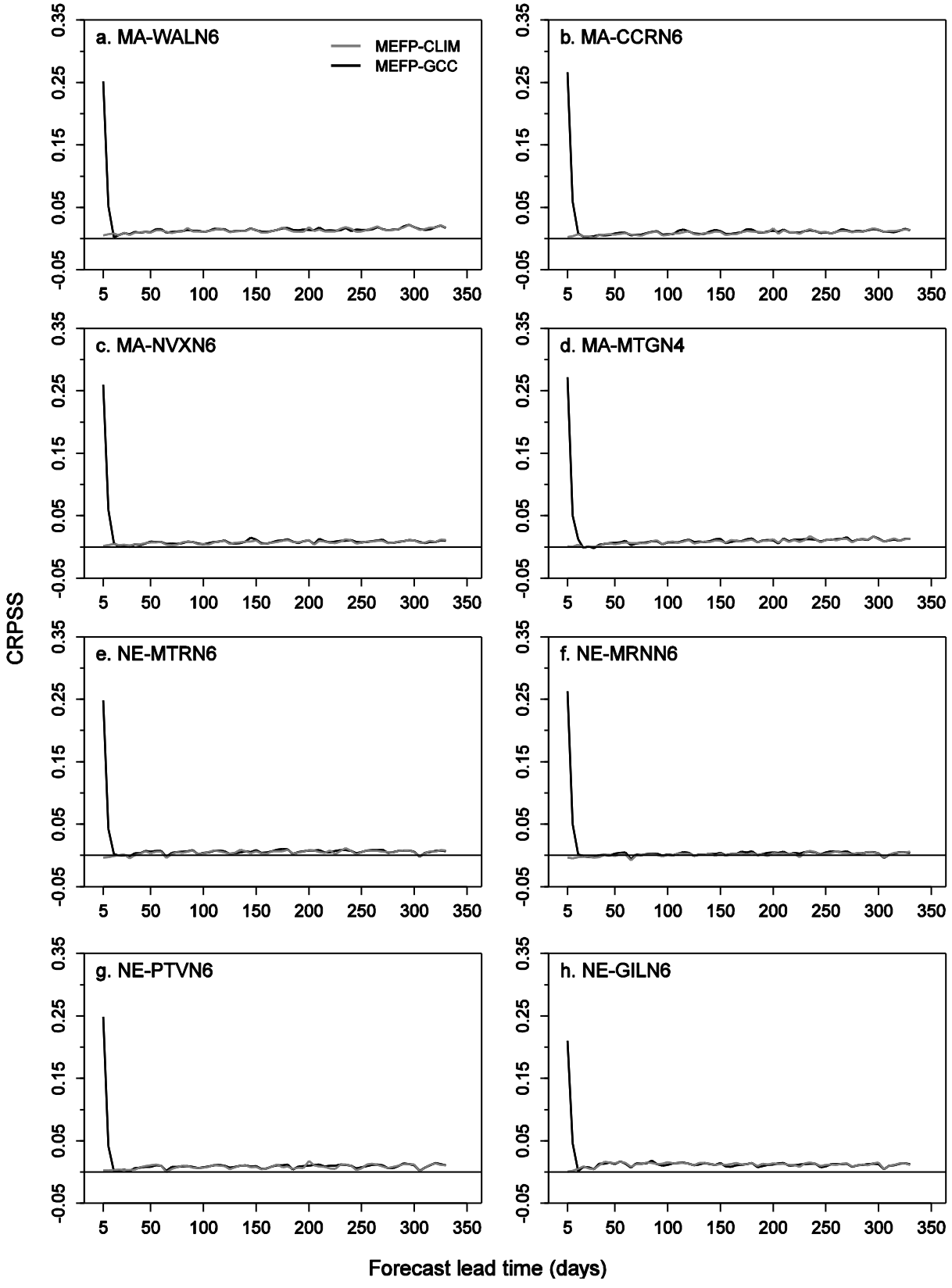

Figure 7: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-CLIM and MEFP-GCC precipitation forecasts relative to sample climatology.

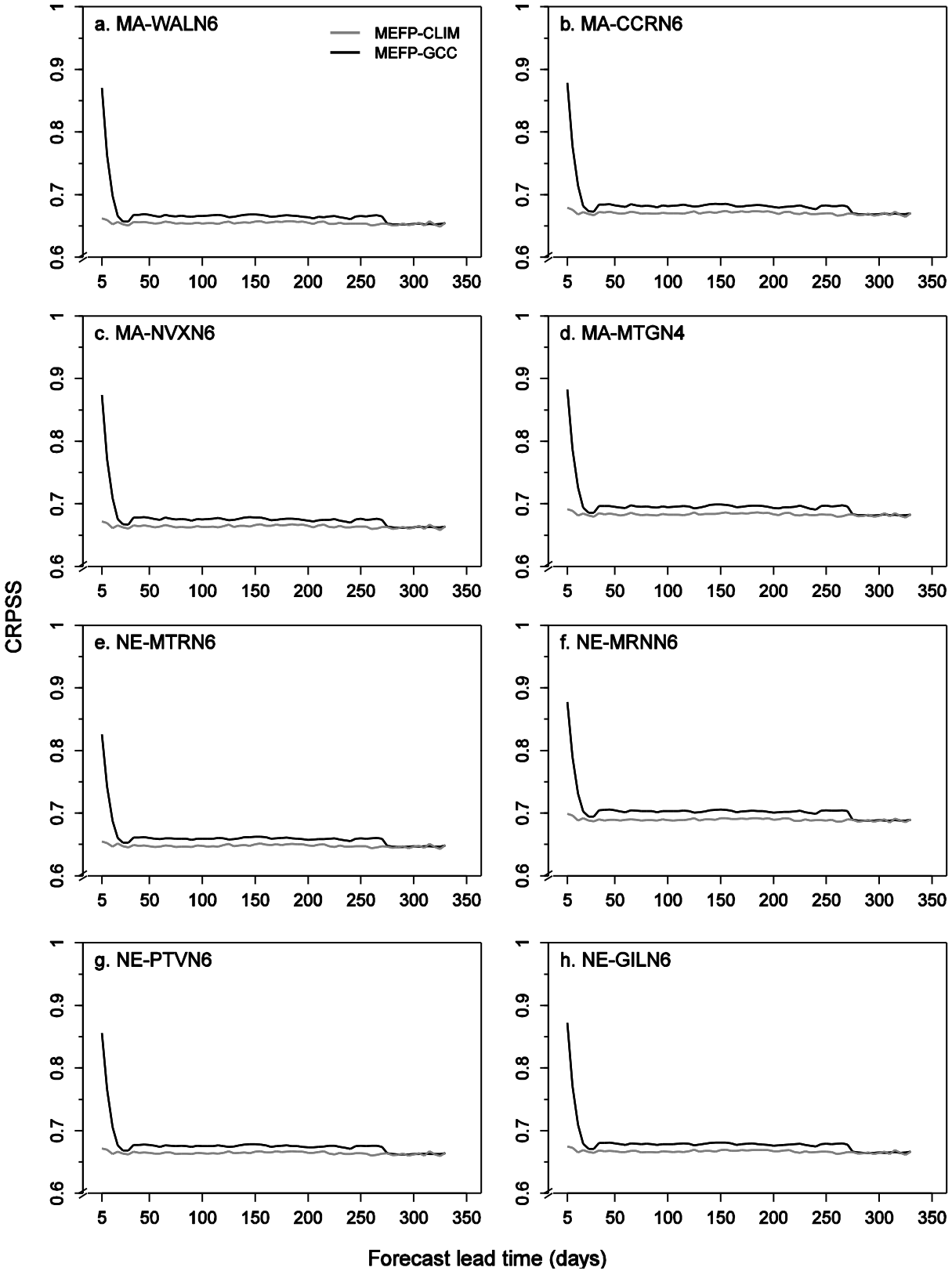

Figure 8: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-CLIM and MEFP-GCC temperature forecasts relative to sample climatology.

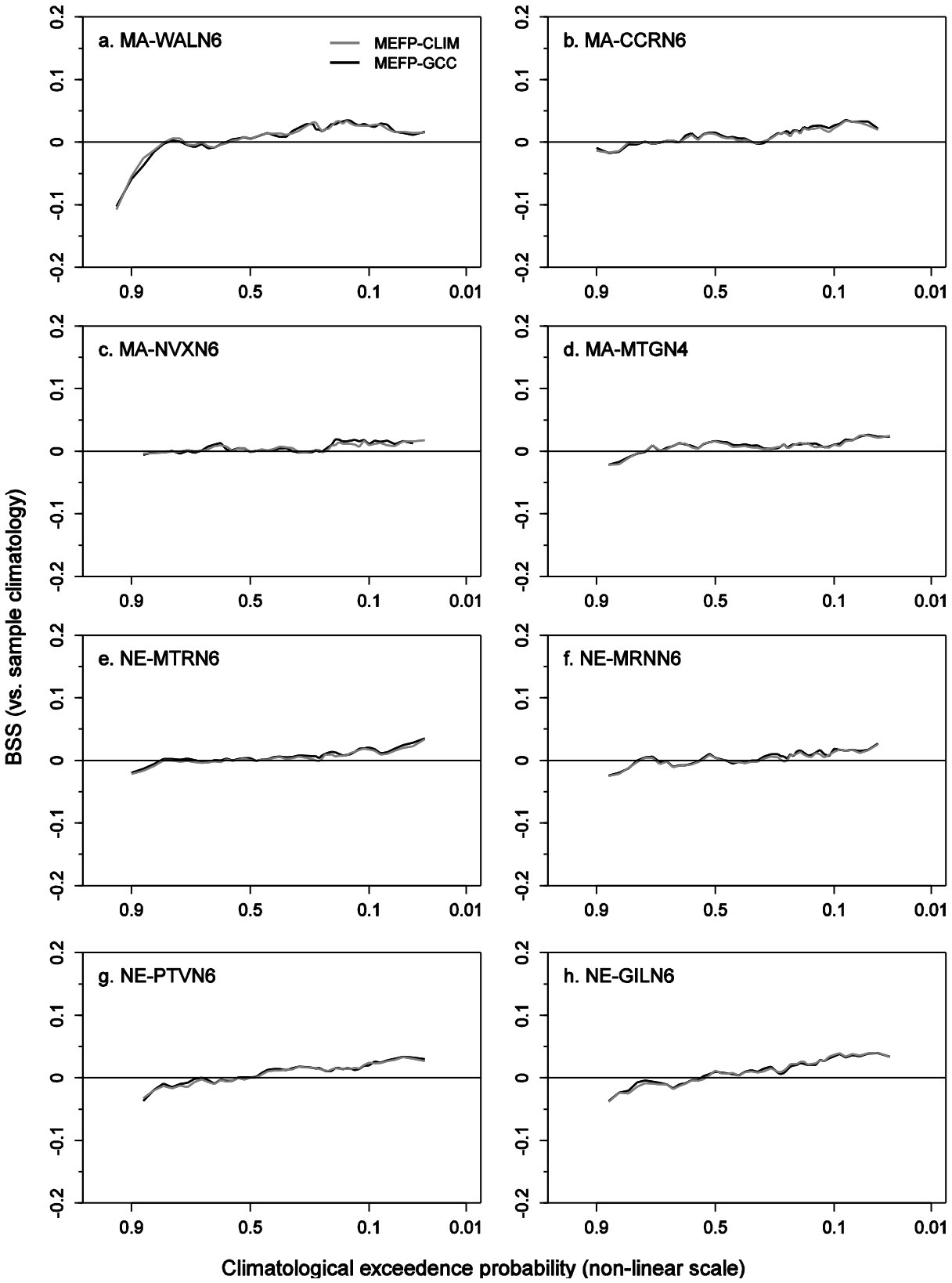

Figure 9a: Brier Skill Score (BSS) of the MEFP-CLIM and MEFP-GCC precipitation forecasts relative to sample climatology. The results are shown for a forecast lead time of 95-100 days and for increasing amounts of observed precipitation. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

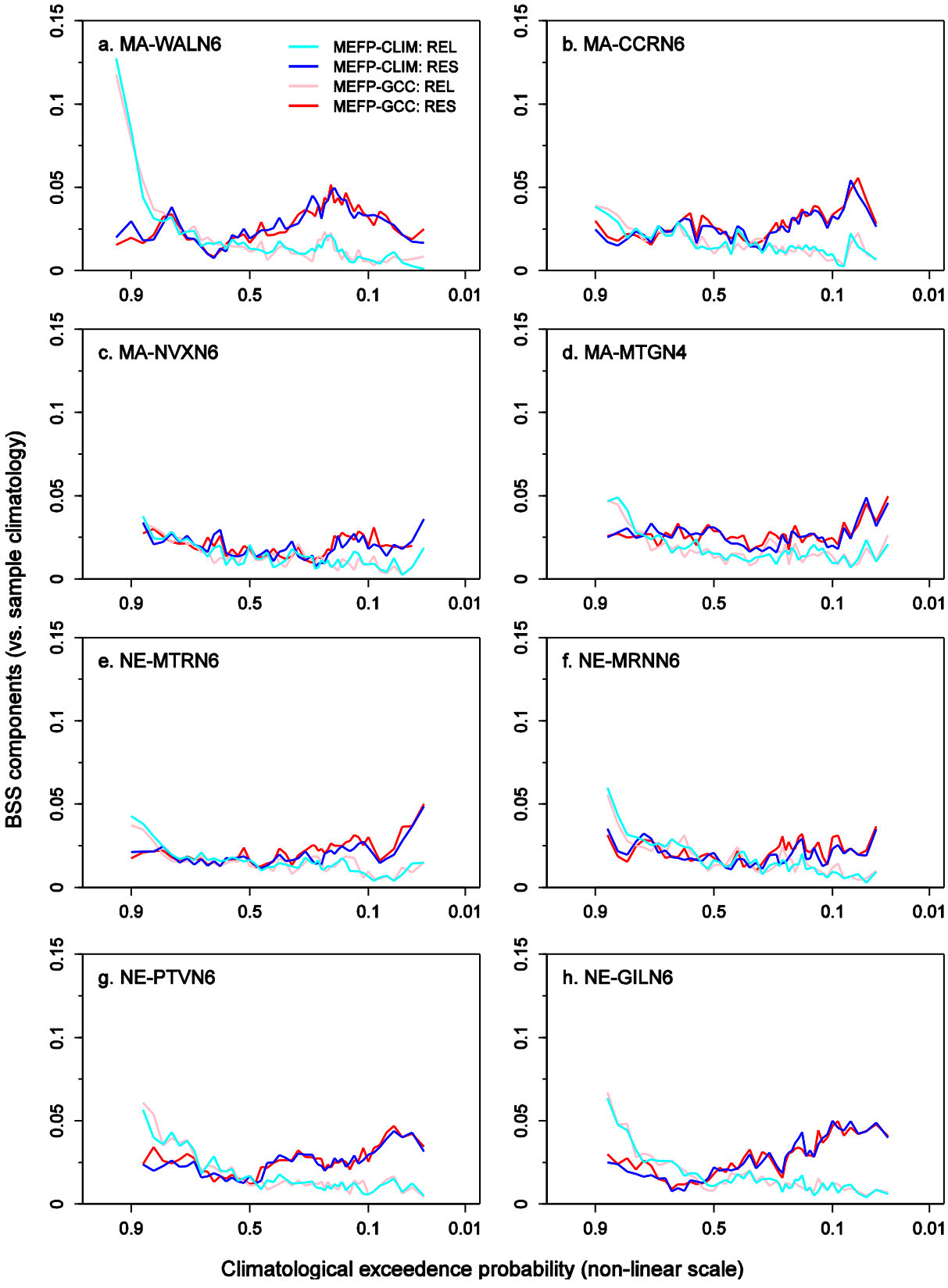

Figure 9b: Calibration-refinement factorization of the Brier Skill Score (BSS) for the MEFP-CLIM and MEFP-GCC precipitation forecasts relative to sample climatology. The results are shown for a forecast lead time of 95-100 days and for increasing amounts of observed precipitation. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale.

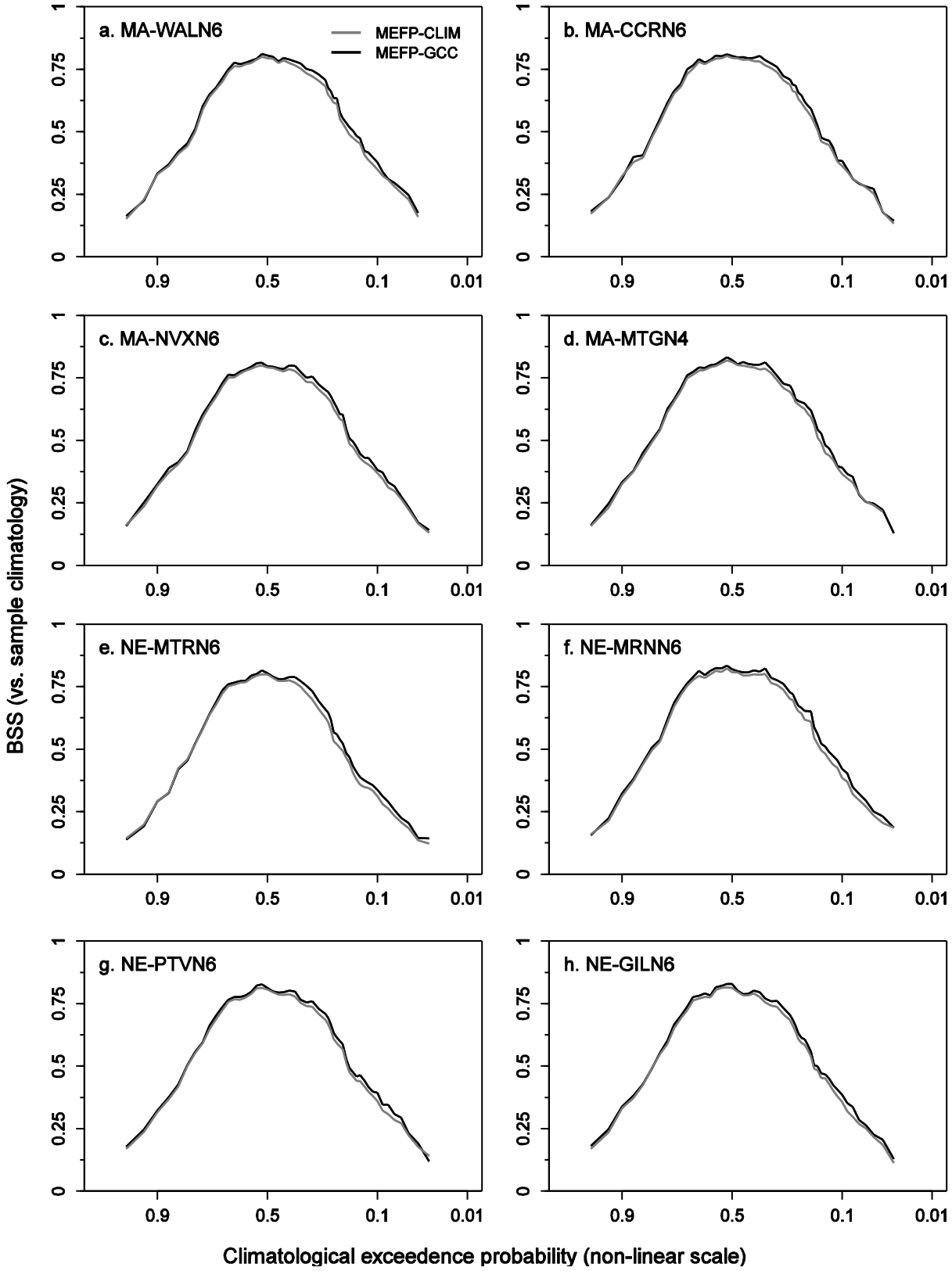

Figure 10: Brier Skill Score (BSS) of the MEFP-CLIM and MEFP-GCC temperature forecasts relative to sample climatology. The results are shown for a forecast lead time of 95-100 days and for increasing amounts of observed precipitation. The precipitation thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

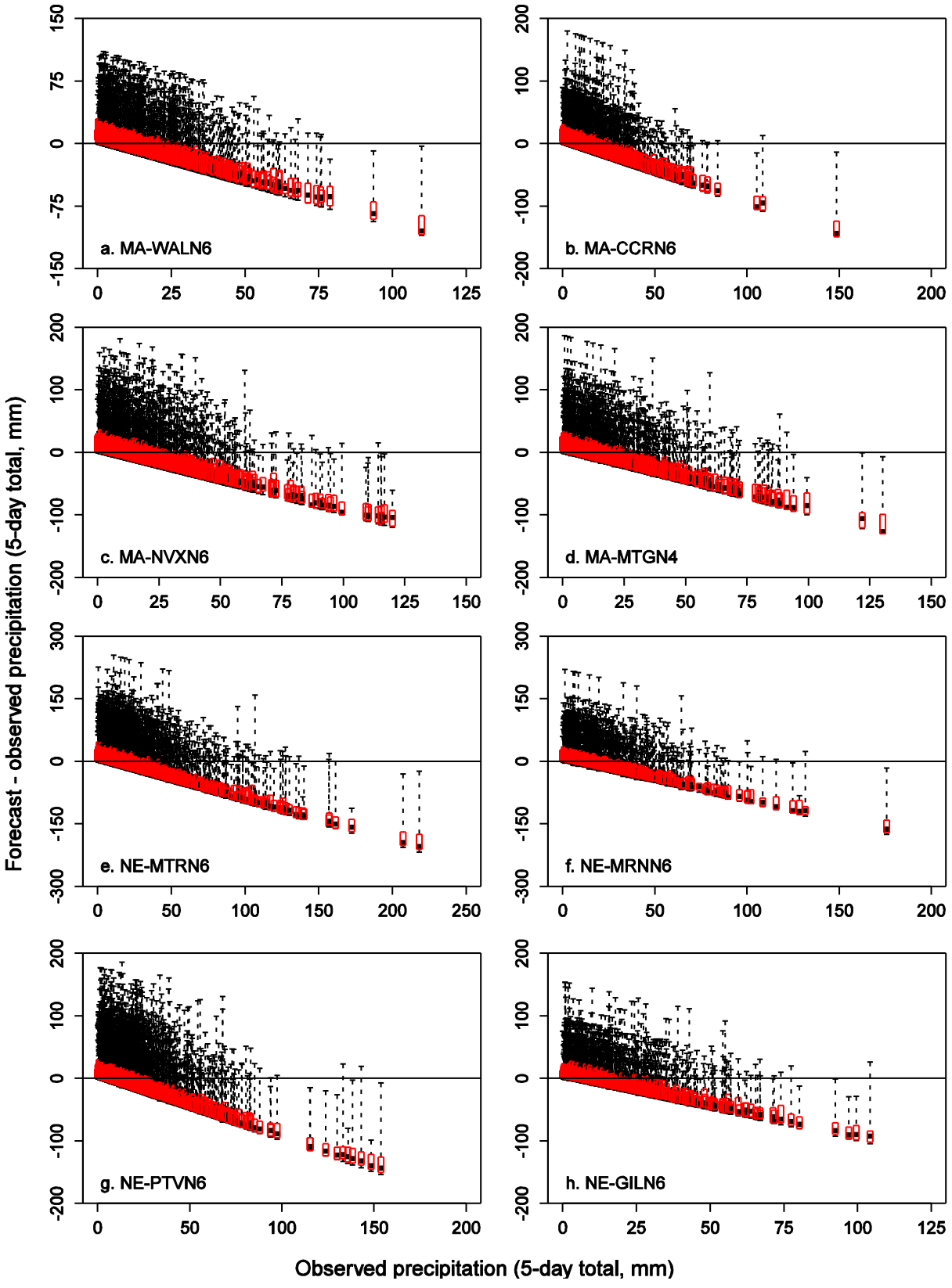

Figure 11: Box plots of errors in the MEFP precipitation forecasts with GCC forcing at a forecast lead time of 95-100 days. The boxes are ordered by increasing amounts of observed precipitation.

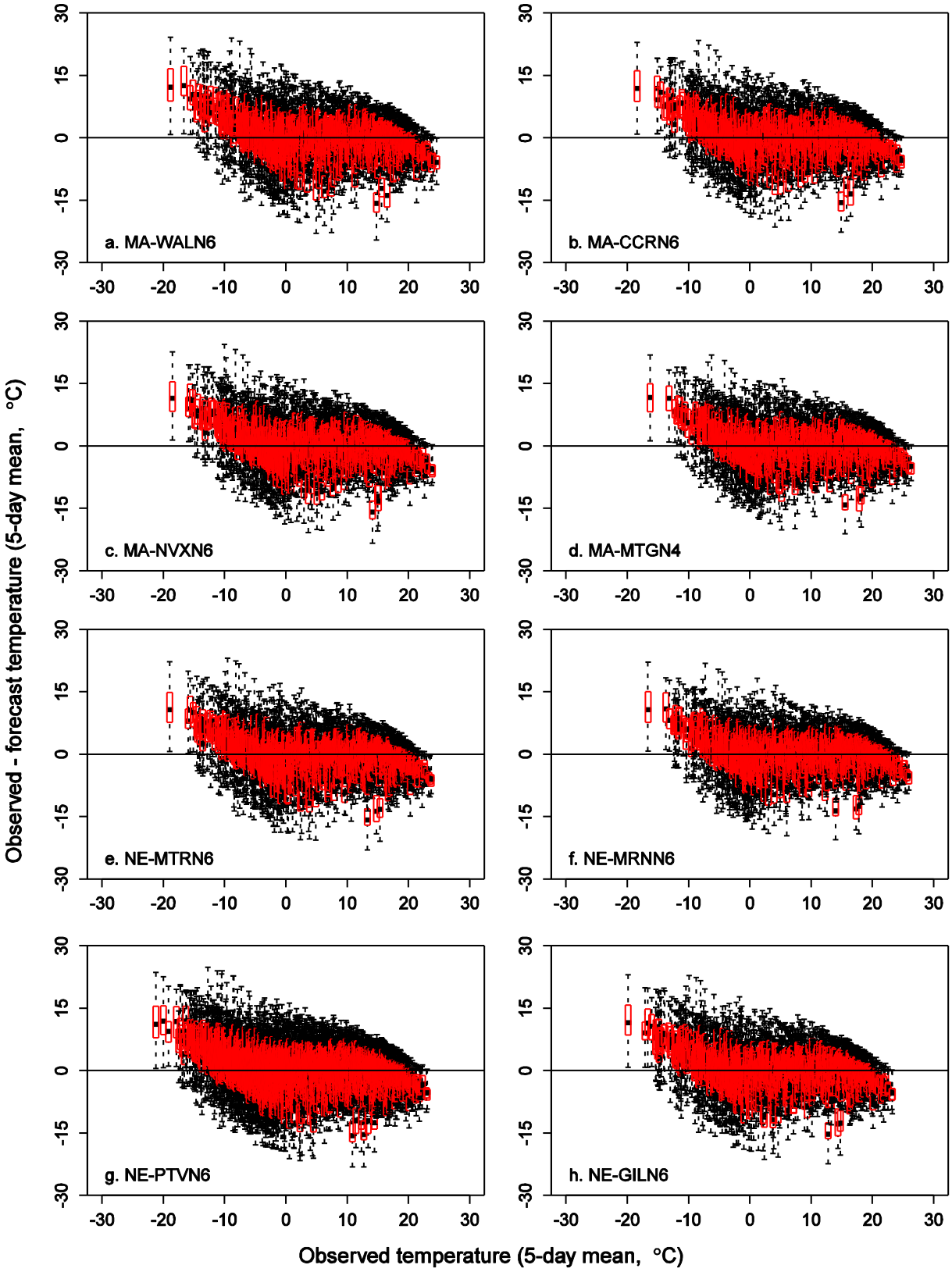

Figure 12: Box plots of errors in the MEFP temperature forecasts with GCC forcing at a forecast lead time of 95-100 days. The boxes are ordered by increasing observed temperatures.

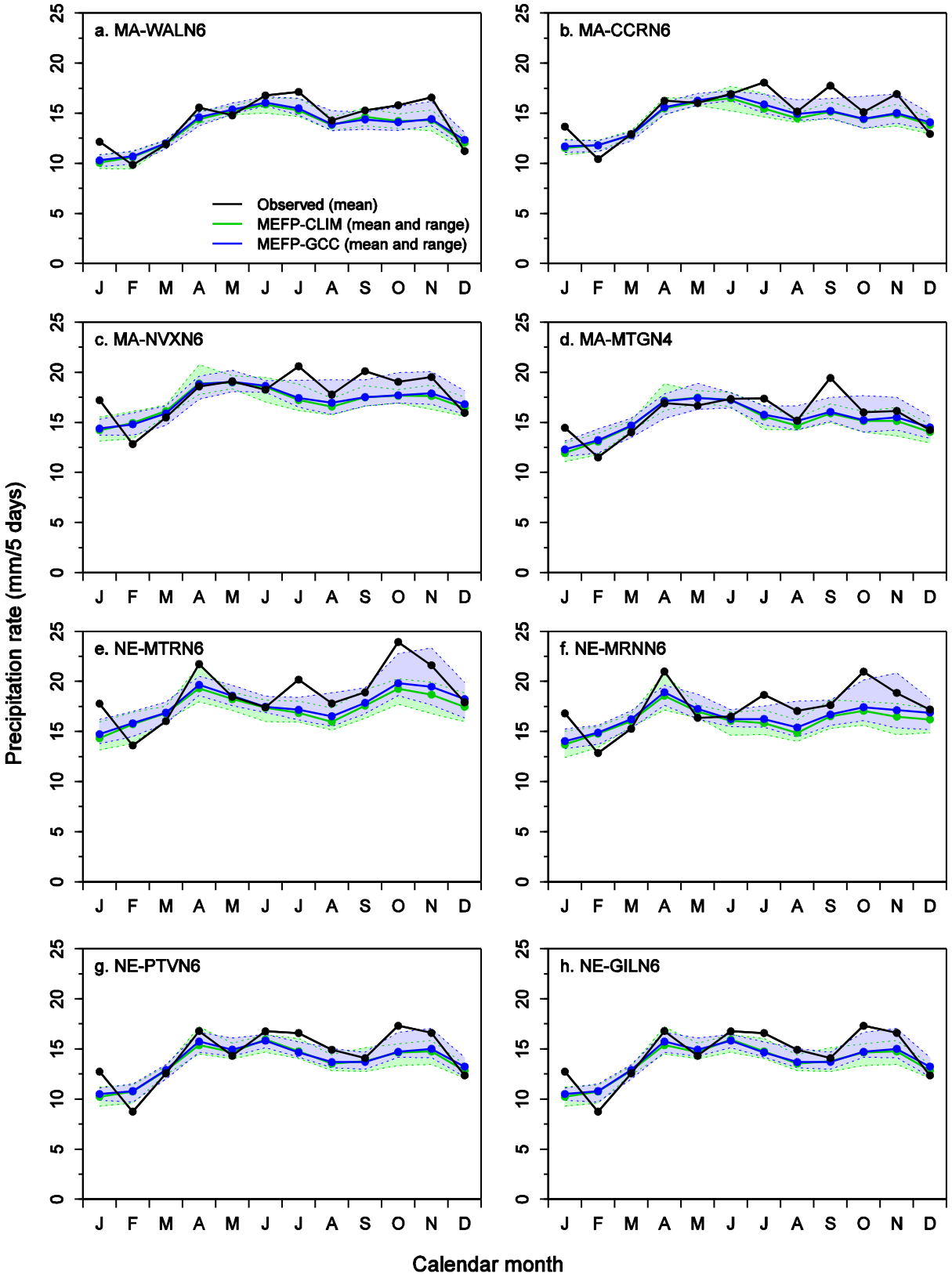

Figure 13: Average observed and forecast precipitation rates by calendar month for each source of forcing from the MEFP (CLIM and GCC). The forecasts comprise the average of the ensemble means by calendar month across all forecast lead times, together with the range of the conditional averages by forecast lead time. The results are shown for basins in MARFC and NERFC.

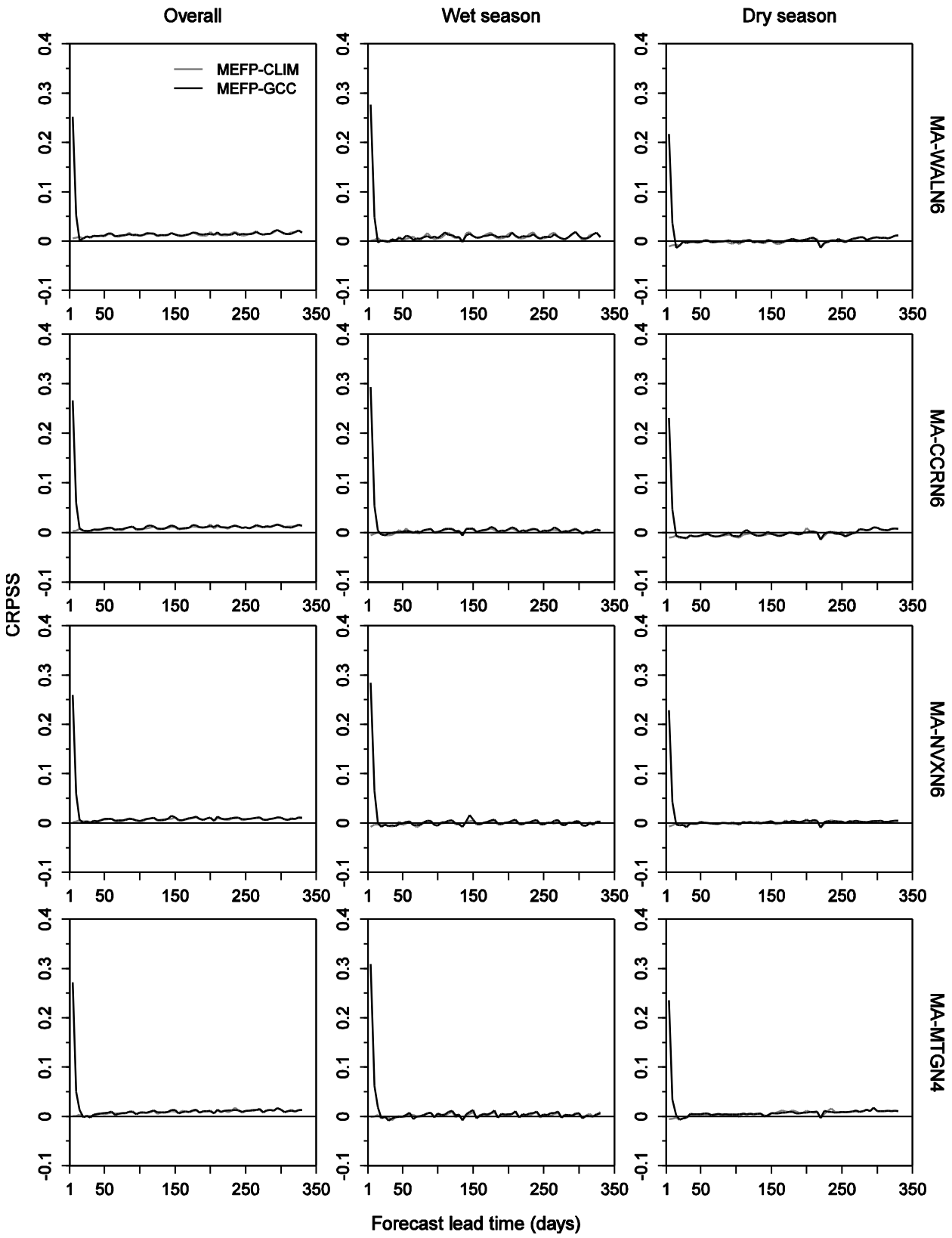

Figure 14a: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-GCC and MEFP-CLIM precipitation forecasts with sample climatology as the baseline. The results are shown for the basins in MARFC with a forecast lead time of 95-100 days and comprise the overall period and the wet and dry seasons separately.

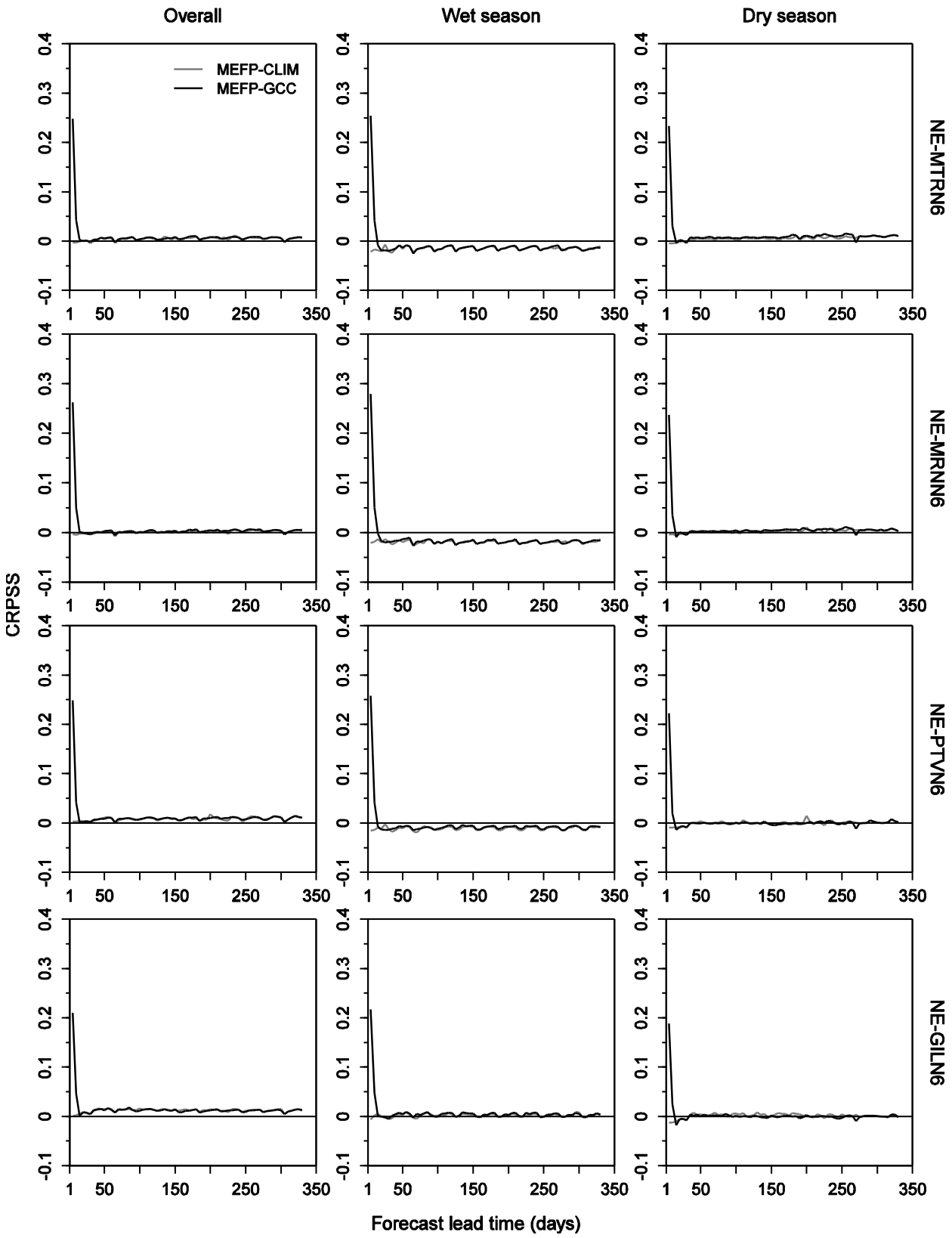

Figure 14b: Mean Continuous Ranked Probability Skill Score (CRPSS) of the MEFP-GCC and MEFP-CLIM precipitation forecasts with sample climatology as the baseline. The results are shown for the basins in NERFC with a forecast lead time of 95-100 days and comprise the overall period and the wet and dry seasons separately.

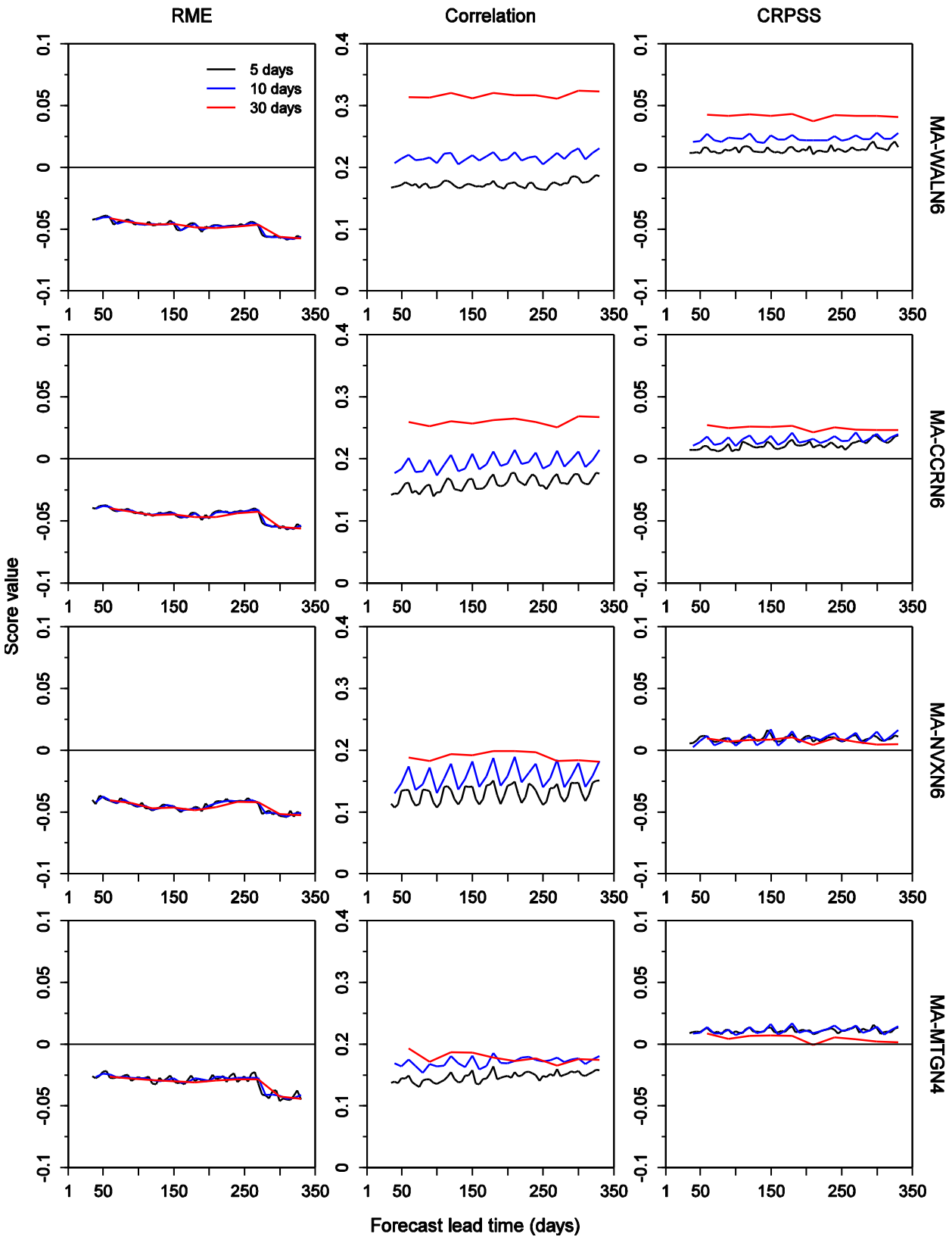

Figure 15a: Selected verification metrics for the MEFP-GCC precipitation forecasts at three aggregation periods (5-, 10- and 30-days) for basins in MARFC.

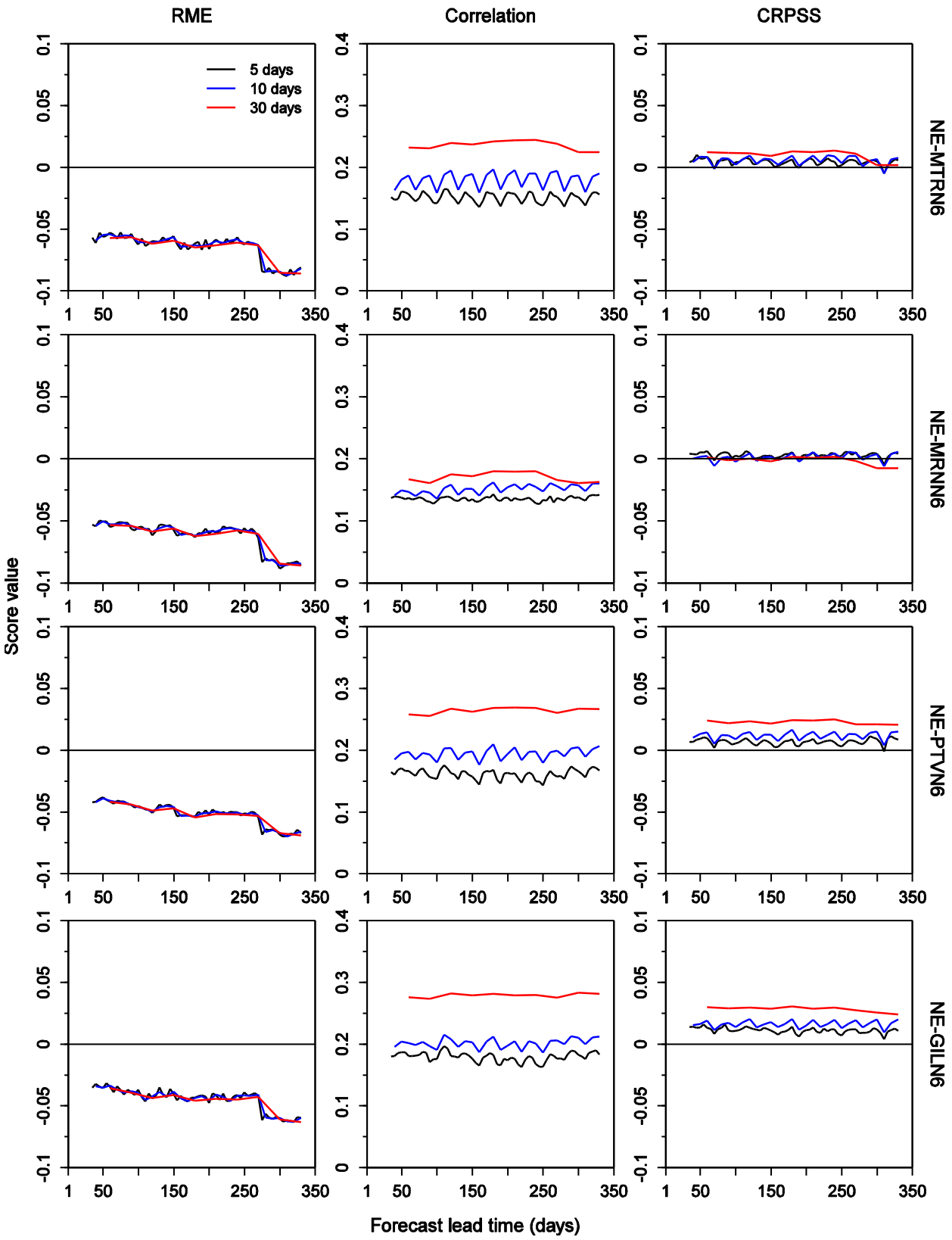

Figure 15b: Selected verification metrics for the MEFP-GCC precipitation forecasts at three aggregation periods (5-, 10- and 30-days) for basins in NERFC.

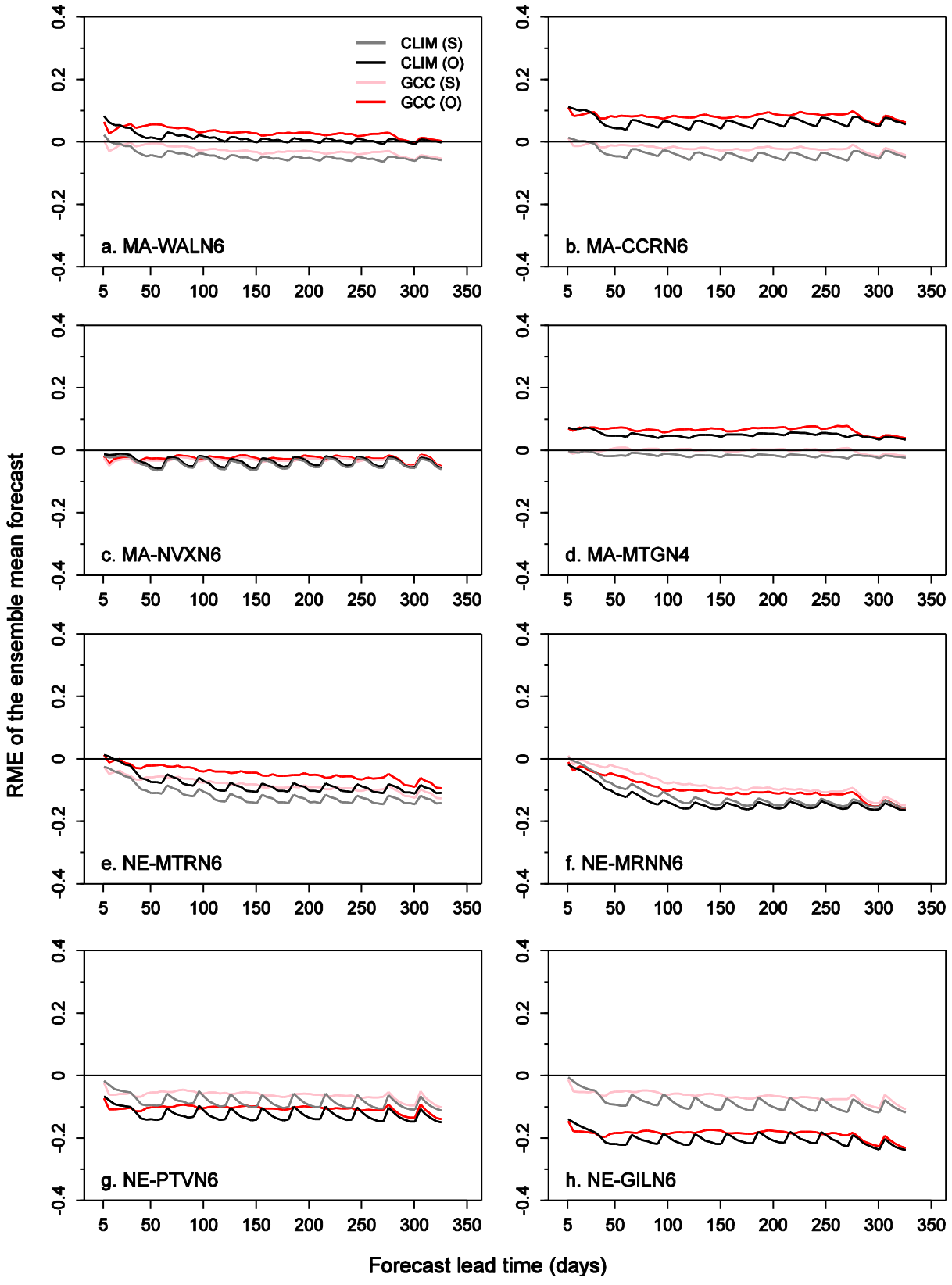

Figure 16: Relative mean error of the streamflow forecasts (ensemble mean) with MEFP-GCC and MEFP-CLIM forcing against observed (O) and simulated (S) flows.

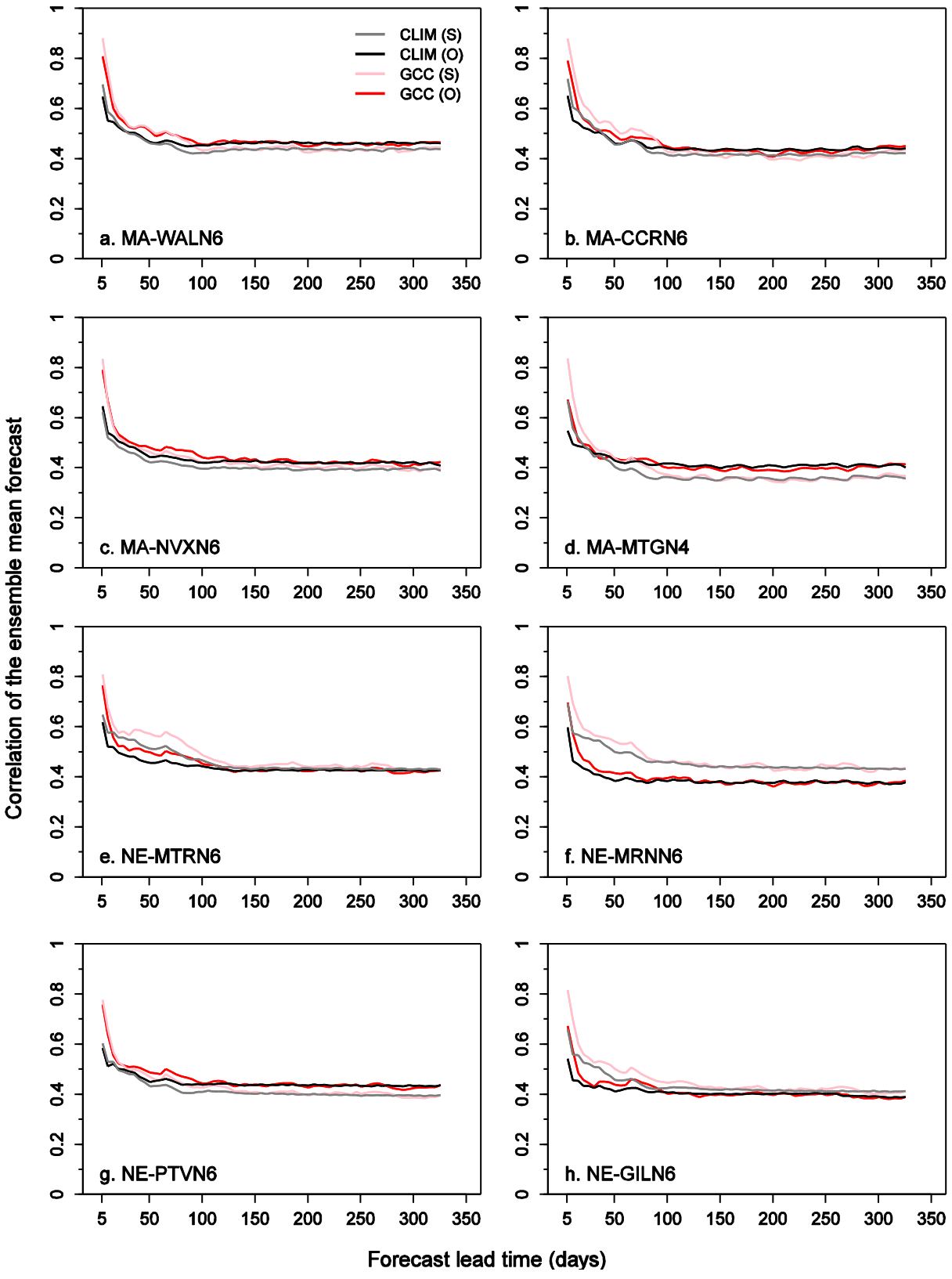

Figure 17: Correlation of the streamflow forecasts (ensemble mean) with MEFP-GCC and MEFP-CLIM forcing against observed (O) and simulated (S) flows.

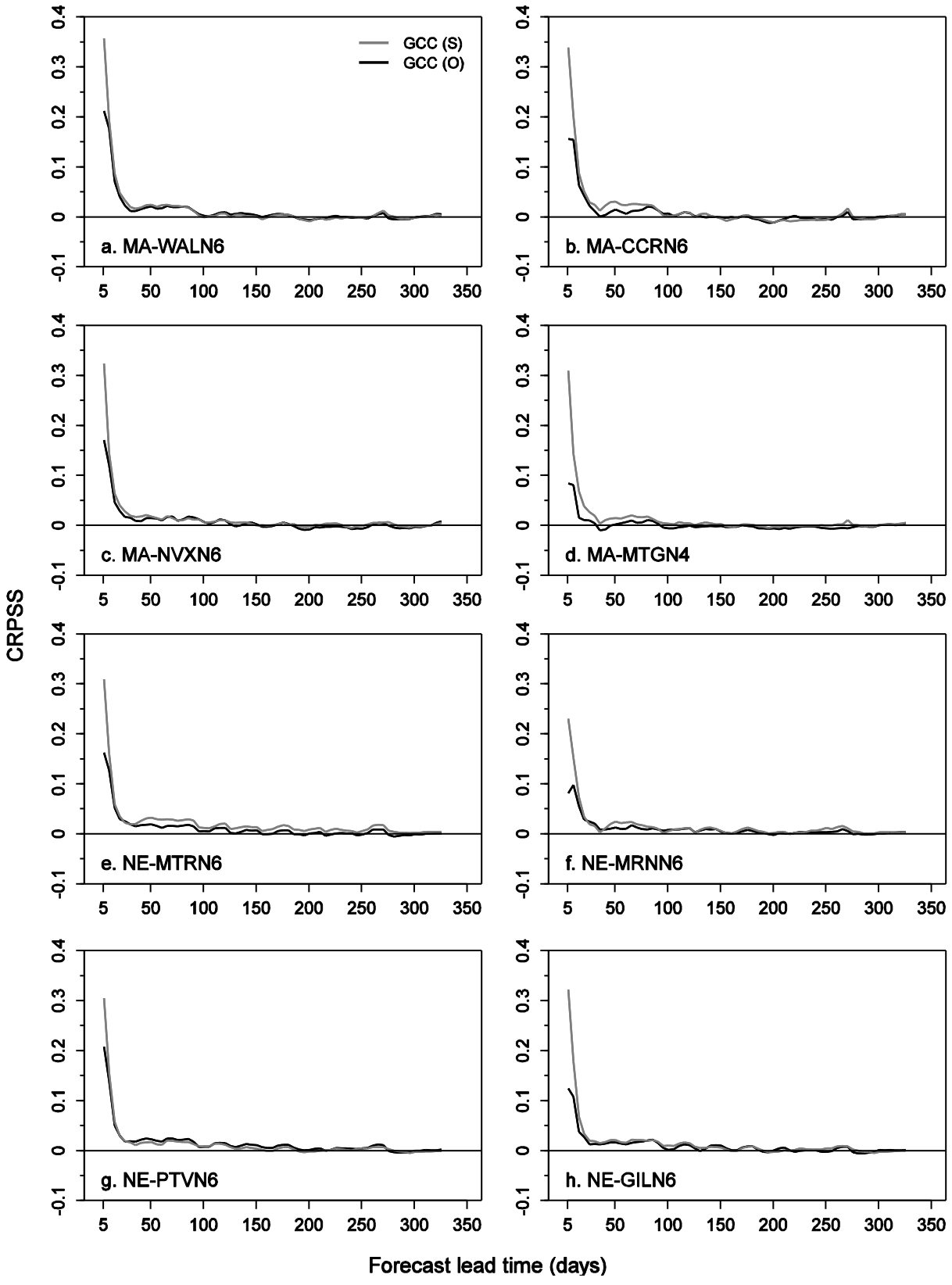

Figure 18: Mean Continuous Ranked Probability Skill score (CRPSS) of the streamflow forecasts with MEFP-GCC forcing (GCC). The forecasts are verified against observed (O) and simulated (S) flows. The reference streamflow forecasts comprise forcing from the MEFP with resampled climatology as input.

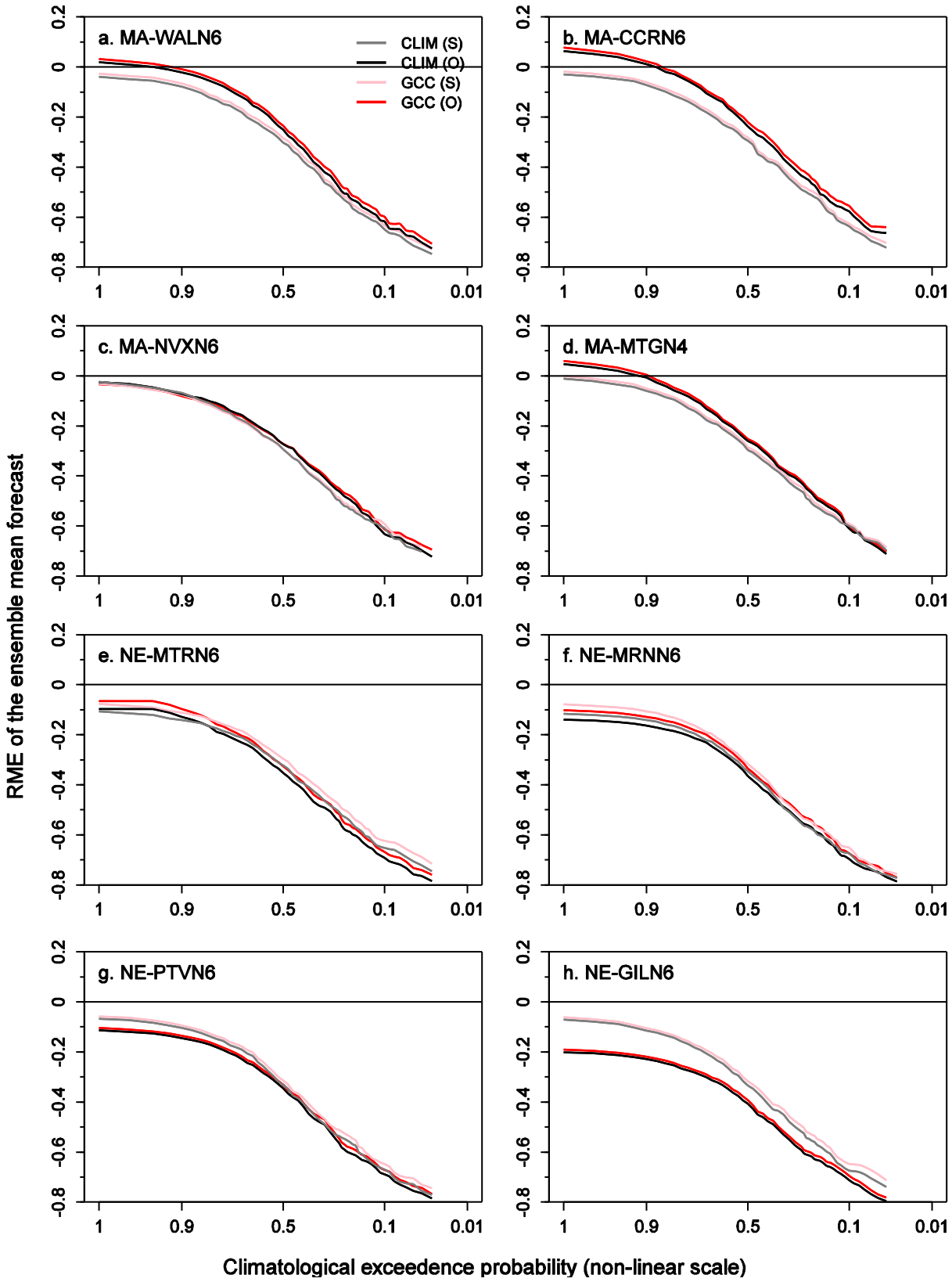

Figure 19: Relative mean error of the MEFP-CLIM and MEFP-GCC streamflow forecasts when verified against observed (O) and simulated (S) flows. The results are shown for a forecast lead time of 95-100 days and for increasing streamflow thresholds. The thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

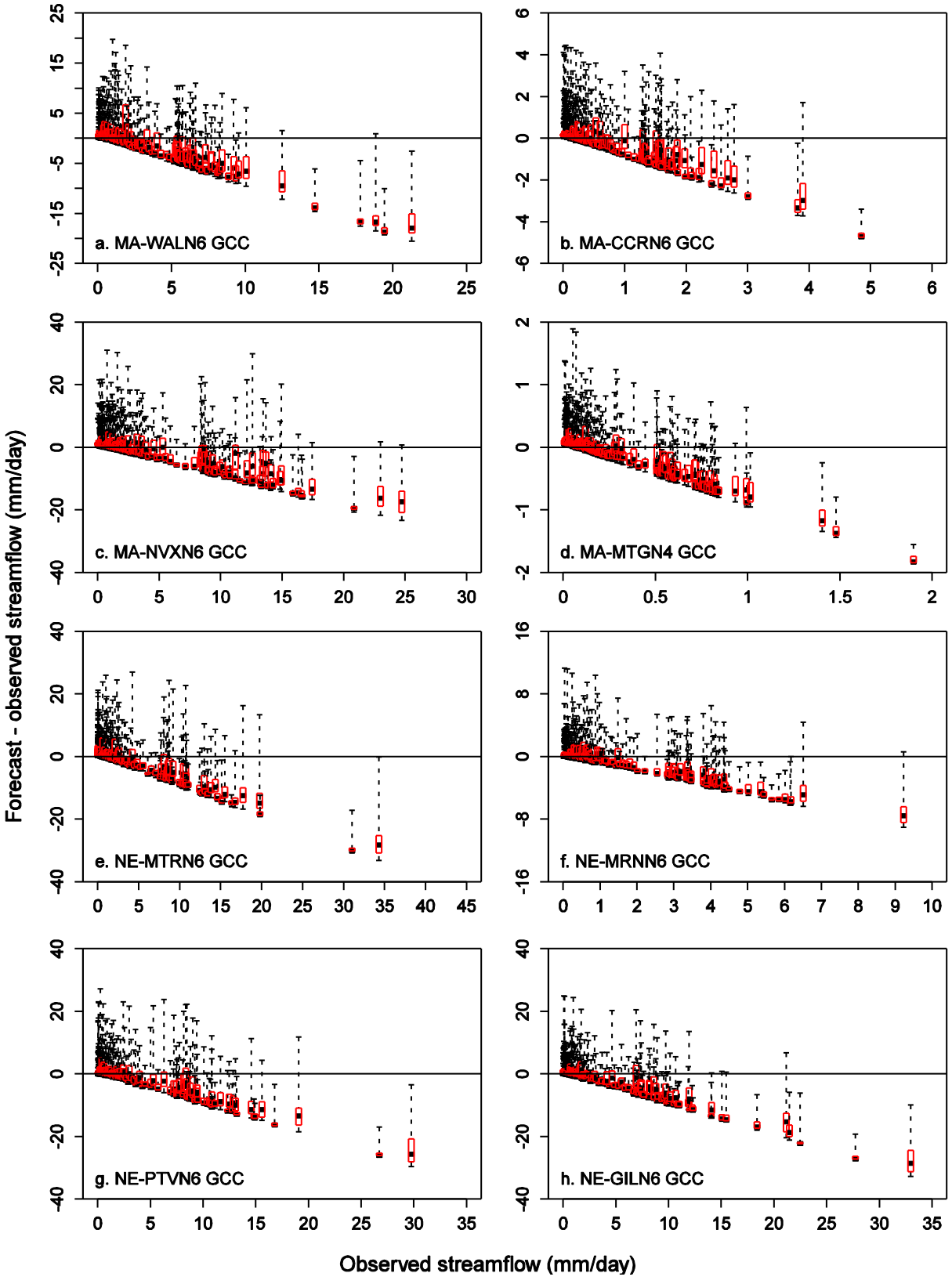

Figure 20: Box plots of errors (forecast - observed) in the MEFP-GCC streamflow forecasts. The results are shown for a forecast lead time of 95-100 days.

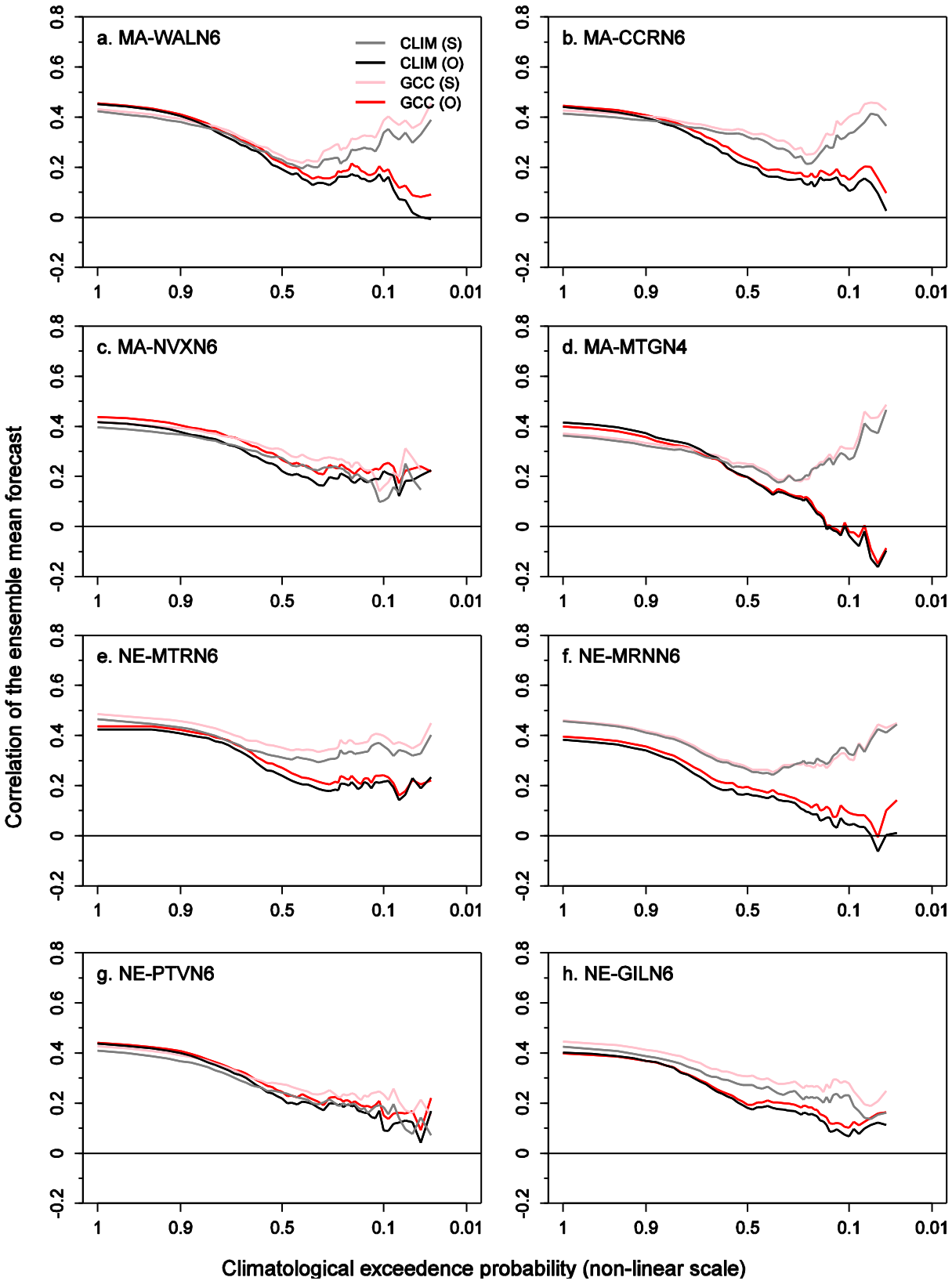

Figure 21: Correlation of the MEFP-CLIM and MEFP-GCC streamflow forecasts (ensemble mean) against observed (O) and simulated (S) flows. The results are shown for a forecast lead time of 95-100 days and for increasing streamflow thresholds. The thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

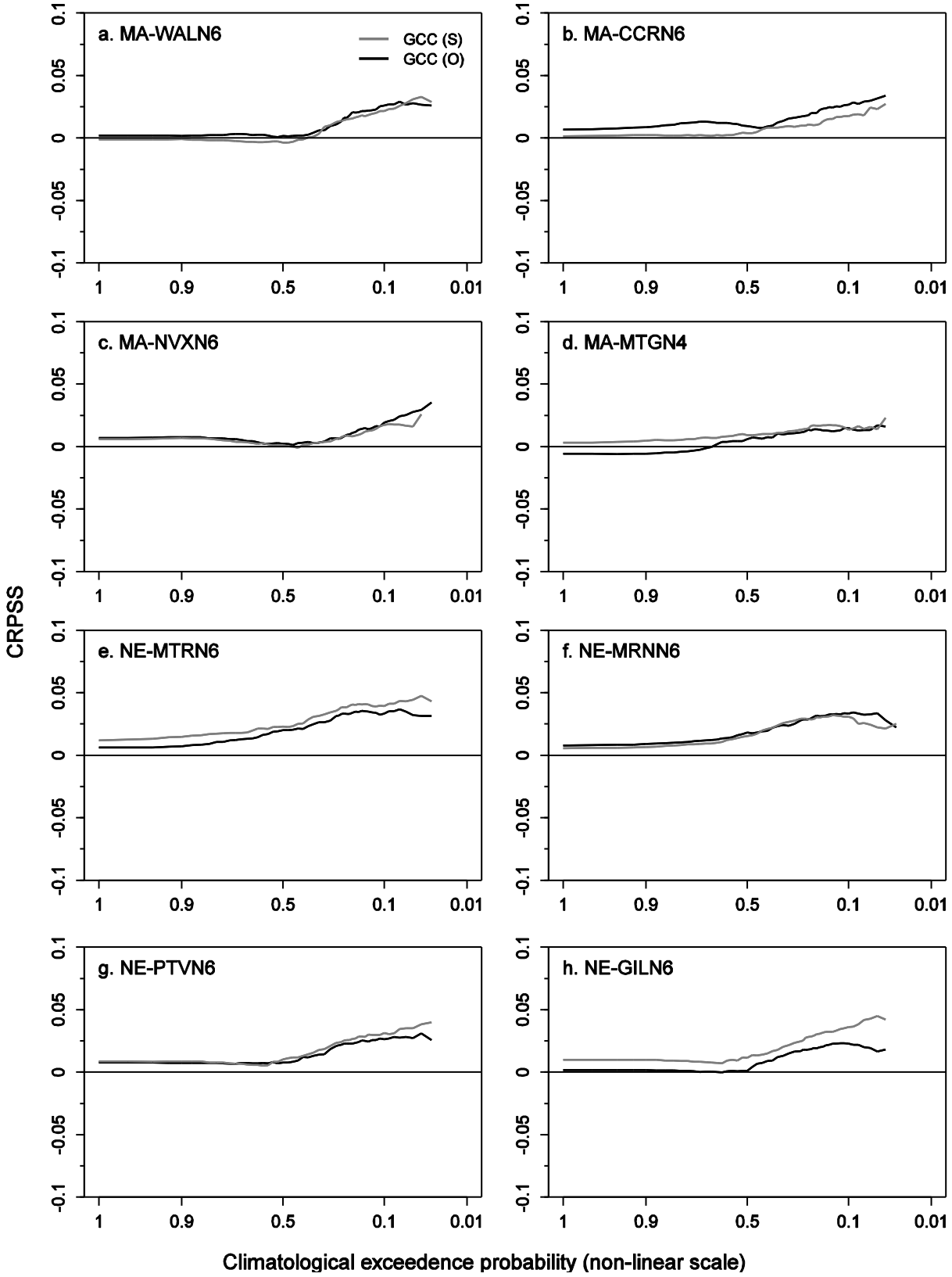

Figure 22: Mean CRPSS of the MEFP-GCC streamflow forecasts against those with MEFP-CLIM forcing using both observed (O) and simulated (S) flows. The results are shown for a forecast lead time of 95-100 days and for increasing streamflow thresholds. The thresholds are expressed as climatological probabilities and plotted on a probit scale (but labeled with actual probability).

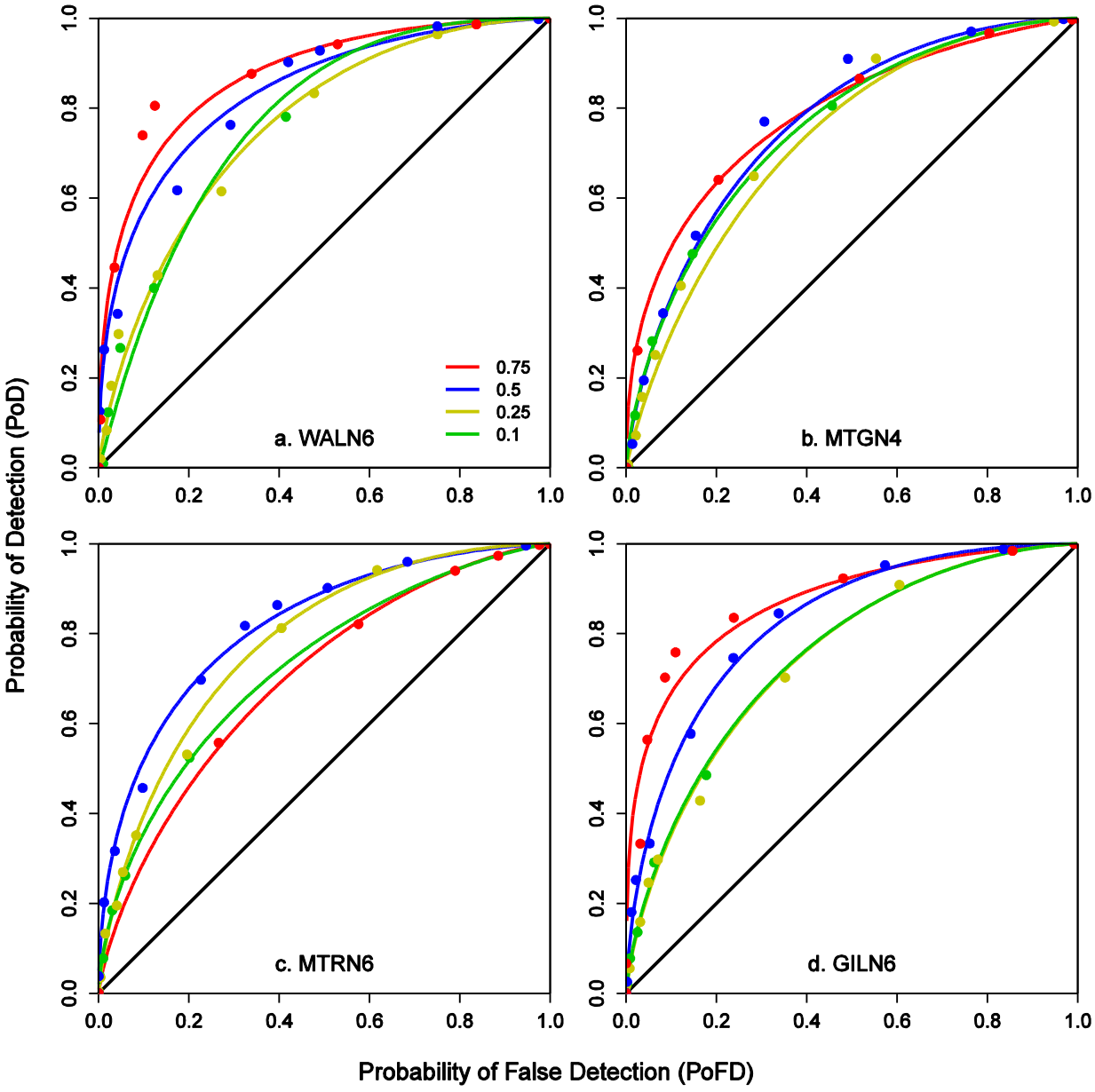

Figure 23: Relative Operating Characteristic (ROC) curves for an upstream basin and an outlet in each RFC. The ROC curves were fitted to the empirical points under an assumption of bivariate normality between the PoD and the PoFD. The results are shown at a forecast lead time of 95-100 days and for selected thresholds, which are denoted by their climatological probabilities of exceedence.

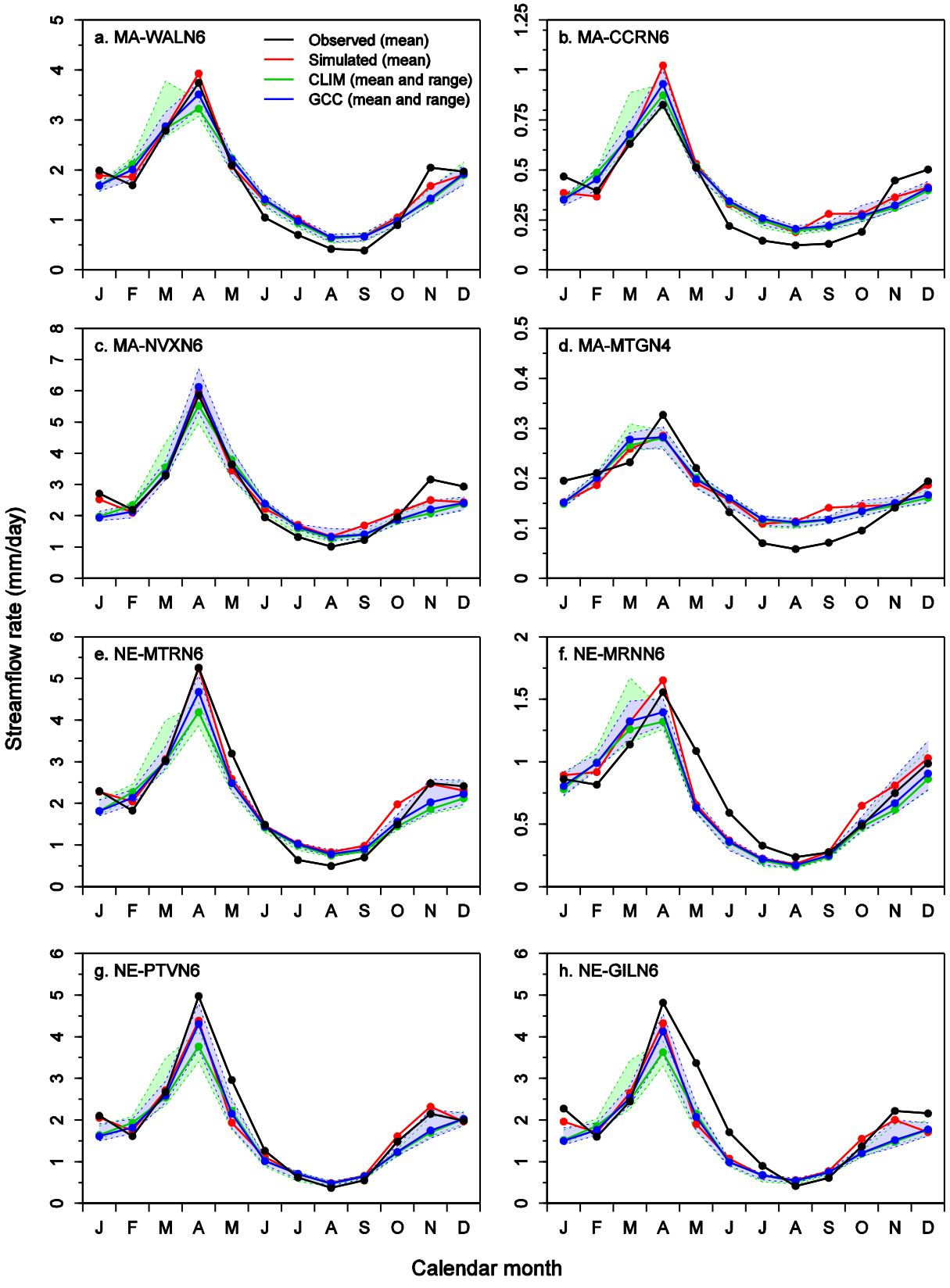

Figure 24: Average observed, simulated, and forecast streamflow by calendar month for each source of forcing from the MEFP (CLIM and GCC). The forecasts comprise the average of the ensemble means by calendar month across all forecast lead times, together with the range of the conditional averages by forecast lead time. The results are shown for basins in MARFC and NERFC.

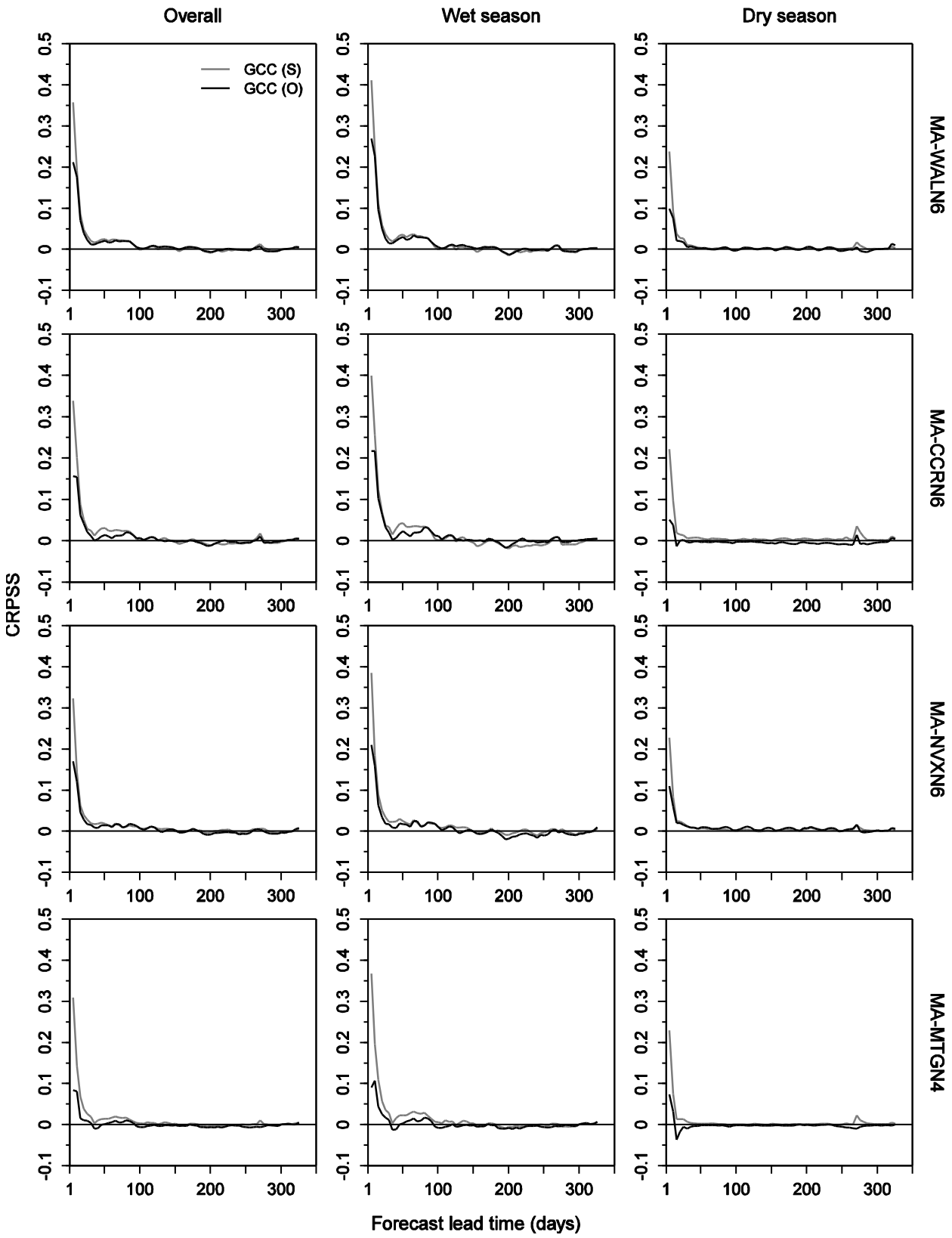

Figure 25a: Mean CRPSS of the MEFP-GCC streamflow forecasts in MARFC against those with MEFP-CLIM forcing using both observed (O) and simulated (S) flows. The results are shown for the overall period and for the "wet" and "dry" seasons separately.

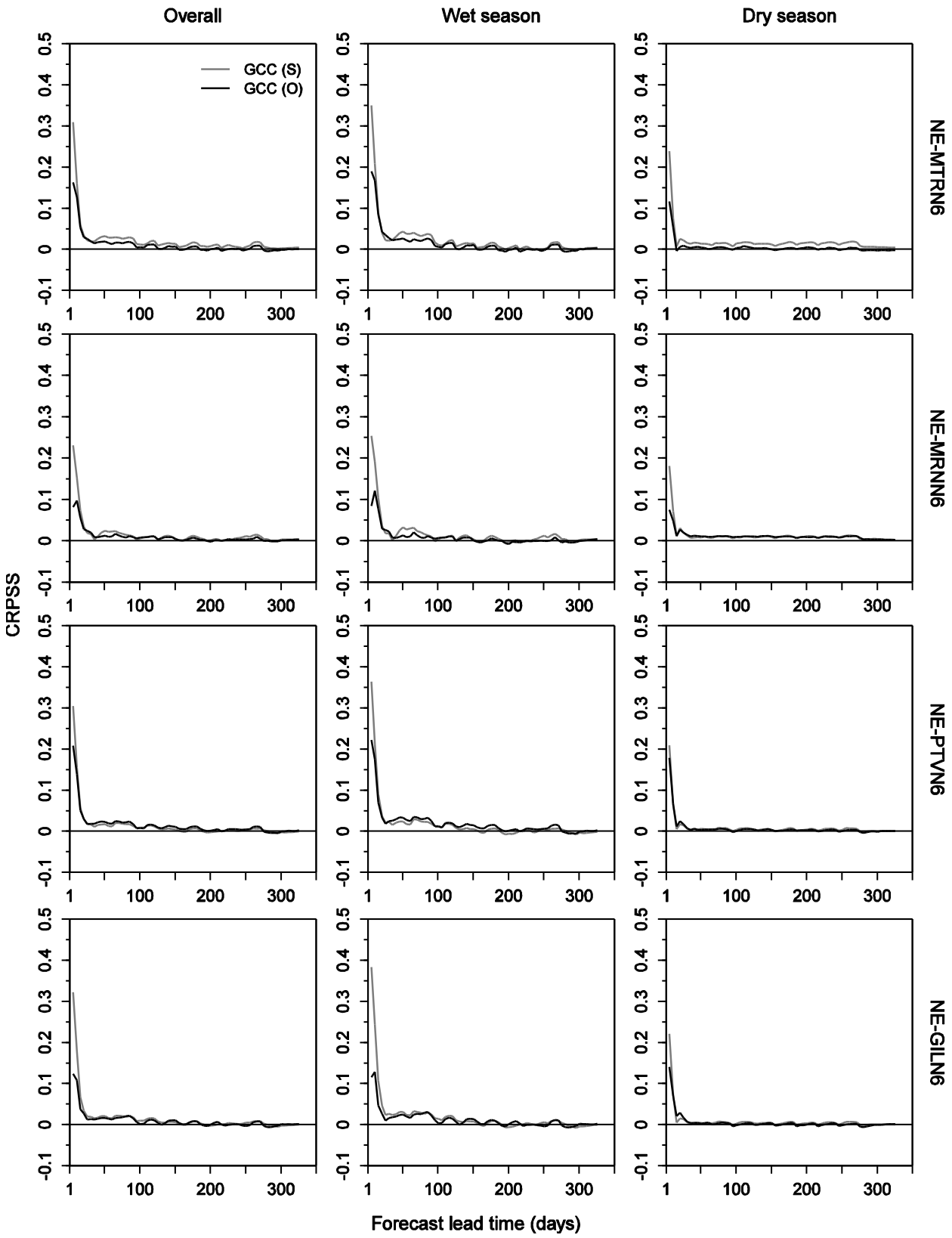

Figure 25b: Mean CRPSS of the MEFP-GCC streamflow forecasts in NERFC against those with MEFP-CLIM forcing using both observed (O) and simulated (S) flows. The results are shown for the overall period and for the "wet" and "dry" seasons separately.

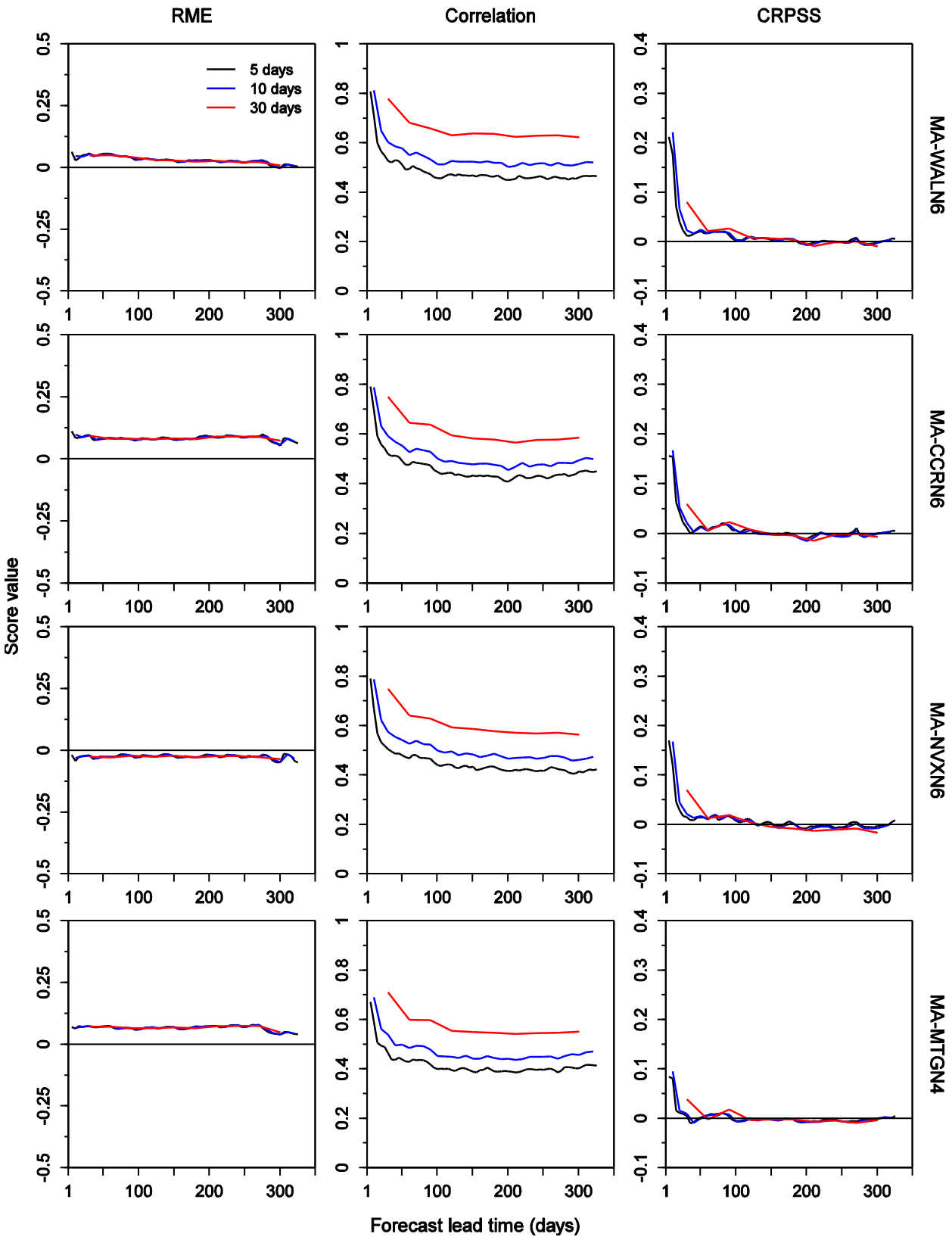

Figure 26a: Selected verification metrics for the MEFP-GCC streamflow forecasts at three aggregation periods (5-, 10- and 30-days) for basins in MARFC.

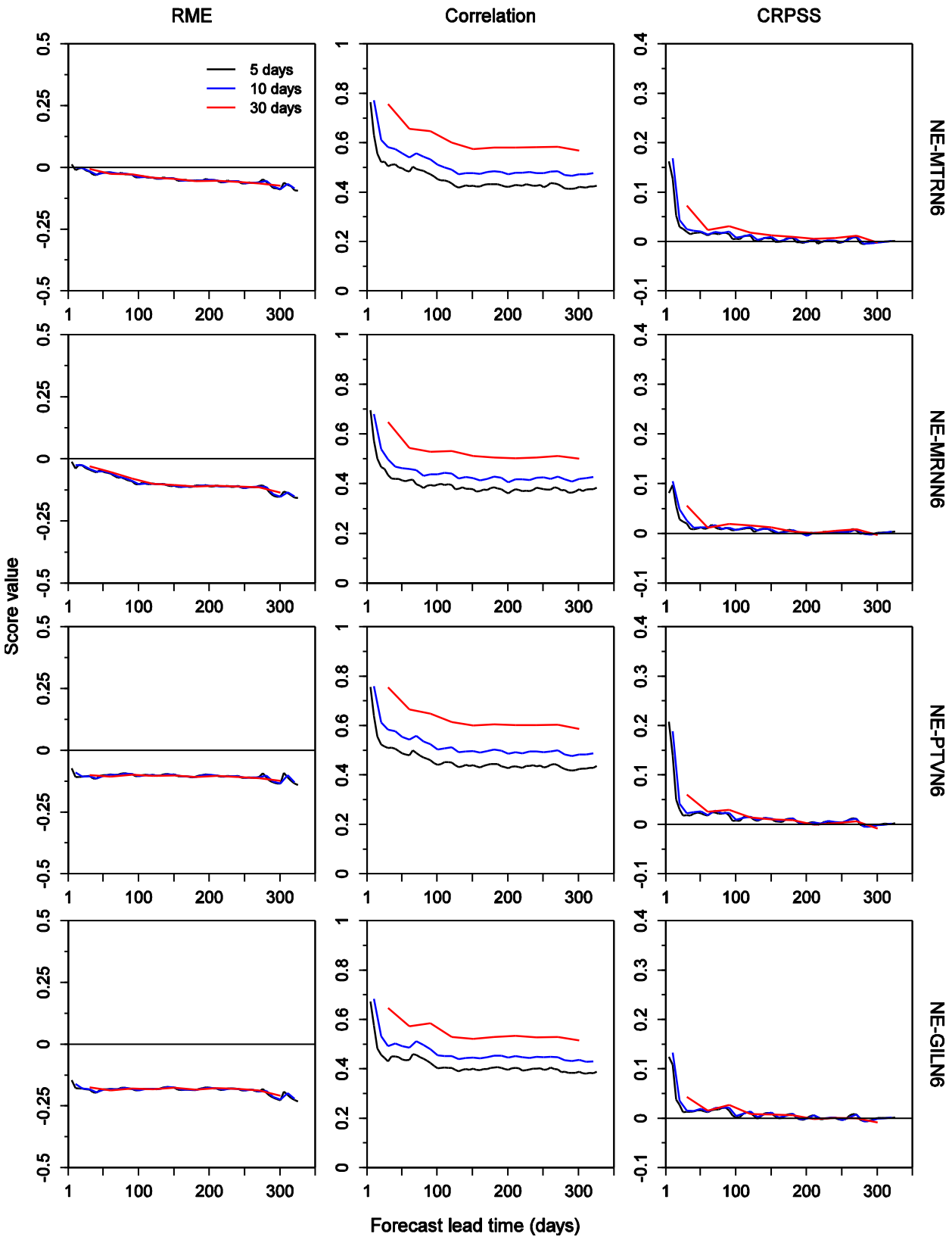

Figure 26b: Selected verification metrics for the MEFP-GCC streamflow forecasts at three aggregation periods (5-, 10- and 30-days) for basins in NERFC.

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,

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

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. Since $\bar{F}_{Y_k}(q)$ is either zero or one, the Type-II CB can only be zero if the forecasts are perfectly sharp. Conditionally upon the observed outcome, the BSS is given by,

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

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

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

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, $\overline{CRPS}_{REF}$

$$CRPSS = \frac{\overline{CRPS}_{REF} - \overline{CRPS}}{\overline{CRPS}_{REF}}. \tag{B8}$$

d. Relative Operating Characteristic

The Relative Operating Characteristic (ROC; Green and Swets, 1966) measures the ability of a forecasting system to correctly predict the occurrence of an event (Probability of Detection or PoD) while avoiding too many incorrect forecasts when it does not occur (Probability of False Detection or PoFD). For probability forecasts, this trade-off is expressed as a probability threshold, d, at which the forecast triggers a decision. The ROC plots the PoD versus the PoFD for all possible values of d in [0,1]. For a particular threshold, the empirical PoD is

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

where I denotes the indicator function. The empirical PoFD is

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

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

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

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

APPENDIX C: Event-based analysis of the streamflow forecasts

Paired streamflow forecasts and observations are presented for selected years in each basin. The results comprise the raw streamflow forecasts with forcing inputs from the MEFP-GCC and MEFP-CLIM. The plots include the single-valued streamflow observations and simulations, 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 18-138 hours, 2298-2418 hours, 4698-4818 hours and 7098-7218 hours and for calendar years 1986 and 1996. 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 above.

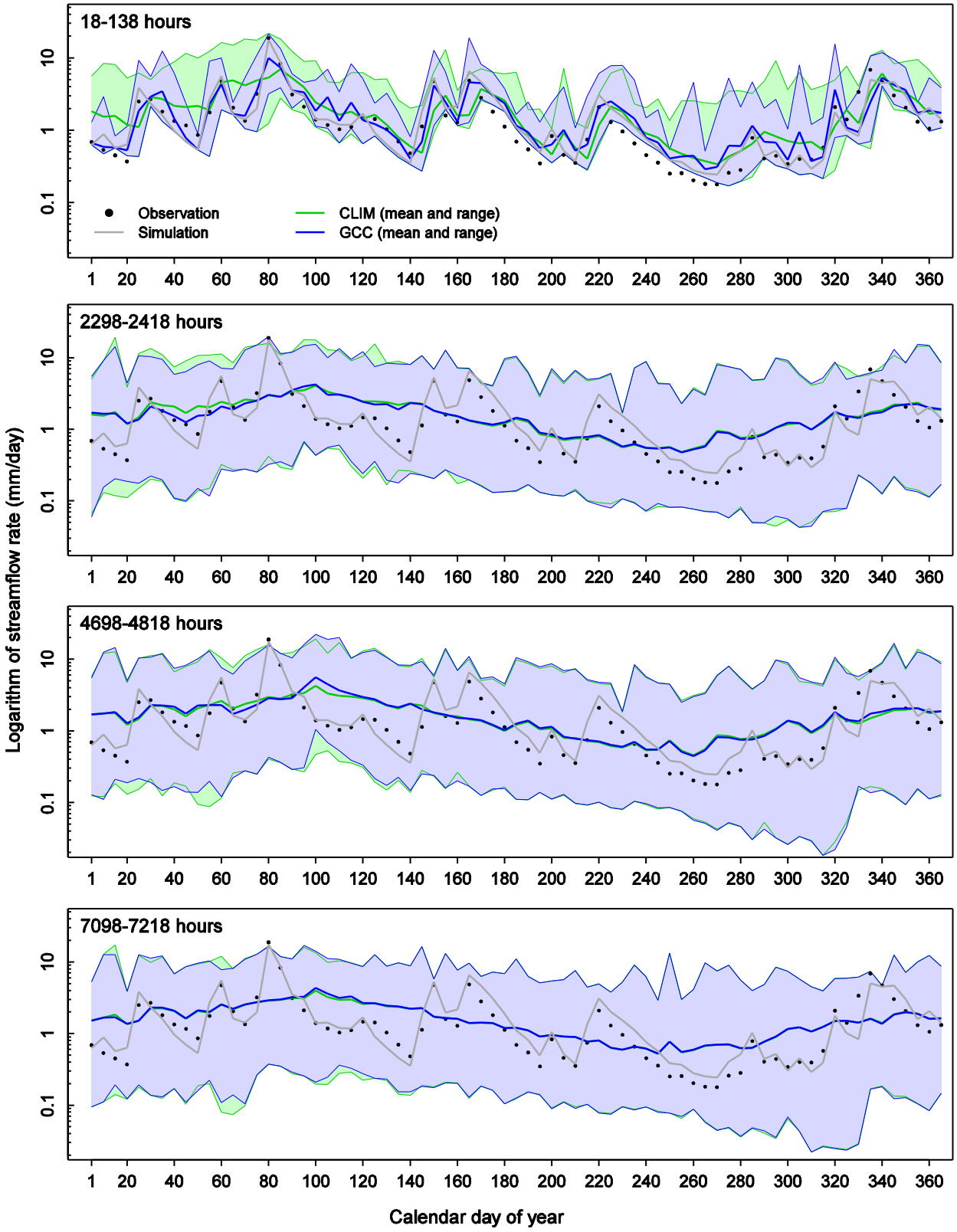

Figure C01: Mean and range of the streamflow forecasts in WALN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

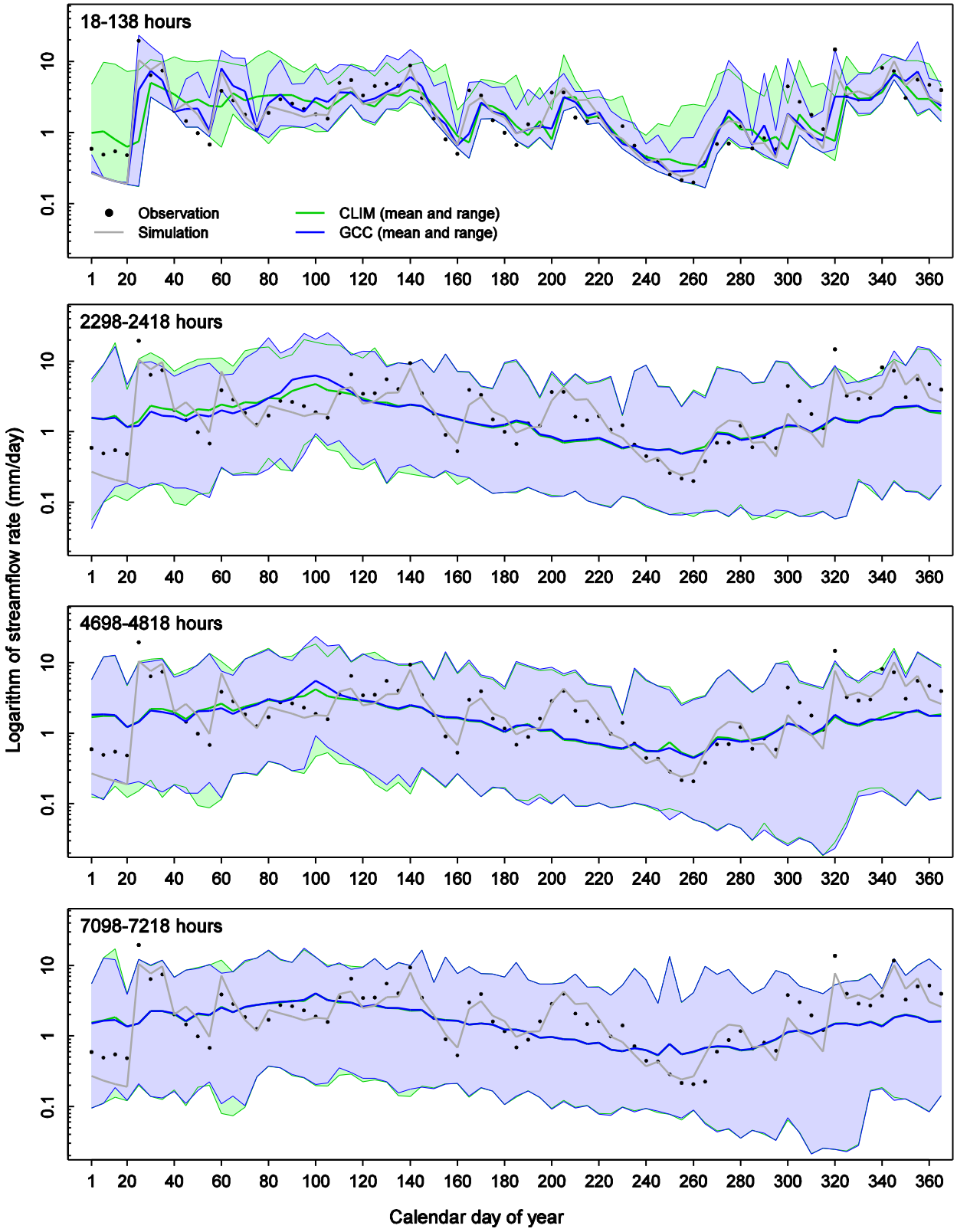

Figure C02: Mean and range of the streamflow forecasts in WALN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

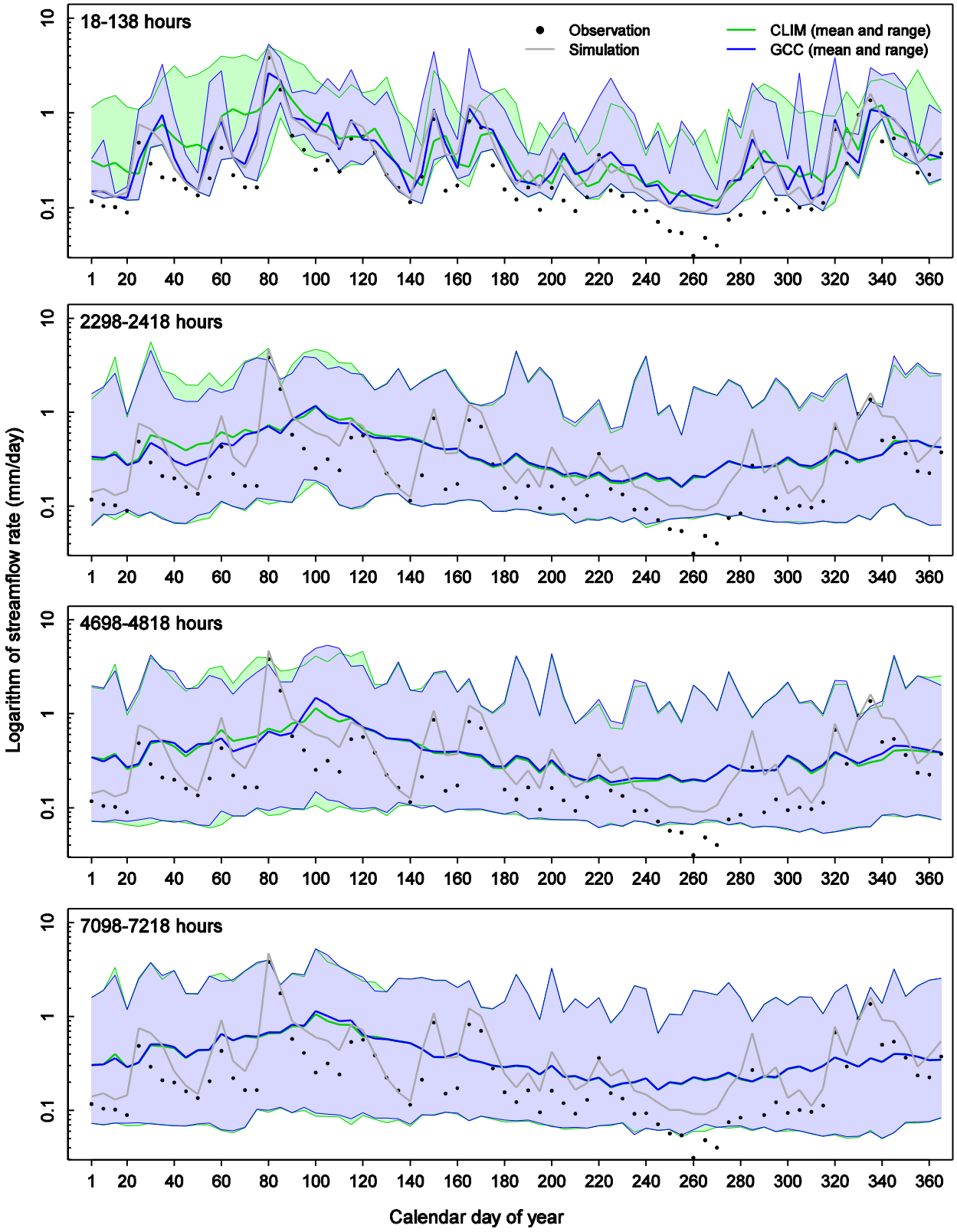

Figure C03: Mean and range of the streamflow forecasts in CCRN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

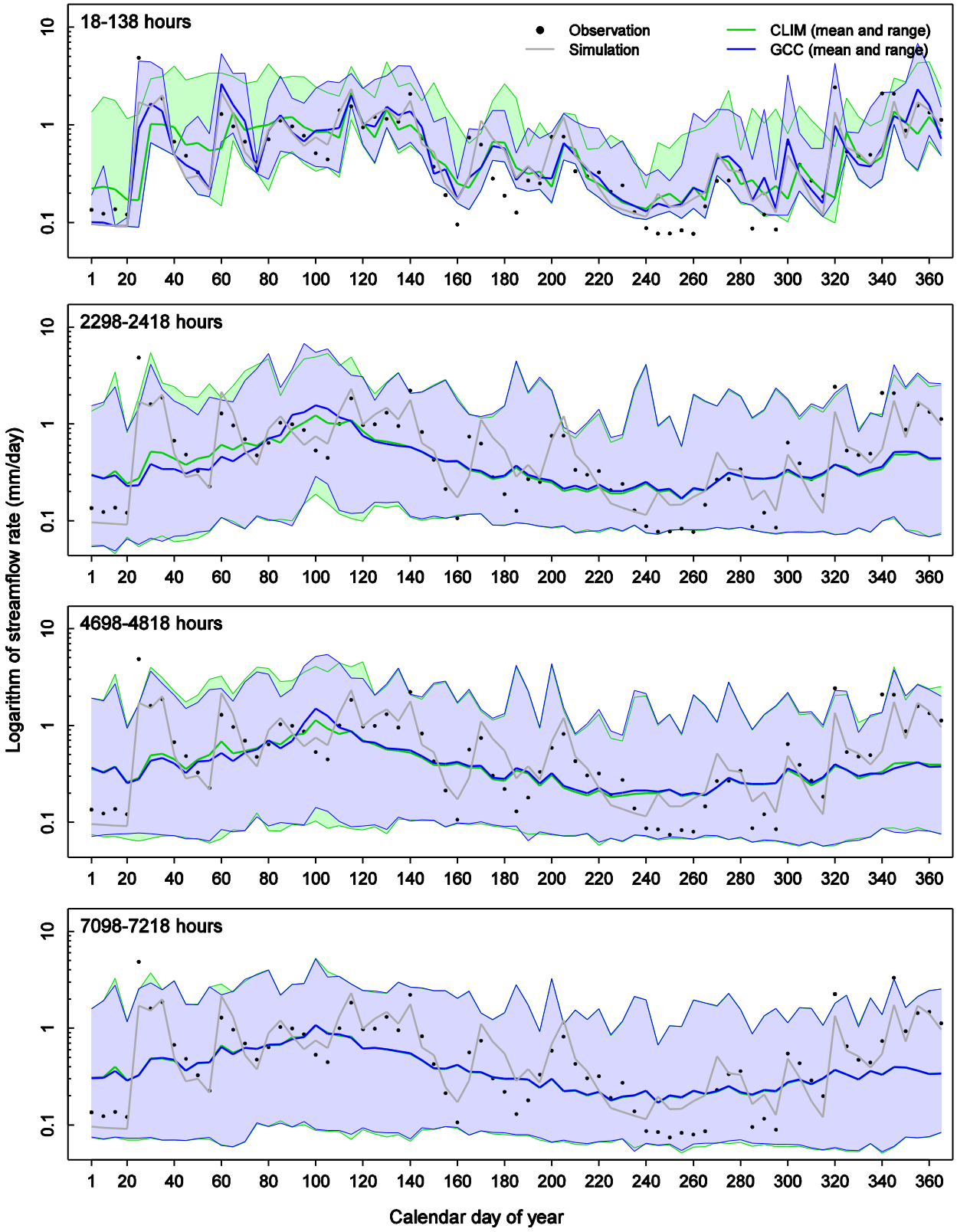

Figure C04: Mean and range of the streamflow forecasts in CCRN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

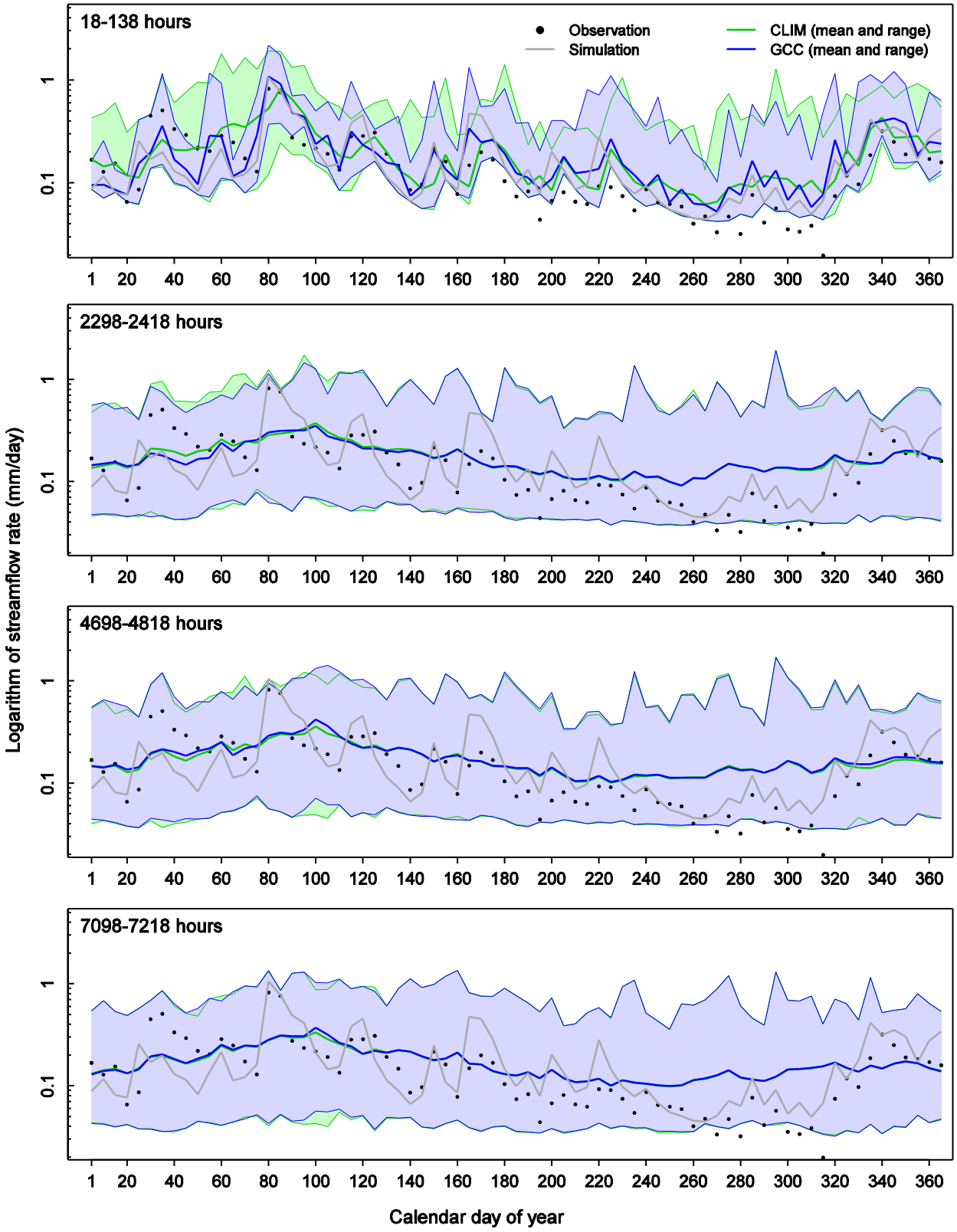

Figure C05: Mean and range of the streamflow forecasts in MTGN4. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

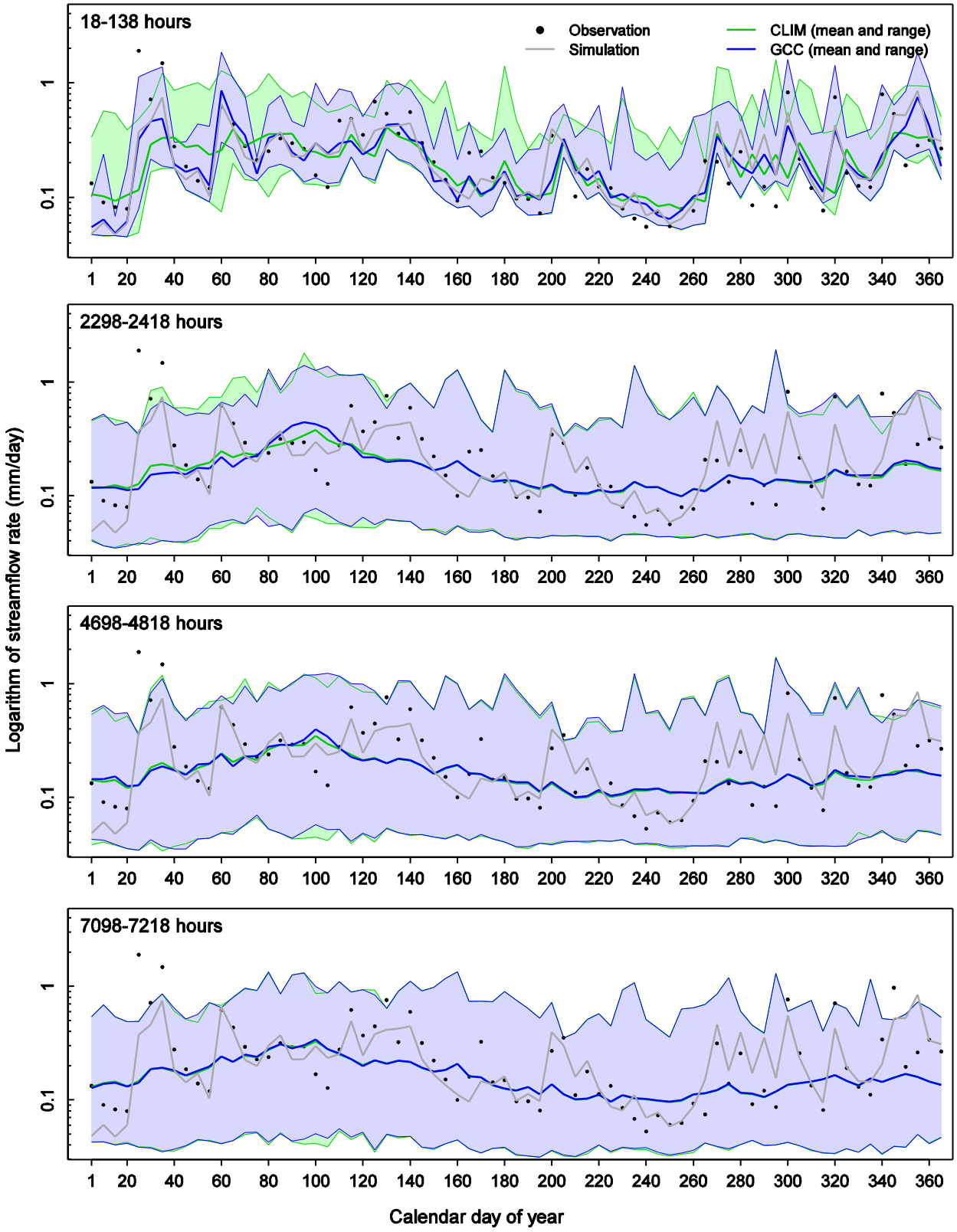

Figure C06: Mean and range of the streamflow forecasts in MTGN4. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

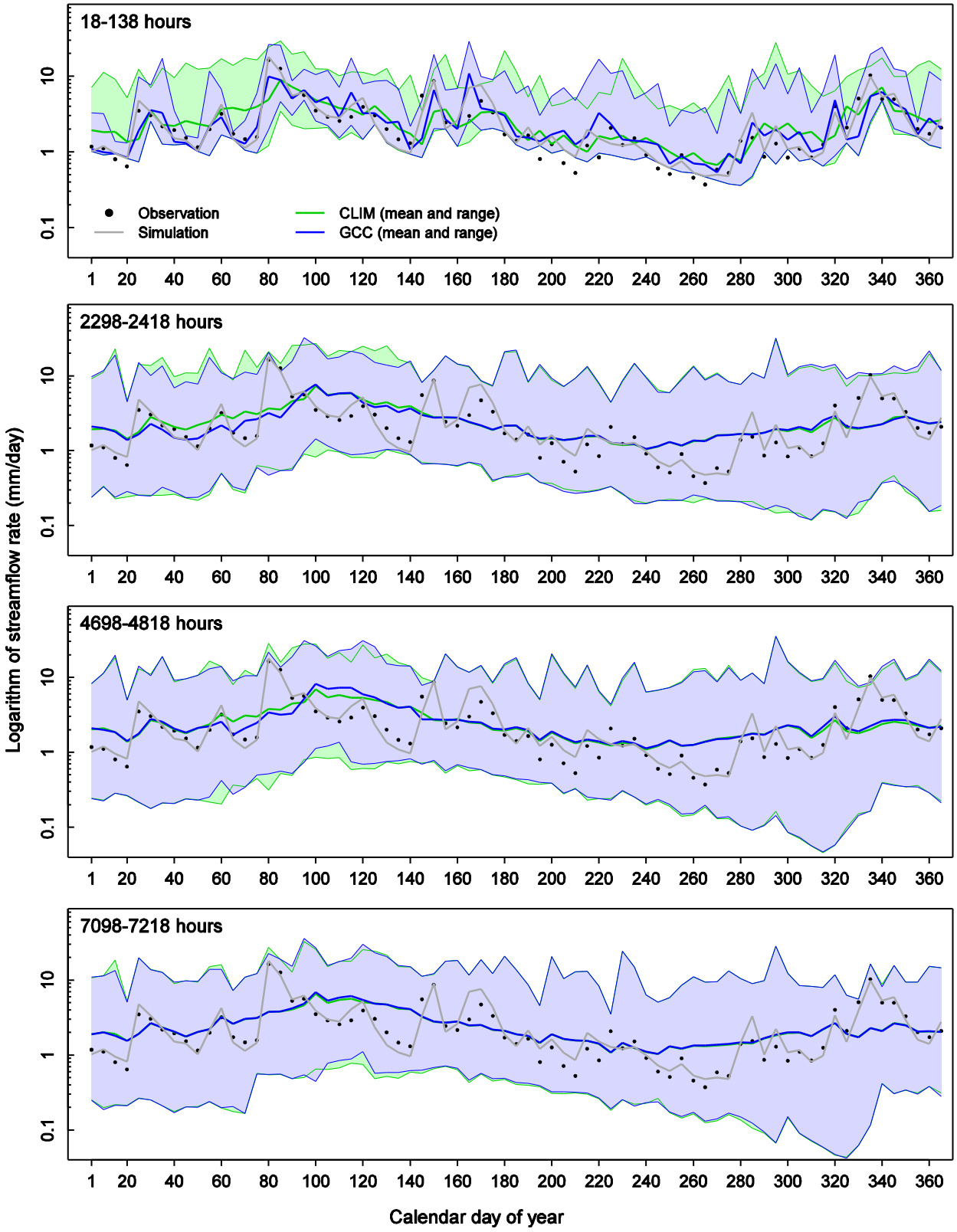

Figure C07: Mean and range of the streamflow forecasts in NVXN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

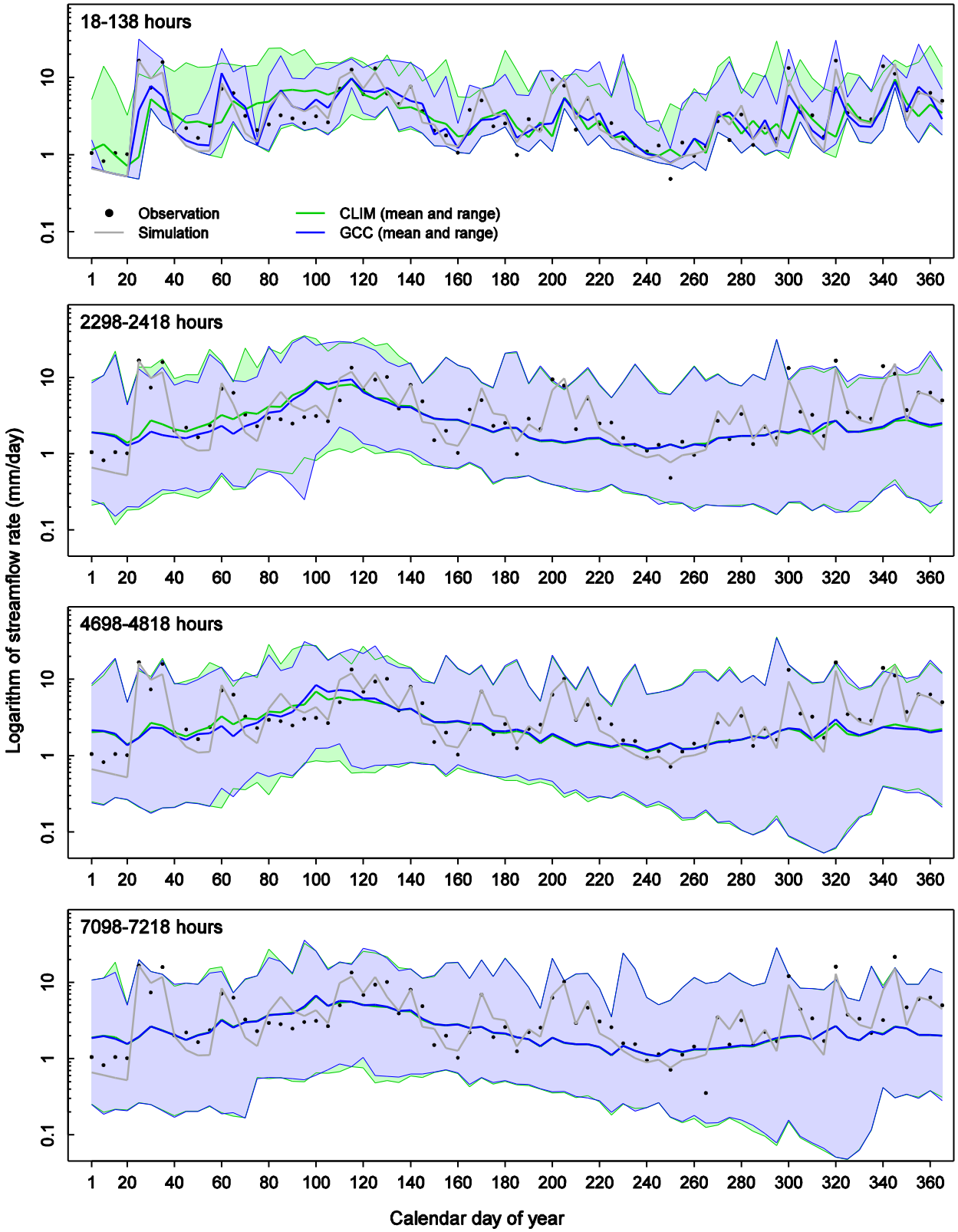

Figure C08: Mean and range of the streamflow forecasts in NVXN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

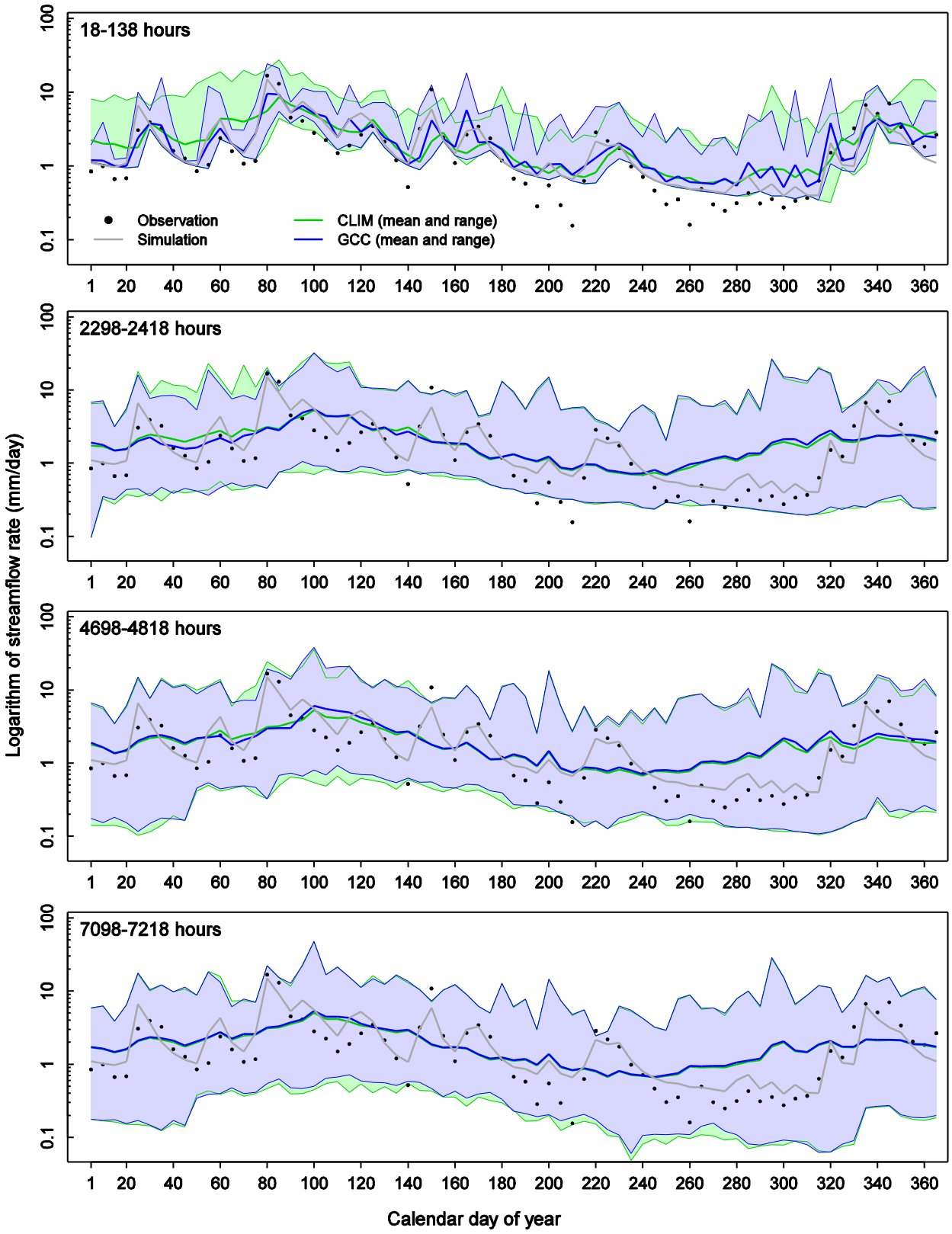

Figure C09: Mean and range of the streamflow forecasts in MTRN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

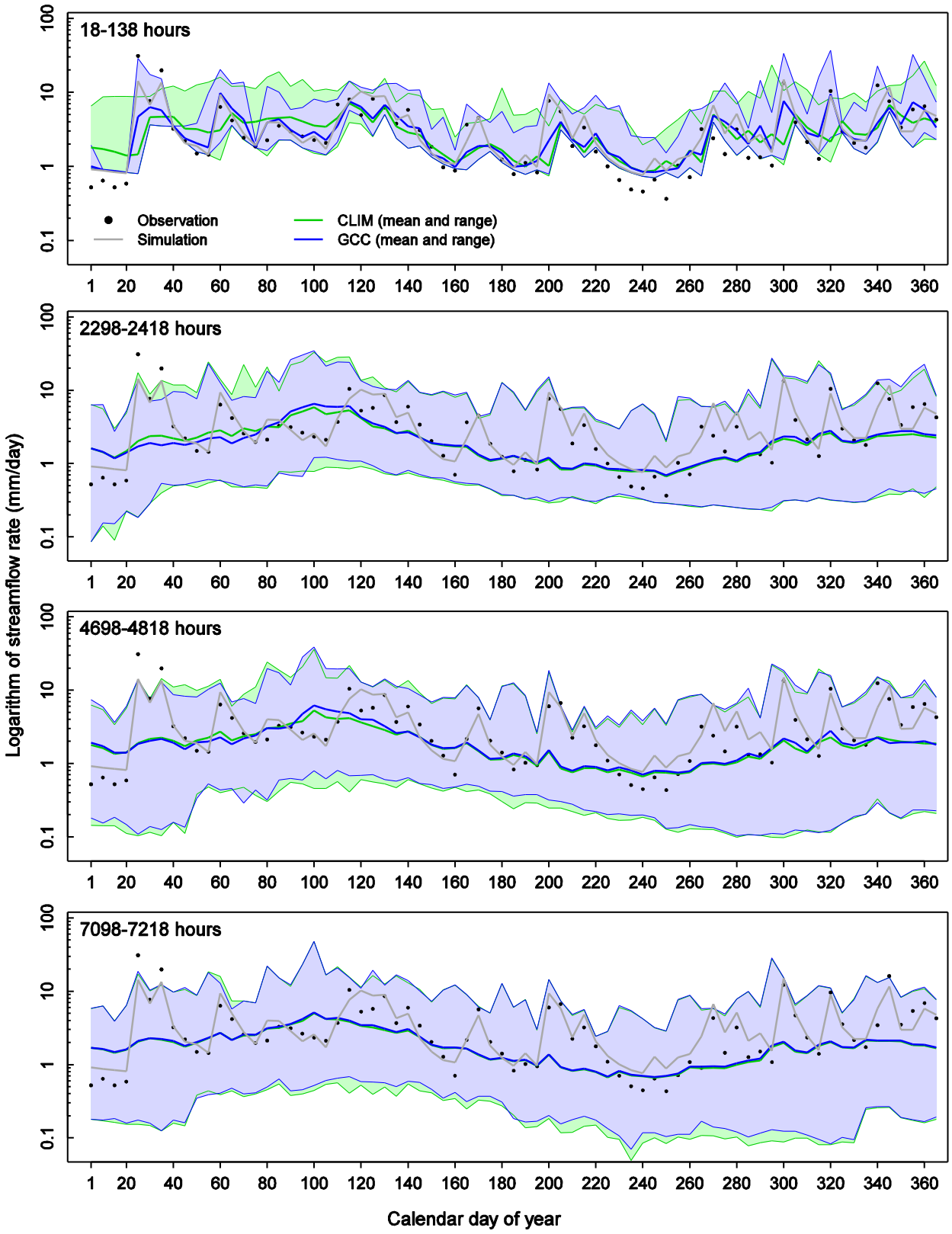

Figure C10: Mean and range of the streamflow forecasts in MTRN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

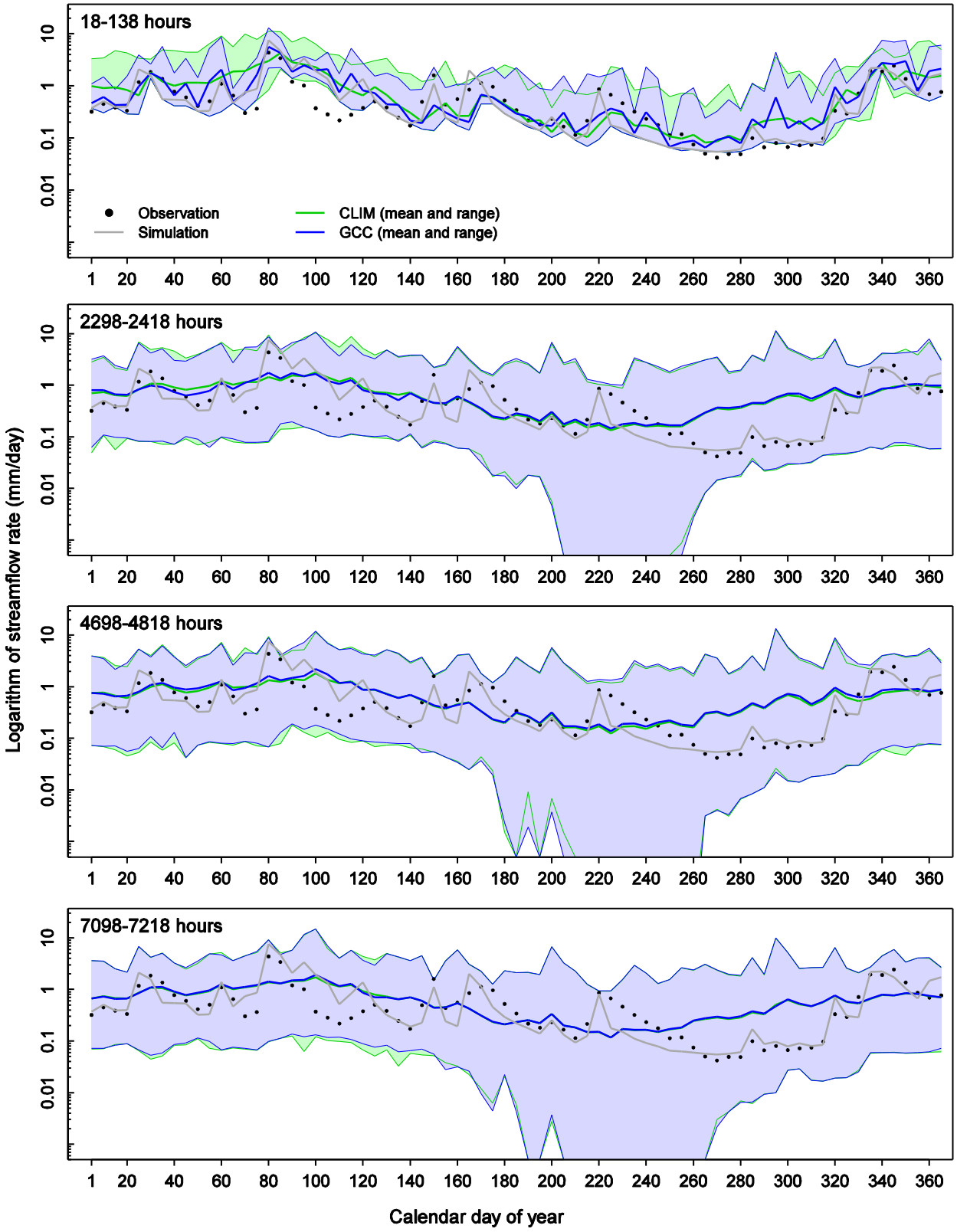

Figure C11: Mean and range of the streamflow forecasts in MRNN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

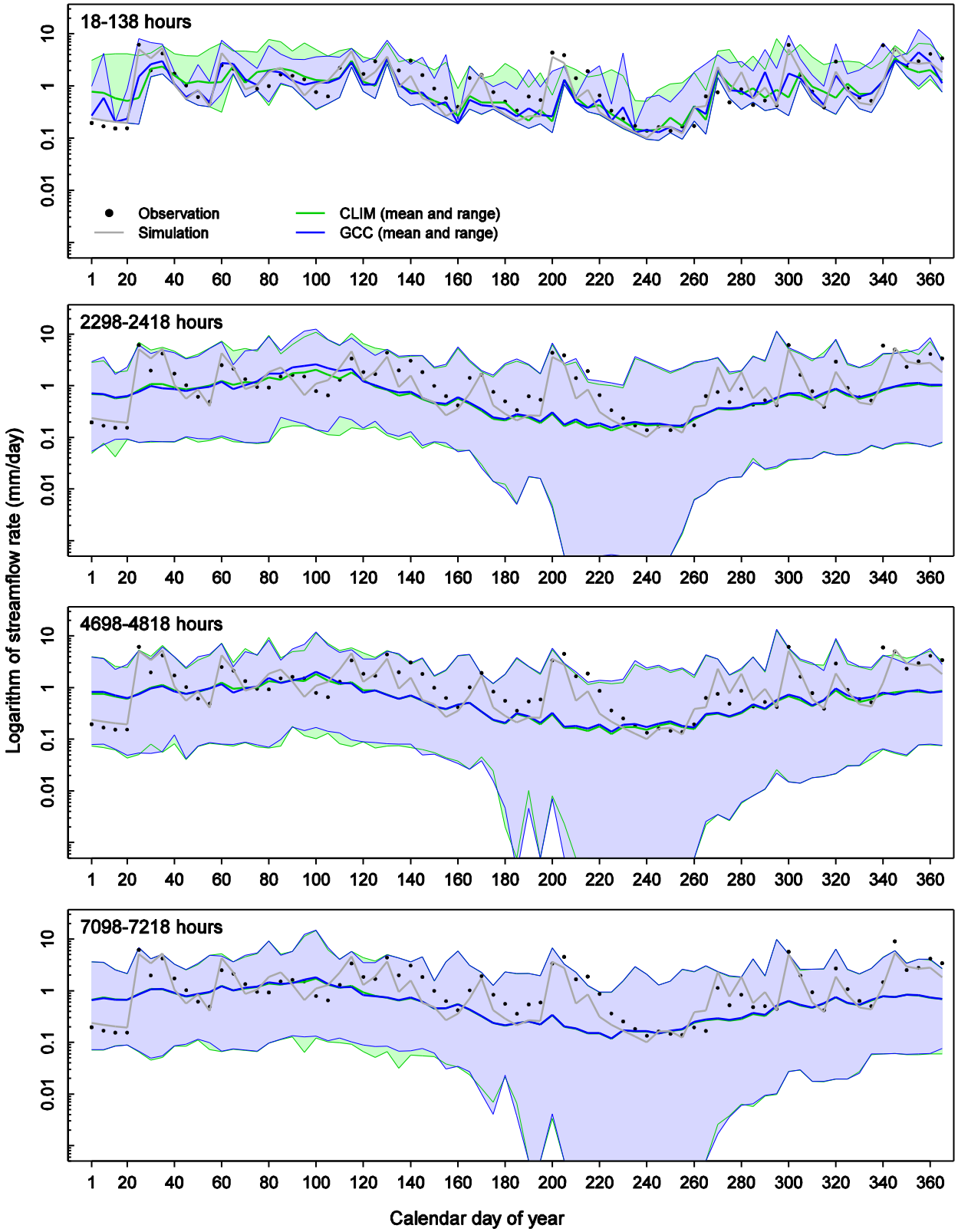

Figure C12: Mean and range of the streamflow forecasts in MRNN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

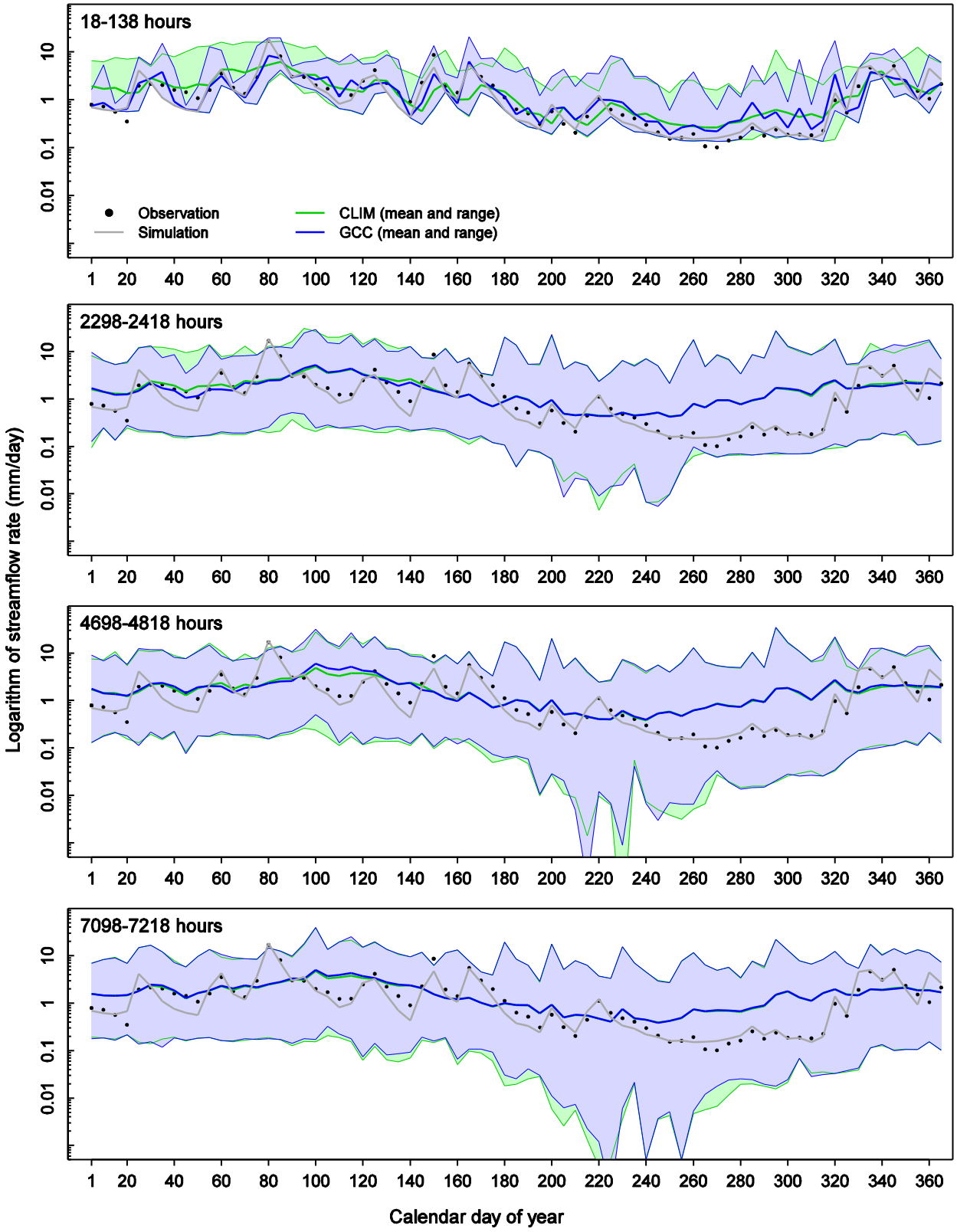

Figure C13: Mean and range of the streamflow forecasts in PTVN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

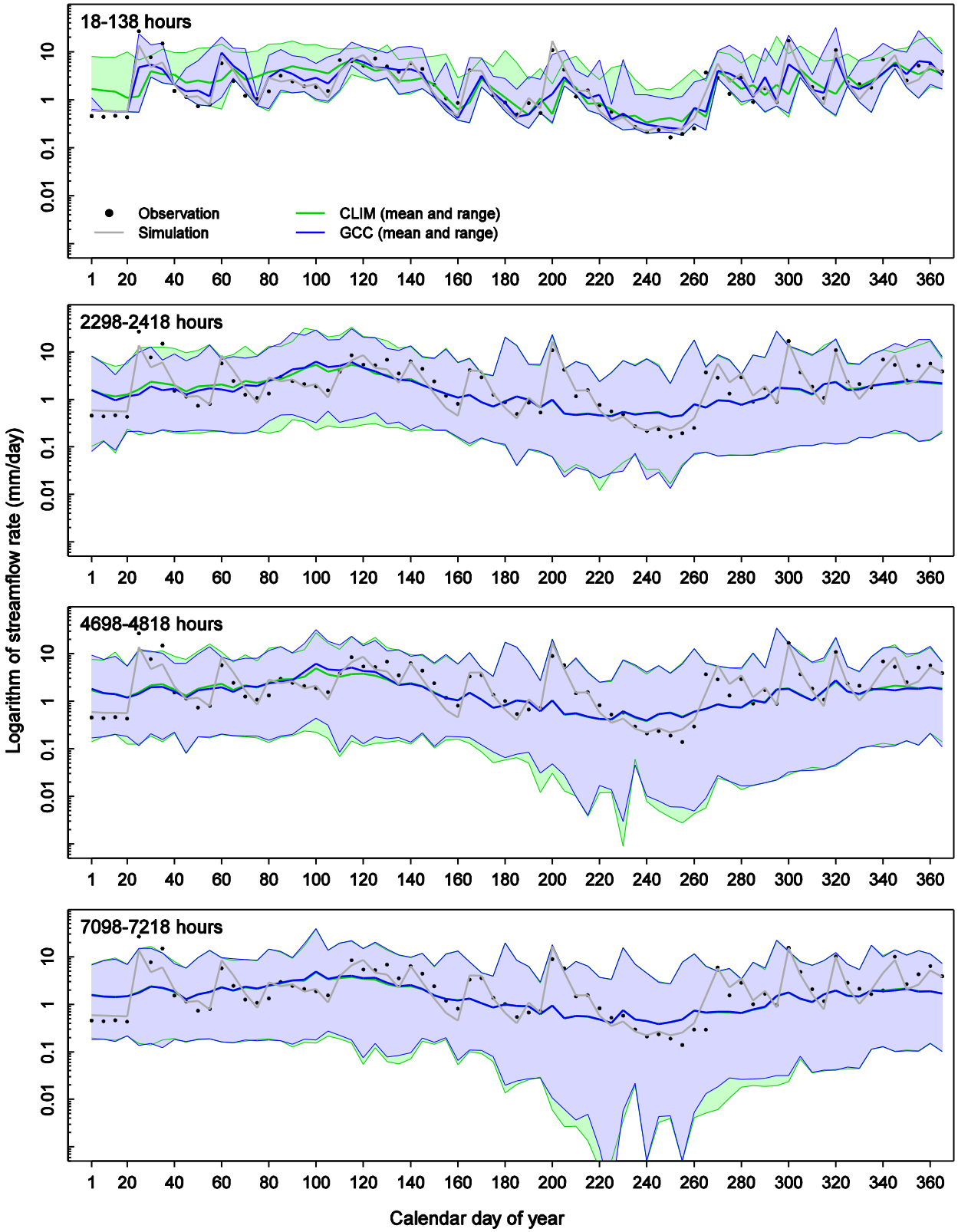

Figure C14: Mean and range of the streamflow forecasts in PTVN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

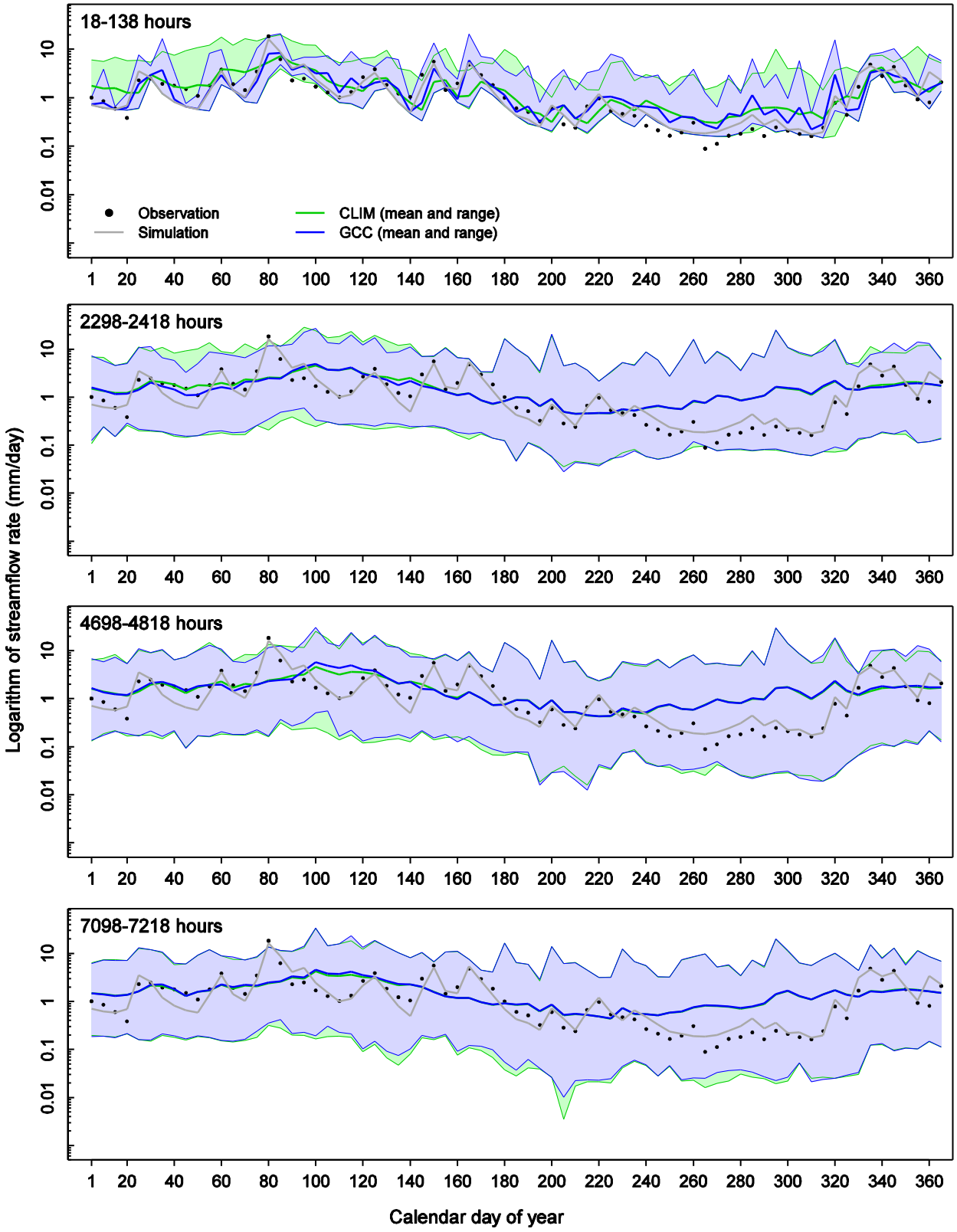

Figure C15: Mean and range of the streamflow forecasts in GILN6. The results are shown by forecast valid date in 1986 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.

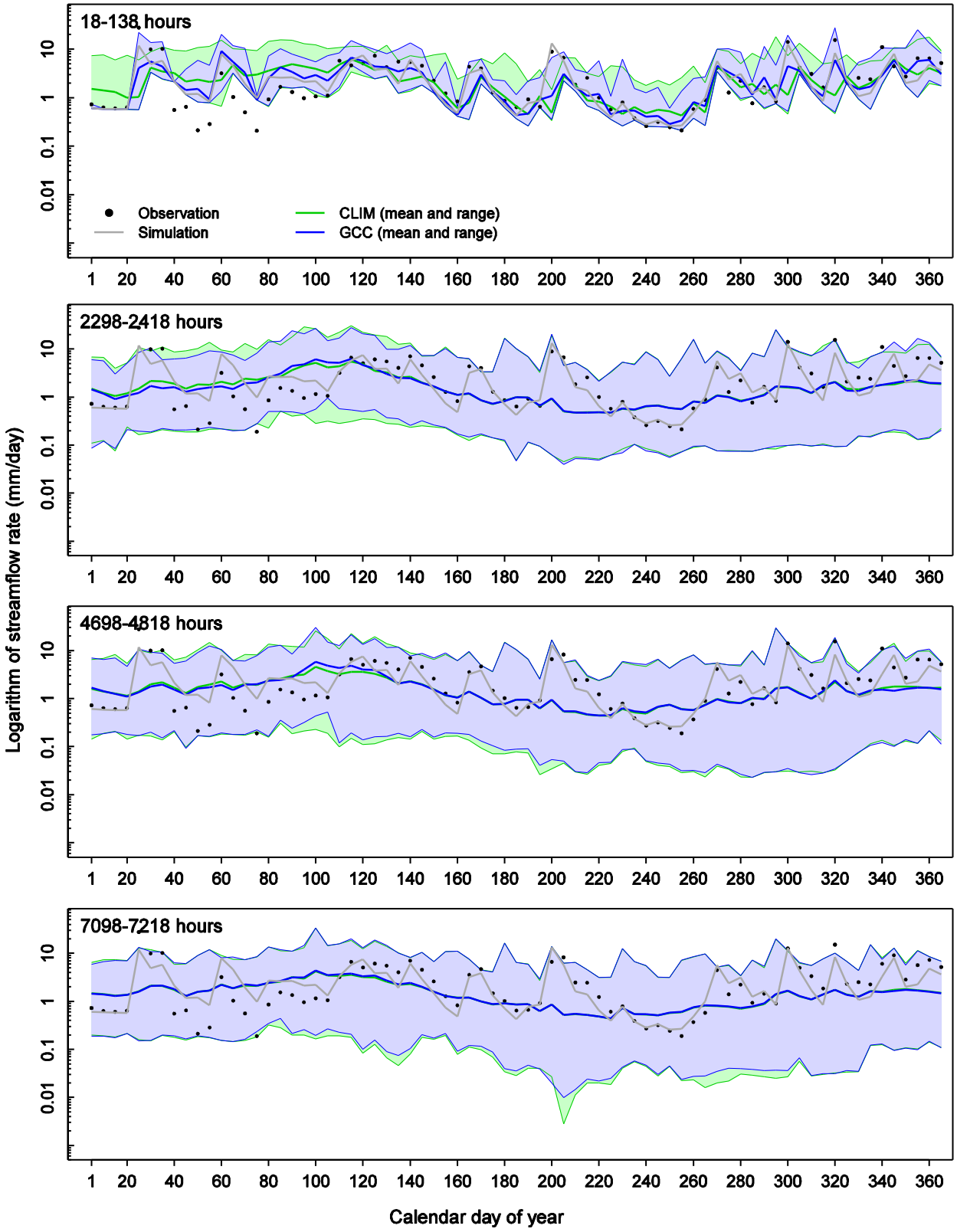

Figure C16: Mean and range of the streamflow forecasts in GILN6. The results are shown by forecast valid date in 1996 and for selected forecast lead times. The raw streamflow forecasts comprise forcing from the MEFP with resampled climatology for 1-330 days (CLIM), together with GEFS (1-15 days), plus CFSv2 (16-270 days), plus CLIM (271-330 days), which is denoted GCC.