An evaluation of the minimum requirements for meteorological reforecasts from the Global Ensemble Forecast System (GEFS) of the U.S. National Weather Service (NWS) in support of the calibration and validation of the NWS Hydrologic Ensemble Forecast Service (HEFS)

Revision number: extended, final

Prepared by Hydrologic Solutions Limited for the U.S. National Weather Service (NWS) under subcontract SubK-2013-1003 with Lynker Technologies LLC in support of NWS Prime Contract DG133W-13-CQ-0042 (in fulfillment of Deliverable No. 4 and, in part, Deliverable No. 9 of Task 3)

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

August 2015

HSL

Contents

i. List of figures

ii. List of tables

1. Executive summary

Motivation

- The Hydrologic Ensemble Forecast Service (HEFS) quantifies the total uncertainty in future streamflow as a combination of the meteorological forcing uncertainties and the hydrologic modeling uncertainties. Reliable and skillful weather and climate forecasting is central to reliable and skillful streamflow forecasting. The HEFS Meteorological Ensemble Forecast Processor (MEFP) quantifies the meteorological uncertainties and corrects for biases in the forcing inputs to the HEFS. For the medium-range (1-15 days), the MEFP uses precipitation and temperature forecasts from the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP).
- The ability of the HEFS to provide useful information for decision making depends upon the accuracy of the forecast probabilities. Crucially, there is a need to demonstrate this accuracy through hindcasting and validation. Hindcasting is necessary to benchmark and improve the HEFS, optimize decision support systems that rely upon the HEFS, and to build confidence among decision makers that the forecasts are accurate, useful, and can lead to better decisions. For example, the New York City Department of Environmental Protection (NYCDEP) is using the HEFS to improve the management of risks to water supply objectives in the NYC area. The NYCDEP has developed an Operational Support Tool (OST), which optimizes the quantity and quality of water stored in the NYC reservoirs and helps to avoid unnecessary, multi-billion dollar, infrastructure costs. The NYCDEP relies on streamflow hindcasts from the HEFS, supported by meteorological reforecasts from the GEFS, in order to optimize and validate the OST.
- Large and extreme hydrologic events are critically important to users of the HEFS, as they pose a significant threat to life and property. Given the manifest uncertainties in forecasting hydrologic extremes, the ability of the HEFS to quantify these uncertainties (and correct for systematic biases) is an important advantage over deterministic forecasting systems. However, validating the HEFS for large and extreme events relies upon an adequate archive of meteorological reforecasts.
- In order to determine the minimum requirement of the HEFS for meteorological reforecasts from the GEFS, this report considers the sensitivity of the HEFS to a limited number of reforecast configuration options. Understanding the minimum requirements for calibrating and validating the HEFS is a *necessary but not a sufficient condition* for understanding the minimum requirements of end users for meteorological and hydrologic reforecasts. The requirements of end users, such as the NYCDEP, will be gathered and presented separately.

Approach

- In order to determine the minimum requirements for meteorological reforecasting in support of calibrating and validating the HEFS, a 26-year reforecast dataset was obtained for the current GEFS. Among other factors, the costs associated with meteorological reforecasting depend on the historical period considered (N years), the interval between reforecasts (M days), and the number of ensemble members in each forecast (C). By sub-sampling the GEFS reforecasts, the MEFP was calibrated for different combinations of N, M and C. The sensitivities of the temperature and precipitation forecasts from the MEFP and the streamflow forecasts from the HEFS were then explored through hindcasting and validation.

- Forcing and streamflow hindcasts were produced and validated at four headwater basins: the Chikaskia River at Corbin, Kansas (AB-CBNK1); the Dolores River at Rico in Colorado (CB-DRRC2); the Middle Fork of the Eel River at Dos Rios in California (CN-DOSC1); and the Wood River at Hope Valley, Rhode Island (NE-HOPR1). The hindcasts were generated at 12Z for each day in the historical period of record. Within this fixed period, the calibration of the MEFP varied according to N, M and C. To ensure that the hindcasting was both practical and statistically reasonable, a combination of dependent and (limited) cross-validation was used.

- In exploring the sensitivities to N, a 24-year validation period was sub-divided into smaller calibration and forecasting periods, namely N={2x12, 3x8, 4x6, and 6x4} years. Dependent validation involved calibrating the MEFP and generating hindcasts for each sub-period and then pooling all of the sub-periods for validation. Independent validation involved borrowing the parameters from an adjacent sub-period. While dependent validation may be regarded as a best case scenario for the expected forecast quality, using parameters from adjacent sub-periods should be regarded as a worst case scenario; in practice, the MEFP would be recalibrated more frequently. In evaluating the sensitivities to M, the MEFP was calibrated for M={1, 3, 5, and 7} days and hindcasts produced daily for the fixed historical period.

- The sensitivities to C were examined by calibrating the MEFP with an ensemble mean derived from C={1, 5, and 11} of the ensemble members from the GEFS reforecasts. In practice, the GEFS reforecasts contain fewer ensemble members (F=11) than the operational forecasts (F=21). Since the operational HEFS forecasts use all available GEFS members (F=21), the HEFS reforecasts were also generated with all available GEFS members (F=11). However, in order to understand the impacts of this discrepancy, a baseline reforecast was generated with the control run only (F=1), using the corresponding MEFP calibration (C=1).

- The minimum requirements for validating the HEFS depend on N and M (not C) are were examined both theoretically and empirically. Theoretically, verification is concerned with the estimation of statistical measures. The quality of these estimates will depend on the number of samples available and their unique information content. Empirically, the effects of reducing the number of reforecasts available is to increase the sampling uncertainty of the verification results and to render some (typically large) events unverifiable, depending on the choice of measure. In order to illustrate the effects of N and M on the uncertainties associated with validating the HEFS, each sub-sample of N and M was verified separately and the results compared to the nominal scores for N=24 and M=1.

Results

- In terms of the quality of the MEFP forcing and HEFS streamflow forecasts, there is no systematic decline in forecast quality as the interval between reforecasts (M) increases from 1 to 7 days or the historical period (N) decreases from 24 to 4 years. However, when considering the *sensitivity* of the verification scores to N and M, measured by the range of scores across these scenarios, there are meaningful differences, particularly at higher thresholds of precipitation and streamflow. In this context, sensitivity is a necessary but not a sufficient condition for a decline in forecast quality. These results imply some sensitivity to N and M, but they do not suggest a consistent decline in forecast quality with increasing M or decreasing N. In practice, a reforecast archive of N=12 years and M=1 day (among other combinations) should be adequate to calibrate the MEFP, but it would not be adequate to validate the HEFS for large events, as described below.
- Against the best available calibration (C) and forecasting (F) scenario (C=11, F=11), there is a material decline in the quality of the HEFS forcing and streamflow forecasts when using the control member only (C=1, F=1). For some basins, metrics and thresholds, this is minimized by using all ensemble members to generate the HEFS forecasts (C=1, F=11). For precipitation and streamflow, the greatest differences occur at CN-DOSC1, particularly in the middle and latter portion of the forecast horizon, where the forecast lead time is increased by 1+ days when using C=11 (F=11) members versus C=1 (F=11). The improvements in temperature are greatest at AB-CBNK1 and NE-HOPR1, particularly at the hottest observed temperatures and during the middle portion of the forecast horizon, where the CRPSS is increased by ~10% in real terms (~30% relative to the baseline CRPSS). In contrast, when calibrating the MEFP with C=5 ensemble members (F=11), the forcing and streamflow forecasts are no more reliable or skillful than those calibrated with C=11 members (F=11). Thus, for the locations, thresholds, and verification metrics considered, 5 ensemble members should be adequate to calibrate the MEFP, while the operational forecasts would benefit from using all available ensemble members.

- The minimum requirements for validating the HEFS are examined both theoretically and empirically. At a daily aggregation, the average number of verification pairs for which the observed value exceeds a climatological probability, C_p, is ~365N(1-C_p). In order to estimate a lumped verification score with reasonably small sampling uncertainty, 30 or more independent samples may be required. Thus, if all of the large (e.g. >C_p=0.995) events in a verification sample are statistically independent, and reforecasts are issued once per day, 16.5 years of reforecasts would be required, on average, to generate a verification sample with 30 "large" events. Clearly, these requirements increase dramatically with increasing C_p; in general, the probability of flooding at a daily aggregation is less than 1-in-200 (C_p=0.995). They also increase when the individual samples are related to each other (e.g. one flood event that spans several days), as described below. More detailed metrics, such as the reliability diagram and Relative Operating Characteristic (ROC), require many more samples than a lumped verification score (perhaps 100-200 samples).
- For two time-series (forecasts and observations) that are both autocorrelated in time, the effective sample size for verification is smaller than the nominal sample size. As the correlations increase, the effective sample size declines. For example, the lag-1 autocorrelation of streamflow at AB-CBNK1 for a daily aggregation is 0.542, while the lag-1 autocorrelation at NE-HOPR1 is 0.897. Based on sampling theory, the effective sample size for computing the cross-correlation between the observed and forecast time-series would be 55% of the nominal sample size at AB-CBNK1 and 11% at NE-HOPR1. In other words, for a given amount of confidence in the streamflow verification, roughly 9x more data would be required at NE-HOPR1 than implied by the nominal sample size and 2x more data at AB-CBNK1. While precipitation is generally autocorrelated over much shorter time-scales than streamflow, this also increases the probability that data thinning (e.g. from M=1 to M=3) would significantly reduce the number of extreme events in the sample. Thus, based on sampling theory alone, reducing the reforecast period and frequency would systematically reduce the precipitation and streamflow thresholds at which the HEFS, and associated decision support, could be validated and optimized, particularly for multi-day aggregations, such as reservoir inflows.
- The sensitivities of the validation results to different configurations of N and M are also explored empirically. Here, the range of verification scores *between* cases of N and M is much smaller than the range of scores *within* cases for different subsets of the validation data. Thus, as anticipated from theory, the minimum requirements for validating the MEFP are much greater than the minimum requirements for calibrating the MEFP, even for relatively simple verification scores. Of the verification scores considered here, the cross-correlation is particularly variable across the subsets of N and M. For example, at AB-CBNK1, precipitation amounts

that exceed C_p=0.995 show correlations of between -0.1 and 0.6 in the three subsets of M=3. Thus, for a 1-in-200 day precipitation amount at AB-CBNK1, forecasts issued every three days over a 24-year period would be unverifiable. For more detailed verification metrics, such as the reliability diagram, the thresholds for which the HEFS remains verifiable are even smaller. Thus, even with daily reforecasts between 1985 and 2008, the sample sizes are too small to evaluate reliability diagrams for moderately large precipitation amounts (C_p=0.99), as these events are rarely forecast with high probability. However, this partly originates from a conditional bias in the precipitation forecasts at high observed thresholds, which would not be addressed by increasing the number of reforecasts alone.

- In summary, therefore, the minimum requirements for meteorological reforecasting in support of the HEFS are determined, primarily, by the need to validate the HEFS with reasonably small sampling uncertainty, including for large events. In general, simple, unconditional, verification measures cannot guide operational practice, because they are not application-specific. For example, a flood warning may be triggered when the forecast probability of flooding exceeds some threshold. In this context, there is trade-off between issuing warnings too regularly (low probability threshold) and failing to warn when floods actually occur (high probability threshold). Given an adequate sample of historical flood occurrences, this trade-off, and hence the triggering threshold, can be defined, objectively, through hindcasting and validation. By way of illustration, the use of a degraded reforecast of M=3 at NE-HOPR1 could lead to flood warnings that are correct on only 40% of occasions, when they could be correct on 58% of occasions for a warning threshold optimized to daily reforecasts (i.e. M=1). For users of the HEFS, such as the NYCDEP, a long and consistent record of historical forecasts is, therefore, essential; it is necessary to optimize and improve decision support systems and to benchmark these systems against historical analogs for future extremes.

Recommendations

- Reforecasting requires both significant human and computational resources. However, unsophisticated approaches to data thinning, such as reducing the number of historical years (N) or increasing the interval between reforecasts (M), will also reduce the value of these reforecasts for hydrologic applications. In terms of the HEFS, the greatest impacts of reducing the sample of historical reforecasts would be to prevent the validation of large events with the necessary statistical confidence. These events are critically important to users of the HEFS, such as the NYCDEP. Thus, any approach to data thinning must accommodate a reasonable sample of large and extreme events. The frequency of reforecasts (M) should also accommodate rapidly evolving hydrometeorological conditions, for which M>1 day would not be appropriate for the short-to-medium range.

- The impacts of reducing the number of ensemble members in the GEFS reforecasts (C) will be to reduce their value for statistical post-processing and other applications. Nevertheless, as C increases, there are diminishing gains for the reliability and skill of the MEFP outputs. This study indicates that C=5 ensemble members should be adequate to calibrate the MEFP with the GEFSv10. However, this cannot be generalized to other techniques or to future implementations of the MEFP. Indeed, the benefits of reforecasting with additional members may vary with location and forecast conditions, and they may be greater for extreme events (for which sample measures of forecast quality are inherently limited). Thus, any compromise should be reviewed as models and applications evolve and diagnostic techniques become more sophisticated.
- While the costs associated with meteorological reforecasting are substantial, the benefits are even more substantial. Thus, a concerted effort should be made to produce reforecasts every day over the maximum historical period for which there is adequate data to initialize the GEFS, rather than compromising on N or M. In the absence of a complete reforecast, more sophisticated approaches to data thinning will be required. Here, emphasis on early forecast lead times and extreme events will increase the utility of a limited reforecast for hydrologic applications.
- Spatial pooling or regionalization may improve the sample sizes for calibration and validation of the MEFP. Studies are underway to establish whether reforecasts from hydrometeorologically similar basins can be used to augment the calibration and validation of the HEFS. However, spatial pooling cannot satisfy user requirements for long historical records at critical forecast locations. Also, in validating the streamflow forecasts, spatial pooling is inherently difficult, as hydrologic state variables, unlike atmospheric state variables, often vary abruptly (over short distances), and with myriad basin characteristics.
- An adequate sample of historical events, including large and extreme events, is only one of several minimum requirements for users of weather and climate reforecasts. Other requirements include the timely communication of development plans, use of transitional arrangements for legacy models (e.g. temporary freezing of models), software version control, coordination of model updates with users, timely access to reforecasts, and consistency of the reforecasts and operational forecasts (including initializations), among others. Collectively, these requirements should contribute to a renewed effort by the NWS and other operational forecasting agencies to deliver weather, climate, and water (re)forecasts for improved decision support. This broader set of requirements must be addressed separately, alongside the minimum requirements of end users, such as the NYCDEP.

2. Introduction

The Hydrologic Ensemble Forecast Service (HEFS) is an operational hydrologic forecasting system that is being implemented by the thirteen River Forecast Centers (RFCs) of the U.S. National Weather Service (NWS). The HEFS quantifies the total uncertainty in future streamflow as a combination of the meteorological forcing uncertainty and the hydrologic modeling uncertainty, while correcting for biases in the forecast probabilities (Seo et al., 2010; Demargne et al., 2010, 2014; Brown et al., 2014a/b). The HEFS ingests weather and climate forecasts from, among other sources, the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP), as well as NCEP's Climate Forecast System Version 2 (CFSv2), and produces ensemble streamflow forecasts for the short- to long- range. The HEFS aims to: 1) span lead times from one hour to one year or more with a seamless transitions between forecast time horizons; 2) issue forecast probabilities that are unbiased for different aggregation periods; 3) be spatially and temporally consistent across RFC domains; 4) capture information from current operational weather and climate forecasting systems, while correcting for biases; 5) be consistent with retrospective forecasts or "hindcasts" that are used for verification and decision support; and 6) be properly validated, in order identify the strengths and weaknesses of the forecasts and to guide forecasting operations and decision support.

By explicitly accounting for the uncertainties inherent in meteorological and hydrologic forecasting, while correcting for biases in the forecast probabilities, the HEFS aims to support improved, risk-based, decision making for a variety of water resources applications, including reservoir operation, flood forecasting, river navigation, and water supply. For example, the New York City Department of Environmental Protection (NYCDEP) is using the HEFS to improve the management of risks to water quantity and quality objectives in the NYC area. In this context, the NYCDEP has developed an Operational Support Tool (OST), which ingests streamflow forecasts from the HEFS that are produced operationally by the Middle-Atlantic RFC and the Northeast RFC. The OST optimizes the quantity and quality of water stored in the NYC reservoirs, while avoiding unnecessary, multi-billion dollar, infrastructure costs, such as water filtration. Elsewhere, the U.S. Army Corps of Engineers (USACE) are redeveloping their water control manual for the Folsom Reservoir and the American River. In this context, the California-Nevada RFC (CNRFC) are evaluating the use of streamflow hindcasts from the HEFS, in order to establish the benefits and risks of using inflow forecasts to manage the flood control space in the Folsom Reservoir. Elsewhere in California, the Yuba County Water Agency (YCWA), together with CNRFC and partners, and exploring the use of probabilistic inflow forecasts to better manage the flood control spaces in Lake Oroville, the Englebright Reservoir and the New Bullards Bar Reservoir.

The ability of the HEFS to provide useful information for decision making depends, crucially, upon the accuracy (unbiasedness and skillfulness) of the forecast probabilities. There is a need to demonstrate this accuracy through retrospective forecasting and verification. Retrospective studies are necessary to guide the development of the HEFS, as well as decision support systems that rely upon the HEFS, and to build confidence among decision makers that the forecasts are accurate, useful, and can lead to better decisions. In order to provide meteorological and streamflow forecasts that are *demonstrably accurate*, the HEFS must be calibrated and validated with historical data. While recent studies have documented the quality of the precipitation, temperature and streamflow forecasts from the HEFS, both for the short-to-medium range (Brown et al., 2014a/b) and for the long-range (Brown, 2013), the minimum requirements for reforecasting have not been evaluated. These requirements are largely driven by the raw meteorological reforecasts used as input to the HEFS and, specifically, by the HEFS Meteorological Ensemble Forecast Processor (MEFP), which aims to correct for biases in the raw forecasts of precipitation and temperature (Schaake et al., 2007; Wu et al., 2011). Observations of precipitation, temperature and streamflow are also required to initialize the HEFS, calibrate the hydrologic models, and to validate the forecasts. Gauge-based observations are typically available for many decades (often 50-100 years) at river forecast locations. However, atmospheric models rely on a best estimate (or a range of possibilities) of the multivariate, spatially distributed, state of the atmosphere-ocean system at the forecast issue time. In order to conduct reforecasting, these estimates must be produced retrospectively. In practice, reliable estimates of the atmosphere-ocean state variables require satellite observations, which are only available since the early 1980s.

Thus, meteorological reforecasting is inherently constrained to the recent past. Also, given the significant cost of conducting reforecasting, a trade-off emerges between expanding reforecasting and improving the underlying weather and climate models. However, for users of the HEFS, such as the NYCDEP and YCWA, hydrometeorological reforecasting is critically important. It is necessary to optimize and improve decision support systems, such as the OST, and to benchmark these systems against historical analogs for future extremes.

In order to support NCEP in determining the requirement of the HEFS for meteorological reforecasting, this report considers the sensitivity of the HEFS to a limited number of reforecast configuration options. Clearly, reforecast configuration is only one of several requirements for users of weather and climate forecasts. Other requirements include the timely communication of development plans, use of transitional arrangements for legacy models, software version control, coordination of model updates with users, timely access to reforecasts, and consistency of the reforecasts and operational forecasts, among others. Collectively, these requirements should contribute to a new business model for NCEP and other operational forecasting agencies in delivering weather, climate, and water (re)forecasts for improved decision support. As indicated above, this report focuses on the minimum technical requirements of the HEFS for meteorological reforecasts. It does not consider the broader set of requirements for delivering an efficient and effective forecasting service, which must be addressed separately.

In terms of the HEFS, the minimum requirements for historical data are driven by: 1) the need for an adequate sample size to estimate the statistical parameters of the HEFS; 2) the need for an adequate sample size to validate the HEFS; and 3) the need for users of the HEFS to calibrate and validate their decision support systems. This report is concerned with the minimum requirements for (1) and (2) only. The requirements of end users, such as the NYCDEP and YCWA, will be gathered and presented separately. In this context, (1) and (2) define the minimum requirements for operating the HEFS, while (3) is necessary to ensure the outputs from the HEFS are useful for decision making. In other words, the minimum requirements associated with (1) and (2) should be regarded as an *incomplete baseline*. In practice, the requirements of users for meteorological and hydrologic reforecasting may exceed those for calibration and validation of the HEFS, and they may evolve as services change and other users adopt the HEFS. Furthermore, this study is concerned with short-to-medium range forecasting only and, specifically, with the minimum requirements for historical data from the Global Ensemble Forecast System (GEFS).

Raw forecasts of temperature and precipitation from the GEFS are used to produce bias-corrected forcing for input to the HEFS. These forecasts are used in water supply decision making for the short-to-medium range, including reservoir management, flood warning, river navigation and recreation. The GEFS uses Version 9.0.1 of the Global Forecast System (GFS), which comprises a horizontal resolution of T254 (~55km) for 1-8 days and T190 (~70km) for 9-16 days, and a vertical resolution of L42 or 42 levels (Wei et al. 2008; Hamill et al. 2011; Hamill et al. 2013). Reforecasts were produced with the GEFS for a ~26-year period between 1985 and 2010 (Hamill et al., 2013). Calibrating and validating the HEFS with a subset of the available reforecasts will identify the sensitivities of the HEFS to a *degraded* reforecast with the *current* GEFS only. Some applications of the HEFS may benefit from a configuration that *improves* upon the available reforecasts, but this cannot be established here. Rather, this study examines the ability to provide accurate forecasts with the HEFS using a degraded calibration sample and the ability to measure that accuracy with a reduced validation sample.

The minimum requirements for calibrating the HEFS include an adequate historical period and frequency of reforecasts from which to estimate the statistical parameters of the HEFS, and sufficient ensemble members to capture the skill in the meteorological forecasts. Since the HEFS relies on statistical modeling, consistency of the reforecasts and operational forecasts is also important. The minimum requirements for validating the HEFS also include an adequate sample (historical period and frequency) of reforecasts under varying basin conditions, again without structural changes that would undermine their interpretation. In slow responding basins, the "effective" sample size is reduced by temporal autocorrelations in streamflow, implying a longer period of record for validation (and calibration of streamflow post-processors). In fast responding basins, conditions evolve rapidly, implying a greater frequency of reforecasts to capture large and extreme events. Assuming the climatology is reasonably stationary, a 25-year reforecast should capture much of this variability. However, at a one-day aggregation, flooding may occur with a climatological frequency of 0.001 (1-in-1000 days) or less. Thus, on average, fewer than ten (0.001*25*365) flood events will occur within a 25-year period. Likewise, for long-range forecasting, where fixed aggregations are often required (e.g. April-July reservoir volumes), a 25-year reforecast will inevitably omit some important variability.

In summary, the aims of this study are twofold, namely to determine the minimum requirements for reforecasting with the GEFS, in order to: 1) calibrate the HEFS adequately; that is without materially reducing the quality of the forecasts, including at high thresholds; and 2) validate the forcing and streamflow forecasts with reasonably small sampling uncertainty. The calibration of the HEFS depends on an adequate sample size, for which the period of record and interval between reforecasts are important. It also depends on the number of ensemble members in the GEFS and the consistency of the reforecasts and operational forecasts. Likewise, the validation of the HEFS depends on an adequate sample size, for which the period of record and interval between reforecasts are important, and a reasonably consistent and representative sample (accepting that these two things may not be aligned). Following a description of the study basins, datasets and approach, the verification results are presented separately for the minimum calibration and validation requirements.

3. Approach

3.1 Study basins

Four headwater basins were considered in this study, namely: the Chikaskia River at Corbin, Kansas (AB-CBNK1); the Dolores River at Rico in Colorado (CB-DRRC2); the Middle Fork of the Eel River at Dos Rios in California (CN-DOSC1); and the Wood River at Hope Valley, Rhode Island (NE-HOPR1). Figure 1 and Table 1 show the location of each basin, its average elevation, area, and the location of the nearest grid node in the GEFS. Table 1 also shows the annual precipitation, the fraction of precipitation that generates runoff (the runoff coefficient), and the ratio of precipitation to potential evaporation (a climate index). The drainage areas range from 188 square kilometers (NE-HOPR1) to 2,057 square kilometers (AB-CBNK1) and the runoff coefficients vary from 0.12 (AB-CBNK1) to 0.55 (NE-HOPR1). The basins were chosen for a combination of practical and hydrological reasons. First, they all originate from RFCs for which the HEFS has been implemented and validated, namely AB-, CB-, CN-, and NE-RFCs, and for which the absolute quality of the forecasts has been documented (Brown, 2013, 2014; Brown et al., 2014a/b). Here, the focus is on the minimum requirements for calibrating and validating of the HEFS; that is, on the *relative* quality of the forecasts for different configurations of the GEFS; and not on the *absolute* quality of the forecasts. Second, headwater basins respond quickly to forcing information and, as the uncertainties and biases propagate from upstream to downstream locations, it is important, initially, to understand the quality of the HEFS in headwater basins. Third, headwater basins are important for operational forecasting of water quantity and quality, including flood warning and reservoir operations. Further downstream, the HEFS will be impacted by additional sources of bias and uncertainty, of which some are inherently difficult to quantify (e.g. the downstream effects of river regulations, simplified hydraulic routing and composite timing errors; see Raff et al., 2013). As part of the ongoing evaluation of the HEFS, more complex regimes, as well as additional sources of forcing, will be considered in future.

Figure 2 shows the daily means of temperature, precipitation, and streamflow for each basin, where CN-DOSC1 and CB-DRRC2 both comprise an average over two sub-basins (see below). The averages are shown by calendar month and were derived from gauged temperature, precipitation, and streamflow over a 24-year period between 1985 and 2008 (see Section 3.3). As indicated in Figure 2, there are marked differences in the seasonality and covariability of precipitation and runoff among these basins.

The Chikaskia River (AB-CBNK1) experiences a warm and humid summer climate. During the late spring and early summer, cool air from Canada and the Rocky Mountains combines with moist air from the Gulf of Mexico and hot air from the Sonoran Desert, leading to intense thunderstorms and tornados in Kansas and Oklahoma. At AB-CBNK1, the relationship between precipitation and runoff is modulated by the shallow terrain and dense vegetation cover, as well as increased evapotranspiration during the summer months.

The Dolores River (CB-DRRC2) is a tributary of the Colorado River and occupies a narrow valley incised into the sandstone of the San Juan Mountains. Precipitation is reasonably constant throughout the year, but falls primarily as snow during the winter months. The snowpack melts in the late spring and early summer, which leads to a sharp increase in runoff between April and July (Figure 2). For the purposes of hydrologic modeling, CB-DRRC2 is separated into two sub-basins, in order to accommodate the varied elevations there. The lower sub-basin accounts for 67% of the total area of CB-DRRC2.

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

The Wood River flows approximately 85km from its source in Sterling, Connecticut, through Hope Valley (NE-HOPR1) in the Arcadia Management Area to Alton, Rhode Island, where it converges with the Pawcatuck River. As indicated in Figure 2, the daily average precipitation at NE-HOPR1 is relatively constant throughout the year, but includes significant snowfall during winter months (the average annual snowfall is 866mm). During the early spring, rising temperatures lead to snowmelt and to a peak in streamflow around March or April, followed by lower flows during the summer months.

3.2 Experimental design

The HEFS quantifies the total uncertainty in future streamflow as a combination of the meteorological and hydrologic uncertainties, while correcting for biases in both the forcing and streamflow (Demargne et al., 2014). Further information about the HEFS methodology can be found in Appendix A. The meteorological uncertainties and biases are quantified with the Meteorological Ensemble Forecast Processor (MEFP). The MEFP produces ensemble forecasts of precipitation and temperature conditionally upon a raw, single-valued, forecast (Wu et al., 2011). For the short- to medium-range, the raw forecasts used by the MEFP include the ensemble mean of the GEFS. In removing the meteorological biases with the MEFP, the hydrologic uncertainties and biases can be modeled independently of the forcing uncertainties and biases (Seo et al., 2006; Demargne et al., 2014). The hydrologic uncertainties and biases are modeled in two stages. First, the meteorological forecasts from the MEFP are used to generate raw streamflow forecasts, which may contain hydrologic biases, but do not explicitly account for any hydrologic uncertainties. Secondly, the raw streamflow forecasts are post-processed with the Ensemble Postprocessor (EnsPost). The EnsPost models the hydrologic uncertainties and biases from the residuals between the observed and simulated streamflows (Seo et al., 2006); that is, streamflow predictions based on observed temperature and precipitation at the forecast issue time.

The simulations and observations used to estimate the hydrologic uncertainties and biases are typically available for several decades at each RFC forecast location. Likewise, the precipitation and temperature observations used to generate the streamflow simulations and to quantify the forcing uncertainties and biases are typically available for several decades. In contrast, the meteorological reforecasts, which are used by the MEFP to estimate the forcing uncertainties and biases, require satellite observations and corresponding reanalysis of the ocean-atmosphere states, in order to initialize the weather and climate models. These datasets are only available from the early 1980s onwards. Thus, as indicated above, the requirements of the HEFS for historical data are primarily constrained by the availability of (appropriate initialization for the) meteorological reforecasts.

As indicated above, the total uncertainty in the streamflow forecasts originates from a combination of uncertainties in the meteorological forecasting and hydrologic modeling. Depending on basin characteristics and antecedent conditions, a large fraction of the total uncertainty can originate from the meteorological uncertainties (Kavetski et al., 2002; Pappenberger et al., 2005; Wu et al., 2011). Thus, the meteorological forecasts are a central component of the HEFS and other hydrologic ensemble prediction systems. When a meteorological model is updated, any changes in the statistical properties of the precipitation and temperature forecasts will, to some degree, impact the streamflow forecasts from the HEFS. For example, the MEFP may be impacted by changes in the spatial or temporal resolution of the model, including the position of grid cells in relation to hydrologic basins, the model physics in different layers, including at the land-surface and ocean boundaries, and the number of (or approach to generating) ensemble members. In terms of calibrating the MEFP, these properties are important insofar as they influence the statistical character of the precipitation and temperature forecasts, including any systematic biases, as well as the information content more generally (e.g. measured in terms of correlation). In general, therefore, the MEFP must be recalibrated when the GEFS is updated in any non-trivial way. Likewise, any non-trivial changes to the HEFS must be accompanied by new streamflow hindcasting and validation. In many cases, this requires further hindcasting and validation by users of the HEFS, such as the NYCDEP, who rely upon streamflow hindcasts to calibrate and validate their own forecasting and decision support systems. Following changes to the operational GEFS, the HEFS requires an “adequate” sample of meteorological reforecasts, in order to recalibrate the MEFP and to produce and validate new forcing and streamflow hindcasts. In this context, the minimum requirements for reforecasting include the number of historical years of data (N), the interval between reforecasts (M), and the number of ensemble members. These and other variables are summarized in Table 2.

In order to evaluate the effects of N and M on the quality of the precipitation and temperature forecasts from the MEFP, the raw GEFS reforecasts (Hamill et al., 2013) were systematically degraded from N=24 years (1985-2008) and M=1 day to combinations of smaller N and larger M. These "thinned" reforecasts were used to calibrate the MEFP and to generate forcing and streamflow hindcasts for a consistent validation period. As indicated above, some applications of the HEFS may benefit from a reforecast configuration that *improves* upon the available reforecasts, but this cannot be established here. In degrading the raw GEFS reforecasts, the hindcasting and validation period was fixed to 24 years (1985-2008), with a forecast issued at 12Z each day. The choice of validation period was motivated by: 1) the need to isolate the effects of N and M on the quality of the MEFP forecasts, independently of any background variability (i.e. from changes in the validation period); and 2) by the choice of experimental design for validation. In terms of the latter, independent validation is always preferred when evaluating statistical techniques, such as the MEFP. Unless the verifying observation is removed from the calibration sample, the statistical parameters will benefit, unfairly, from "seeing" the outcome in advance of predicting it. Depending upon the number of parameters to estimate and their sampling properties, among other factors, this advantage can be important. The results from dependent validation should, therefore, be regarded as a "best case scenario" of the actual forecast quality. In practice, however, the MEFP is relatively parsimonious (Wu et al., 2011). In other words, a single observation should not greatly influence the estimated parameters. Furthermore, independent validation poses significant practical challenges, as the HEFS is an operational forecasting system; it is not well-suited to automatic calibration, and hindcasting is extremely time-consuming.

In evaluating the sensitivities to N, both dependent and (limited) cross-validation were employed. Specifically, the 24-year validation period was sub-divided into smaller calibration periods, N={2x12, 3x8, 4x6, and 6x4} years. Dependent validation involved estimating the parameters for each sub-period, issuing forecasts for that sub-period, and collating the forecasts from *all* sub-periods for validation (i.e. 24 years in total). Independent validation involved borrowing the parameters from an adjacent sub-period. In practice, this should be regarded as a "worst case scenario" for the expected forecast quality, because independent forecasting is conducted for multiple years (i.e. 12 years, for N=12) without recalibrating the MEFP. Table 3 summarizes the dependent and independent calibration scenarios for N. In evaluating the sensitivities to M, the MEFP was calibrated for M={1, 3, 5, and 7} days and forecasts were issued at 12Z each day between 1985 and 2008. In this context, M=1 represents dependent validation, whereas M={3, 5 and 7} involves a mixture of dependent and independent validation. Specifically, for M=3, 5, and 7 days, 1/3rd, 1/5th and 1/7th of the validation sample appears in the calibration sample, respectively. The calibration scenarios for M are summarized in Table 4. Alongside the precipitation and temperature forecasts from the MEFP, streamflow forecasts were produced at the outlet of each basin (see below).

As a post-processing technique, the MEFP aims to improve skill by reducing bias in the raw GEFS forecasts, but does not introduce any new predictors. Thus, the quality of the MEFP outputs is sensitively dependent on the quality of the raw forcing inputs from the GEFS. The MEFP uses the ensemble mean from the GEFS to capture the information content in these (re)forecasts. In order to examine the sensitivity of the MEFP outputs to the number of ensemble members in the GEFS inputs, the GEFS reforecasts were systematically degraded by using only a subset of the ensemble members to derive the ensemble mean. These “thinned” reforecasts were used to calibrate the MEFP and to generate forcing and streamflow hindcasts for a consistent validation period (i.e. 26 years, from 1985-2010). In practice, the GEFS reforecasts contain fewer ensemble members (C) than the operational forecasts (F). Specifically, the GEFSv10 reforecasts comprise only 11 ensemble members (10 + control), while the operational forecasts comprise 21 members (20 + control). Hindcasting and validation was conducted with the all available members (11). For example, when calibrating the MEFP with an ensemble mean derived from C=5 members, the hindcasts were generated with an ensemble mean derived from F=11 members. However, in order to better understand the impacts of this discrepancy, a baseline scenario was included. Here, the control run was used to both estimate the MEFP parameters (C=1) and to derive the forcing and streamflow hindcasts (F=1). The scenarios for C and F are summarized in Table 5.

3.3 Datasets

For each scenario of N and M, hindcasts of mean areal temperature (MAT) and mean areal precipitation (MAP) were generated with the MEFP for a 24-year period between 1985 and 2008. For each combination of C and F, the hindcasts were generated for the full GEFS reforecast period (1985-2010); unlike N and M, the historical period was not integral to the validation design for C and F (see below). The hindcasts of MAP and MAT each comprise ~60 ensemble members (the precise number varying between basins, as described in Wu et al., 2011), with lead times varying from 6 to 360 hours in six-hourly increments. In order to evaluate the skill of the MEFP forecasts with GEFS inputs (MEFP-GEFS), precipitation and temperature forecasts were also generated with a conditional or "resampled" climatology (MEFP-CLIM). The latter involves resampling the historical observations of MAP and MAT in a moving window of, respectively, 61 days and 31 days around the forecast valid date.

Raw streamflow hindcasts were generated with the Community Hydrologic Prediction System (CHPS) using the precipitation and temperature forecasts from the MEFP. The hindcasts were produced with the hydrologic models and parameter settings used operationally in each RFC. For all RFCs considered here, the Snow Accumulation and Ablation Model (SNOW-17; Anderson, 1973) is used together with the Sacramento Soil Moisture Accounting Model (SAC-SMA; Burnash, 1995). The models are executed at a six-hourly timestep, but interpolated to an hourly timestep at CB-DRRC2 and CN-DOSC1. Routing from the headwater to the downstream basins is conducted with Lag/K using constant or variable lag and attenuation. Historical simulations were generated with observed forcing for each basin and used to examine the sensitivities of the hydrologic predictions to the meteorological forcing (see below).

Observations of precipitation and temperature were obtained from each RFC and comprise areal averages (MAP, MAT) of the gauged precipitation and temperature in each basin. The data comprise six-hourly observations, recorded in local time, and covering the period ~1948-2010. In order to pair the meteorological observations and forecasts, the observed values were chosen from the nearest available synoptic times in

{0Z, 6Z, 12Z, 18Z}. This introduced a timing error into the observations of +1 hours, 0 hours, -1 hours and -2 hours for NE-HOPR1, AB-CBNK1, CB-DRRC2 and CN-DOSC1, respectively. As the forecasts were verified at an aggregated scale of one day or larger (see below), this timing error was deemed acceptable. The hydrologic forecasts and simulations were paired without any timing errors.

3.4 Verification strategy

Verification was conducted with the NWS Ensemble Verification Service (EVS; Brown et al., 2010). The temperature and precipitation forecasts were verified against observed temperature and precipitation, respectively. In order to establish the sensitivities of the hydrologic forecasts to different calibrations of the MEFP, the raw streamflow forecasts were verified against simulated streamflows. Differences between the hydrologic forecasts and simulations reflect the contribution of the MEFP-GEFS forcing to the quality of the streamflow forecasts, independently of any hydrologic errors and biases (which are ordinarily removed by the HEFS Ensemble Postprocessor, EnsPost). Aside from eliminating these hydrologic biases, simulated streamflows avoid the timing and other errors associated with pairing streamflow forecasts and observations. For example, the streamflow observations are only available as daily averages and in different time zones to the forecasts. No streamflow post-processing was conducted in this study, as the EnsPost uses hydrologic simulations and observations only and is, therefore, insensitive to the meteorological reforecasting. In this context, the aim is to establish the sensitivity of the HEFS forcing and streamflow forecasts to different calibrations of the MEFP, and not to examine the absolute quality of the forecasts, which is considered elsewhere (Brown, 2013, 2014; Brown et al., 2014a/b).

Verification was conducted both unconditionally (i.e. for all data) and conditionally upon observed and forecast amount. Unconditional bias and skill are important, as the HEFS is an operational forecasting system for which many applications are anticipated. However, “average conditions”, particularly the ensemble mean, generally favor dryer weather and lower flows, as precipitation and streamflow are both skewed variables. In order to compare the verification results between basins, for different forecast lead times and valid times, and for specific aggregation periods, common thresholds were identified for each basin. Specifically, for each aggregation period, a, and basin, b, a climatological distribution function, $\hat{F}_{n,a,b}(x)$, was computed from the n values of the hydrometeorological variable, x, between 1985 and 2008. Real-valued thresholds were then determined for $k \approx 100$ non-exceedence probabilities, c_p, $\hat{F}^{-1}_{n,a,b}(c_p)$, where $c_p \in [0,1]$ and $p = 1,\ldots,k$. These non-exceedence probabilities provide a consistent mapping between the likelihood of a particular hydrometeorological occurrence and its corresponding real value across different basins and aggregation periods.

As indicated above, verification was performed for different magnitudes of the observed and forecast variables. When conditioning on observed amount, the quality of the forecasting system is evaluated for the full range of historical occurrences, including extreme events that were forecast inadequately (as small or moderate events). When conditioning on forecast amount, the verification results may discount important observed extremes. However, since the observed amount is unknown when a forecast is issued, conducting verification by forecast amount is useful for guiding operational forecasting and real-time decision making. While some verification metrics provide integral measures of error across multiple thresholds (e.g. the mean error), others are defined for discrete occurrences (e.g. the probability of detection). Integral measures, such as the mean error, were derived from the subsample in which the prescribed condition was met (e.g. the observation exceeded the threshold). Measures defined for discrete events were computed from the observed and forecast probabilities of exceeding the threshold.

4. Results and analysis

4.1 Minimum requirements for estimating the parameters of the MEFP

4.1.1 Sensitivity to the historical period and interval between reforecasts

The precipitation and temperature forecasts from the MEFP were verified against observed MAP and MAT, respectively. The results are shown for a daily aggregation, as this is a representative volume for short-to-medium range forecasting. The results are presented by forecast lead time and magnitude of the forcing variable for each scenario of N (the number of years of reforecasts) and M (the interval between reforecasts). The analysis focuses on the sensitivity of the forecasts to N and M in terms of bias, skill, and other attributes of forecast quality, and not on the absolute quality of the forecasts. Figure 3 provides selected verification scores (in the rows) at three climatological probabilities (in the columns), for the MEFP-GEFS precipitation forecasts. Here, C_p=0.0 denotes the Probability of Precipitation (PoP), while C_p=0.995 represents a daily precipitation amount that is exceeded, on average, once every 200 days. The scores were derived from the subsample of verification pairs in which the *observed* precipitation amount exceeded the threshold. Here, the verification statistics for the daily accumulations were averaged over the first three days of forecast lead time. The results are shown for each calibration scenario, N={24, 12, 8, 6, and 4 years}, and for the two validation scenarios, namely dependent validation (all scenarios of N) and cross-validation, i.e. N={12, 8, 6, 4} (see Table 3). The verification measures are summarized in Appendix B. The correlation coefficient measures the degree of association between the ensemble mean of the MEFP-GEFS precipitation forecasts and the observed precipitation amount. The relative mean error (RME) measures the fractional bias of the ensemble mean forecast, where a negative RME denotes an under-forecasting bias. The Continuous Ranked Probability Skill Score (CRPSS) measures the fractional improvement of the MEFP-GEFS precipitation forecasts when compared to the MEFP-CLIM forecasts, where 1.0 denotes a perfect score. The Brier Skill Score (BSS) also provides a lumped measure of skill relative to the MEFP-CLIM forecasts. However, unlike the CRPSS, the BSS measures the ability the forecasting system to predict the exceedence (or non-exceedence) of a discrete threshold.

Figures 4-7 show selected verification scores by forecast lead time for the MEFP-GEFS precipitation forecasts at AB-CBNK1, CB-DRRC2, CN-DOSC1, and NE-HOPR1, respectively. Again, the results are shown for subsamples in which the observed precipitation amount exceeded C_p={0.0,0.9,0.995}. Unlike Figure 3, the results are shown separately for each one-day accumulation, and with a separate curve for each scenario of N. While Figures 4-7 shows the verification results for selected thresholds at all forecast lead times, Figures 8-11 shows the results for all thresholds at selected forecast lead times. In Figures 8-11, the climatological probabilities are plotted on a non-linear scale, in order to emphasize the larger thresholds. The origin of each curve in Figures 8-11 is the climatological PoP, i.e. the zero-precipitation threshold. The BSS denotes the ability of the MEFP-GEFS forecasts to predict the exceedence of this threshold. The correlation, RME and CRPSS denote the quality of the MEFP-GEFS forecasts for *wet conditions,* i.e. for the subsample that exceeds the threshold, with the lowest threshold being zero.

As indicated in Figure 3, the sensitivities of the MEFP-GEFS precipitation forecasts to the number of years of calibration data (N) are relatively small, both for the dependent and independent validation scenarios. In general, the forecast quality is slightly reduced under independent validation. However, as indicated above, independent forecasting for multiple years (up to 12 years) should be regarded as a “worst case scenario” for the expected forecast quality, as the MEFP should be recalibrated more frequently. The greatest differences between dependent and independent validation occur in CB-DRRC2, particularly for light and moderate precipitation amounts, where the forecast quality is generally lower. This is understandable because CB-DRRC2 lies in the San Juan Mountains of Colorado, where the steep terrain leads to reduced predictability and increased climatological variability on inter-annual timescales. While the MEFP assumes that the joint distribution of forecasts and observations is reasonably stationary, any climatological non-stationarities may introduce a trade-off between larger N (smaller sampling uncertainty) and smaller N (greater climatological specificity). As indicated in Figure 3, for most verification scores, locations and thresholds, there is no systematic increase in forecast quality with increasing N. Indeed, in some cases, the forecast quality increases slightly with *decreasing* N. Given the sampling uncertainties, this should not be overstated. However, it may originate from climatological variability over the validation period and thus a greater specificity of the estimated parameter at smaller N. As indicated in Figures 4-7, the sensitivities to N are relatively small at all forecast lead times, although some erratic behaviors are seen at N=4 in AB-CBNK1 and CB-DRRC2, where the absolute forecast quality is also lower. Similarly, Figures 8-11 suggest that the MEFP is relatively insensitive to N across a broad range of precipitation thresholds. However, at CB-DRRC2, there is a material decline in BSS for N=4, particularly for light and moderate precipitation amounts, while the CRPSS is higher (Figure 9). These differences originate from the structure of the BSS and CRPSS. The CRPSS is sensitive to biases in the ensemble mean forecast, which are also smaller for N=4. The BSS is sensitive to these biases only insofar as they impact the forecast probability (of exceeding C_p), and not to their absolute magnitude.

Figure 12 shows the quality of the MEFP-GEFS precipitation forecasts for different scenarios of M. The verification scores include the correlation coefficient and the RME, together with the BSS and CRPSS. They were computed at a daily accumulation for C_p={0.0, 0.99, 0.995}, and averaged over the first three days of forecast lead time. Figures 13-16 show the verification results at all precipitation thresholds for AB-CBNK1, CB-DRRC2, CN-DOSC1 and NE-HOPR1, respectively. Here, the results comprise daily accumulations at forecast lead times of 1, 2, and 3 days. In terms of data thinning, the scenarios of N are broadly comparable to M, with M=7 comprising 1/7th of the original calibration sample, versus 1/6th for N=4. In principle, for atmospheric variables that are statistically dependent over multiple days, thinning by M should have a smaller impact than an equivalent N. In practice, however, except for large-scale systems, such as atmospheric rivers, precipitation varies over short periods and at small spatial scales, as evidenced by the majority of forecast skill occurring in the first 1-7 days (or less at AB-CBNK1, which is located in the Central Plains). Thus, depending on forecast lead time and location (among other factors), thinning by M may be more or less aggressive than an equivalent N.

As indicated in Figure 12, when averaged across forecast lead times of 1-3 days, there is no systematic decrease in forecast quality with increasing M at any location or precipitation threshold considered. Similarly, when considering forecast lead times of 1, 5, and 10 days separately (Figures 13-16), the quality of the MEFP-GEFS precipitation forecasts is relatively insensitive to M at most locations. However, at AB-CBNK1, where the forecast skill declines rapidly over the first week (Figure 4), there is a non-trivial sensitivity to M from 0-24 hours across a range of precipitation thresholds, particularly for the correlation coefficient, RME and CRPSS (Figure 13). This is evidenced by the range of verification scores for different scenarios of M. To further illustrate these sensitivities, Figure 17 shows the range of verification scores across all scenarios of M at selected forecast lead times. Figure 18 shows the equivalent range of scores for N. Clearly, the range of scores is not indicative of a systematic dependence of forecast quality on N or M (see above). However, it is indicative of a sensitivity to the amount of calibration data available. In general, AB-CBNK1 shows the greatest sensitivities to M and N, while CB-DRRC2 is only sensitive to N (and specifically to N=4, as indicated above).

In order to illustrate the effects of N and M on the largest observed and forecast precipitation amounts, box plots were computed from the MEFP-GEFS precipitation forecasts. Figure 19 shows box plots of the forecast errors for each basin (in the rows) and for two scenarios of N (in the columns), namely N={24, 12}. The results are plotted against observed precipitation amount and are shown at a forecast lead time of 0-24 hours. Figure 20 shows the corresponding results against forecast precipitation amount, specifically the ensemble mean forecast. Selected quantiles of the forecasting errors are plotted together with the median error and range (extreme residuals) as whiskers. The verifying observation is denoted by the zero-error line. Verification pairs for which the observation falls outside the ensemble range are denoted as “misses”. However, each forecast comprises only a small number of ensemble members (60). Thus, some misses should be expected, even if the forecasts are conditionally unbiased. Figures 21 and 22 show box plots of the errors in the MEFP-GEFS precipitation forecasts for two scenarios of M, namely M={1, 3}, again ordered by observed and forecast precipitation amount, respectively. Here, each box represents one ensemble forecast from the period 0-24 hours. As indicated in Figures 19 and 21, there is no systematic decline in forecast quality at N=12 or M=3 for the most extreme observed precipitation amounts. Rather, any differences between scenarios are consistent with sampling uncertainty. While there are some differences in the largest precipitation forecasts (by ensemble mean) for N=12 (Figure 20) and M=3 (Figure 22), these differences are again consistent with sampling uncertainty and do not translate into additional skill for N=24 or M=1 (e.g. Figures 8-11).

Figure 23 shows the quality of the temperature forecasts at selected thresholds for each scenario of N, while Figure 24 shows the corresponding results for each scenario of M. The results for N include both validation scenarios, namely dependent validation, i.e. N={24, 12, 8, 6, 4}, and cross-validation, i.e. N={12, 8, 6, 4} (see Table 3). The verification metrics include the mean error of the ensemble mean forecast (°C), the correlation coefficient, BSS and CRPSS. The metrics were computed at a daily aggregation and averaged over the first three days of forecast lead time. Figures 25-28 show the same verification metrics for selected daily aggregations at all temperature thresholds for AB-CBNK1, CB-DRRC2, CN-DOSC1, and NE-HOPR1, respectively. The results are shown for each scenario of N. In keeping with the MEFP-GEFS precipitation forecasts, there is no systematic increase in forecast quality with increasing N. As indicated in Figure 23 and 24, when averaged over the first three days of forecast lead time, the MEFP-GEFS temperature forecasts show similar mean error, correlation, BSS and CRPSS for all scenarios of N and M, respectively. While the independent validation results differ from the dependent validation results for some basins and thresholds (Figure 23), there is no systematic increase in forecast quality with increasing N in either case. In practice, these differences between dependent and independent validation are likely to originate from climatological variability over the calibration period (and not from the influence of a single, dependent, sample on the quality of the calibration). Indeed, for the same reason, there is an increase in forecast quality with *decreasing* N in some basins, notably at CN-DOSC1 (Figure 27), where the BSS and CRPSS are materially higher for N=4 than N=24. Thus, particularly in basins that experience significant climatological variability, a more recent or otherwise similar calibration period (relative to the forecast period) may improve the quality of the MEFP temperature forecasts. However, rather than using a smaller calibration sample, which may eliminate important historical extremes, the prior observed temperature could be integrated into the MEFP, alongside the raw temperature forecast. In this context, the MEFP would comprise an autoregressive model with the raw temperature forecast as an exogenous predictor (akin to the EnsPost; Regonda et al., 2013).

In order to examine the sensitivities of the HEFS to the calibration of the MEFP, independently of any hydrologic errors and biases, the raw streamflow forecasts were verified against simulated streamflow. In practice, hydrologic models respond non-linearly to precipitation and temperature forcing. Thus, depending on basin conditions (e.g. thunderstorms in the summer versus snowmelt in the spring), as well as the spatial and temporal consistency of the forcing, the outputs from the HEFS may show materially different sensitivities than either of the forcing inputs separately. Figure 29 shows the fractional bias of the ensemble mean forecast, together with the correlation coefficient, BSS and CRPSS (in the rows), at three streamflow thresholds (in the columns), namely C_p={0.75, 0.9, 0.995}. In Figure 29, the results are shown for each scenario of N, while, in Figure 30, the corresponding results are shown for each scenario of M.

When averaged over the first three days of forecast lead time, the HEFS streamflow forecasts show only limited sensitivities to the N (Figure 29) and M (Figure 30). Furthermore, there is no systematic decline in forecast quality with increasing M or decreasing N. Thus, in terms of average forecast quality, the outputs from the HEFS are no more sensitive to N or M than either of the forcing inputs separately. As with the precipitation forcing, there are some differences in dependent versus independent validation at AB-CBNK1 and CB-DRRC2 (Figure 29). However, as indicated above, a multi-year period was used to calibrate the MEFP and to conduct independent validation, which should be considered a “worst case scenario”. In contrast, dependent validation provides a “best case scenario” (although probably closer to operational performance). Thus, rather than a dependent validation advantage, which should not vary much with location, the differences between dependent and independent validation for AB-CBNK1 and CB-DRRC2 are likely to originate from climatological variability in these two basins, possibly exaggerated by sampling uncertainty. Indeed, for daily streamflow amounts that are exceeded on average, once every four days (C_p=0.75), there is no difference in BSS at CN-DOSC1 for dependent versus independent validation when N=12 (Figure 29). In this context, N=12 implies operational forecasting with the HEFS for a 12-year period without recalibration, which is unlikely in practice. In contrast, at CB-DRRC2 the BSS is materially lower for N=12 when using independent validation, while at N=6 and N=4, there are no differences between dependent and independent validation.

4.1.2 Sensitivity to the number of ensemble members in the GEFS

The GEFS reforecasts comprise 11 ensemble members, from which C=1, C=5 and C=11 ensemble members were used to calibrate the MEFP. Hindcasts of temperature, precipitation and streamflow were generated for each calibration scenario using an ensemble mean derived from all 11 ensemble members (F=11), together with a baseline comprising the control member only (F=1). Figure 31 shows the residual quality of the MEFP precipitation forecasts at a daily accumulation when calibrated with all ensemble members (C=11, F=11) versus the control run only (C=1, F=11). The results are shown at all forecast lead times, but for selected amounts of observed precipitation, namely: wet conditions (> C_p=0); the top 10% of observed precipitation amounts (> C_p=0.9); and the top 1% of observed precipitation amounts (> C_p=0.99). By way of contrast, Figure 32 shows the residual quality of the precipitation forecasts for all (observed) precipitation amounts, but only selected forecast lead times. In both cases, the results are shown for selected verification measures. Positive values of the correlation coefficient and skill scores imply an improvement in forecast quality when using additional ensemble members. In contrast, the RME is a measure of directional bias. Thus, any departure from zero in the residual RME is indicative of a *sensitivity* to the number of ensemble members, and not necessarily to an advantage of using more members. However, as indicated above, the MEFP precipitation forecasts contain a systematic dry bias under (observed) wet conditions. In practice, therefore, a positive residual RME can be interpreted as a reduction in the dry bias when using C=11 members versus C=1 member, while a negative value denotes an increase in the dry bias.

Figure 33 summarizes the quality of the MEFP-GEFS precipitation forecasts for all calibration and forecasting scenarios. The results are shown for selected verification measures and precipitation thresholds. Again, the results comprise a daily aggregation, but the scores for the individual aggregations are averaged over the middle portion of the forecast horizon (4-8 days), where the differences between the calibration scenarios are generally greatest. The bold lines show the calibration scenarios with F=11 forecast members. The dashed lines shows the calibration scenario with the control run only (C=1, F=1). Although composed of a single scenario, the dashed lines connect the (C=1, F=1) scenario to the (C=5, F=11) scenario for each basin. Figure 34 shows the absolute values of the correlation, RME, CRPSS, and BSS for selected calibration and forecasting scenarios at CN-DOSC1. The results are shown at all forecast lead times and for selected amounts of observed precipitation. Figure 35 shows the residual values of the correlation, mean error, CRPSS and BSS for the MEFP temperature forecasts when calibrated with C=11 ensemble members (C=11, F=11) versus the control run only (C=1, F=11). The verification results comprise a daily aggregation and are shown at all forecast lead times.

As indicated in Figures 31 and 32, there is a systematic increase in the quality of the MEFP precipitation forecasts when using all C=11 ensemble members to derive the GEFS ensemble mean (versus the control run only). In general, the benefits are greatest at CN-DOSC1 and NE-HOPR1, where the underlying correlations and skill are greatest. The most noticeable increases in correlation, BSS, and CRPSS occur during the middle portion of the forecast horizon (Figures 31). The largest increases in CRPSS occur at higher thresholds (Figure 32). Unsurprisingly, the patterns in the CRPSS and RME are closely aligned. For example, at high precipitation thresholds, there is a decline in the dry bias when using C=11 members, which leads to an improvement in the residual CRPSS (Figure 32). However, at low and moderate precipitation thresholds, there is an increase in the dry bias with C=11 members, particularly at CN-DOSC1, where the residual CRPSS is also subdued. As indicated in Figure 34, the improvements in BSS, in particular, translate into some improvement in forecast lead time at CN-DOSC1. However, for several metrics and precipitation thresholds, these improvements are much greater against the baseline of (C=1, F=1). For example, Figure 33 shows a large improvement in CRPSS during the middle portion of the forecast horizon when using F=11 (versus F=1) ensemble members, particularly for low and moderate precipitation amounts. In short, an increase the number of forecast members will, to some extent, offset a decrease in the number of calibration members. However, these benefits are not consistent across all basins, forecast lead times, thresholds, or measures of forecast quality (Figure 33). For example, at CN-DOSC1, there is no meaningful improvement in the BSS when using F=11 members in the operational MEFP-GEFS forecast, unless that is accompanied by C=11 members (Figure 33 and Figure 34).

In general, the temperature forecasts are less sensitive to the number of ensemble members used to calibrate the MEFP than the precipitation forecasts (Figure 31). However, for the hottest observed temperatures in AB-CBNK1 and NE-HOPR1, there is a large and systematic increase in the CRPSS when using C=11 members, particularly during the middle portion of the forecast horizon. In practice, this amounts to a gain in forecast lead time of up to ~2 days (not shown). In other words, when using C=11 members, the MEFP predicts the hottest 5% of observed temperatures with similar skill between 168-192 hours as those predicted between 120-144 hours when using only C=1 member. However, the cool and moderate (winter) temperatures are relatively insensitive to the number of ensemble members used to calibrate the MEFP. As such, the impacts on forecasting snowmelt in CB-DRRC2 or NE-HOPR1 are unlikely to be important.

The sensitivity of the HEFS streamflow forecasts to the quality of the MEFP forcing depends jointly on the degree of variability in the forcing quality and the sensitivity of the hydrologic model outputs to the forcing inputs. These sensitivities are expected to be greatest at CN-DOSC1, where the forcing sensitivities are greatest (see above) and the hydrologic models outputs are driven largely by the precipitation inputs. In contrast, the sensitivities should be smallest at CB-DRRC2, because the forcing inputs are less sensitive to the number of ensemble members in the GEFS (see above) and the hydrologic response at CB-DRRC2 is driven by snow accumulation and melting, which moderates the impacts of any forcing sensitivities. Figure 36 shows the residual quality of the HEFS streamflow forecasts when calibrating the MEFP with C=11 ensemble members versus C=1 (with F=11 in both cases). The streamflow forecasts are verified against simulated flows, in order to capture the hydrologic sensitivities to the forcing scenarios independently of any hydrologic biases. The results comprise a daily aggregation and are shown at all forecast lead times for selected thresholds of the simulated streamflow. As described above, the hydrologic response is greatest at CN-DOSC1, where the forecast quality is materially lower when calibrating the MEFP with the control run only, C=1 (Figure 36). At a forecast lead time of 10 days and for observed flows that exceed the median climatological flow (> C_p=0.5), the BSS is improved by around 10% in real terms (or 33% in relative terms) when using C=11 (versus C=1) ensemble members. In keeping with the precipitation forcing, this translates to around ~1 additional day in forecast lead time (not shown). In contrast, the hydrologic response is weakest at CB-DRRC2, where there is little or no advantage of using C=11 members to calibrate the MEFP (Figure 36). At NE-HOPR1 and AB-CBNK1, the hydrologic response is more variable, but the correlation, BSS and CRPSS are systematically higher with C=11 members, particularly during the middle portion of the forecast horizon and at low and moderate streamflow thresholds.

Figure 37 shows the residual quality of the MEFP precipitation forecasts at a daily accumulation when calibrated with C=11 members versus C=5 members (with F=11 in both cases). FigureS 38 and 39 show the equivalent results for the MEFP temperature forecasts and the HEFS streamflow forecasts, respectively. Figure 40 shows the reliability of the streamflow forecasts at a daily aggregation. The results are shown at a forecast lead time of 5 days and for several thresholds of the simulated streamflow, as well as the two calibration scenarios, namely C=5 members (open lines) and C=11 members (closed lines). The cumulative rank histogram measures the (cumulative) fraction of observations that fall within different portions of the ensemble forecast distribution. If the streamflow forecasts are perfectly reliable, there is an equal probability that the simulated streamflow will fall between any two forecast ensemble members when those members are arranged in ascending order (Appendix B). In other words, a perfectly reliable forecast is denoted by a cumulative rank histogram that approaches the diagonal line.

As indicated above, when calibrating the MEFP with a control run only, there is a material decline in the quality of the precipitation, temperature and streamflow forecasts for some (but not all) locations, forecast lead times, and thresholds. However, when calibrating the MEFP with C=5 ensemble members, there is no material decline in the quality of the precipitation, temperature or streamflow forecasts for any of the locations, forecast lead times or thresholds considered. Indeed, the MEFP precipitation forecasts show no meaningful improvement with C=11 ensemble members (Figure 37), while the MEFP temperature forecasts show only a slight improvement for the hottest observed temperatures (Figure 38). Even at CN-DOSC1, the streamflow forecasts show similar BSS and CRPSS when calibrating the MEFP with C=5 and C=11 ensemble members (Figure 39), and there is no discernable improvement in the cumulative rank histograms with C=11 members (Figure 40). In summary, when using a control run to calibrate the MEFP (C=1), there is a material loss of forecast quality, at least for some basins, forecast lead times and thresholds. This is partially recovered when using all F=11 members to generate the forecasts (i.e. C=1, F=11), but only for some basins, thresholds and verification measures. Nevertheless, using all available forecast members is preferred, as no *loss* of quality was identified. In contrast, when calibrating the MEFP with C=5 ensemble members (C=5, F=11), the forcing and streamflow forecasts are no more reliable or skillful than those calibrated with C=11 members (C=11, F=11). As such, for the locations, thresholds, and metrics considered, five ensemble members would be adequate to calibrate the MEFP. Operationally, the HEFS is likely to benefit from all available forecast members (currently F=21).

4.2 Minimum requirements for verifying the HEFS forecasts

Validation of the HEFS relies on a suitably long and consistent archive of streamflow hindcasts, which in turn depends on meteorological reforecasting. In general, the minimum requirements for meteorological reforecasts (in terms of N and M) will be much greater for validation of an operational forecasting system than calibration. Indeed, the HEFS aims is to provide a *parsimonious* description of the meteorological and hydrologic uncertainties (Seo et al., 2006; Wu et al., 2011). In contrast, validation must consider several attributes of forecast quality, including various conditional biases and skill. Crucially, the HEFS must be validated for different thresholds of the observed and forecast variables, including for large events, and for different temporal aggregations. As an operational forecasting system, the HEFS is intended for a broad range of decision contexts, with varying sensitivities to streamflow amount (and temporal aggregations thereof). These vary from short-range flood prediction in headwater basins to seasonal water supply forecasting in large river basins (e.g. April-July volumes), for which sample sizes may be extremely small. Also, streamflow forecasts (unlike many forcing variables) are sensitively dependent upon local conditions, both geographical and antecedent. For example, in snow-dominated basins, the hydrologic response may persist for months or even years through intra- and inter-annual snow accumulation and melting. Here, much of the information content in the forcing is effectively aggregated in time and, therefore, shared between individual forecasts. For a given amount of confidence in the verification results, these statistical dependencies increase the minimum sample size required for verification of the HEFS. Since the focus here is on short-to-medium range forecasting, the minimum requirements depend on shorter (e.g. hourly to multi-day) aggregations. At these timescales, applications include river flood forecasting (sub-hourly to daily) and reservoir inflow forecasting (multi-day volumes). The minimum requirements for long-range forecasting, such as seasonal water supply forecasting, must be determined separately (e.g. for the CFSv2).

Verification is concerned with the statistical properties of paired forecasts and observations. Table 6 shows the average number of verification pairs that might be expected for different scenarios of N and M in which the observed value exceeds a prescribed threshold. Here, the verification pairs comprise a daily aggregation and the threshold is expressed in terms of a climatological probability, C_p. For example, C_p=0.995 denotes an observed amount that is exceeded, on average, once every 200 days. Ignoring leap years, the average number of verification pairs is 365N(1-C_p). However, this is simply the nominal sample size for verification purposes. If the verification pairs are statistically dependent, the *effective* sample size is smaller than the nominal sample size. As the statistical dependencies increase, the effective sample size declines. These effects depend on the choice of measure and, for certain types of statistical dependence and measures thereof (e.g. cross-correlation), they can be quantified analytically. For example, when computing the cross-correlation of two time-series, x and y, that are first-order autocorrelated, the effective sample size, S', is related to the nominal sample size, S (Dawdy and Matalas, 1964):

$$S' = S\frac{1-r_{1,x}r_{1,y}}{1+r_{1,x}r_{1,y}}, \qquad (1)$$

where $r_{1,x}$ and $r_{1,y}$, are the first-order (lag-1) autocorrelations of x and y, respectively. Thus, for a daily streamflow amount, where the forecasts and observations both comprise a lag-1 autocorrelation of 0.9, a nominal sample size of 8760 verification pairs (24 years) amounts to an effective sample size of 8769(1-0.9^2)/(1+0.9^2)=920 verification pairs (2.5 years). The lag-1 autocorrelation of streamflow at AB-CBNK1 for a daily aggregation is 0.542, while the lag-1 autocorrelation at NE-HOPR1 is 0.897. Thus, according to Eqn. 1, the effective sample size for computing the cross-correlation between the ensemble mean forecast and observed variable would be 11% of the nominal sample size at NE-HOPR1 and 55% at AB-CBNK1.

As indicated in Table 6, the nominal sample sizes can decline rapidly with increasing M and decreasing N. For example, on average, a streamflow reforecast with N=25 and M=1 would produce 44 samples above the flood threshold at AB-CBNK1 and 74 samples at NE-HOPR1, while N=10 and M=5 would produce only 4 and 6 samples, respectively. However, the probability of flooding is relatively high at both AB-CBNK1 and NE-HOPR1. At other forecast locations, it may be less than 0.001 and flooding may persist for several days. Thus, in practice, some of these verification pairs will be statistically dependent. Of course, when compared to average conditions, flood events are more likely to be separated in time (M>>1) and, hence, less likely to share information. However, they are unlikely to be statistically independent. Assuming independence as a best case scenario and, specifically, that ~30 independent samples are required to estimate the cross-correlation with "reasonably small" sampling uncertainty, the number of years (N) required to validate flood forecasts at AB-CBNK1 would be: N=30/[0.00484x365], namely ~17 years of reforecasts. In practice, given the temporal and spatial variability of extreme precipitation (except, perhaps, for atmospheric rivers and other large-scale systems), meteorological reforecasts with a longer interval than M=1 could significantly reduce the number of flood events in the sample. Finally, several important verification measures, such as the reliability diagram (Hsu and Murphy, 1986) and relative operating characteristic (ROC; Green and Swets, 1966), require many more samples than lumped verification scores. For example, each bin in the reliability diagram must contain several forecast events, yet flood events are rarely forecast with high probability (e.g. between 0.8 and 1.0). While these estimates are purely indicative, they illustrate the manifest problems associated with validating the HEFS unless an adequate sample of meteorological reforecasts is available to conduct streamflow hindcasting. In this context, a reforecast period much shorter than 25 years of daily reforecasts would limit the validation of the HEFS to relatively moderate streamflow thresholds at most locations, particularly for multi-day aggregations (such as reservoir inflow volumes).

In order to illustrate the effects of N and M on the sampling uncertainties associated with validating the HEFS, each sub-sample of N and M was verified separately and the results compared with the nominal scores for N=24 years and M=1 day. In this context, the sub-periods of N comprise not only sampling variability, but climatological variability on inter-annual timescales. While systematic changes in climate or basin hydrology may undermine the use of verification statistics as guidance, other types of intra- and inter-annual variability, such as teleconnection cycles and seasonality, are important to capture for verification purposes, as they allow for more targeted guidance. Figure 41 shows selected verification scores for the MEFP-GEFS precipitation forecasts at three climatological probabilities. The results are shown for the dependent validation scenarios of N, together with the range of scores across the sub-periods. Thus, for example, the range at N=4 comprises the minimum and maximum scores from the six, four-year, sub-periods, namely, 1985-1988; 1989-1992; 1993-1996; 1997-2000; 2001-2004; and 2005-2008. Each threshold comprises a verification score for all sub-periods; in other words, the range was not computed from fewer scores than the number of sub-periods. The precipitation thresholds are fixed across all sub-periods and were derived from the daily precipitation amounts between 1985 and 2008. Figure 42 shows the corresponding results for M. In this context, the range at M=3 comprises the minimum and maximum scores from the three sub-periods between 1985 and 2008 in which the forecast issue times are separated by three days (offset by one day each time).

As indicated in Figures 41 and 42, the range of verification scores *between* cases of N and M is much smaller than the range of scores *within* cases for different sub-periods. Thus, even for these relatively simple scores, the minimum requirements for validating the MEFP are much greater than the minimum requirements for calibrating the MEFP. Of the scores considered here, the correlation coefficient shows the greatest variability across sub-periods, particularly for smaller N, larger M, and larger precipitation thresholds. Thus, depending on the choice of sub-period within (N=12, M=1) or (N=24, M=3), precipitation amounts that are exceeded, on average, only once every 200 days (C_p=0.995), show substantial variations in correlation. For example, at CN-DOSC1, the two sub-periods of (N=12, M=1) show correlations of -0.04 and 0.36, while the overall period shows a correlation of 0.16. Elsewhere, at AB-CBNK1, the correlations range from -0.1 to 0.6 for (N=24, M=3) at C_p=0.995. Thus, even for relatively simple scores, such as the correlation coefficient, the ability to verify the MEFP for large and extreme events would be significantly reduced by data thinning. Furthermore, the impacts of thinning by M would be no less severe than thinning by N. Indeed, large precipitation events typically occur over hours to days, as evidenced by the verification results here. For more detailed verification metrics, such as the reliability diagram and ROC, which are widely used in operational forecasting, the HEFS would be unverifiable at even moderate thresholds. Figure 43 shows the reliability of the forecast probabilities at selected thresholds for N=12, together with the range of probabilistic biases (plotted as error bars) across the two sub-periods. Figure 44 shows the corresponding results for M=5. As indicated in Figures 43 and 44, the differences between sub-periods of N=12 and M=5 are substantial. In practice, therefore, even at moderate precipitation amounts, such as C_p=0.9, it would be difficult to establish the reliability of the forecasts except, perhaps, in a lumped sense (e.g. using a score decomposition). For larger precipitation amounts, such as C_p=0.99, the reliabilities cannot be determined at high forecast probabilities, because large precipitation amounts are rarely forecast with high probability (see the sample plots, inset). This partly originates from a conditional bias in the precipitation forecasts at large observed thresholds (Figure 19), which would not be addressed by increasing the sample size alone.

Alongside reliability, the extent to which the HEFS can discriminate between observed occurrences and non-occurrences of streamflow events is also important. The ROC 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 a triggering event, such as flooding, a forecasting system will provide skillful warnings if the PoD exceeds the PoFD; that is, when the warnings are more accurate than chance. Furthermore, for a given tolerance to false alarms (PoFD), the ROC shows the forecast probability at which to issue warnings, in order to maximize the PoD, while ensuring that false alarms do not exceed the prescribed PoFD.

Figure 45 shows the PoD and PoFD for each forecast probability associated with the 48 ensemble members in the HEFS streamflow forecasts at NE-HOPR1. Here, the streamflow forecasts were verified against observed flow, rather than simulated flow, and the results are shown for a forecast lead time of 18-42hrs. The period 18-42hrs represents the first 24hr aggregation at which a daily streamflow observation (5Z-5Z) can be paired with a streamflow forecast (6Z-6Z), in order to minimize the timing error between them (1 hour). As the forecasts are issued at 12Z, the first 6Z-6Z period occurs between 18-24hrs. For a given forecast probability and a prescribed tolerance to false alarms (PoFD), Figure 45 shows the ability of the HEFS to classify observed streamflows that exceed the flood threshold at NE-HOPR1. The results are shown for (N=24, M=3) and for each sub-period thereof. For illustrative purposes, a PoFD (and corresponding forecast probability) is highlighted that allows for no more than 1.5% of flood warnings to be incorrect, i.e. PoFD=0.015. As indicated in Figure 45, the PoFD is less than or equal to 1.5% at a forecast probability of 2/48, except for the first and second sub-periods of M=3, when the forecast probability is 4/48. Thus, a degraded reforecast could lead to a poor choice of probability threshold for issuing flood warnings; that is, a threshold for which some true positives are sacrificed or more false positives are allowed than prescribed. For example, in choosing a probability threshold of 4/48 for the overall period (Figure 45a), the desired PoFD would be met, but flood warnings would be correct on only 40% of occasions, when they could be correct on 58% of occasions. In practice, the sampling uncertainties associated with flooding would undermine such a prescriptive approach to optimizing the warning threshold. However, these results provide a decision context in which a degraded reforecast could materially impact the validation of the HEFS and, hence, decisions about water resources that are guided by validation results.

5. Summary and recommendations

The Hydrologic Ensemble Forecast Service (HEFS) quantifies the total uncertainty in future streamflow as a combination of the meteorological forcing uncertainties and the hydrologic modeling uncertainties, while correcting for biases in both the raw forcing and streamflow forecasts. Reliable and skillful weather and climate forecasting is central to reliable and skillful streamflow forecasting and, in many cases, accounts for the majority of uncertainty in hydrologic forecasting. The HEFS Meteorological Ensemble Forecast Processor (MEFP) quantifies the meteorological uncertainties and corrects for biases in the raw forcing inputs to the HEFS. For medium-range forecasting, the MEFP relies on forcing inputs from the Global Ensemble Forecast System (GEFS) of the National Centers for Environmental Prediction (NCEP). The operational viability of the HEFS and similar hydrologic ensemble forecasting services depends on the availability of an adequate sample of meteorological reforecasts. These meteorological reforecasts are required to calibrate and validate the MEFP, generate streamflow hindcasts for validation, and to allow end users of the HEFS to calibrate and validate their own forecasting and decision support systems. End users of the HEFS, such as the New York City Department of Environmental Protection (NYCDEP) and the Yuba County Water Agency (YCWA), rely on the HEFS for critical water supply decisions. For example, the NYCDEP is using the HEFS to improve the management of risks to water quantity and quality objectives in the NYC area. In this context, the NYCDEP has developed an Operational Support Tool (OST), which ingests streamflow forecasts from the HEFS that are produced operationally by the Middle-Atlantic RFC and the Northeast RFC. The OST optimizes the volume and quality of water stored in the NYC reservoirs, while avoiding unnecessary, multi-billion dollar, infrastructure costs, such as water filtration.

Extreme hydrologic events are particularly important to users of the HEFS, including those of drought and flooding whose consequences (e.g. for ecosystems, recreation and water quality) may be regionally or nationally significant. For example, Deutsche Bank Securities estimated that the 2012-14 drought reduced the U.S. Gross Domestic Product by 0.5-1 percentage point in 2012 alone (Richter, 2012), while flooding on the Red River of the North during the winter of 1996-97 caused over $5 billion in damage (Perry, 2005). At smaller scales, hydrologic extremes are profoundly important for local economies. For example, according to Scott and Lemieux (2010), the 2002 drought on the Colorado River substantially curtailed the rafting season and led to $50 million in lost revenue, with some outfitters losing over 40% of their normal business. Given the manifest uncertainties associated with forecasting extreme weather, the ability of the HEFS to provide not only a central forecast, but a range of possible outcomes, is an important advantage over deterministic forecasting systems. However, the availability of an adequate archive of meteorological reforecasts is critically important to the NYCDEP and other users of the HEFS, both in calibrating their existing decision support systems and improving them for future extremes. Thus, understanding the minimum requirements for calibrating and validating the HEFS; that is, the statistical or sampling requirements; is a *necessary but not a sufficient* condition for understanding the minimum requirements of end users for meteorological and hydrologic reforecasts. This report focuses on the minimum requirements for calibrating and validating the HEFS. The requirements of end users, such as the NYCDEP and YCWA, will be gathered and presented separately.

In order to establish the minimum requirements for meteorological reforecasting in support of calibrating and validating the HEFS, a 26-year reforecast dataset was obtained for the GEFSv10 (Hamill et al., 2013). The minimum requirements for calibrating the HEFS are determined by the MEFP, for which several statistical parameters must be estimated. Among other factors, the costs associated with meteorological reforecasting depend on the historical period considered (N years), the interval between reforecasts (M days) and the number of ensemble members (C). By sub-sampling the GEFSv10 reforecasts, the MEFP was calibrated for different combinations of N, M and C. The sensitivities of the temperature and precipitation forecasts from the MEFP and the streamflow forecasts from the HEFS were then explored through hindcasting and validation. Specifically, the forcing and streamflow hindcasts were produced and validated for a fixed historical period (based on daily reforecasting). Within this fixed period, the calibration of the MEFP varied according to N, M and C. In principle, for atmospheric variables that are statistically dependent over multiple days, thinning by M should have a smaller impact than an equivalent N. In practice, however, except for large scale systems, such as atmospheric rivers, precipitation generally varies over short time periods and at small spatial scales, particularly in variable terrain. Thus, thinning a reforecast by M may not be less aggressive than an equivalent N.

In order to ensure that the hindcasting was both practical and statistically reasonable, a combination of dependent and (limited) cross-validation was used. Specifically, in exploring the sensitivities to N, a 24-year period between 1985 and 2008 was sub-divided into smaller calibration and forecasting periods, namely N={2x12, 3x8, 4x6, and 6x4} years. Dependent validation involved calibrating the MEFP and generating hindcasts for each sub-period and then pooling all of the sub-periods for validation. Independent validation involved borrowing the parameters from an adjacent sub-period. In this context, dependent validation may be regarded as a “best case scenario” for the expected forecast quality, while borrowing the parameters from adjacent sub-periods may be regarded as a “worst case scenario”, as the MEFP is likely to be re-calibrated more frequently than even N=4 years. In evaluating the sensitivities to M, the MEFP was calibrated for M={1, 3, 5, and 7} days and hindcasts produced and validated between 1985 and 2008. In this context, M=1 represents dependent validation, whereas M={3, 5 and 7} include a mixture of dependent and independent samples. The sensitivities to C were examined by calibrating the MEFP with an ensemble mean derived from C={1, 5, and 11} ensemble members and forecasting with a mean derived from F=11 ensemble members. In practice, the GEFSv10 reforecasts contain fewer ensemble members (11) than the operational GEFS forecasts (21). This leads to a discrepancy between the calibration and operational use of the MEFP. In order to examine the impacts of this discrepancy, a baseline reforecast was also produced. The baseline involved calibrating the MEFP and forecasting with the control run only (C=1, F=1). In examining the sensitivities to C and F, the MEFP was calibrated and validated for a 26-year period between 1985 and 2010.

Overall, there is no systematic decline in forecast quality with increasing M or decreasing N, either for the MEFP precipitation and temperature forecasts or for the HEFS streamflow forecasts. While the forecast quality is slightly lower under independent validation than dependent validation, these differences are likely to originate from climatological variability over the sub-periods considered, rather than a meaningful advantage from dependent validation. In this context, climatological non-stationarities may introduce a trade-off between larger N (smaller sampling uncertainty) and smaller N (greater climatological specificity). Indeed, rather than a systematic increase in forecast quality with increasing N, the forecast quality generally *decreases* with increasing N. Given the sampling uncertainties, this should not be overstated. However, it may originate from greater specificity of the estimated parameters at smaller N. In keeping with the results for N, there is no systematic decline in forecast quality with increasing M. Nevertheless, when considering the *sensitivity* of the verification scores to N and M, measured by the range of scores across these scenarios, there are meaningful differences, particularly at higher thresholds of precipitation and streamflow. In this context, sensitivity is a necessary but not a sufficient condition for a decline in forecast quality, and these results imply some sensitivity to N and M, but they do not suggest a systematic decline in forecast quality with increasing M or decreasing N (and generally the opposite for N). In practice, some of this variability is likely to originate from sampling uncertainty in the verification statistics. Indeed, for the same reason, the minimum requirements for reforecasting will be determined largely by the need to validate the HEFS with reasonably small sampling uncertainty (see below). Thus, a multi-year reforecast archive (e.g. 12+ years) should be adequate to calibrate the MEFP, but it would not be adequate to validate the HEFS, at least for high thresholds.

When calibrating the MEFP with one ensemble member (C=1), there is a systematic decline in the quality of the MEFP forcing, as well as the HEFS streamflow forecasts. This is partially offset by using all ensemble members to generate the MEFP forecasts (F=11). However, the best available calibration (C=11, F=11) remains materially better in most cases. For precipitation and streamflow, the greatest improvements occur at CN-DOSC1, particularly in the middle and latter portion of the forecast horizon, where the calibration with C=11 members (F=11) produces a similar BSS to the control run (C=1, F=11) at a forecast lead time of 1+ additional days. The improvements in temperature are greatest at AB-CBNK1 and NE-HOPR1, particularly at the hottest observed temperatures and during the middle portion of the forecast horizon, where the CRPSS increases by ~10% in real terms (~30% relative to the baseline CRPSS). In contrast, when calibrating the MEFP with C=5 ensemble members (C=5, F=11), the forcing and streamflow forecasts do not materially improve on those calibrated with C=11 members (C=11, F=11). A gradual decline in the benefits of adding ensemble members should also be expected from sampling theory. Indeed, under certain conditions, the standard error of the ensemble mean is inversely related to the square root of the sample size and, therefore, declines only gradually as the sample size increases (more gradually as the autocorrelation increases). Empirically, for the locations, thresholds, and verification metrics considered, five ensemble members should be adequate to calibrate the MEFP, while the operational forecasts would benefit from using all available ensemble members. A similar conclusion was reached by Hamill et al. (2014) who recommended, as an acceptable compromise between cost and accuracy, that the GEFS reforecasts should contain five ensemble members for statistical post-processing. However, this cannot be generalized to all post-processing techniques or, indeed, to future implementations of the

MEFP. Furthermore, the benefits of reforecasting with additional ensemble members will vary with forecast conditions, and they may be greater for extreme events (for which sample measures of forecast quality are inherently limited). Thus, any reduction in C should be reviewed as models and applications evolve and diagnostic techniques become more sophisticated.

In general, more historical data will be required to validate an operational forecasting system than to estimate its statistical and other parameters. In this context, the parameters of the MEFP do not vary with threshold, whereas forecasting applications (and thus validation) are strongly threshold-dependent. Likewise, validation must consider several attributes of forecast quality, including various conditional biases and skill, which are important for guiding operational practice, but generally require much larger sample sizes. The minimum requirements for validation of the HEFS can be examined both theoretically and empirically.

Theoretically, verification is concerned with the sampling properties of statistical measures. Among other factors, these sampling properties depend on the number of samples available and their unique information content (statistical independence). However, they also depend on the consistency (stationarity) of the verification measure over the times and locations from which the samples are pooled. Thus, changes in meteorological or basin conditions may alter the trade-off between sample size and representativeness. Even for short-to-medium range forecasting, which is generally concerned with hourly to multi-day aggregations, the sample sizes required to verify extreme events may be prohibitive. For example, at a daily aggregation, flooding may occur at a climatological probability of C_p=0.995 or larger. A reasonable estimate of a lumped verification score, such as the correlation coefficient or CRPSS, may require 30 or more independent samples. More complex, but operationally valuable, metrics, such as the reliability diagram and ROC, require many more samples (perhaps 100-200). If all of the large (e.g. >C_p=0.995) events in a verification sample are statistically independent (a flattering assumption) and reforecasts are issued once per day, approximately 30/365(1-0.995)=16.5 years of reforecasts would be required, on average, to sample 30 events larger than C_p=0.995. Clearly, these requirements increase dramatically with increasing C_p. They also increase as the amount of unique information in the sample declines, which is inevitable if flooding lasts for multiple days. For example, when estimating the cross-correlation between two time-series, any autocorrelations in the individual time-series will increase the sample size required for a given amount of statistical confidence (Dawdy and Matalas, 1964). If the forecasts and observations each comprise a lag-1 autocorrelation of 0.9 (not uncommon in snow basins), a nominal sample size of 8760 verification pairs (24 years) amounts to an effective sample size of $8769(1-0.9^2)/(1+0.9^2)=920$ verification pairs (2.5 years). While precipitation is typically correlated over much shorter time-scales than streamflow, this also increases the probability that data thinning (e.g. from M=1 to M=3) would significantly reduce the number of extreme events in the sample. In practice, a reforecast period much shorter than 25 years with daily forecasts would limit the validation of the HEFS to relatively moderate precipitation and streamflow thresholds.

Empirically, the effects of reducing the number of reforecasts available is to increase the sampling uncertainty of the verification results and to render some events unverifiable, typically those (larger) events that are most important for decision making. In order to illustrate the effects of N and M on the sampling uncertainties associated with validating the HEFS, each sub-sample of N and M was verified separately and the results compared with the nominal scores for N=24 years and M=1 day. Here, the range of verification scores *between* cases of N and M is much smaller than the range of scores *within* cases for different sub-periods. Thus, as anticipated, the minimum requirements for validating the MEFP are much greater than the minimum requirements for calibrating the MEFP, even for relatively simple verification scores. Of the verification scores considered here, the correlation coefficient is particularly variable across the sub-periods of N and M. For example, at AB-CBNK1, precipitation amounts that exceed $C_p=0.995$ show correlations of between -0.1 and 0.6 in the three sub-periods of M=3. Thus, for a 1-in-200 day precipitation amount at AB-CBNK1, forecasts issued every three days over a 24-year period would be unverifiable. For more detailed verification metrics, such as the reliability diagram, the thresholds for which the HEFS remains verifiable are even smaller. For example, at N=12 or M=5, the sample sizes are too small to evaluate reliability diagrams for even moderately large precipitation amounts ($C_p=0.99$). This is evidenced by the broad range of results across the two sub-periods of N=2 and the five sub-periods of M=5. It is also exacerbated by the small sample sizes at high forecast probabilities; in practice, high thresholds are rarely forecast with high probability (even if the smaller bins are well-populated), partly because the precipitation forecasts are conditionally biased at large observed amounts.

In summary, therefore, the minimum requirements for meteorological reforecasting in support of the HEFS are determined, primarily, by the need to validate the HEFS with reasonably small sampling uncertainty, including for large events. As the MEFP provides a relatively parsimonious description of the forecasting errors, a much shorter, multi-year, period (of perhaps 12 or more years) is required for calibration. Among other factors, the minimum requirements for validation depend on the thresholds of interest (greater for higher thresholds), the statistical dependencies between samples (greater when the dependencies are larger), the aggregation periods of interest (greater for larger aggregations), the verification measures used (larger for more detailed measures) and any other conditions that reduce the verification sample size (e.g. seasonal verification). In practice, however, as an operational forecasting system, the HEFS is intended for a broad range of applications, including for drought and flood prediction. In this context, simple, unconditional, measures cannot guide operational practice, because they are not application-specific. For example, a flood warning may be issued when the forecast probability of flooding exceeds a prescribed threshold. In this context, there is trade-off between issuing warnings too regularly (low probability threshold) and failing to warn when floods actually occur (high probability threshold). Both of these risks involve costs and benefits, which should be quantifiable. Crucially, given an adequate sample of historical flood occurrences (and non-occurrences), this trade-off, and hence the probability threshold at which to trigger flood warnings, can be defined, objectively, through hindcasting and validation. By way of illustration, the use of a degraded reforecast of M=3 at NE-HOPR1 could lead to flood warnings that are correct on only 40% of occasions, when they could be correct on 58% of occasions for a warning threshold optimized to daily reforecasts. For users of the HEFS, such as the NYCDEP, a long and consistent record of historical forecasts is, therefore, essential; it is necessary to optimize and improve decision support systems and to benchmark these systems against historical analogs for future extremes.

Clearly, reforecasting requires both significant human and computational resources. However, unsophisticated approaches to data thinning, such as reducing the number of historical years (N) or increasing the period between reforecasts (M), will also reduce the value of these datasets for hydrologic applications. Rather, any approach to data thinning must accommodate a reasonable sample of large and extreme events (i.e. sufficient historical years or targeted sampling of extremes) and allow for rapidly evolving hydrometeorological conditions (i.e. produce sufficiently frequent reforecasts). Alongside approaches to data thinning, spatial pooling or regionalization may be used to improve the sample sizes for calibration and validation of the MEFP. However, spatial pooling cannot satisfy user requirements for long historical records at individual forecast locations. Furthermore, in validating streamflow forecasts, spatial pooling would be fraught with difficulty, as hydrologic state variables, unlike atmospheric state variables, are “geographically embedded.” In other words, they vary over short distances, and with myriad basin characteristics.

6. 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

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)

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

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

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 Service. 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

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

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

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

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

Reforecast interval – The interval between consecutive reforecasts (e.g. 1 day)

Reforecast period – The historical period for which reforecasts are available (e.g. 20 years)

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

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

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

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

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

7. References

Anderson, T.W. 1962. On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, **33** (3), 1148–1159.

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

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

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

Brown, J. D., Demargne, J., Seo, D-J, and Liu, Y. 2010. 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. 2013. Verification of long-range temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service. Technical Report prepared by Hydrologic Solutions Limited for the U.S. National Weather Service, Office of Hydrologic Development, 128pp. [Available at:

http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/Contract_2012-04-HEFS_Deliverables_03_05_Phase_II_report_FINAL.pdf, accessed 07/12/2014).

Brown, J.D. 2014. Verification of temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service: an evaluation of the medium-range forecasts with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS) and a comparison to the frozen version of NCEP's Global Forecast System (GFS). Technical Report prepared by Hydrologic Solutions Limited for the U.S. National Weather Service, Office of Hydrologic Development, 139pp. [Available at: http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/Contract_2013-09-HEFS_Deliverable_02_Phase_III_report_FINAL.pdf, accessed 07/12/2014).

Brown, J.D., Wu, L., He, M., Regonda, S., Lee, H. and Seo, D-J. 2014a. Verification of temperature, precipitation and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS). 1. Experimental design and forcing verification. *Journal of Hydrology* **519D**, 2869–2889, doi:10.1016/j.jhydrol.2014.05.028

Brown, J.D., He, M., Regonda, S., Wu, L., Lee, H. and Seo, D-J. 2014b. Verification of temperature, precipitation and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS). 2. Streamflow verification. *Journal of Hydrology* **519D**, 2847-2868, doi:10.1016/j.jhydrol.2014.05.030

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

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

Dawdy, D.R., and Matalas, N.C., 1964. Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series, in Ven Te Chow, ed., *Handbook of Applied Hydrology*: A Compendium of Water-Resources Technology. McGraw-Hill Book Company: New York, pp. 8.68-8.90.

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* **95**(1), 79-98, doi: 10.1175/BAMS-D-12-00081.1

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

Hamill, T.M., Alcott, T., Antolik, A., Brown, J.D., Charles, M., Collins, D.C., Fresch, M., Gilbert, K., Guan, H., Herr, H., Hogsett, W., Novak, D., Ou, M., Rudack, D., Schafer, P., Scheuerer, M., Wagner, G., Wagner, J., Workoff, T., Veenhuis, B., and Zhu, Y. 2014. White paper on a recommended reforecast configuration for the NCEP Global Ensemble Forecast System. Unpublished. March, 2014.

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* **94**, 1553–1565. doi: 10.1175/BAMS-D-12-00014.1

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

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

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

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

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

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

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

Perry, C.A. 2005. Summary of significant floods in the United States and Puerto Rico, 1994 through 1998 water years. U.S. Geological Survey Scientific Investigations Report 2005-5194. Washington, DC.

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 12/07/14].

Raff, D., Brekke, L., Werner, K., Wood, A. and White, K. 2013. *Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information*. A report of the U.S. Army Corps of Engineers (USACE), Bureau of Reclamation, and the National Oceanic and Atmospheric Administration (NOAA), CWTS 2013-1 [Available at: http://www.ccawwg.us/docs/Short-Term_Water_Management_Decisions_Final_3_Jan_2013.pdf, accessed 07/12/14].

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.

Richter, J. 2012. U.S. Drought May Cut GDP by 1 Percentage Point, Deutsche Says. Bloomberg.com, November 12th, 2012. [Available at: http://www.bloomberg.com/news/2012-11-12/u-s-drought-may-cut-gdp-by-one-percentage-point-deutsche-says.html, accessed 07/12/2014].

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

Scott, D. and Lemieux, C. 2010. Weather and climate information for tourism. *Procedia Environmental Sciences* **1**, 146-183. doi:10.1016/j.proenv.2010.09.011.

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.

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

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

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

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.

8. Tables

Table 1: Characteristics of the study basins

Characteristic	**CBNK1**	**DRRC2**	**DOSC1**	**HOPR1**
USGS basin identifier	07151500	09165000	11473900	01118000
Latitude (outlet)	37.1292	37.6389	39.71	41.4981
Longitude (outlet)	-97.6017	-108.06	-123.32	-71.7169
Latitude (GEFS, week 1)	37.2171	37.6853	39.5578	41.0
Longitude (GEFS, week 1)	-97.5	-108.2812	-123.2812	-72.0
Latitude (GEFS, week 2)	37.123	37.7469	39.6186	41.0
Longitude (GEFS, week 2)	-97.5	-108.125	-123.125	-72.0
Area (total, km^2)	2057	275	1930	188
Mean elevation (m)	115	2567	340	19
Annual P (mm)	935.68	961.94	1682.36	1339.17
Annual PE (mm)	1264.43	1034.34	876.23	765.7
P/PE	0.74	0.93	1.92	1.75
$\mathrm{F_P^{-1}(0.9)}$ (mm)	6.47	7.17	13.62	11.8
$\mathrm{F_P^{-1}(0.95)}$ (mm)	14.75	12.29	26.28	21.87
$\mathrm{F_P^{-1}(0.995)}$ (mm)	47.61	29.27	71.63	58.38
Runoff coefficient	0.12	0.45	0.42	0.55
Q_{action} (m^3/s^{-1})	85.23	N/A	N/A	17.56
Q_{flood} (m^3/s^{-1})	148.38	N/A	N/A	21.52
$1 - \mathrm{F_Q(Q_{action})}$	0.0117	N/A	N/A	0.01546
$1 - \mathrm{F_Q(Q_{flood})}$	0.00484	N/A	N/A	0.00815
$\mathrm{F_Q^{-1}(0.1)}$ (m^3/s^{-1})	0.63	0.41	0.39	0.95
$\mathrm{F_Q^{-1}(0.95)}$ (m^3/s^{-1})	23.14	18.1	169.99	11.93
$\mathrm{F_Q^{-1}(0.995)}$ (m^3/s^{-1})	152.83	36.71	646.56	25.43

P = total precipitation
PE = potential evaporation
Q = streamflow
Q_{action} = streamflow at action stage
Q_{flood} = streamflow at flood stage
F = climatological probability distribution of the subscripted variable at a one-day aggregation

Table 2: Reforecast configuration parameters

Reforecast variable	**Significance for MEFP**	**Assessed here?**
Number of historical years (N)	An important control on the amount of “unique” data available (i.e. forecasts of events that are unrelated or only minimally related to each other). More years of historical data should improve the calibration of the MEFP (to a point). The precise trade-off between the number of years of data and the quality of the MEFP outputs will depend on many factors, including the importance of extreme events, forecast location, and season.	Yes, for limited combinations of years.
Interval between reforecasts (M)	There are two separate controls here, namely the number of daily cycles (e.g. 00UTC) and the interval (in days) between cycles. In future, reforecasts may be restricted to a daily cycle (00UTC) every 5 or 7 days. Of these two factors, the interval between cycles is likely to be more important. Particularly for heavy and extreme precipitation, a cycle every 5 or 7 days would substantially reduce the probability of capturing rapidly evolving extremes (e.g. hurricanes; atmospheric rivers). However, under moderate and dry conditions, this should be less important. Also, depending on the forcing variable considered (e.g., temperature versus precipitation), correlations generally persist for several days, reducing the need for more frequent runs.	Yes, partly. The interval (in days) between each 00UTC cycle is considered. The reforecasts only include the 00UTC cycle.
Number of ensemble members for calibration (C) and forecasting (F)	The number of ensemble members will impact the ensemble mean. Based on optimal estimation theory, the ensemble mean forecast will out-perform any single-valued forecast from the same model, on average. However, if the ensemble mean is based on a small number of ensemble members, this advantage is reduced. Also, differences between the number of ensemble members in the reforecasts and operational forecasts could potentially impact the statistical properties of the forecasts.	Yes, for limited scenarios, constrained by the total number of ensemble members in the GEFS reforecasts (11).
Horizontal resolution	Ideally, the reforecasts would use the same configuration as the operational forecasts. A higher horizontal resolution should improve forecast skill in smaller basins and variable terrain. Differences between the reforecast and operational configuration could potentially impact the statistical properties of the forecasts.	No.
Vertical resolution	Ideally, the reforecasts would use the same configuration as the operational forecasts. A higher vertical resolution should improve forecast skill. Differences between the reforecast and operational configuration could potentially impact the statistical properties of the forecasts.	No.
Other variables (model physics etc.)	There are many other controls on the configuration of the forecast model. Ideally, the reforecasts would use the same options and parameters as the operational forecasts. For example, the operational and reforecast models should be initialized with the same or similar analysis (in terms of methodology). Differences between the reforecast and operational configuration could potentially impact the statistical properties of the forecasts.	No.

Table 3: Calibration and validation scenarios for N

Calibration years (N)	Dependent calibration and (hindcast/validation) years	Independent calibration and (hindcast/validation) years	Validation years
24	1985-2008 (1985-2008)	N/A	1985-2008
12	1985-1996 (1985-1996); 1997-2008 (1997-2008)	1985-1996 (1997-2008); 1997-2008 (1985-1996)	1985-2008
8	1985-1992 (1985-1992); 1993-2000 (1993-2000); 2001-2008 (2001-2008)	1985-1992 (2001-2008); 1993-2000 (1985-1992); 2001-2008 (1993-2000)	1985-2008
6	1985-1990 (1985-1990); 1991-1996 (1991-1996); 1997-2002 (1997-2002); 2003-2008 (2003-2008)	1985-1990 (1991-1996); 1991-1996 (1985-1990); 1997-2002 (2003-2008); 2003-2008 (1997-2002)	1985-2008
4	1985-1988 (1985-1988); 1989-1992 (1989-1992); 1993-1996 (1993-1996); 1997-2000 (1997-2000); 2001-2004 (2001-2004); 2005-2008 (2005-2008)	1985-1988 (1989-1992); 1989-1992 (1985-1988); 1993-1996 (1997-2000); 1997-2000 (1993-1996); 2001-2004 (2005-2008); 2005-2008 (2001-2004)	1985-2008

Table 4: Calibration and validation scenarios for M

Interval (M days)	Calibration frequency and (period)	Validation frequency and (period)
1	Every 1 day (1985-2008)	Every 1 day (1985-2008)
3	Every 3 days (1985-2008)	Every 1 day (1985-2008)
5	Every 5 days (1985-2008)	Every 1 day (1985-2008)
7	Every 7 days (1985-2008)	Every 1 day (1985-2008)

Table 5: Calibration and validation scenarios for the number of ensemble members

Scenario ID	Calibration members (C)	Forecast/validation members (F)	Validation frequency and (period)
(C=1, F=11)	1	11	Every 1 day (1985-2010)
(C=5, F=11)	5	11	Every 1 day (1985-2010)
(C=11, F=11)	11	11	Every 1 day (1985-2010)
(C=1, F=1)	1	1	Every 1 day (1985-2010)

Table 6: Average sample sizes by climatological probability (C_p) and reforecast scenario

Threshold (C_p)	**Reforecast configuration scenario (N years, M days)**								
	(30,1)	**(25,1)**	**(20,1)**	**(15,1)**	**(10,1)**	**(25,3)**	**(25,5)**	**(25,7)**	**(10,5)**
0.05	10403	8669	6935	5201	3468	2890	1734	1238	694
0.1	9855	8213	6570	4928	3285	2738	1643	1173	657
0.25	8213	6844	5475	4106	2738	2281	1369	978	548
0.5	5475	4563	3650	2738	1825	1521	913	652	365
0.75	2738	2281	1825	1369	913	760	456	326	183
0.95	548	456	365	274	183	152	91	65	37
0.99	110	91	73	55	37	30	18	13	7
0.995	55	46	37	27	18	15	9	7	4
0.999	11	9	7	5	4	3	2	1	1
0.9995	5	5	4	3	2	2	1	1	0
$Q>Q_{action}$ (AB-CBNK1)	128	107	85	64	43	36	21	15	9
$Q>Q_{flood}$ (AB-CBNK1)	53	44	35	26	18	15	9	6	4
$Q>Q_{action}$ (NE-HOPR1)	169	141	113	85	56	47	34	20	11
$Q>Q_{flood}$ (NE-HOPR1)	89	74	59	45	30	25	18	11	6

Q = streamflow
Q_{action} = action stage
Q_{flood} = flood stage

9. Figures

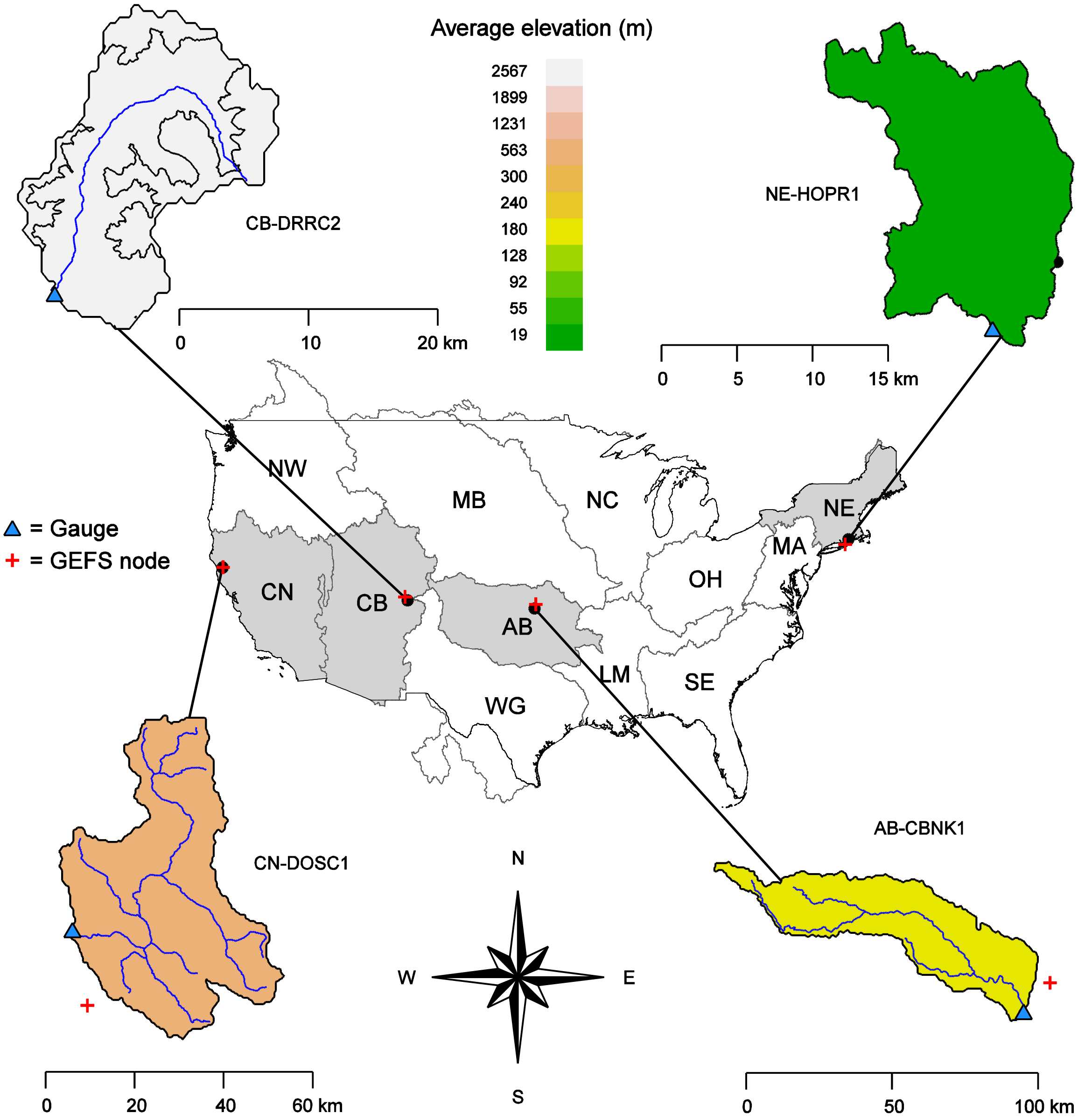

Figure 1: The four study basins, including their average elevation, the location of each outlet (gaging station), and the positions of the nearest grid nodes in the GEFS.

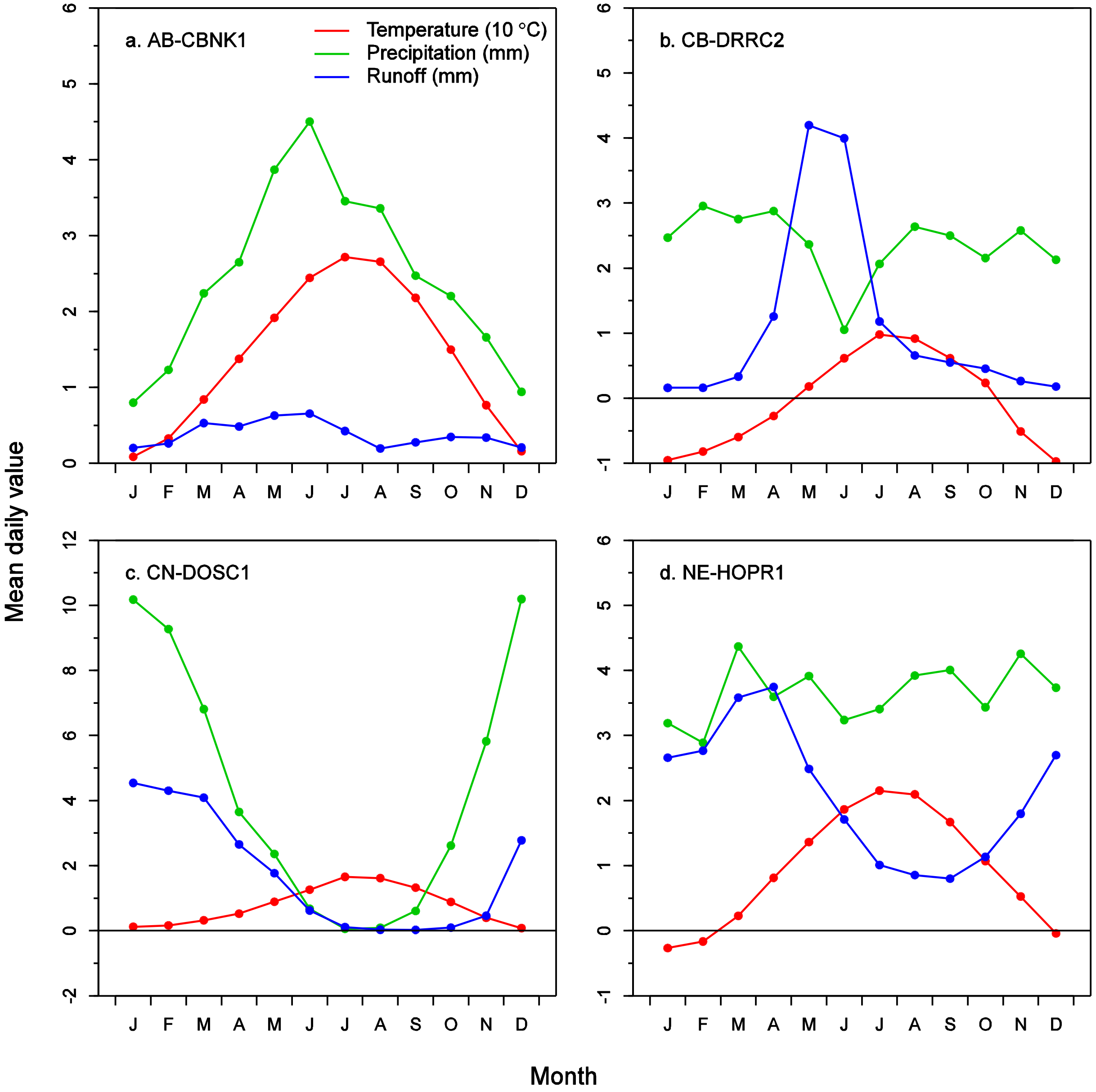

Figure 2: Daily average temperature, total daily precipitation and daily average streamflow by calendar month for each study basin.

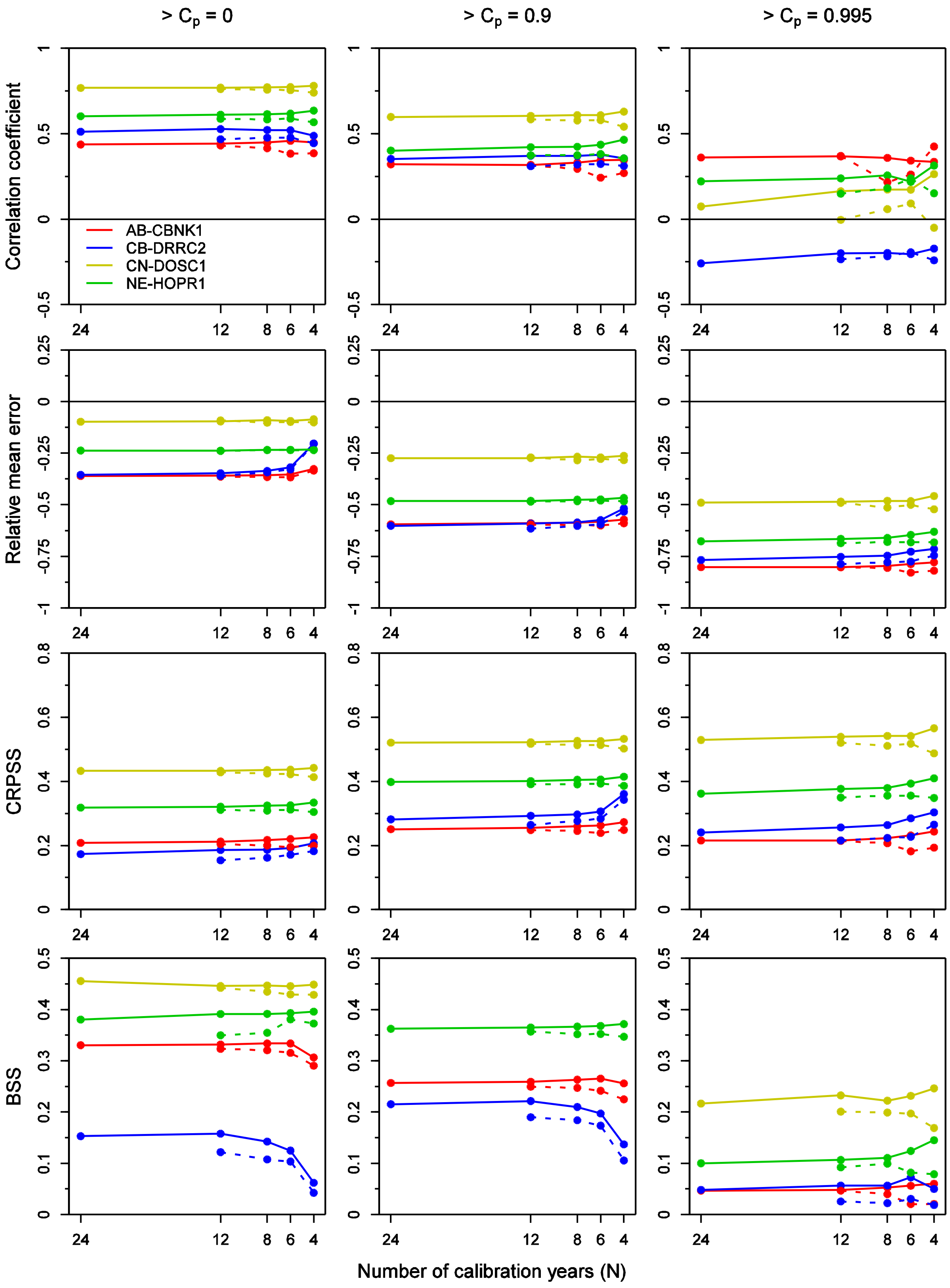

Figure 3: Selected verification metrics for the MEFP-GEFS precipitation forecasts. The results are shown for the dependent (solid) and independent (dashed) validation scenarios of N (the number of years of calibration data), and include several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

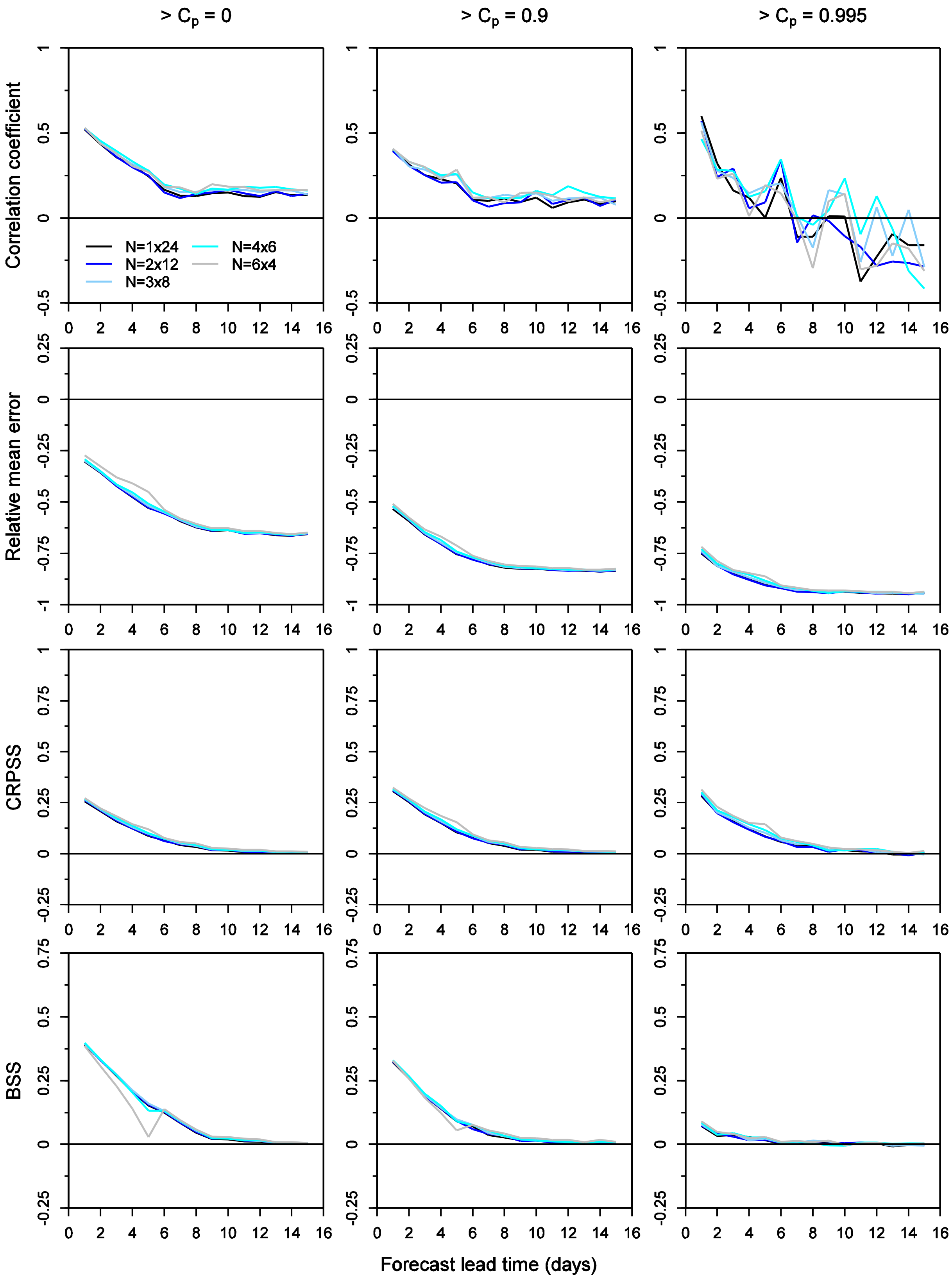

Figure 4: Selected verification metrics for the MEFP-GEFS precipitation forecasts at AB-CBNK1. The results are plotted against forecast lead time for each scenario of N (the number of years of calibration data), and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

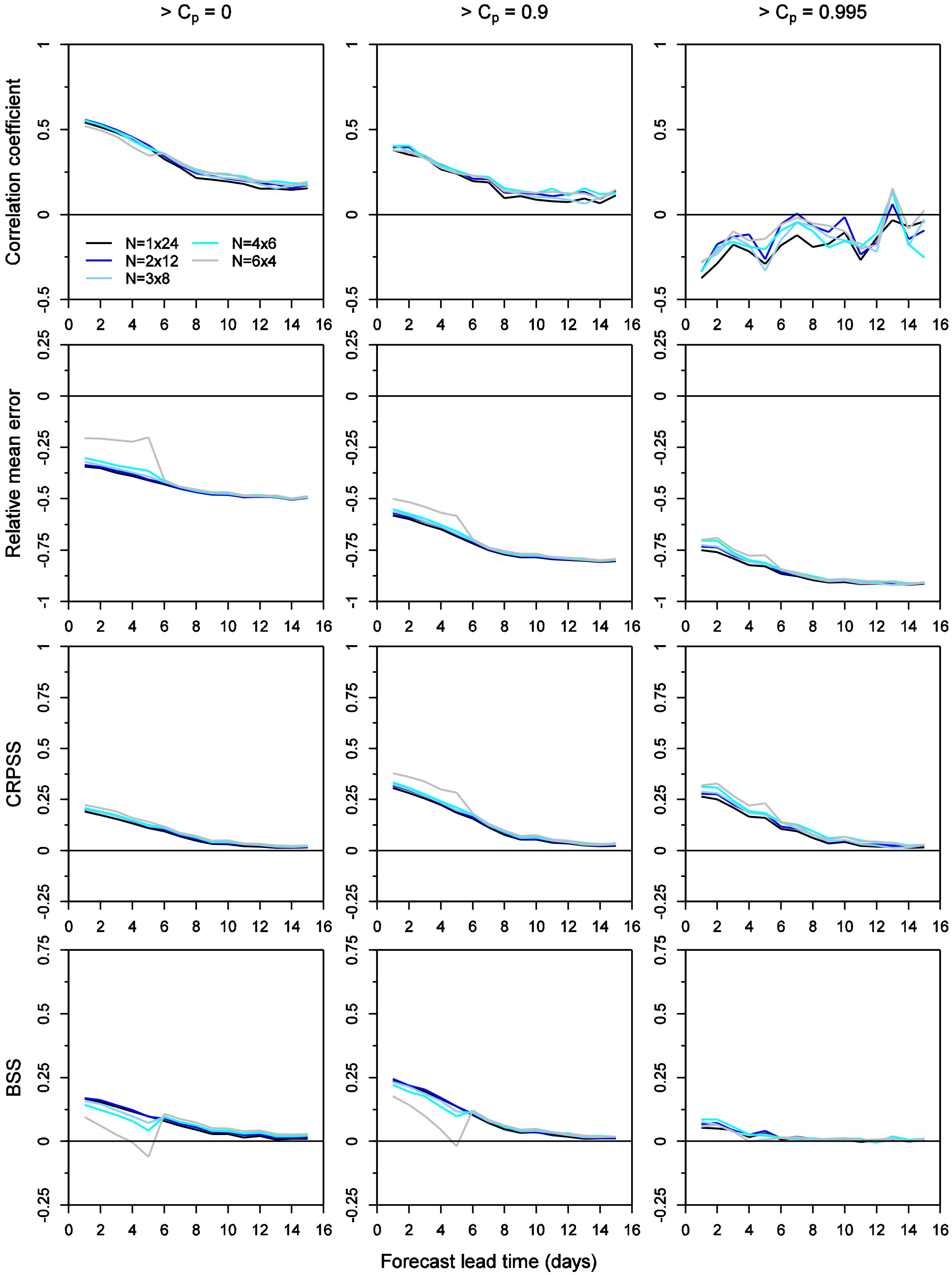

Figure 5: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CB-DRRC2. The results are plotted against forecast lead time for each scenario of N (the number of years of calibration data), and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

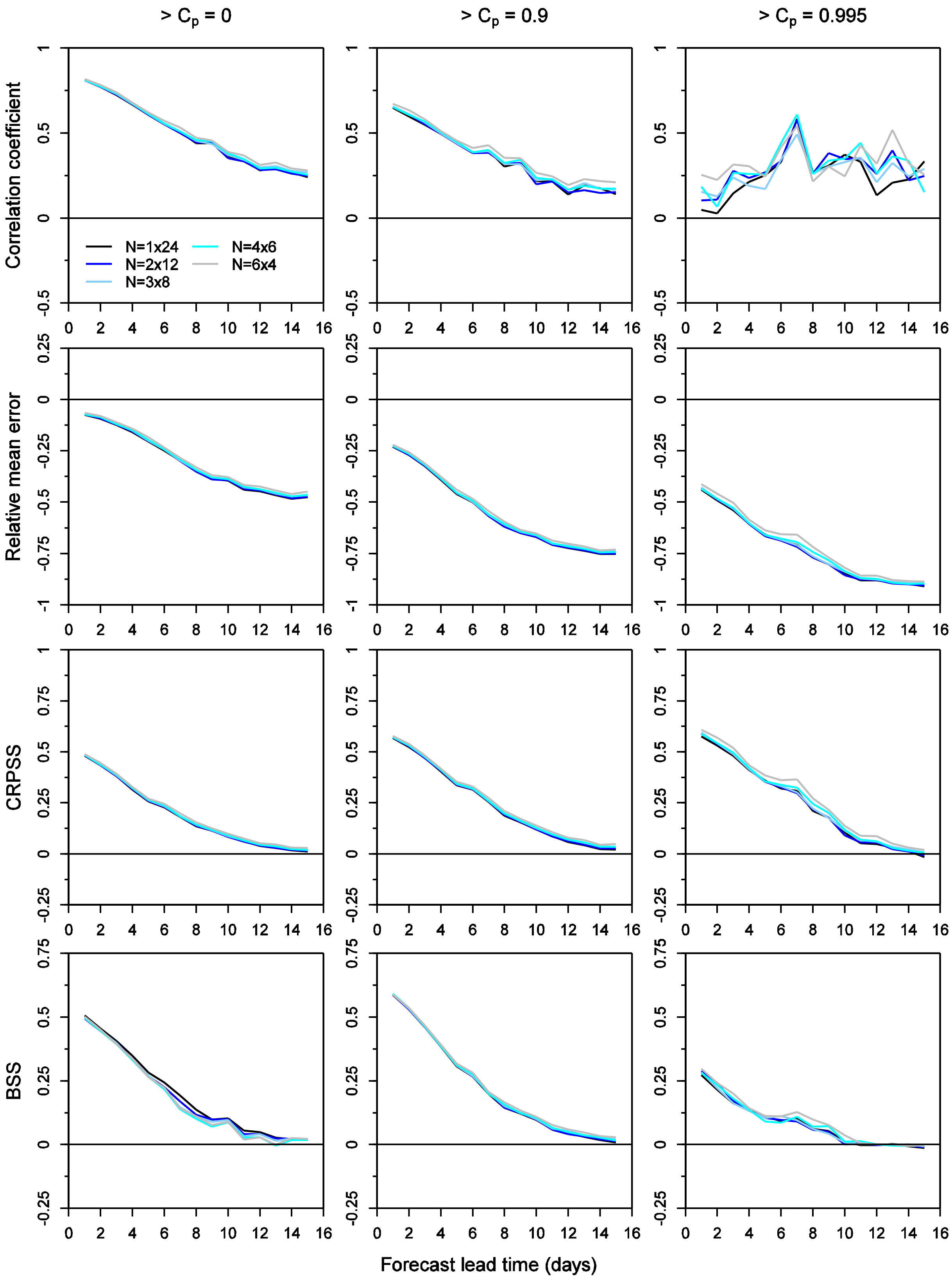

Figure 6: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CN-DOSC1. The results are plotted against forecast lead time for each scenario of N (the number of years of calibration data), and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

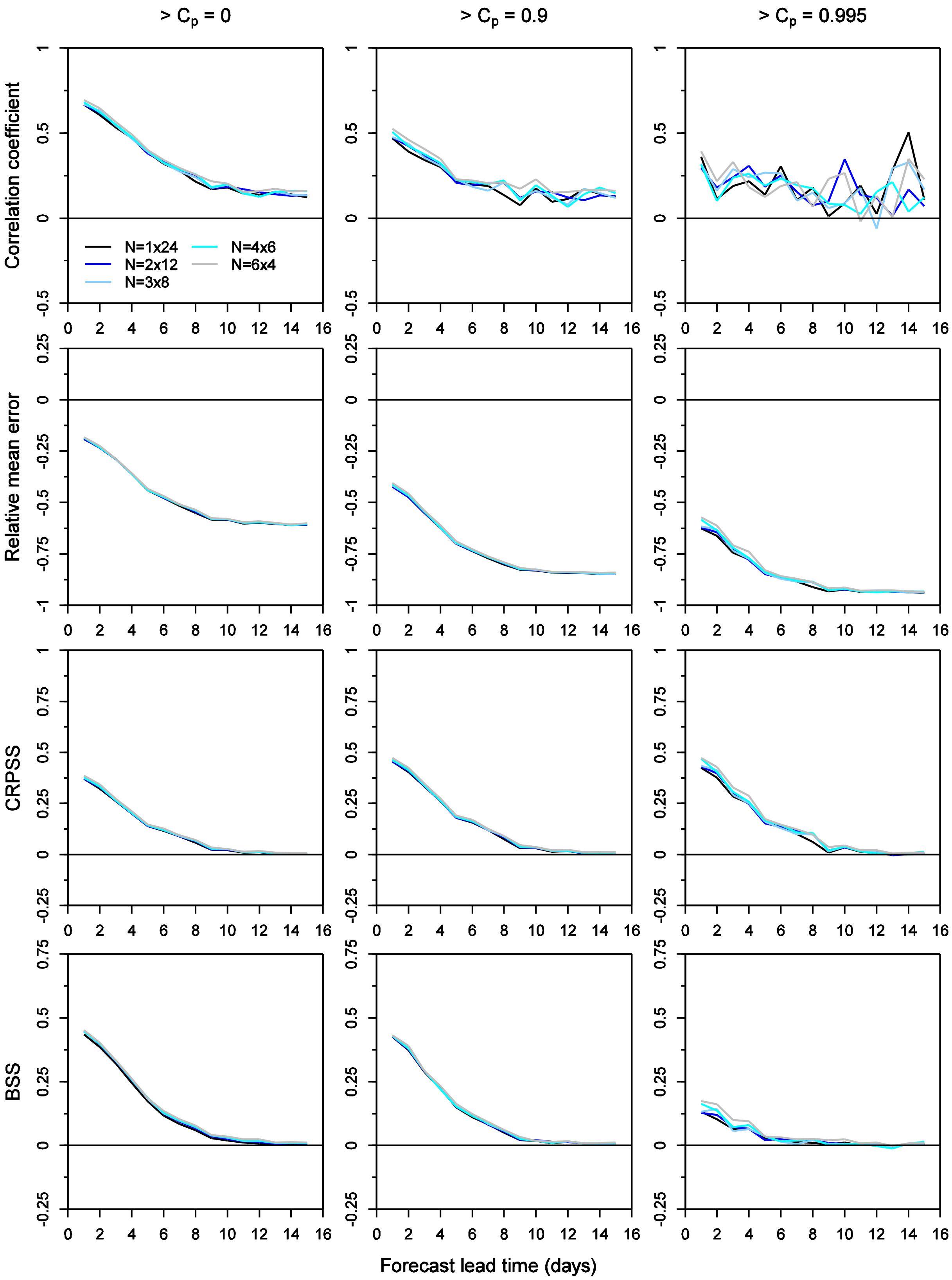

Figure 7: Selected verification metrics for the MEFP-GEFS precipitation forecasts at NE-HOPR1. The results are plotted against forecast lead time for each scenario of N (the number of years of calibration data), and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

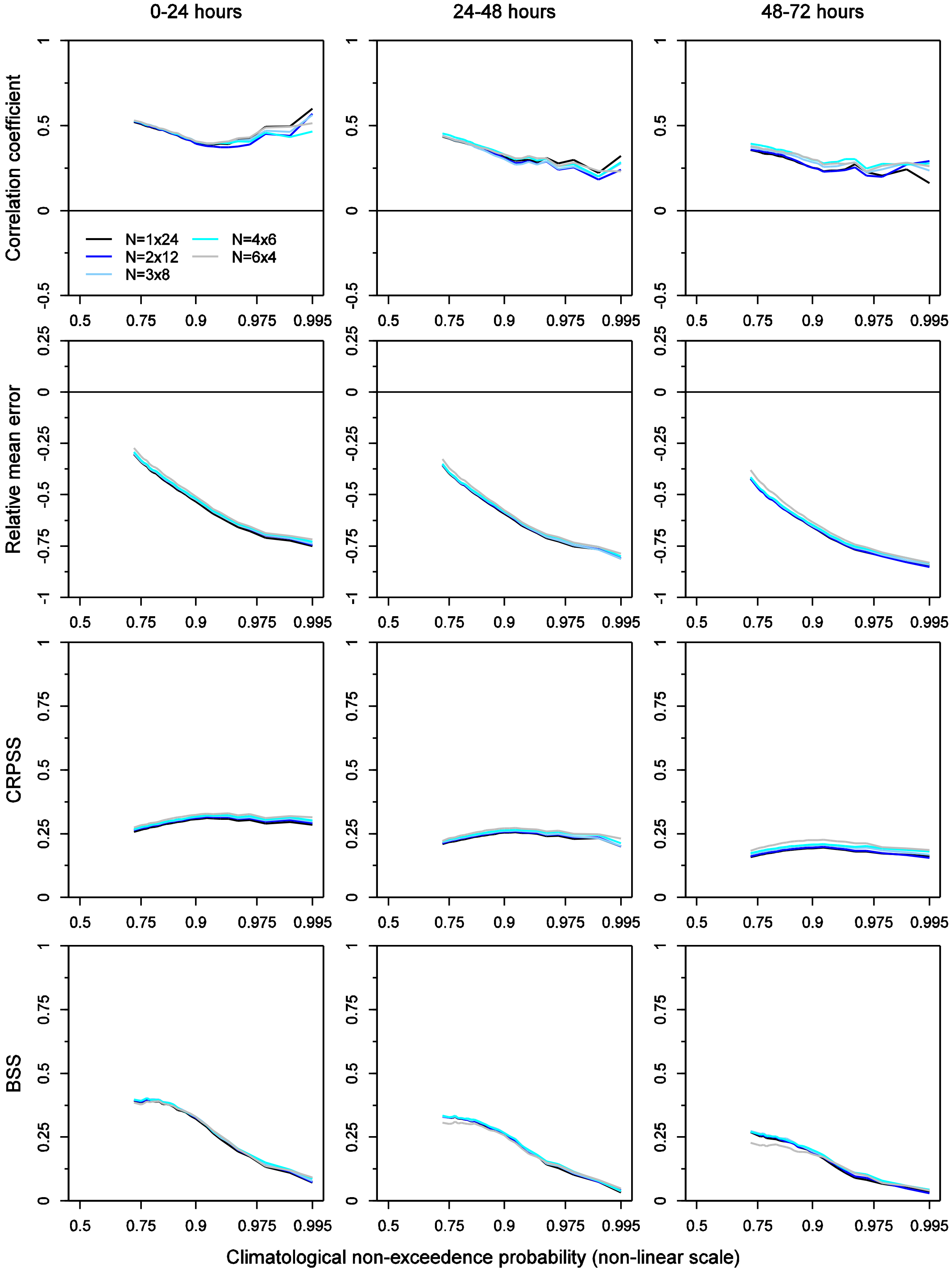

Figure 8: Selected verification metrics for the MEFP-GEFS precipitation forecasts at AB-CBNK1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

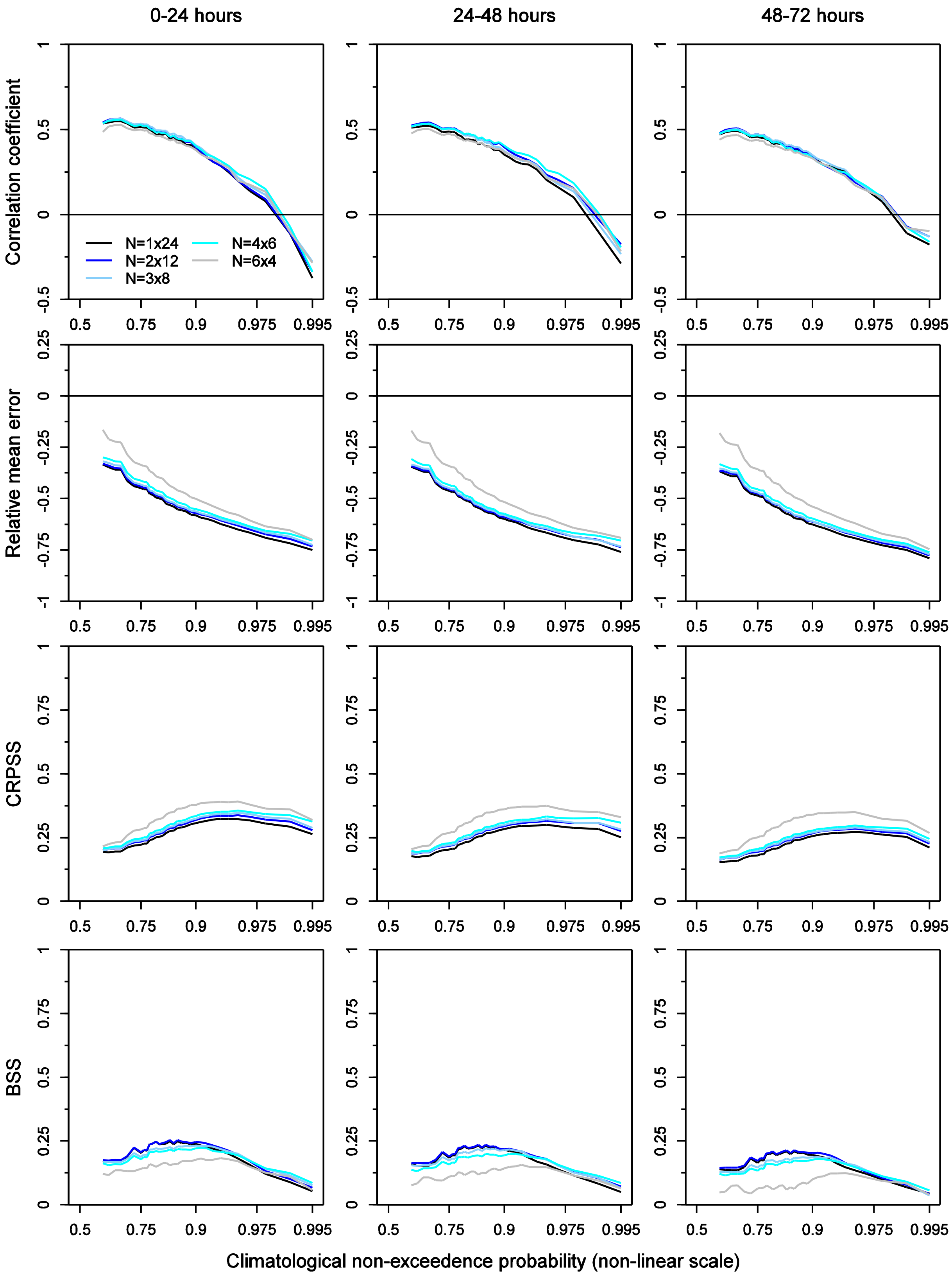

Figure 9: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CB-DRRC2. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

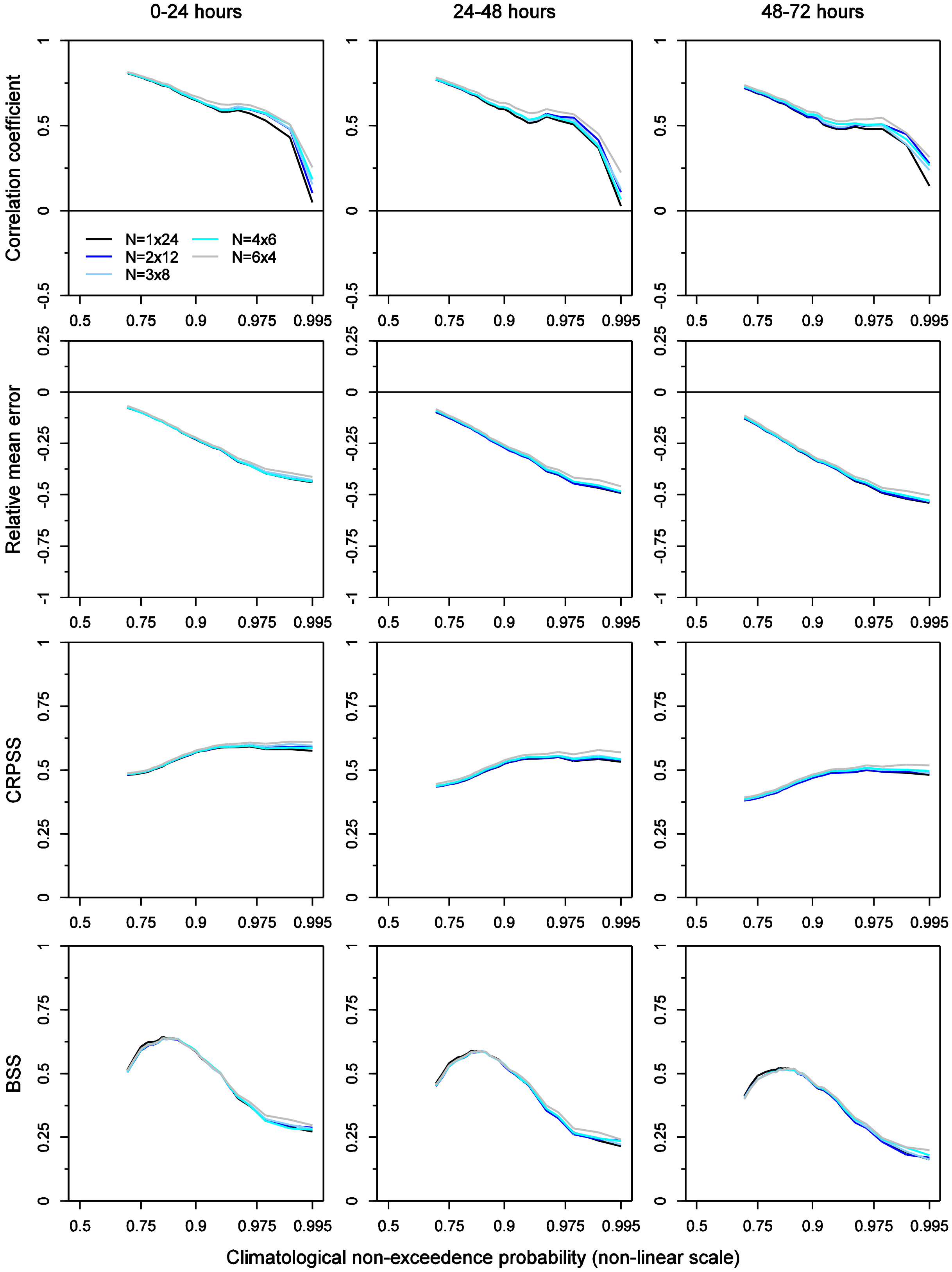

Figure 10: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CN-DOSC1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

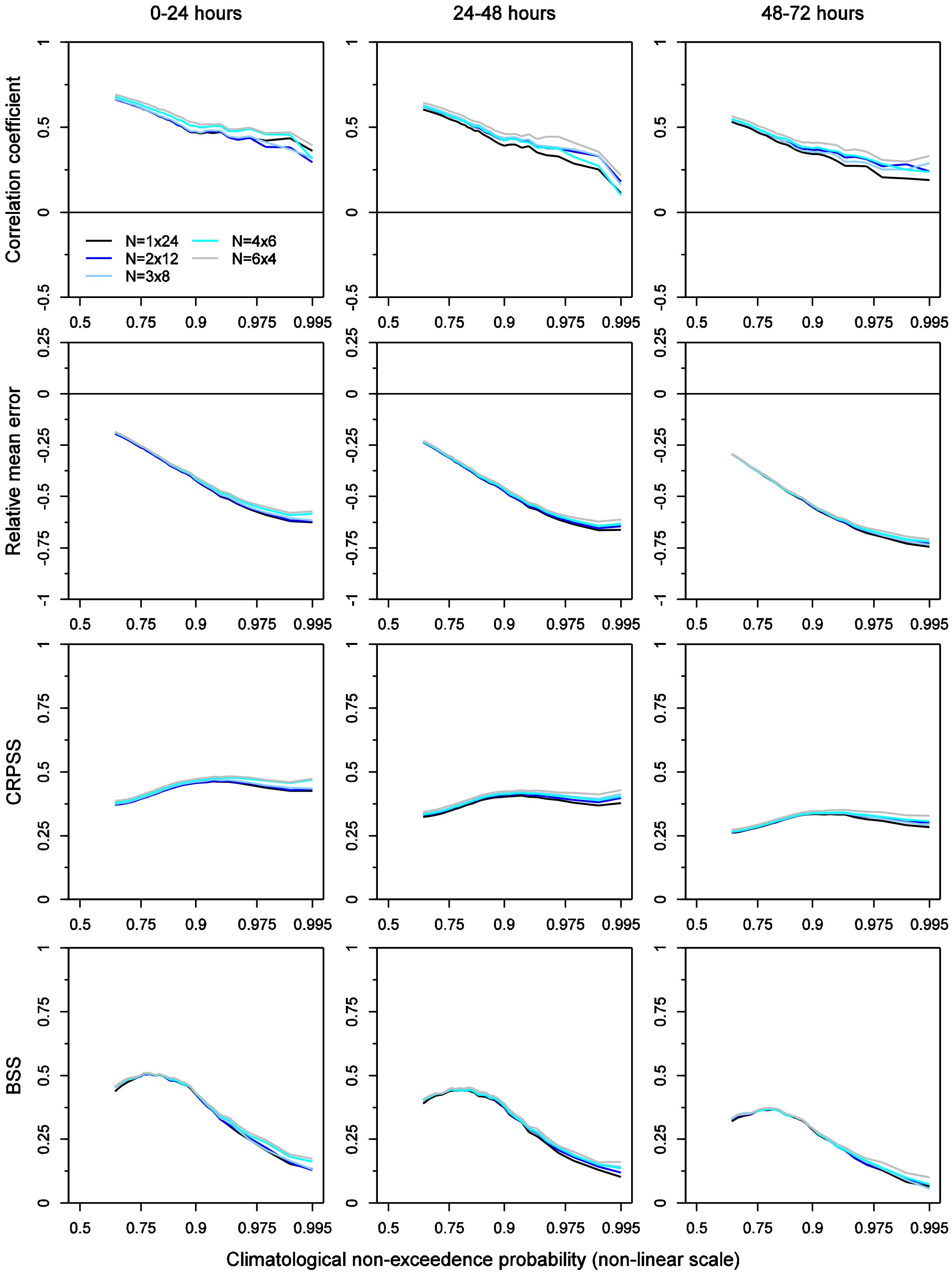

Figure 11: Selected verification metrics for the MEFP-GEFS precipitation forecasts at NE-HOPR1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

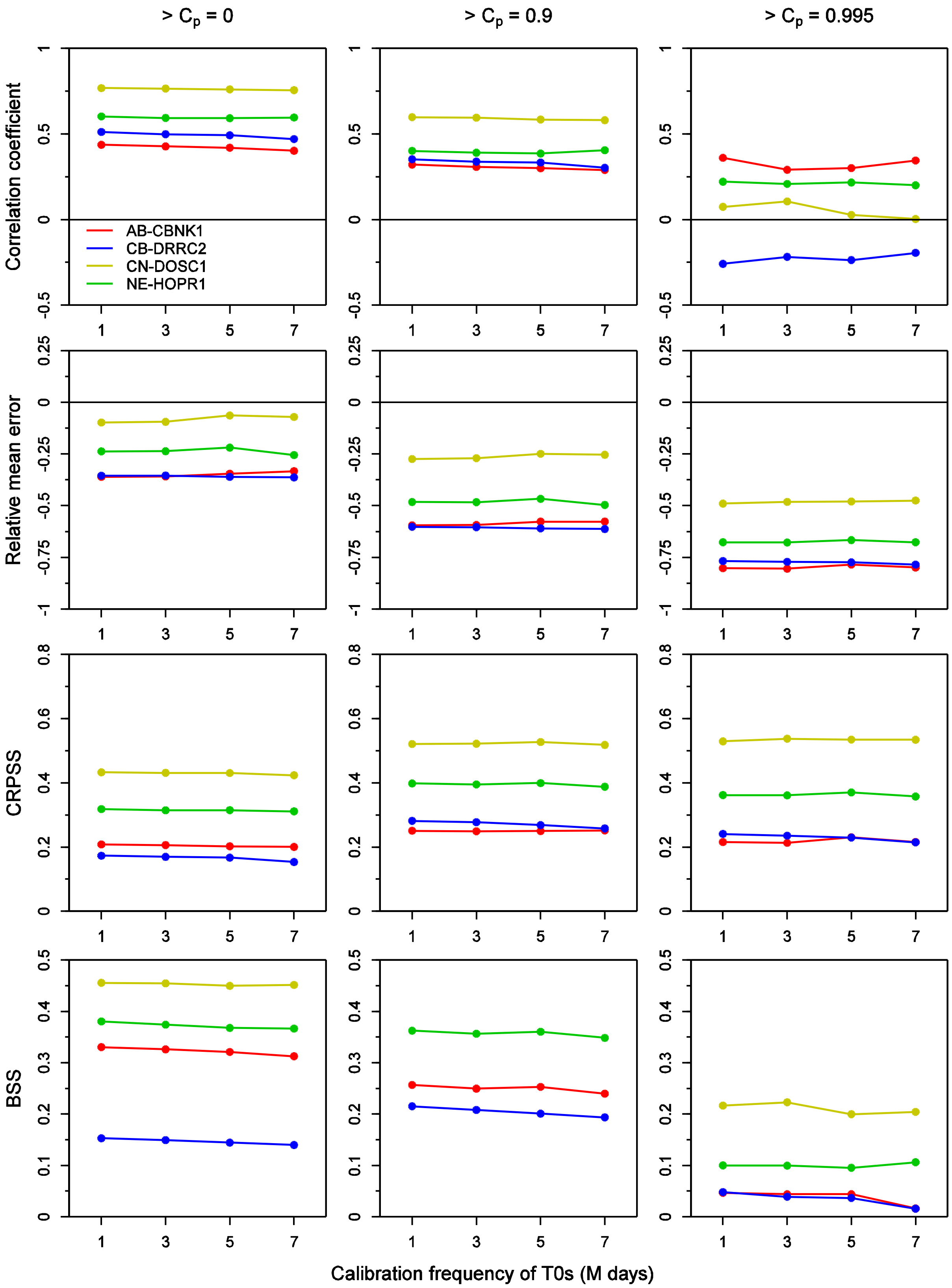

Figure 12: Selected verification metrics for the MEFP-GEFS precipitation forecasts. The results are plotted against the interval between reforecasts (M days) used to calibrate the MEFP, and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

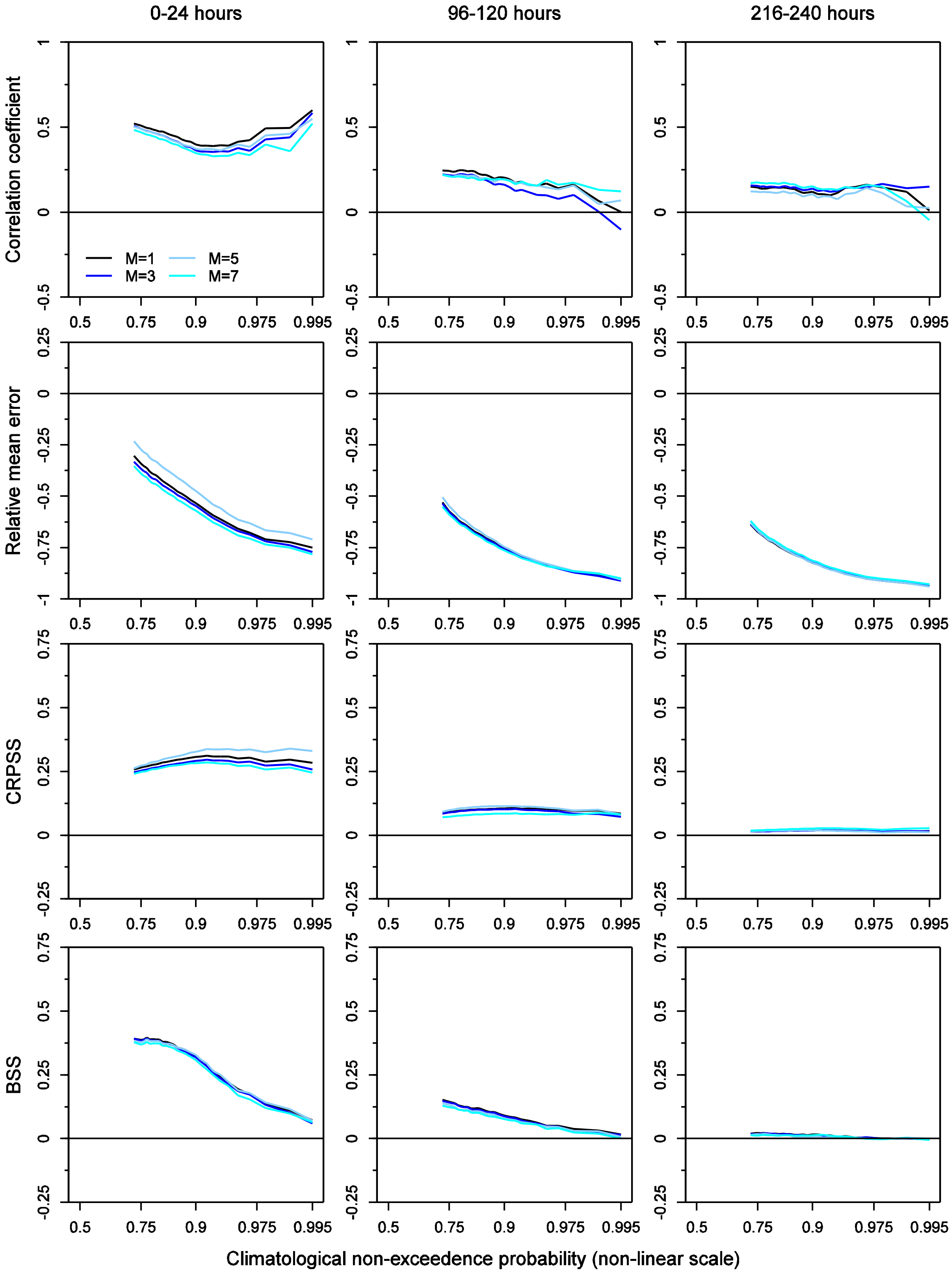

Figure 13: Selected verification metrics for the MEFP-GEFS precipitation forecasts at AB-CBNK1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of M (the interval between reforecasts in days), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

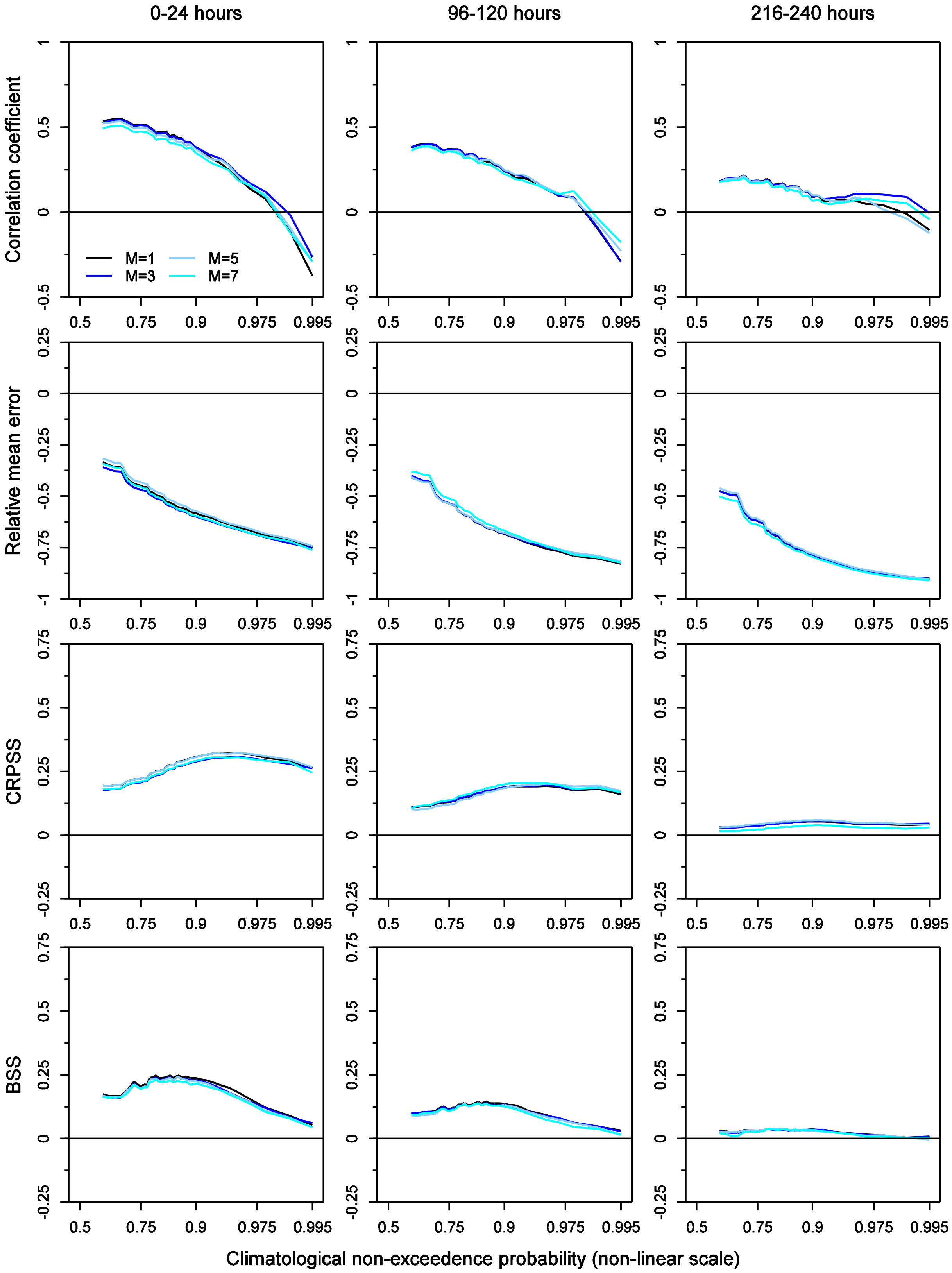

Figure 14: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CB-DRRC2. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of M (the interval between reforecasts in days), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

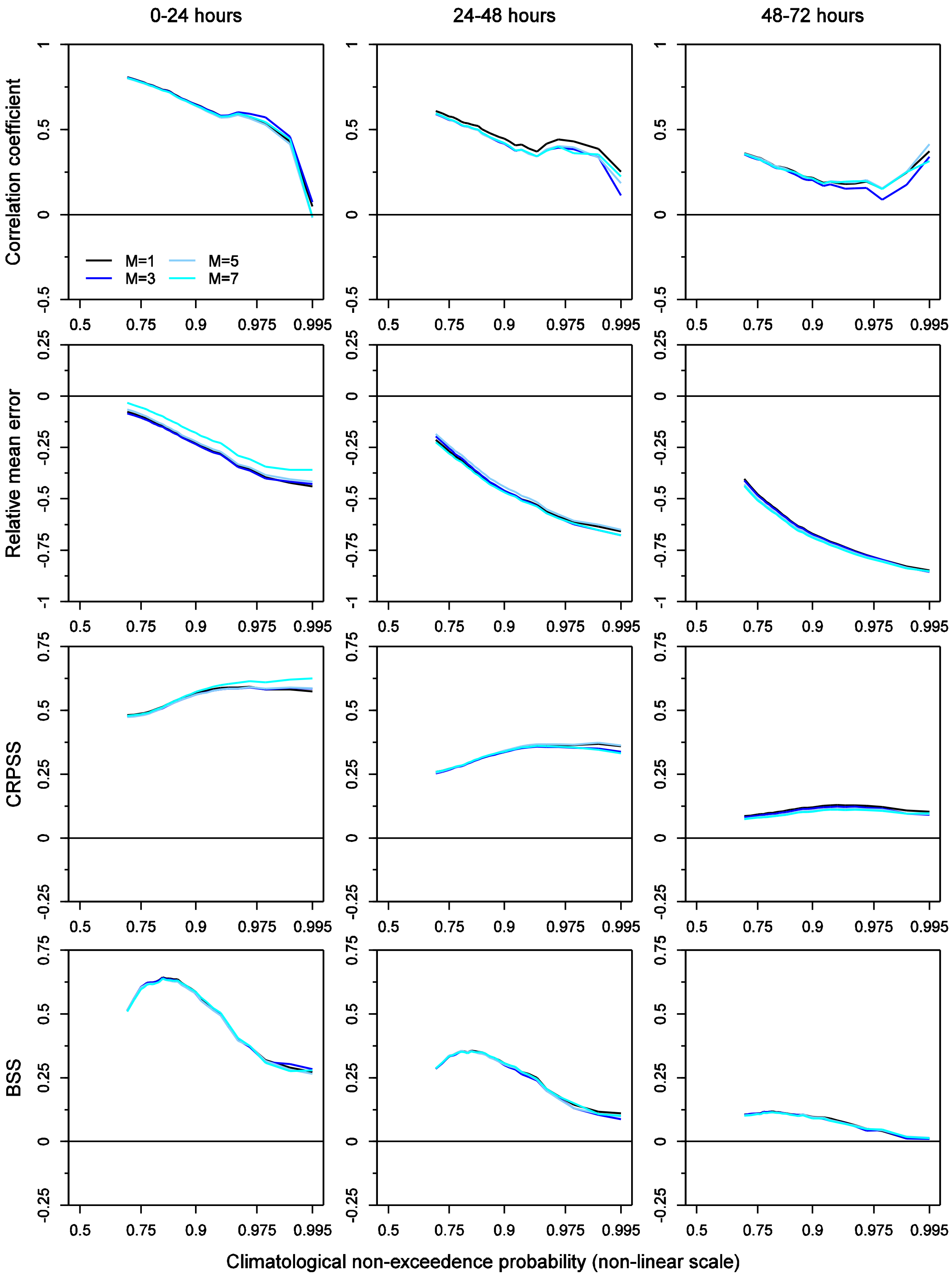

Figure 15: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CN-FTSC1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of M (the interval between reforecasts in days), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

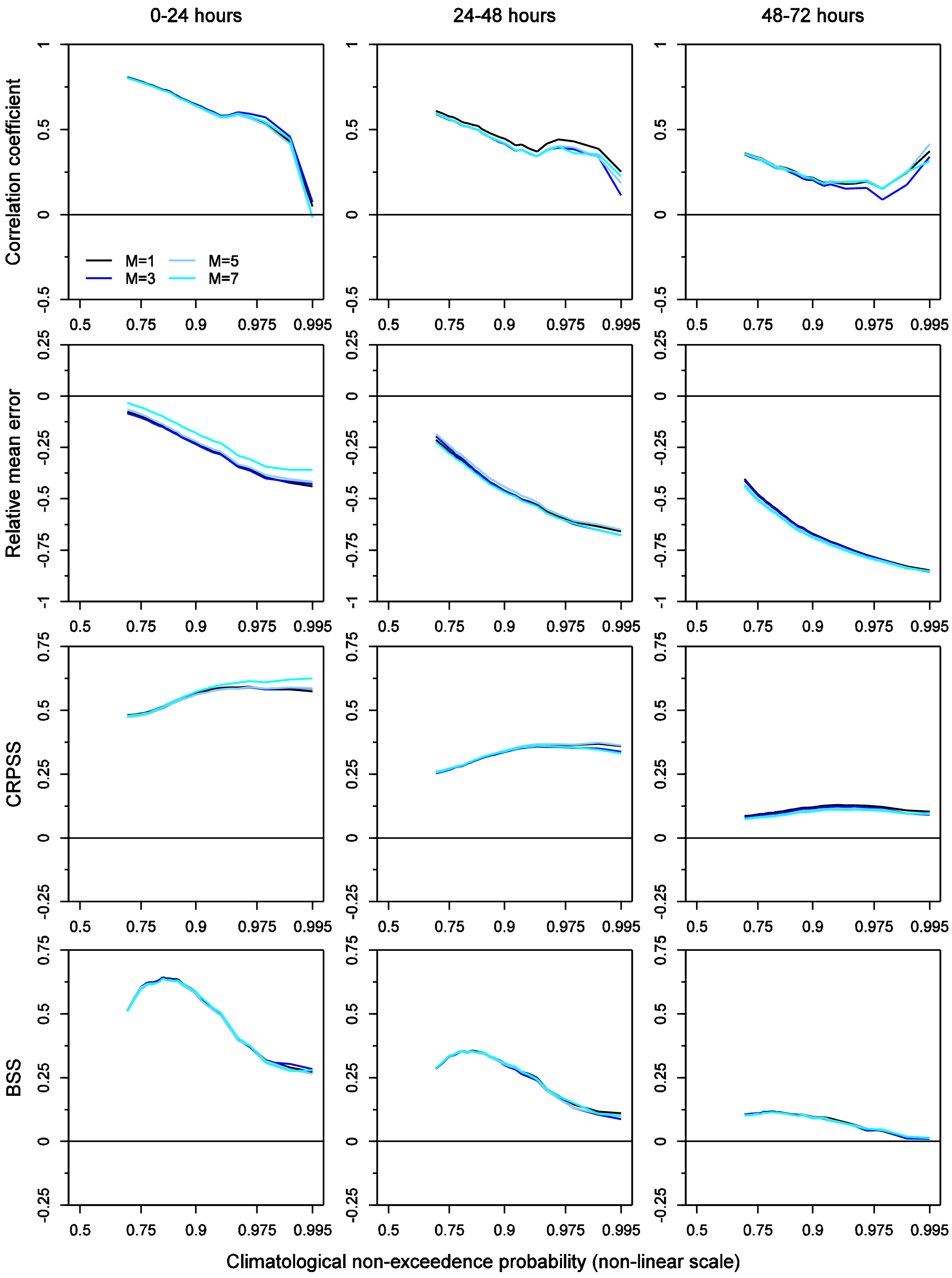

Figure 16: Selected verification metrics for the MEFP-GEFS precipitation forecasts at NE-HOPR1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of M (the interval between reforecasts in days), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

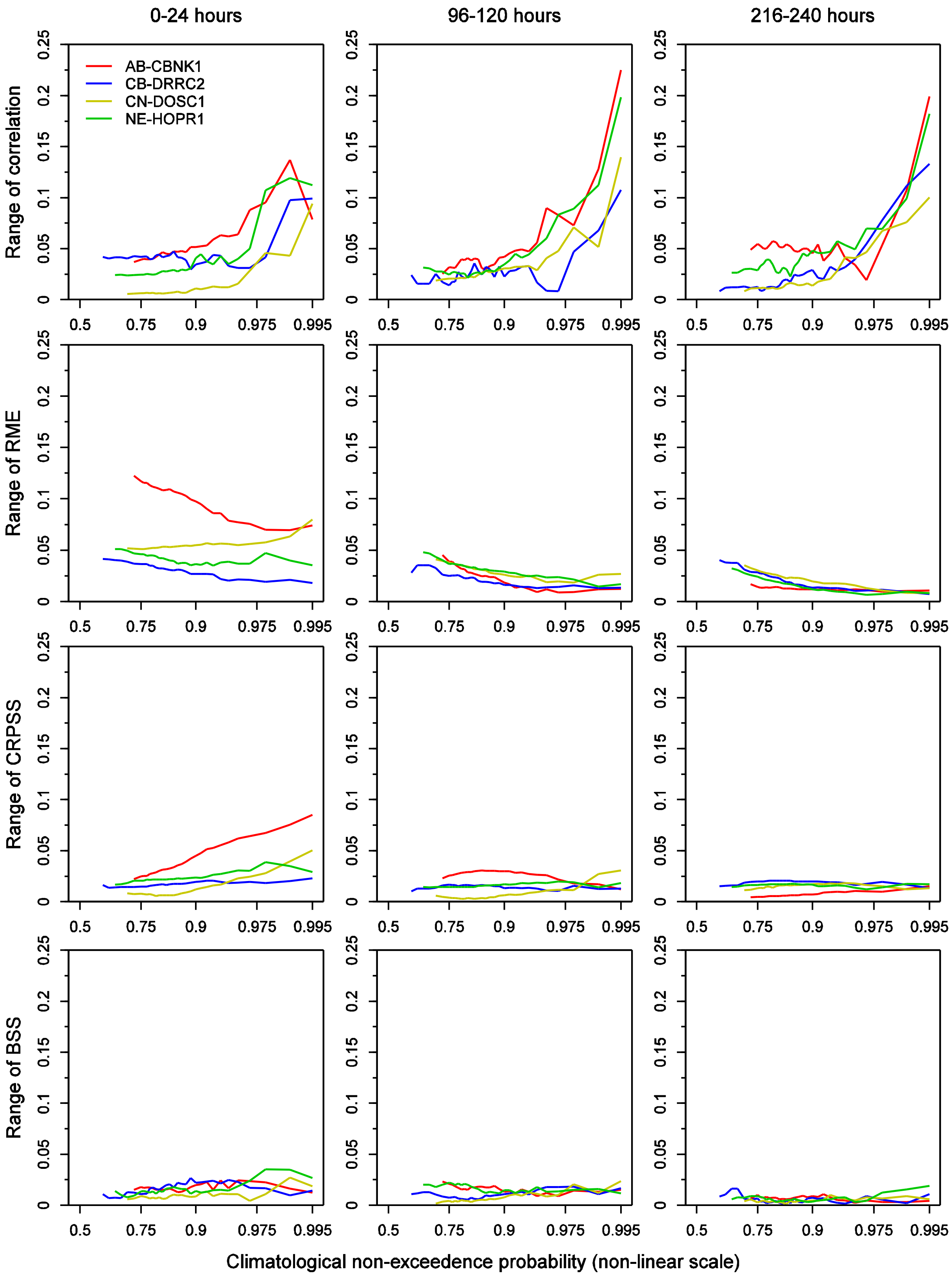

Figure 17: Range (maximum-minimum) of selected verification metrics for the MEFP-GEFS precipitation forecasts. The results are plotted against climatological non-exceedence probability (C_p) across all scenarios of M (interval between reforecasts in days), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

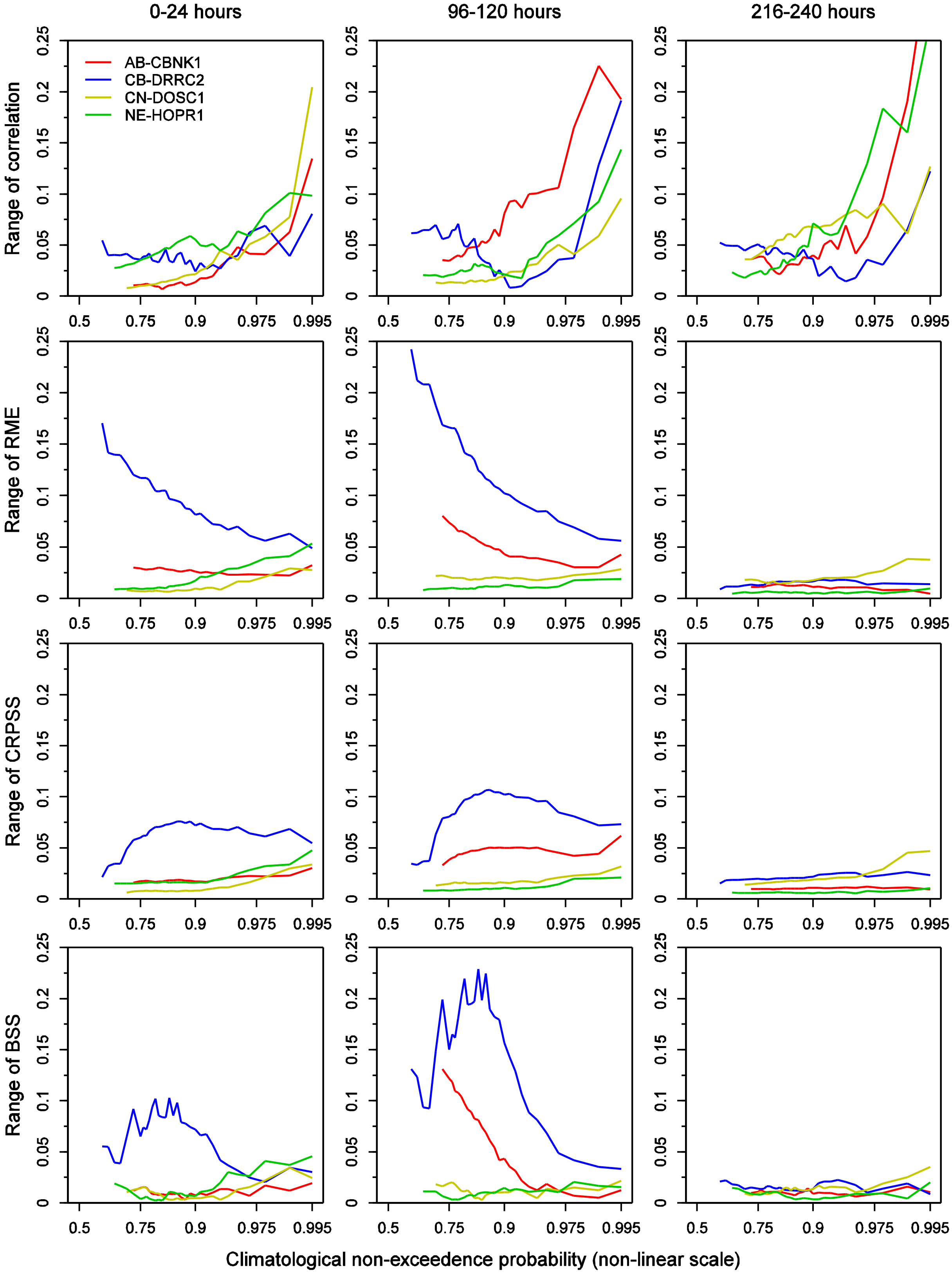

Figure 18: Range (maximum-minimum) of selected verification metrics for the MEFP-GEFS precipitation forecasts. The results are plotted against climatological non-exceedence probability (C_p) across all scenarios of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

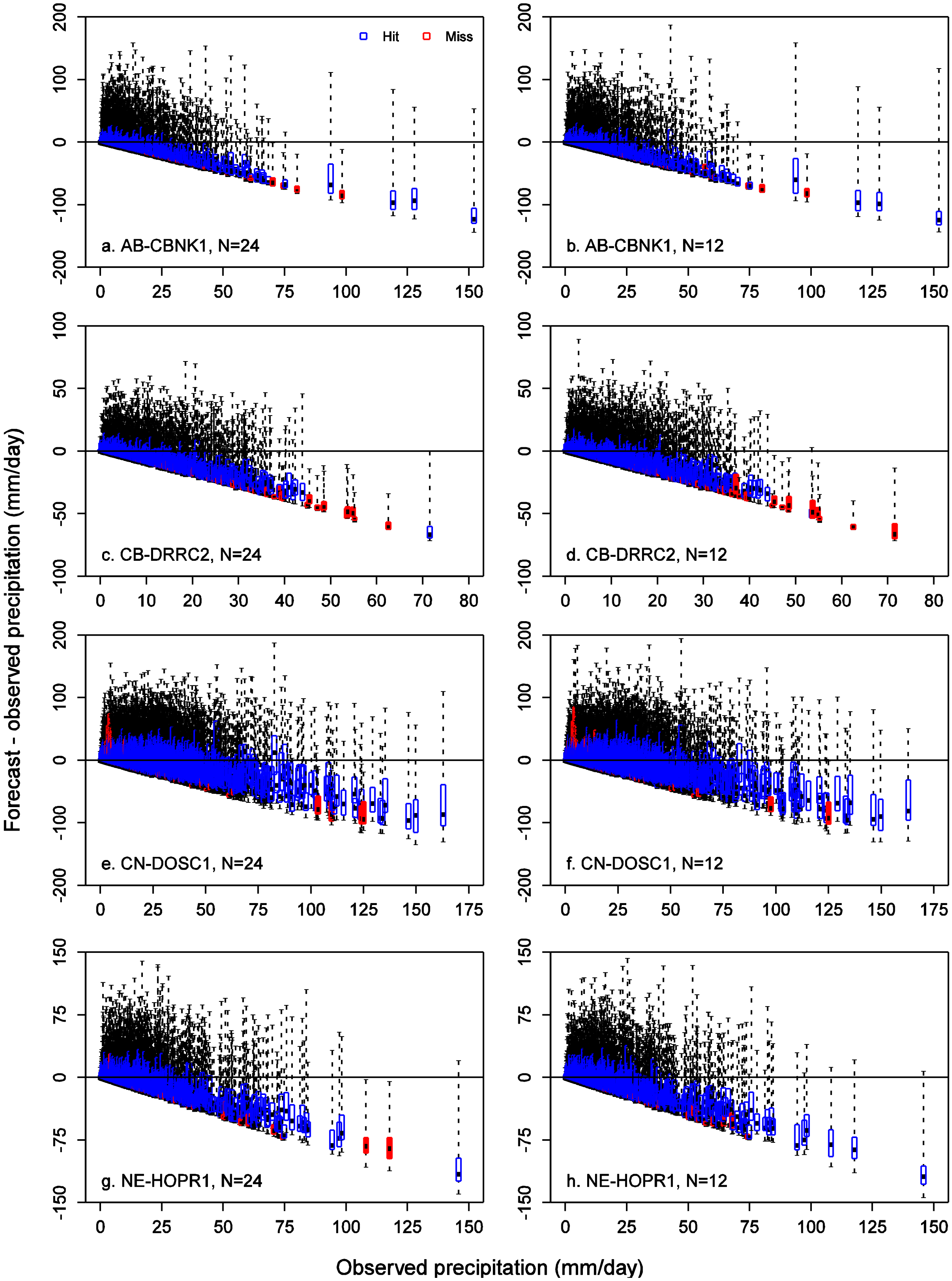

Figure 19: Box plots of forecast errors against observed precipitation amount for N={24 and 12} years of calibration data. The results are shown at a forecast lead time of 0-24 hours.

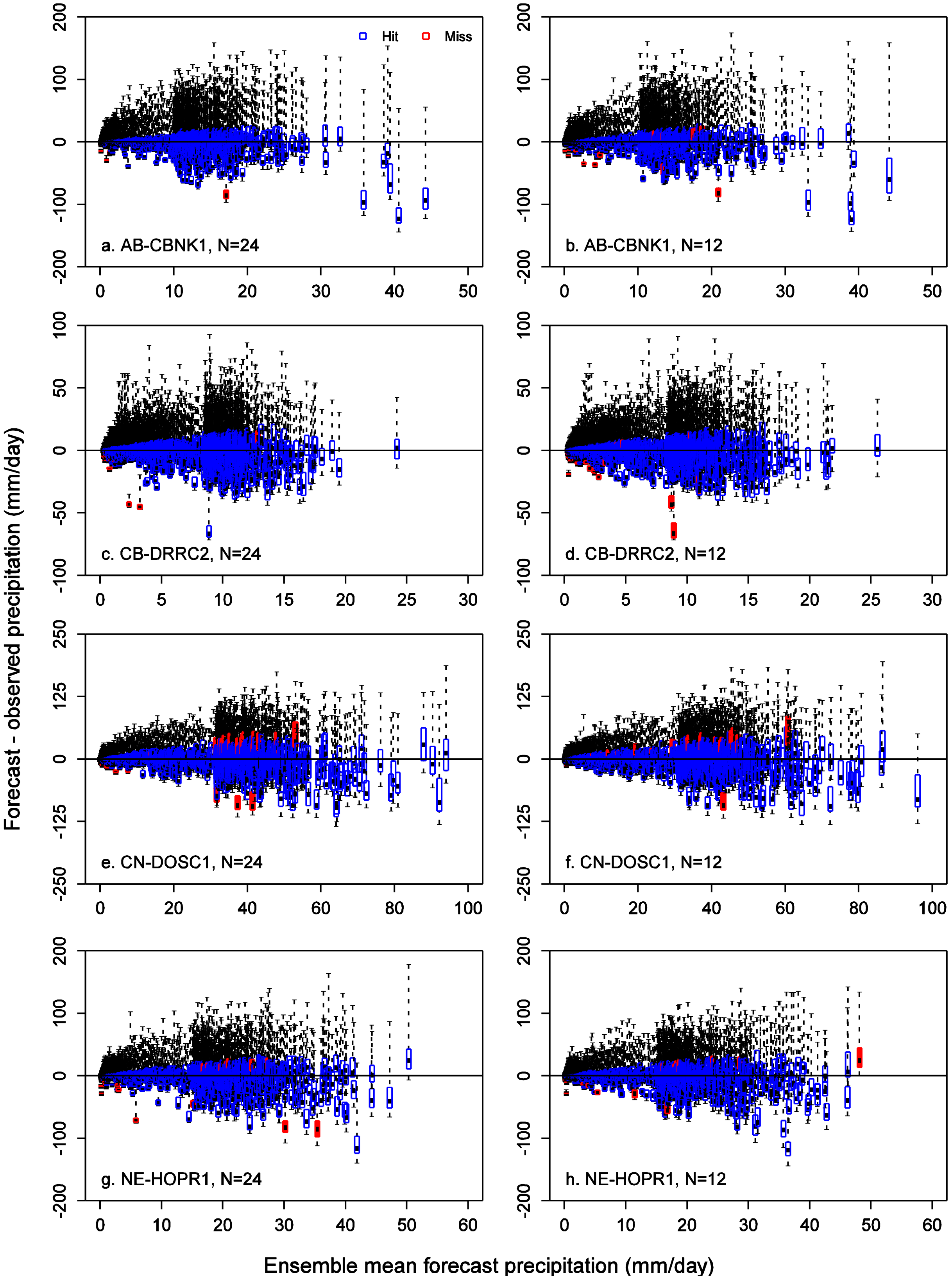

Figure 20: Box plots of forecast errors against forecast precipitation amount (ensemble mean) for N={24 and 12} years of calibration data. The results are shown at a forecast lead time of 0-24 hours.

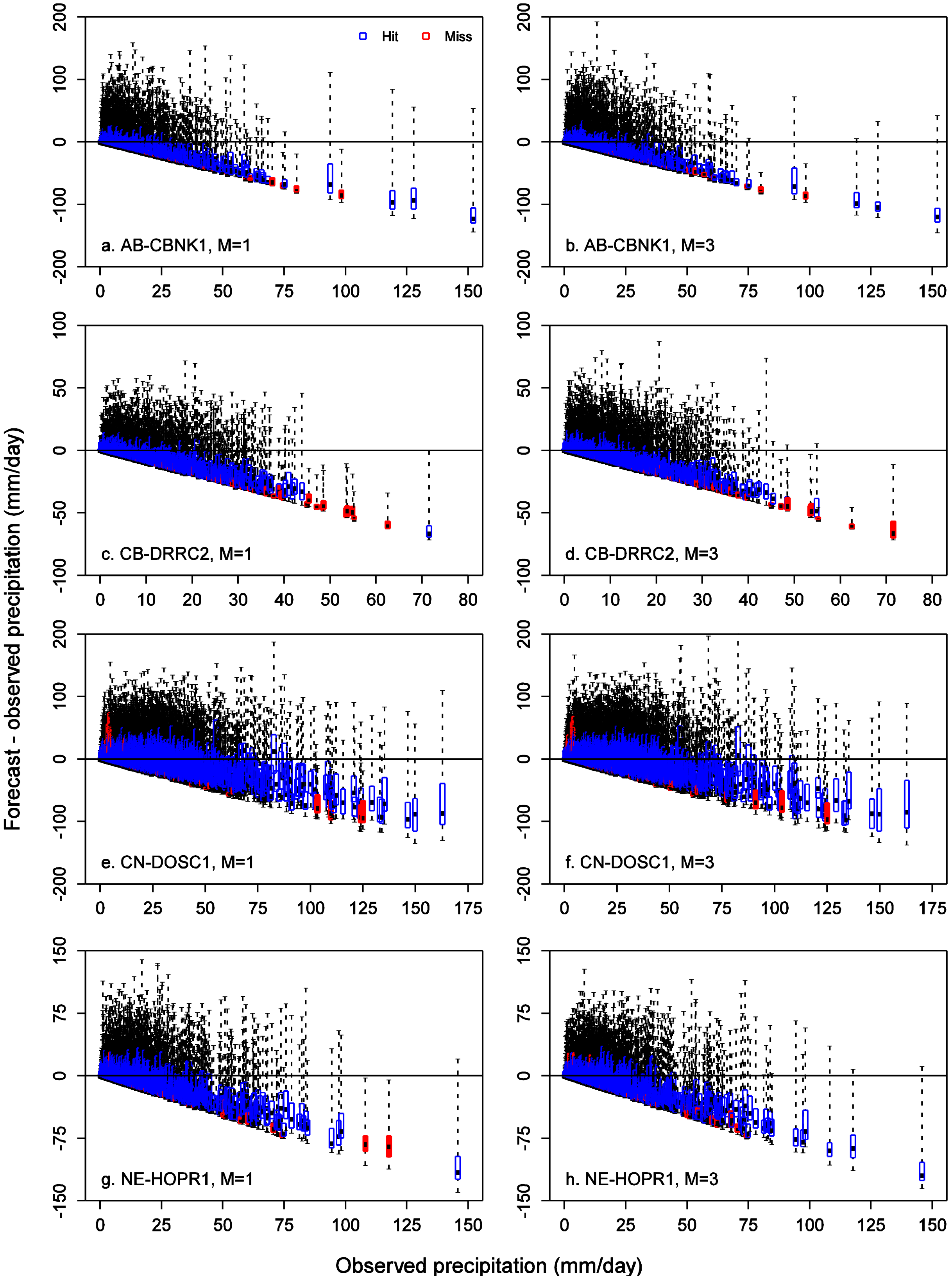

Figure 21: Box plots of forecast errors against observed precipitation amount for calibration scenarios of M={1 and 5} days between reforecasts. The results are shown at a forecast lead time of 0-24 hours.

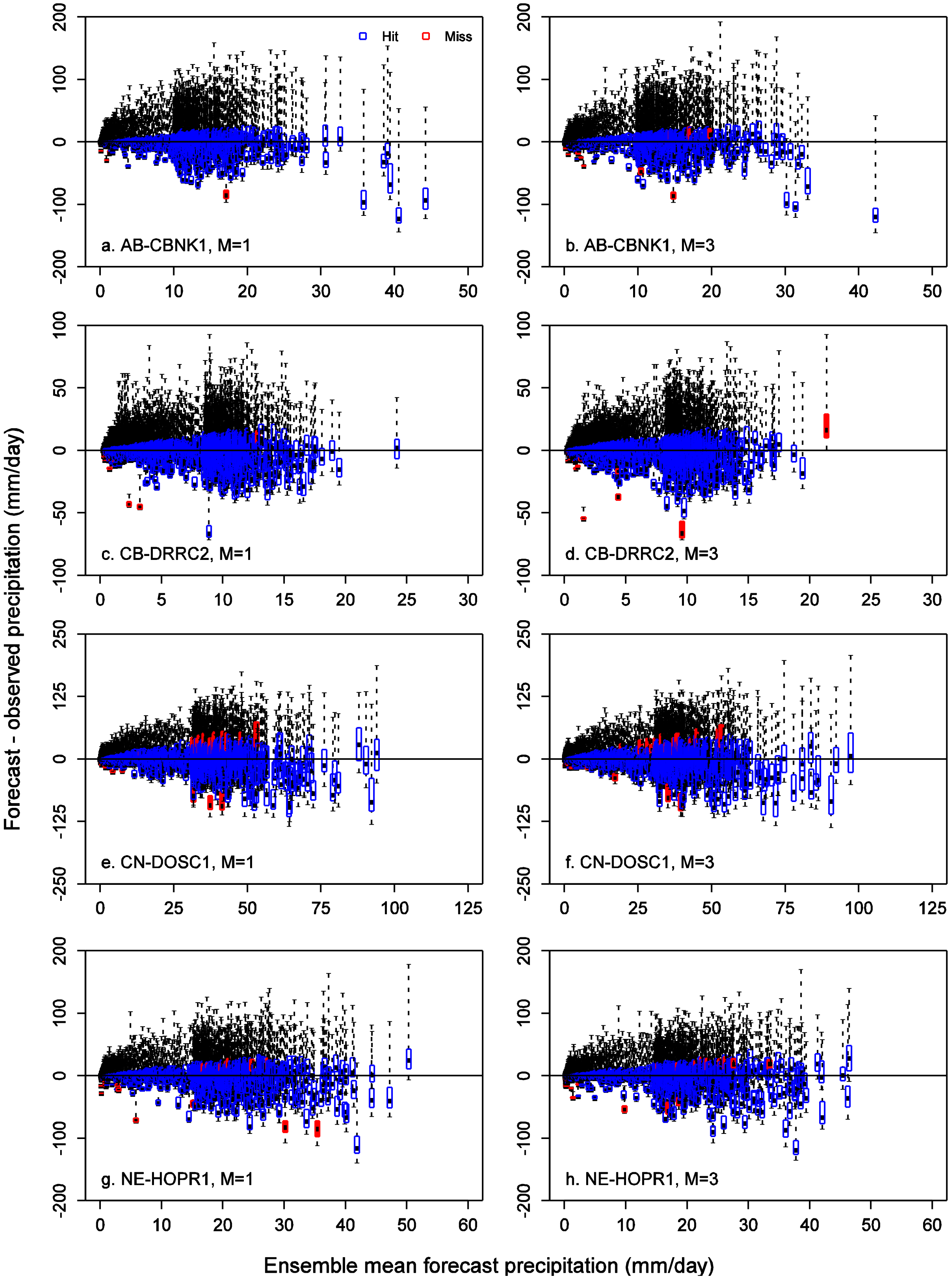

Figure 22: Box plots of forecast errors against forecast precipitation amount (ensemble mean) for calibration scenarios of M={1 and 5} days between reforecasts. The results are shown at a forecast lead time of 0-24 hours.

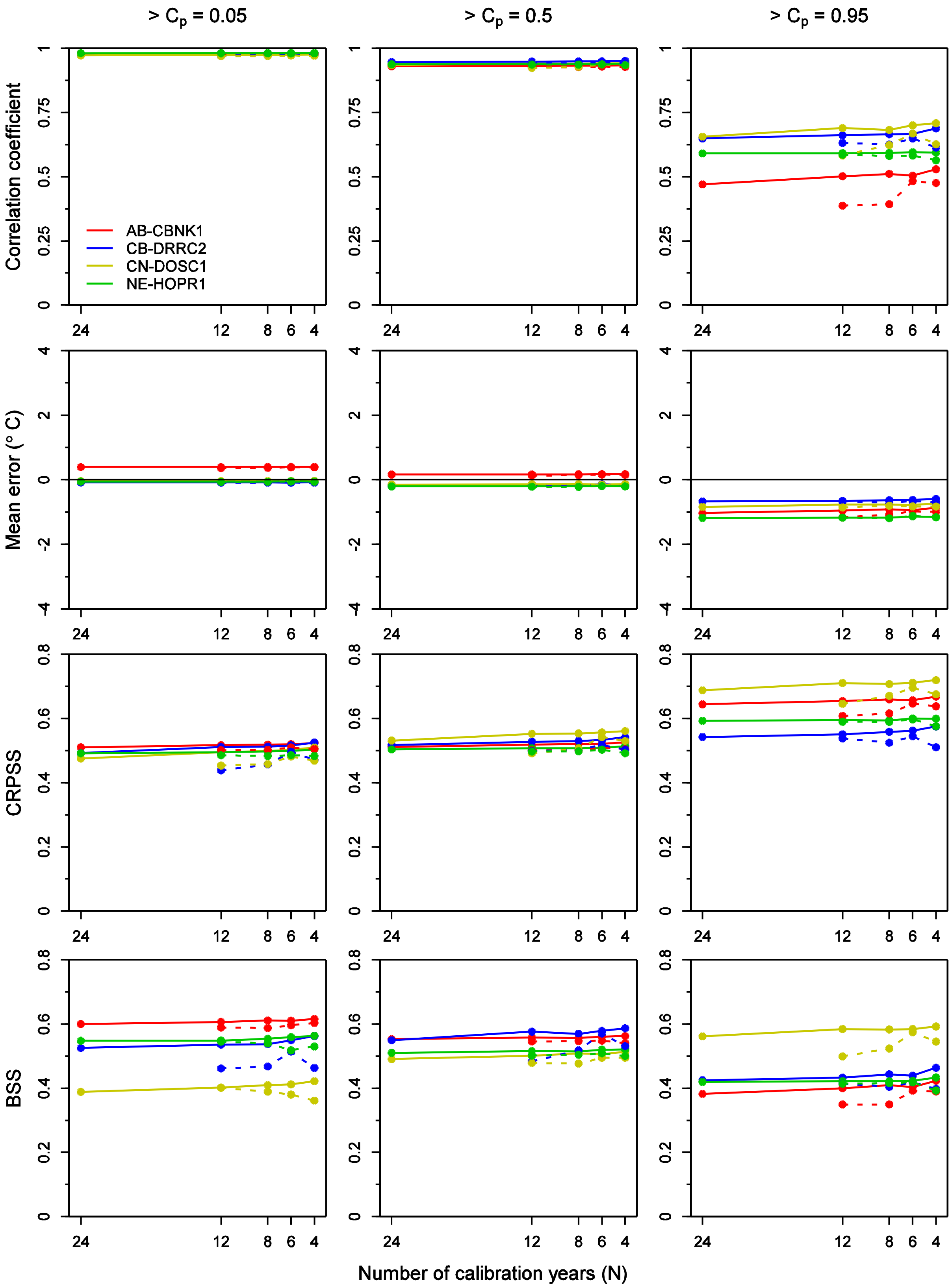

Figure 23: Selected verification metrics for the MEFP-GEFS temperature forecasts. The results are shown for the dependent (solid) and independent (dashed) validation scenarios of N (the number of years of calibration data), and include several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

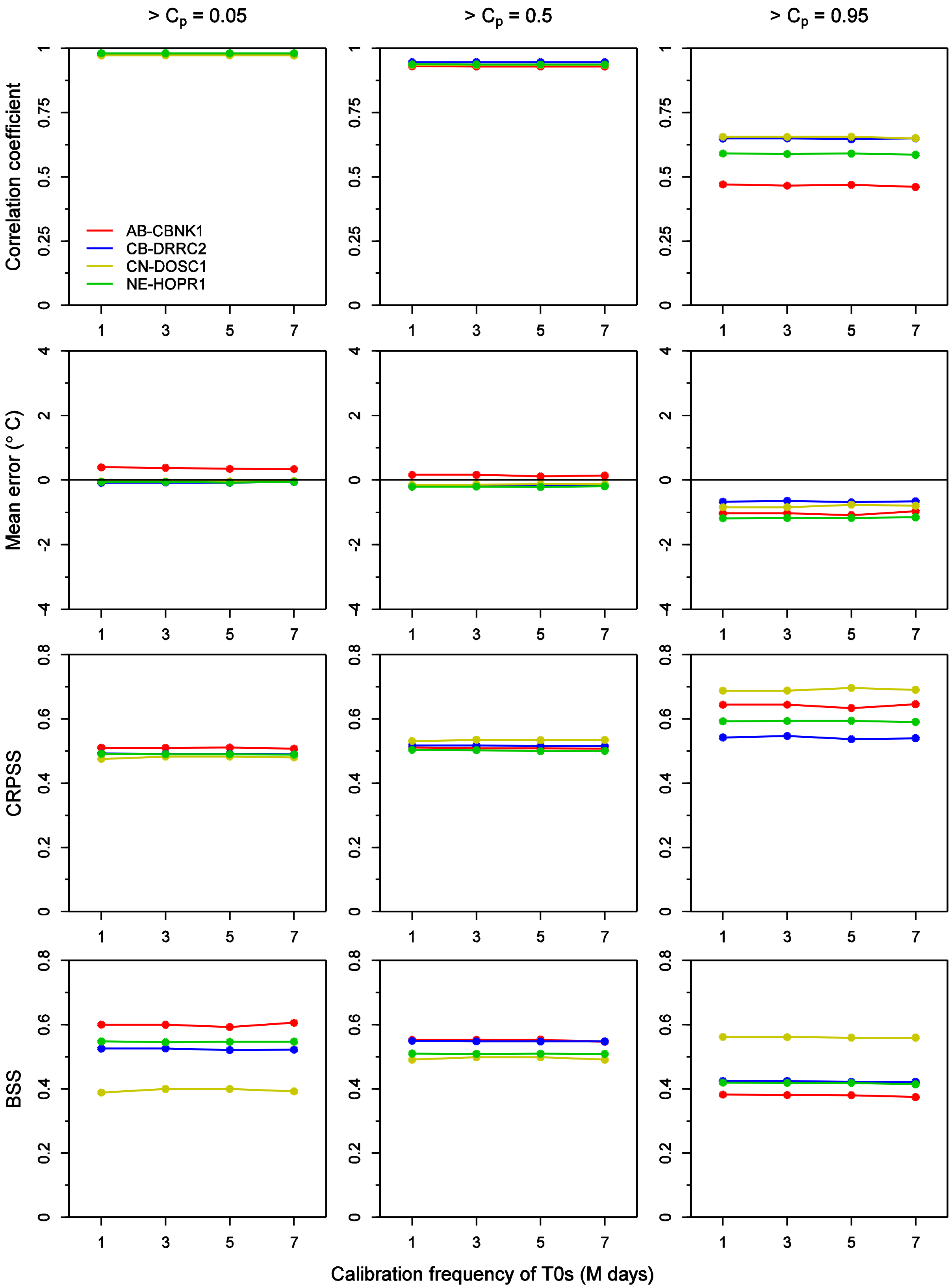

Figure 24: Selected verification metrics for the MEFP-GEFS temperature forecasts. The results are plotted against the interval between reforecasts (M days) used to calibrate the MEFP, and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

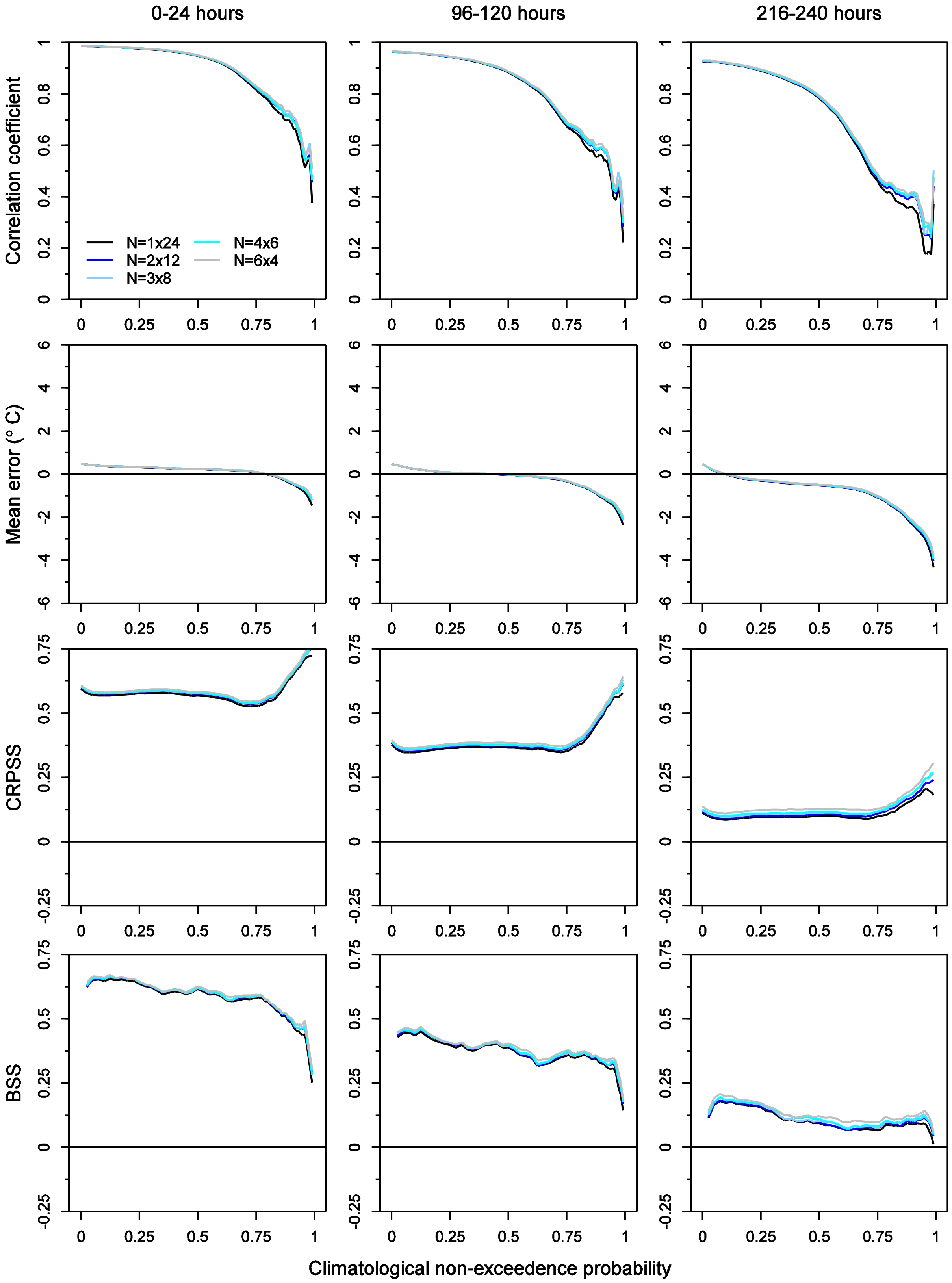

Figure 25: Selected verification metrics for the MEFP-GEFS temperature forecasts at AB-CBNK1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

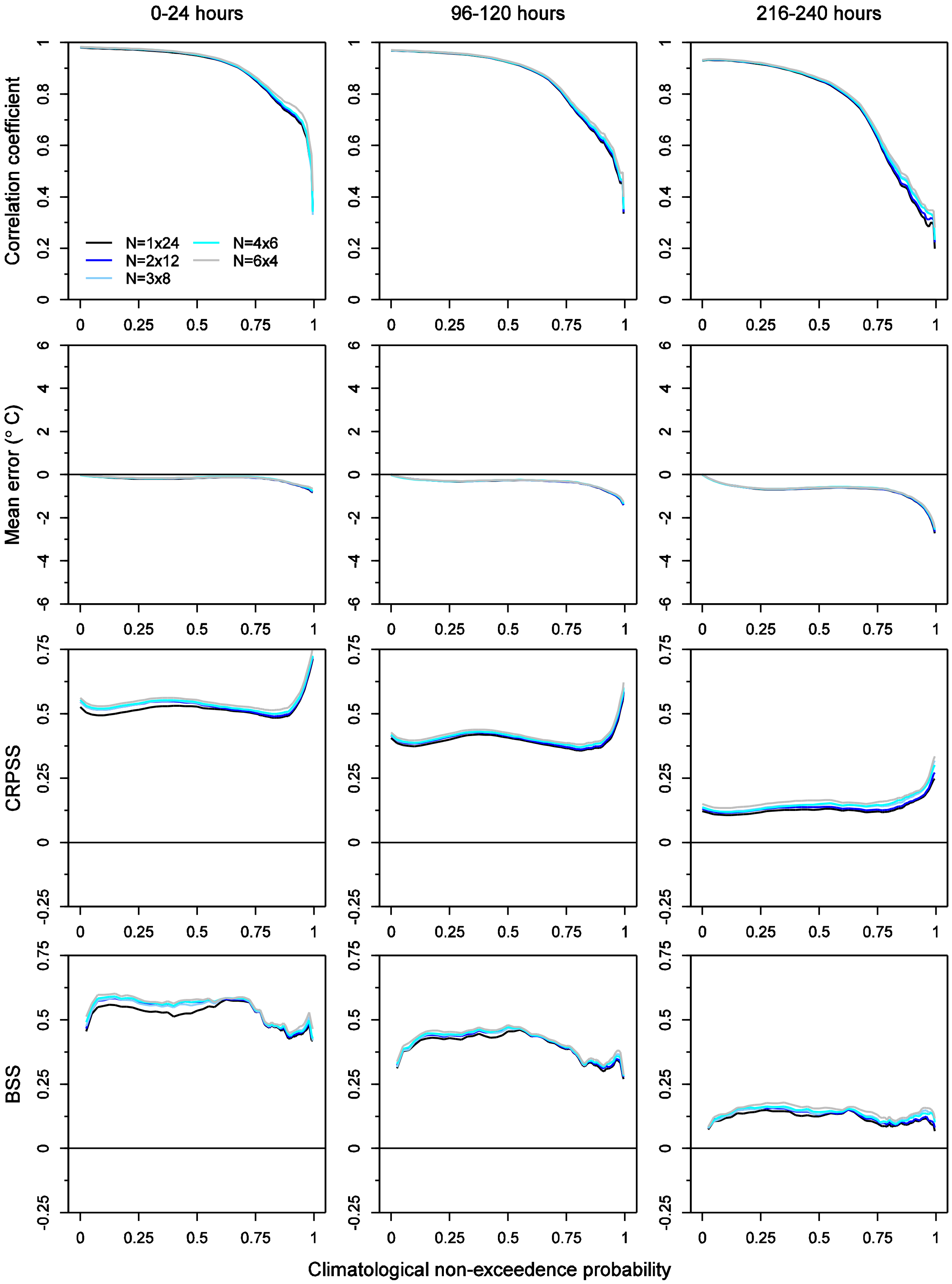

Figure 26: Selected verification metrics for the MEFP-GEFS temperature forecasts at CB-DRRC2. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

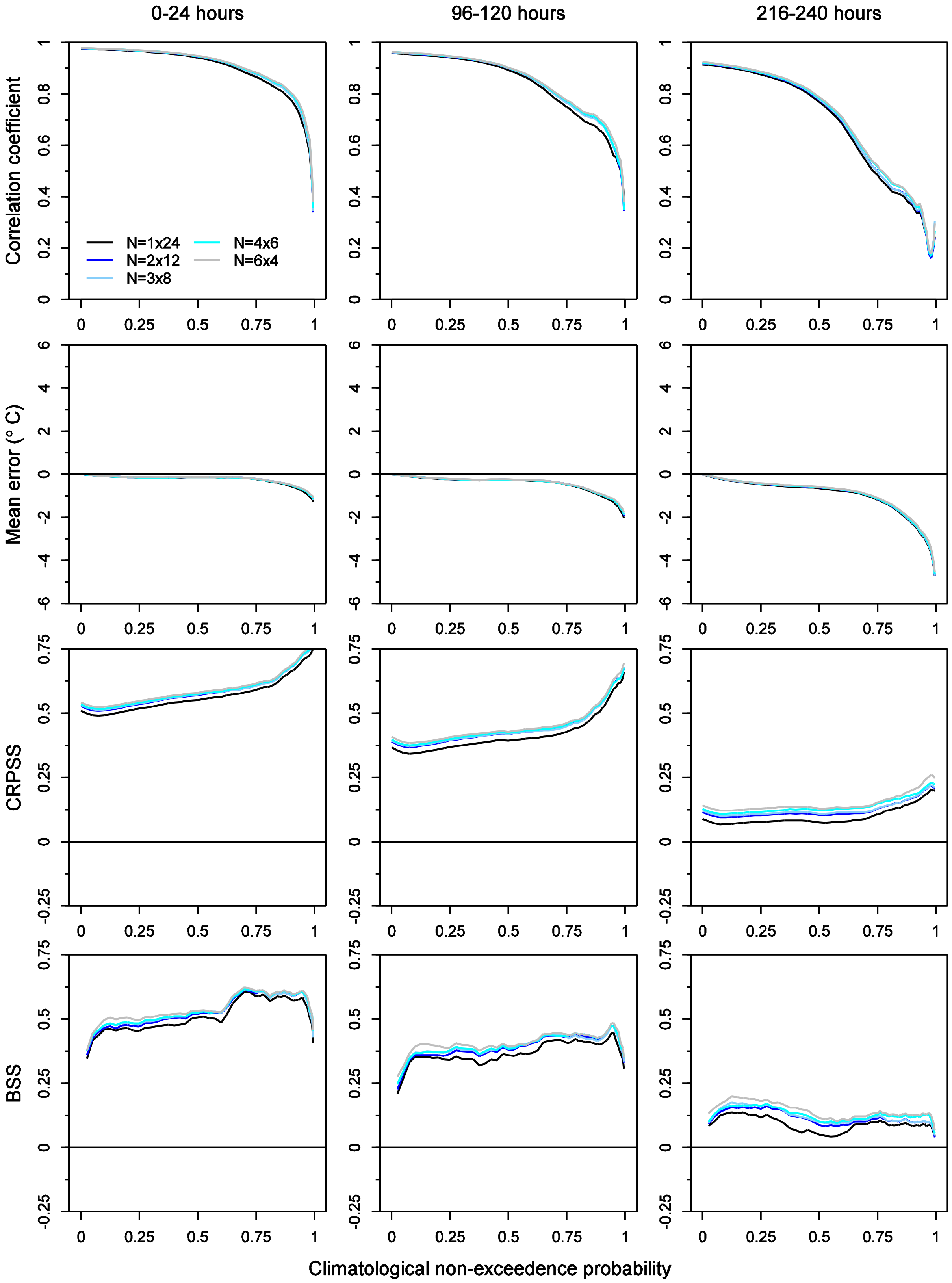

Figure 27: Selected verification metrics for the MEFP-GEFS temperature forecasts at CN-DOSC1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

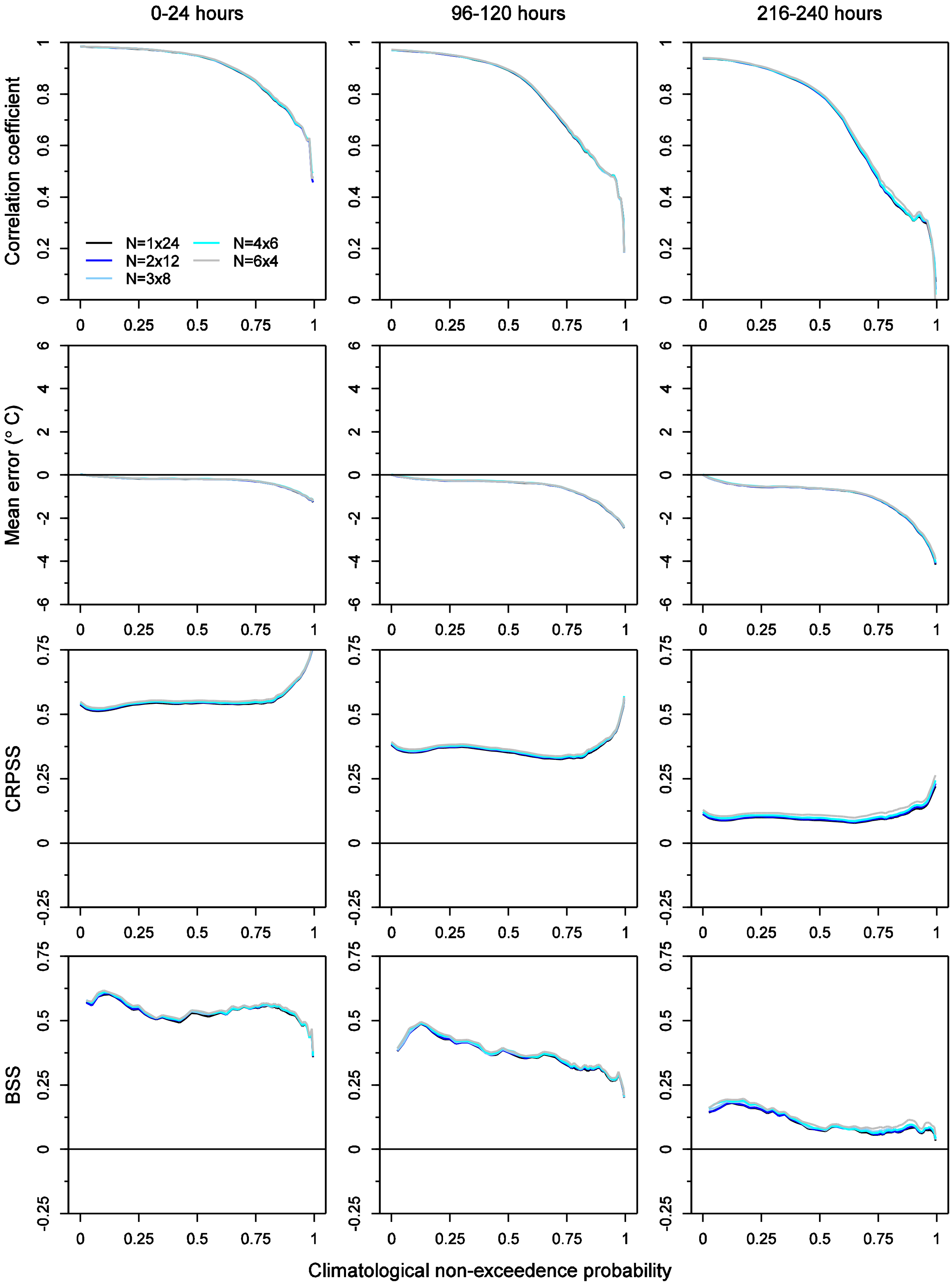

Figure 28: Selected verification metrics for the MEFP-GEFS temperature forecasts at NE-HOPR1. The results are plotted against climatological non-exceedence probability (C_p) for each scenario of N (the number of years of calibration data), and are shown for several forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

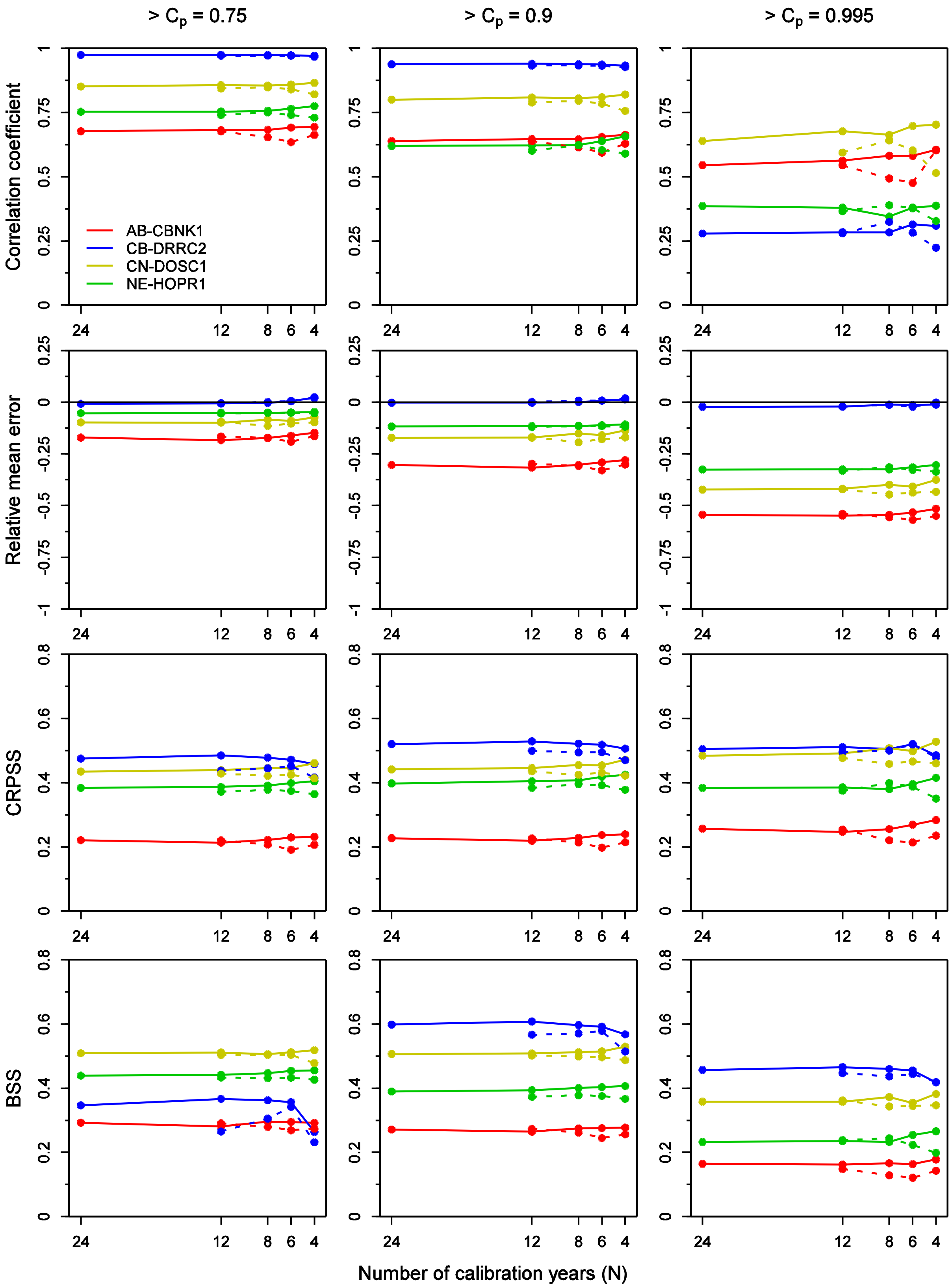

Figure 29: Selected verification metrics for the MEFP-GEFS streamflow forecasts. The results are shown for the dependent (solid) and independent (dashed) validation scenarios of N (the number of years of calibration data), and include several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

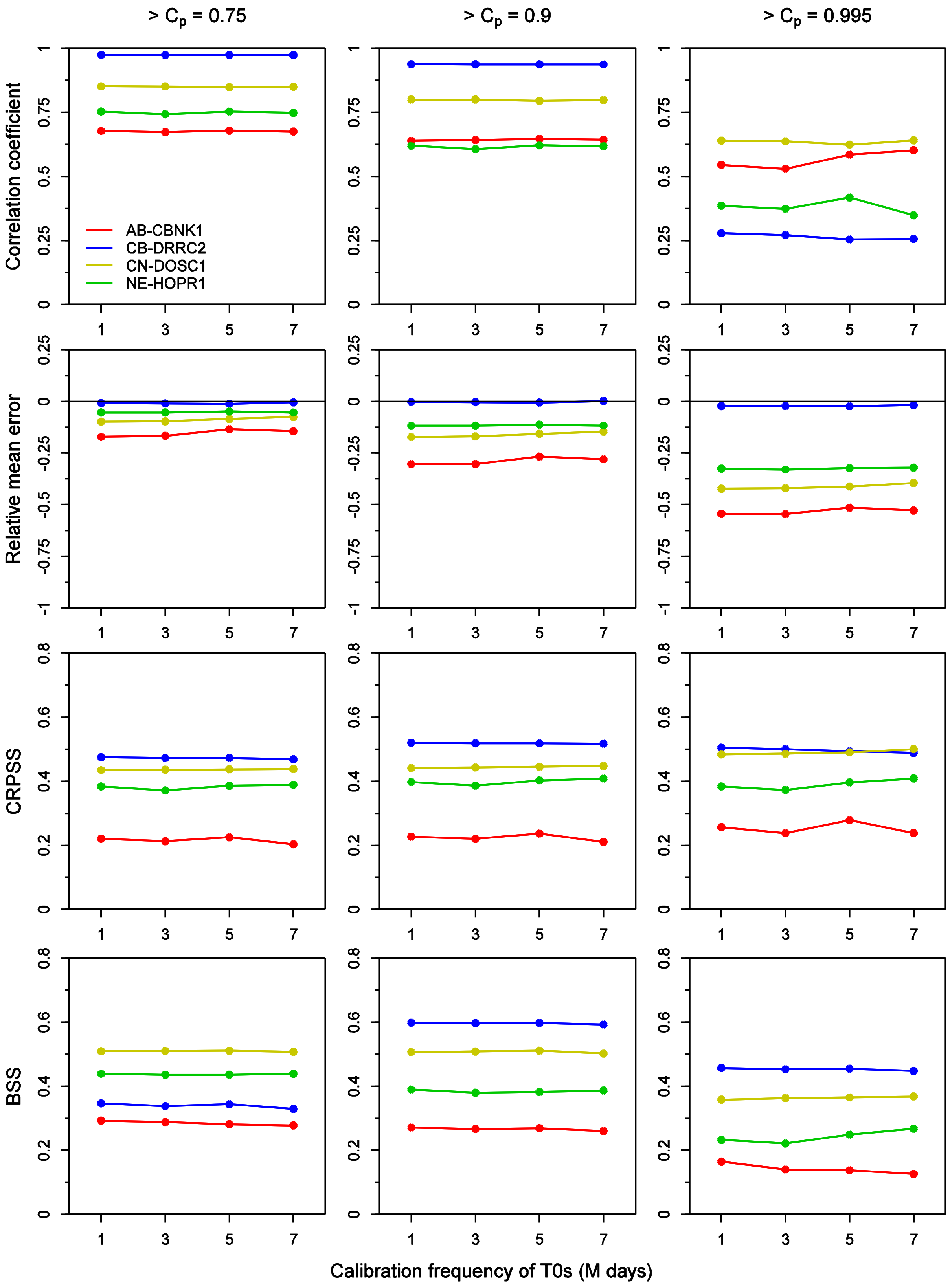

Figure 30: Selected verification metrics for the MEFP-GEFS streamflow forecasts. The results are plotted against the interval between reforecasts (M days) used to calibrate the MEFP, and are shown for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

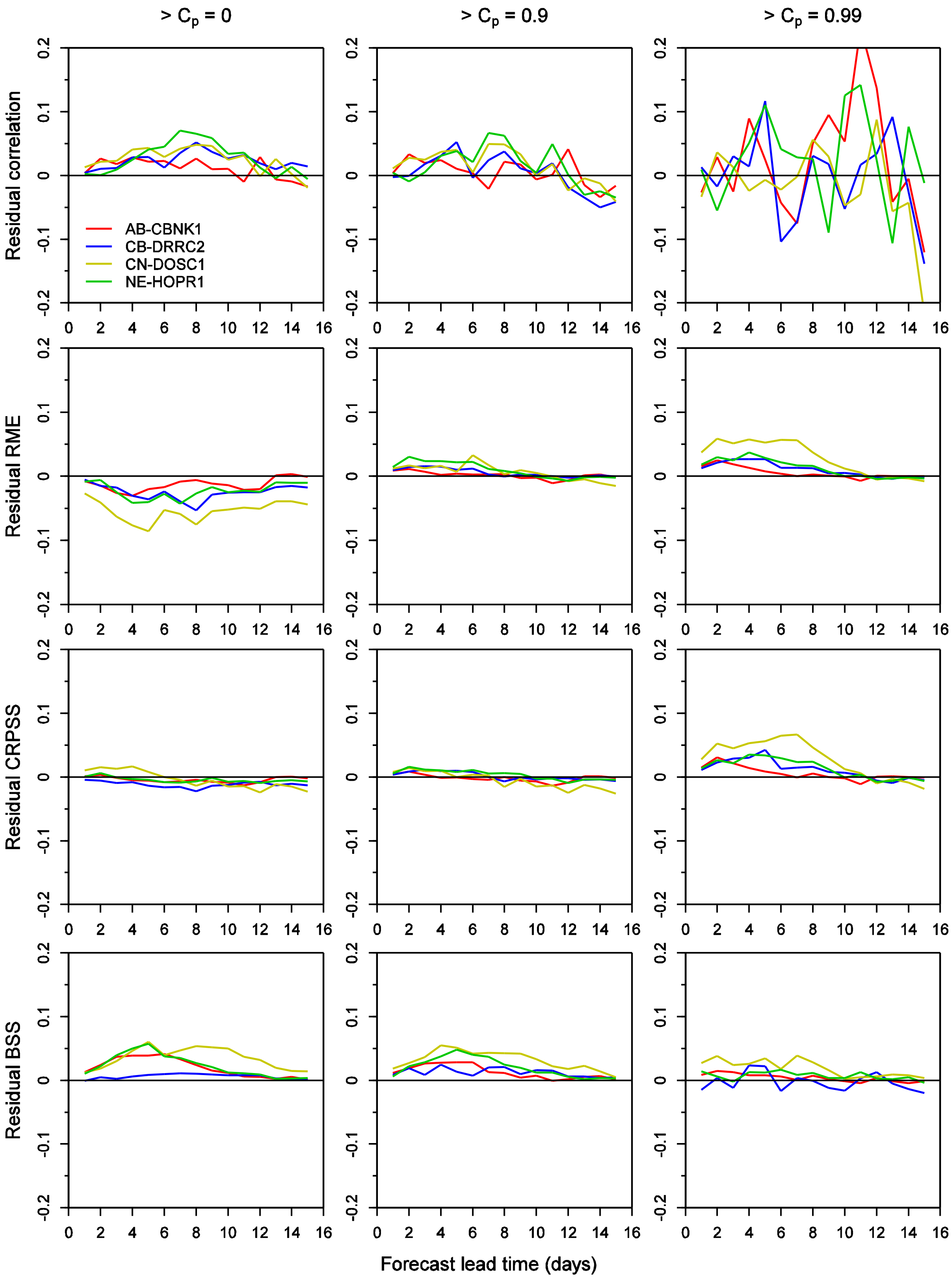

Figure 31: Residuals of selected verification metrics for the MEFP-GEFS precipitation forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=1 (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

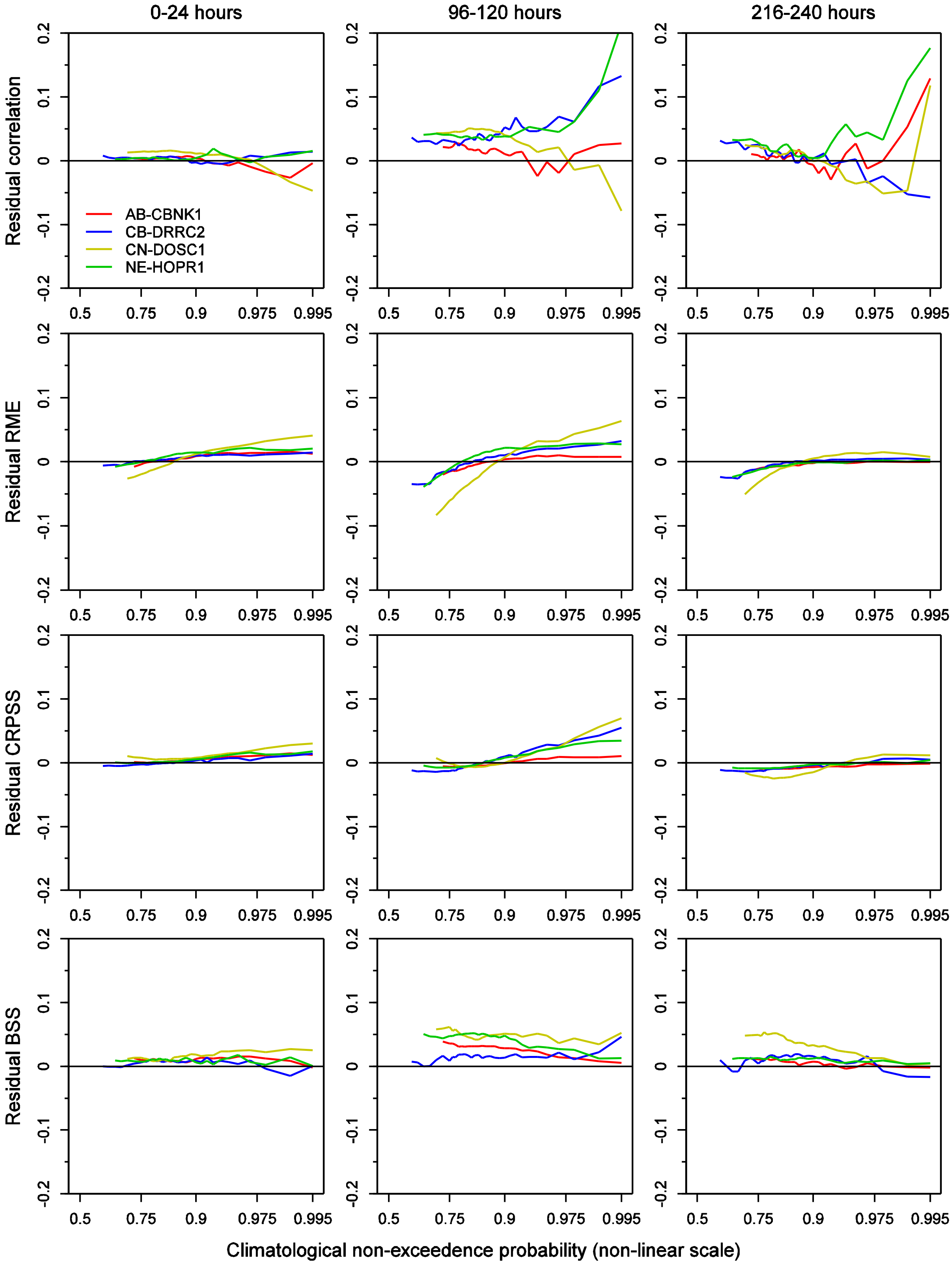

Figure 32: Residuals of selected verification metrics for the MEFP-GEFS precipitation forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=1 (F=11). The results are shown by climatological non-exceedence probability at selected forecast lead times. The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

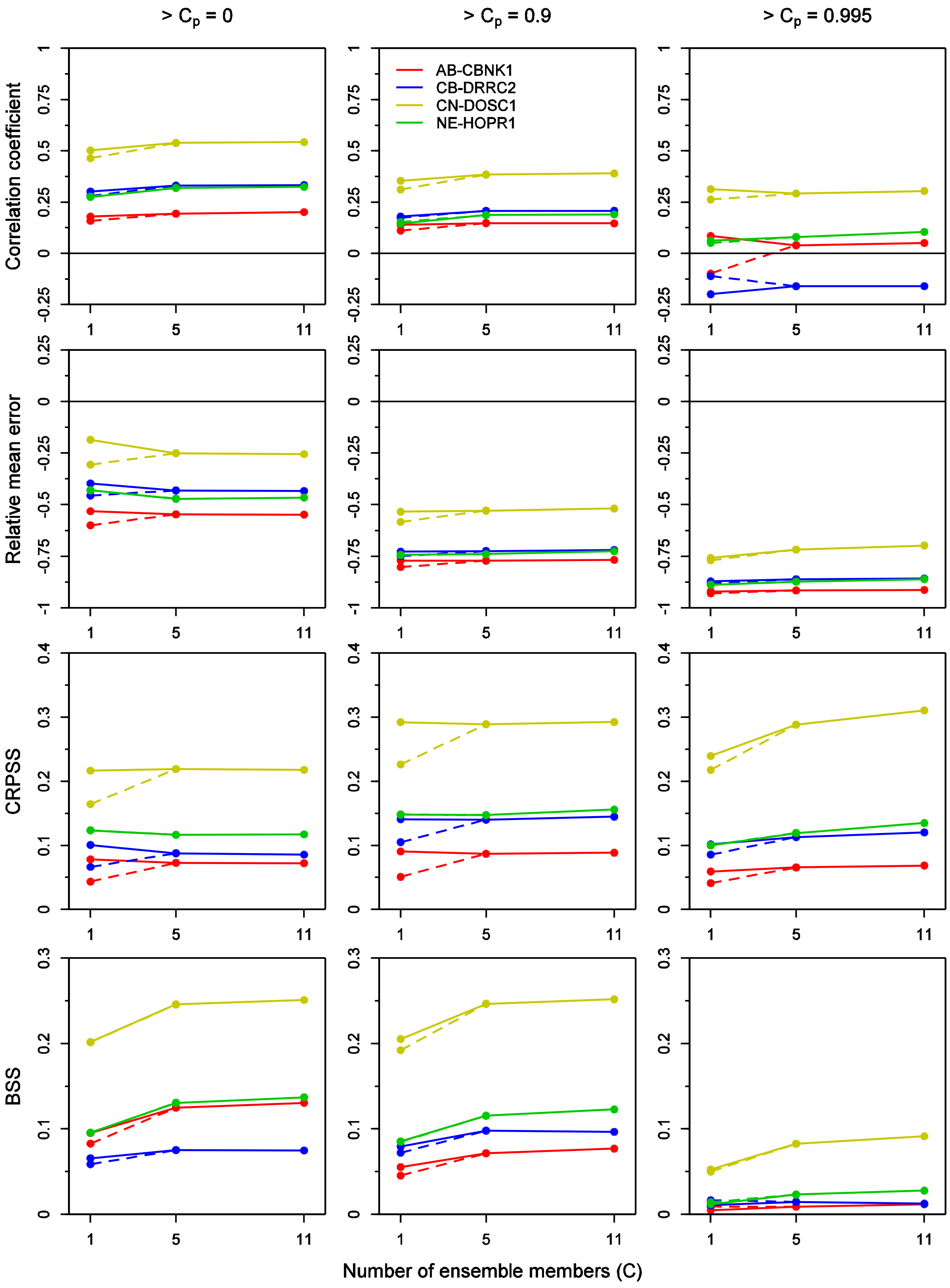

Figure 33: Sensitivity of the MEFP-GEFS precipitation forecasts to the number of members (C) used to calibrate the MEFP. The results comprise an average over the middle portion of the forecast horizon (4-8 days) for selected climatological probabilities (C_p). The bold lines show the calibration scenarios with F=11 forecast members. The dashed line shows the (C=1, F=1) scenario.

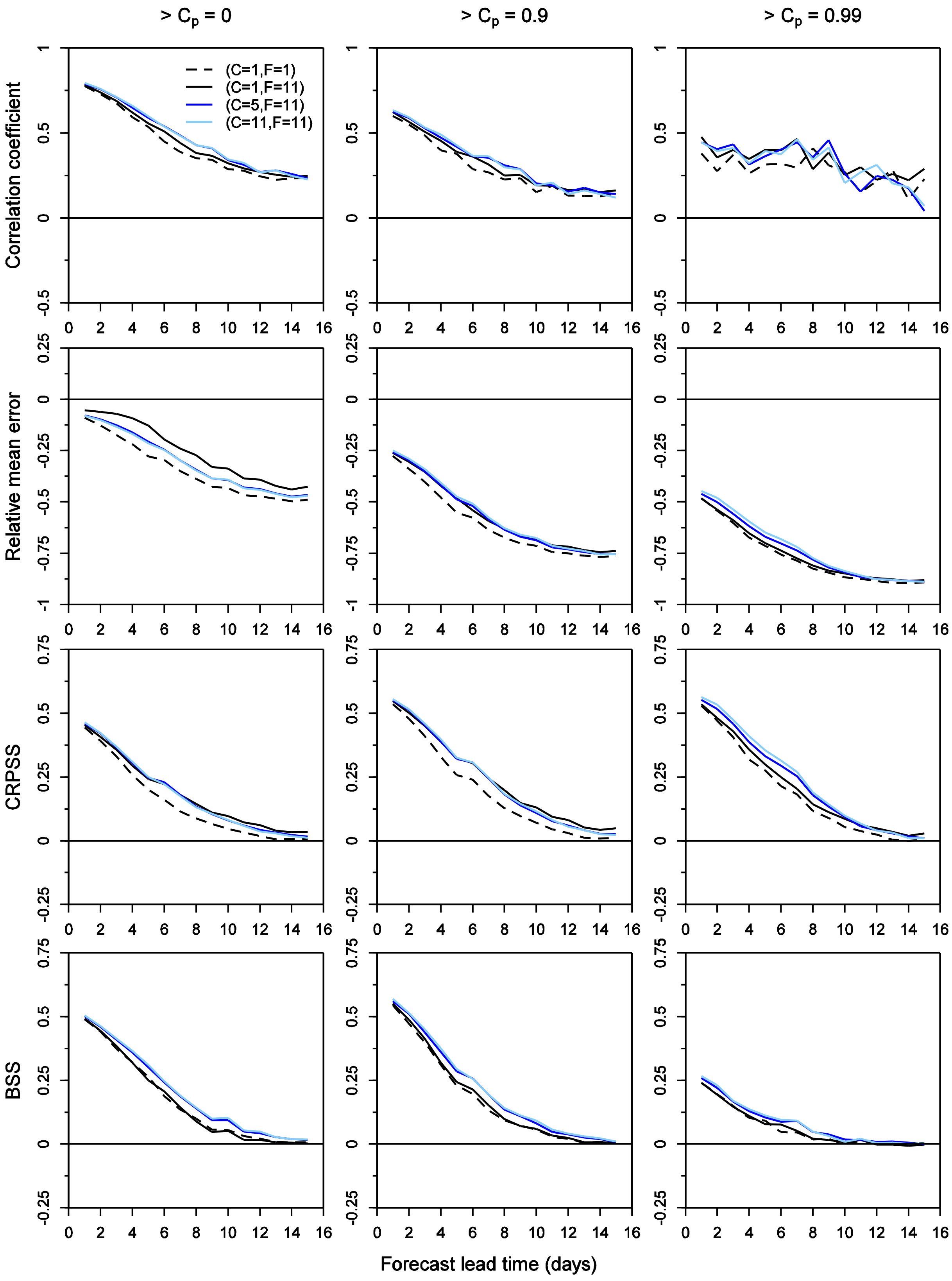

Figure 34: Selected verification metrics for the MEFP-GEFS precipitation forecasts at CN-DOSC1. The results are shown by forecast lead time for multiple calibration (C) and forecasting (F) scenarios and for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

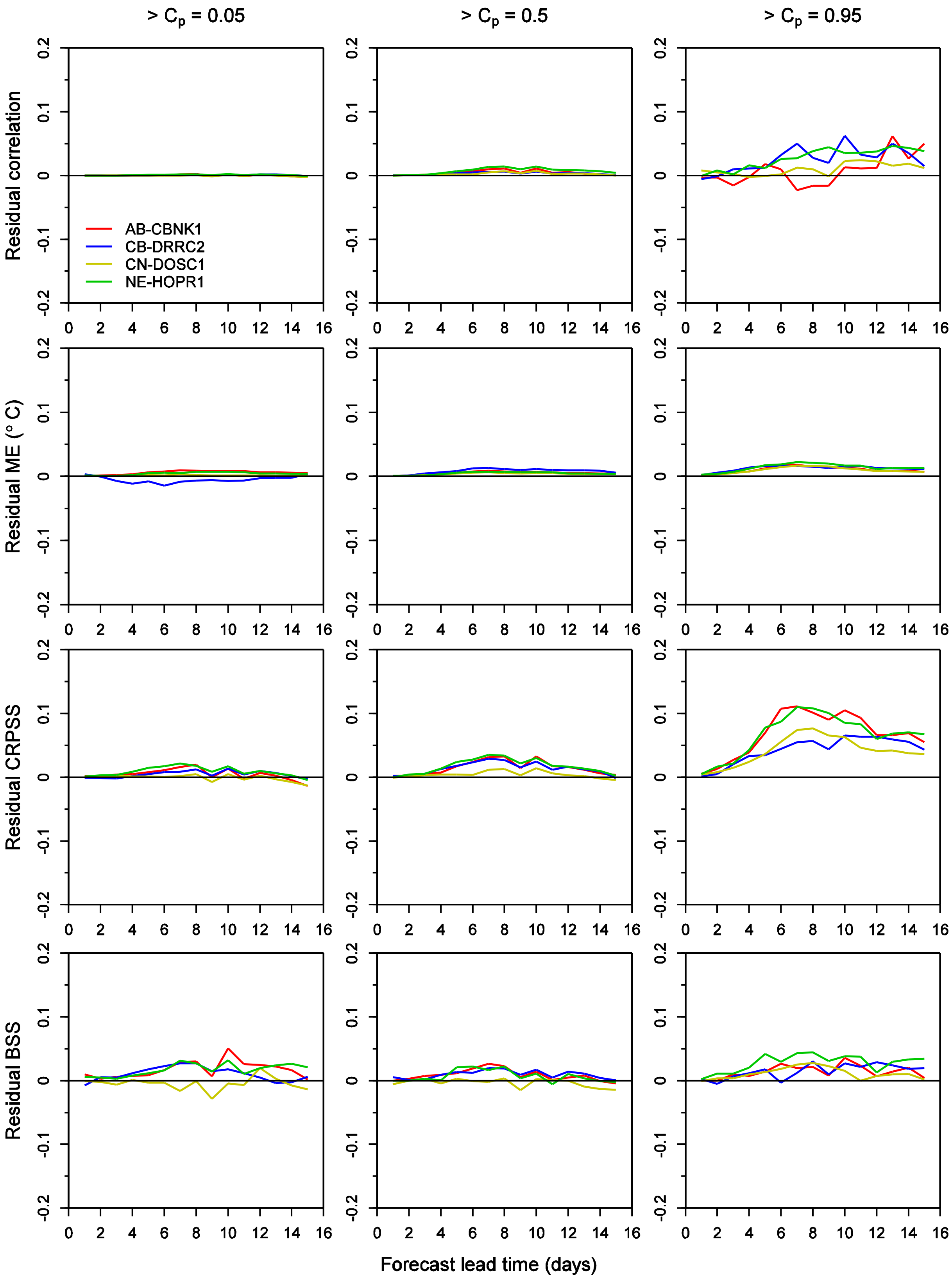

Figure 35: Residuals of selected verification metrics for the MEFP-GEFS temperature forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=1 (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

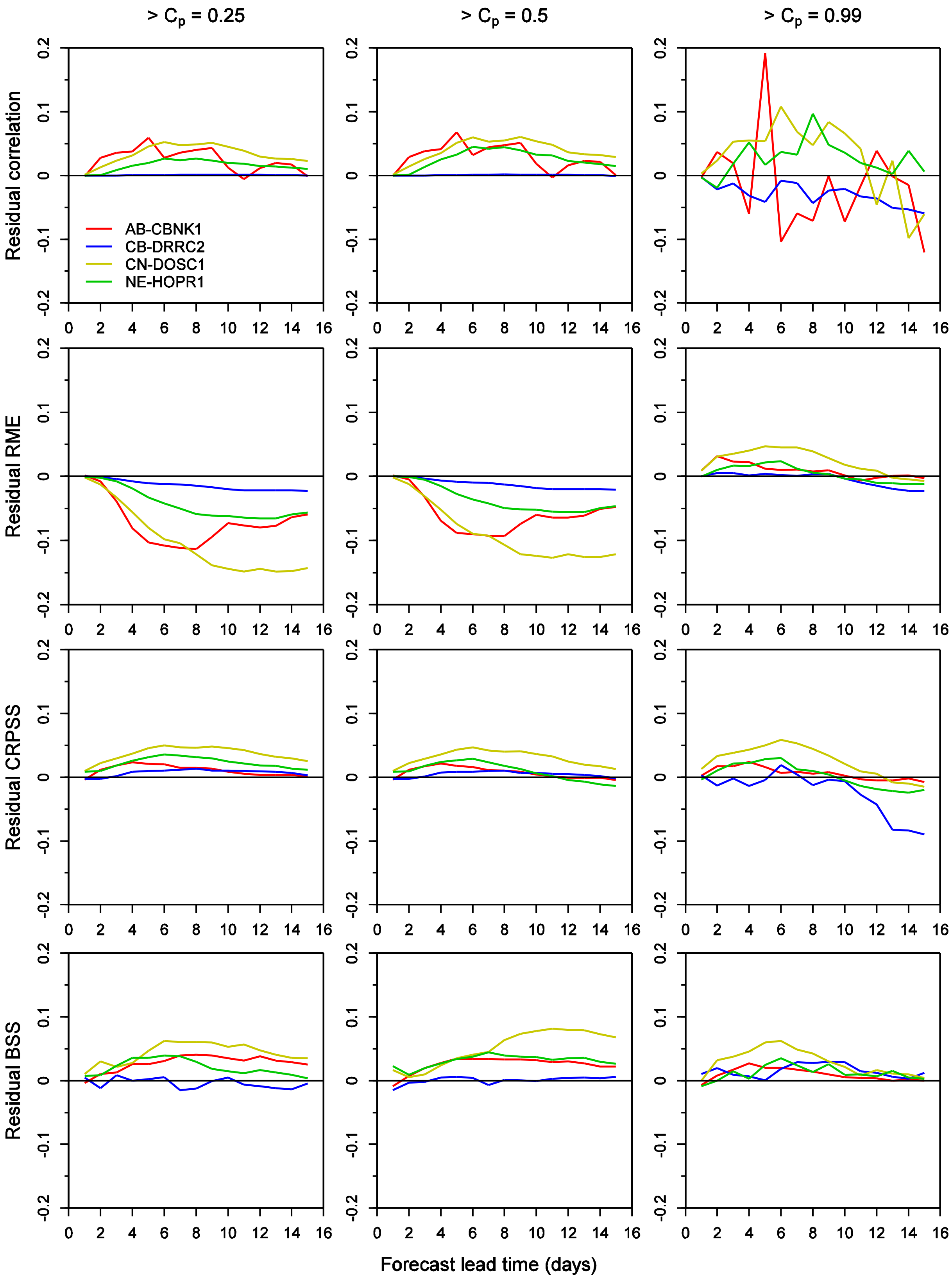

Figure 36: Residuals of selected verification metrics for the HEFS streamflow forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=1 member (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

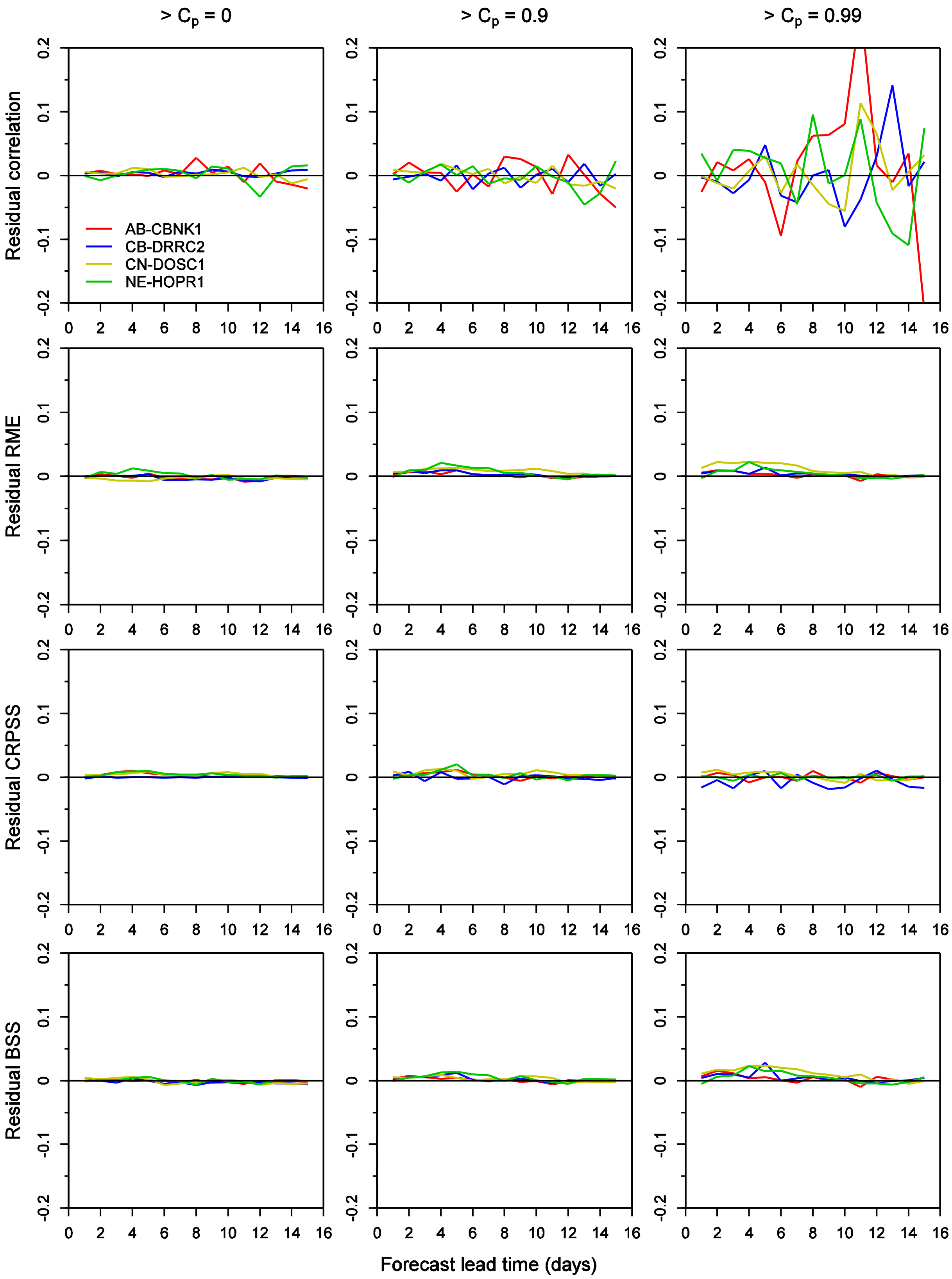

Figure 37: Residuals of selected verification metrics for the MEFP-GEFS precipitation forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=5 (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

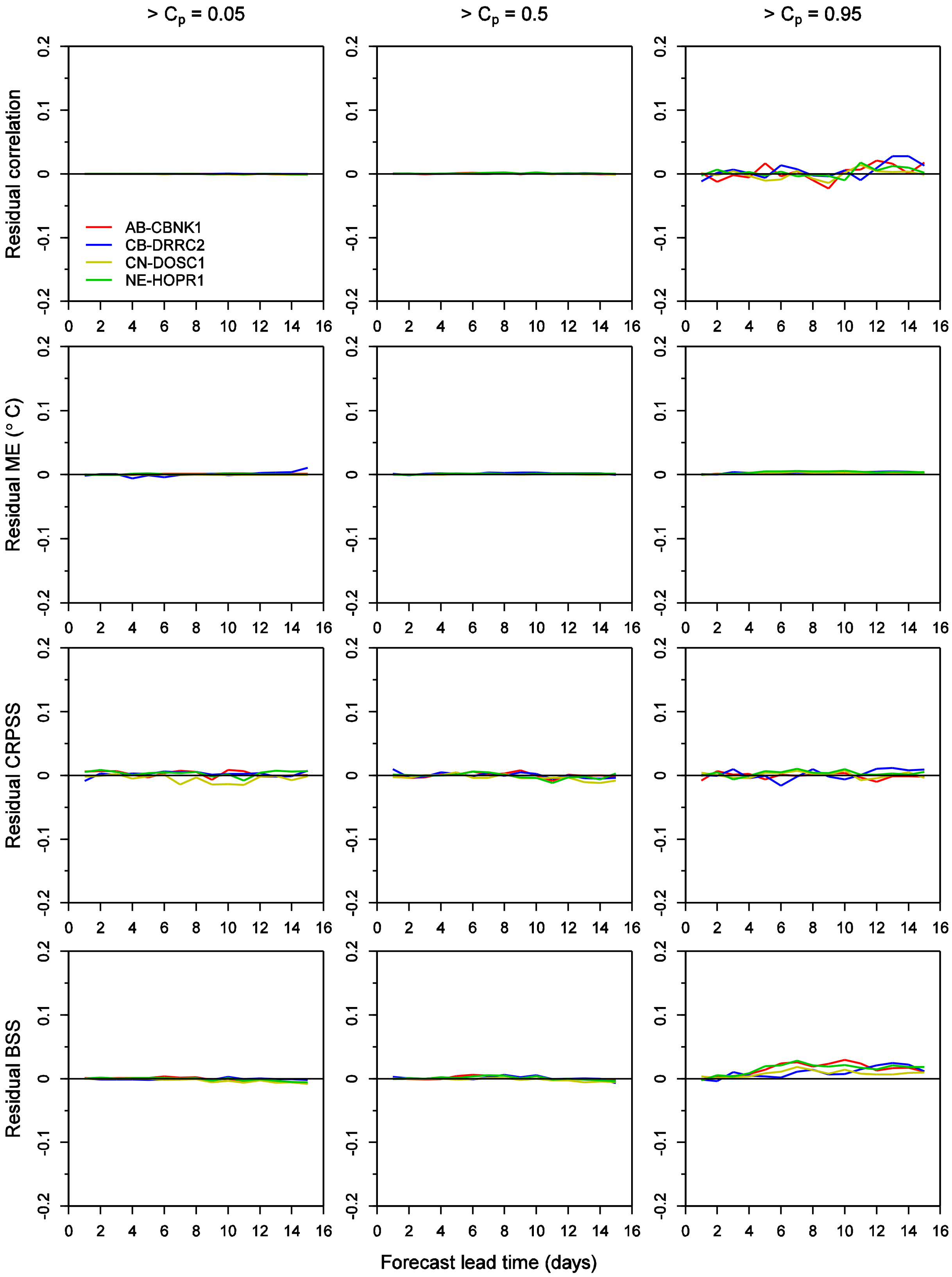

Figure 38: Residuals of selected verification metrics for the MEFP-GEFS temperature forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=5 (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

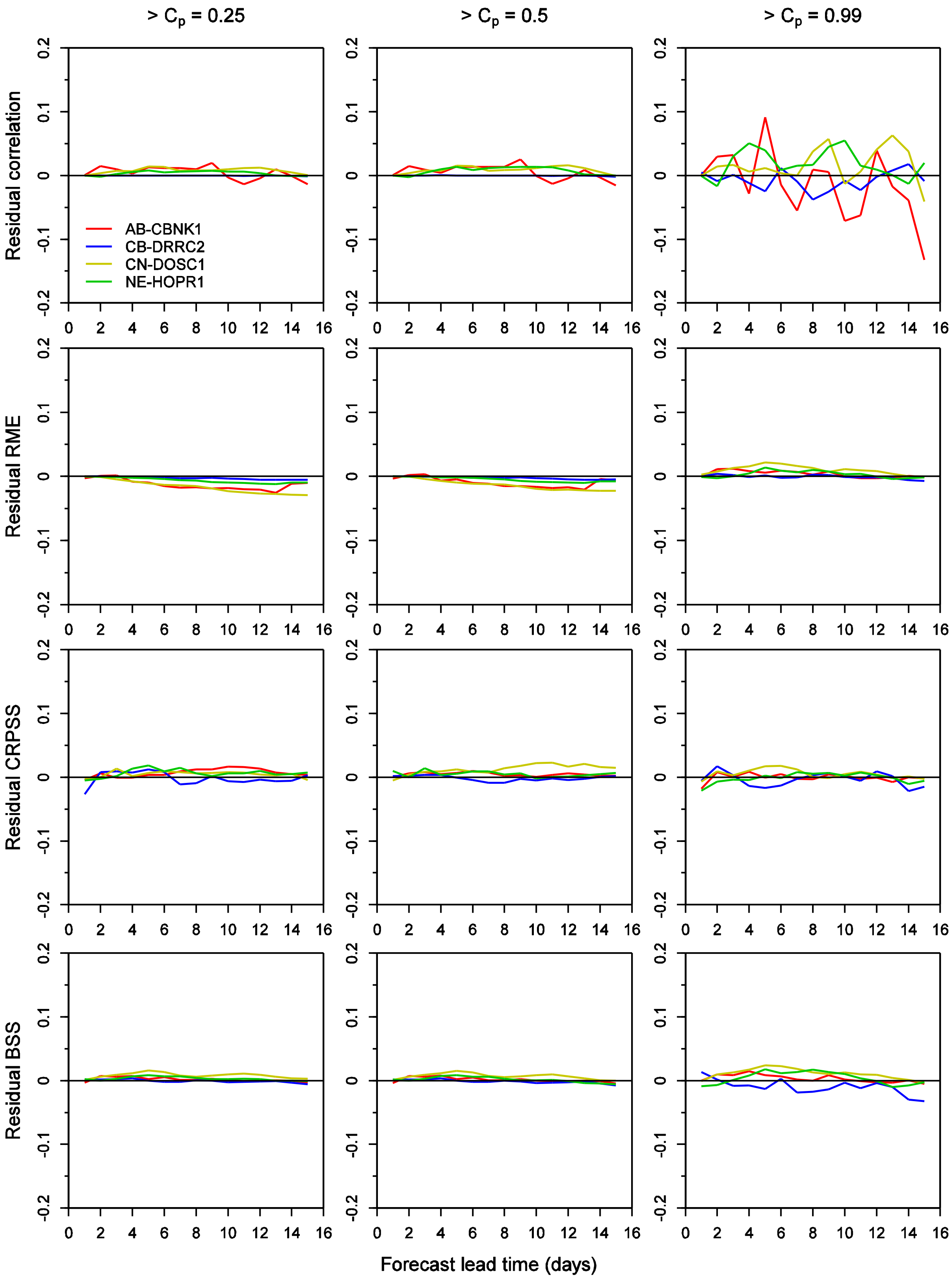

Figure 39: Residuals of selected verification metrics for the HEFS streamflow forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members versus C=5 member (F=11). The results are shown by forecast lead time for several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

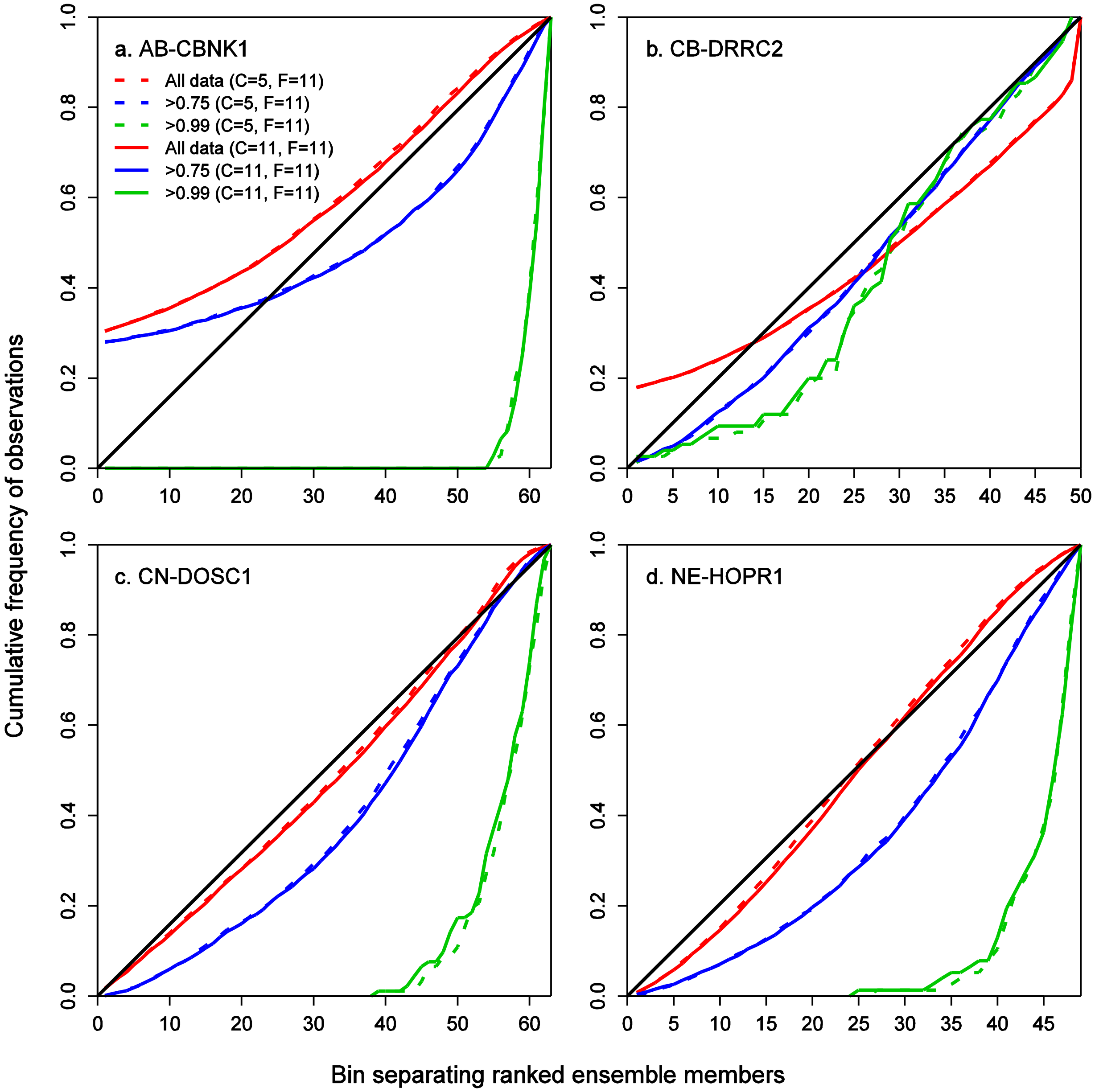

Figure 40: Cumulative rank histograms for the HEFS streamflow forecasts when calibrating the MEFP with an ensemble mean derived from C=11 members (solid) and C=5 members (dashed). The results are shown at a forecast lead time of 96-120 hours and for observed streamflow volumes that exceed several (non-exceedence) climatological probabilities.

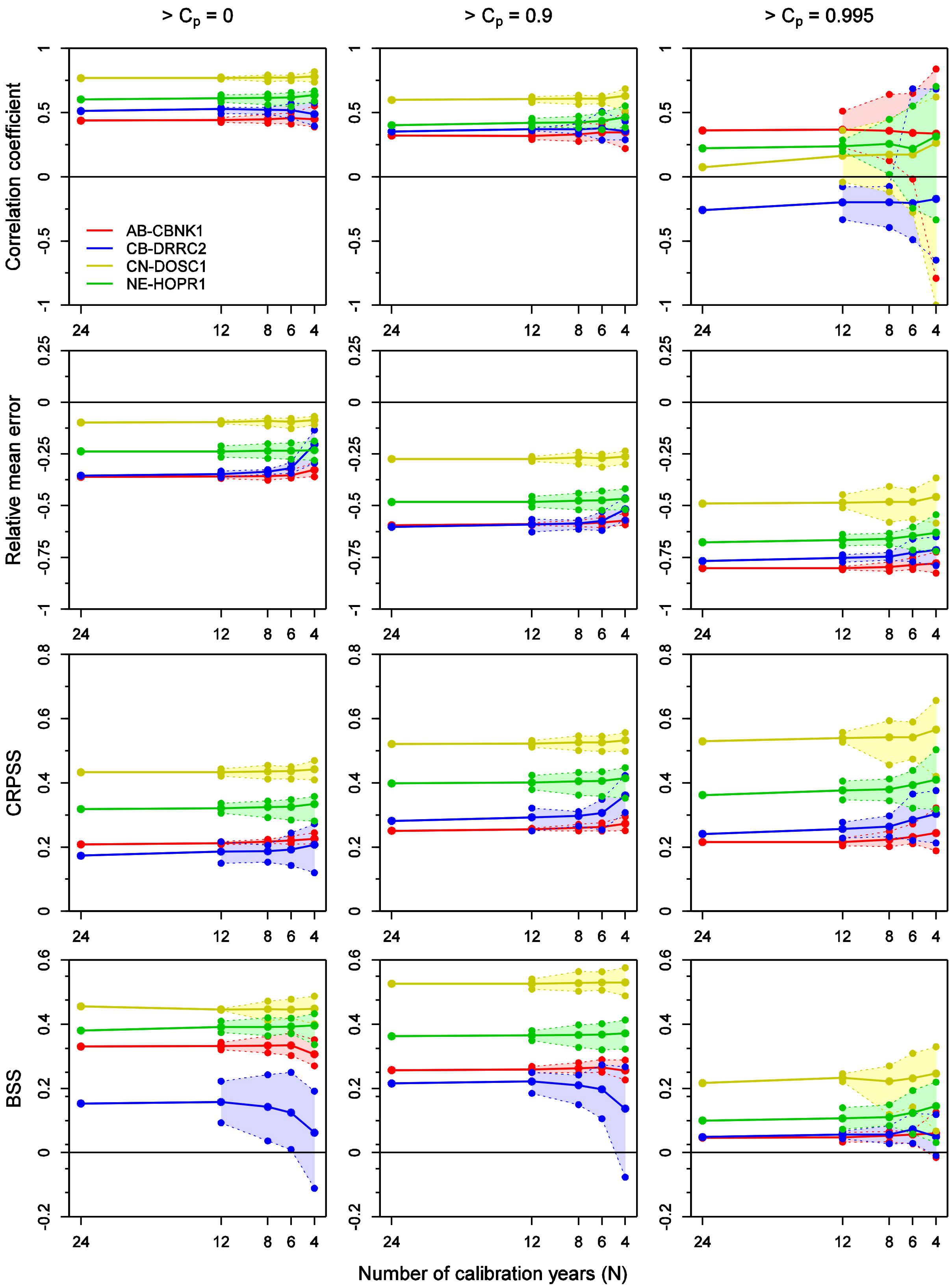

Figure 41: Selected verification scores for the MEFP-GEFS precipitation forecasts. The nominal scores are shown for each scenario of N (solid lines), together with the range of scores across the subcases of each scenario. The results include several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

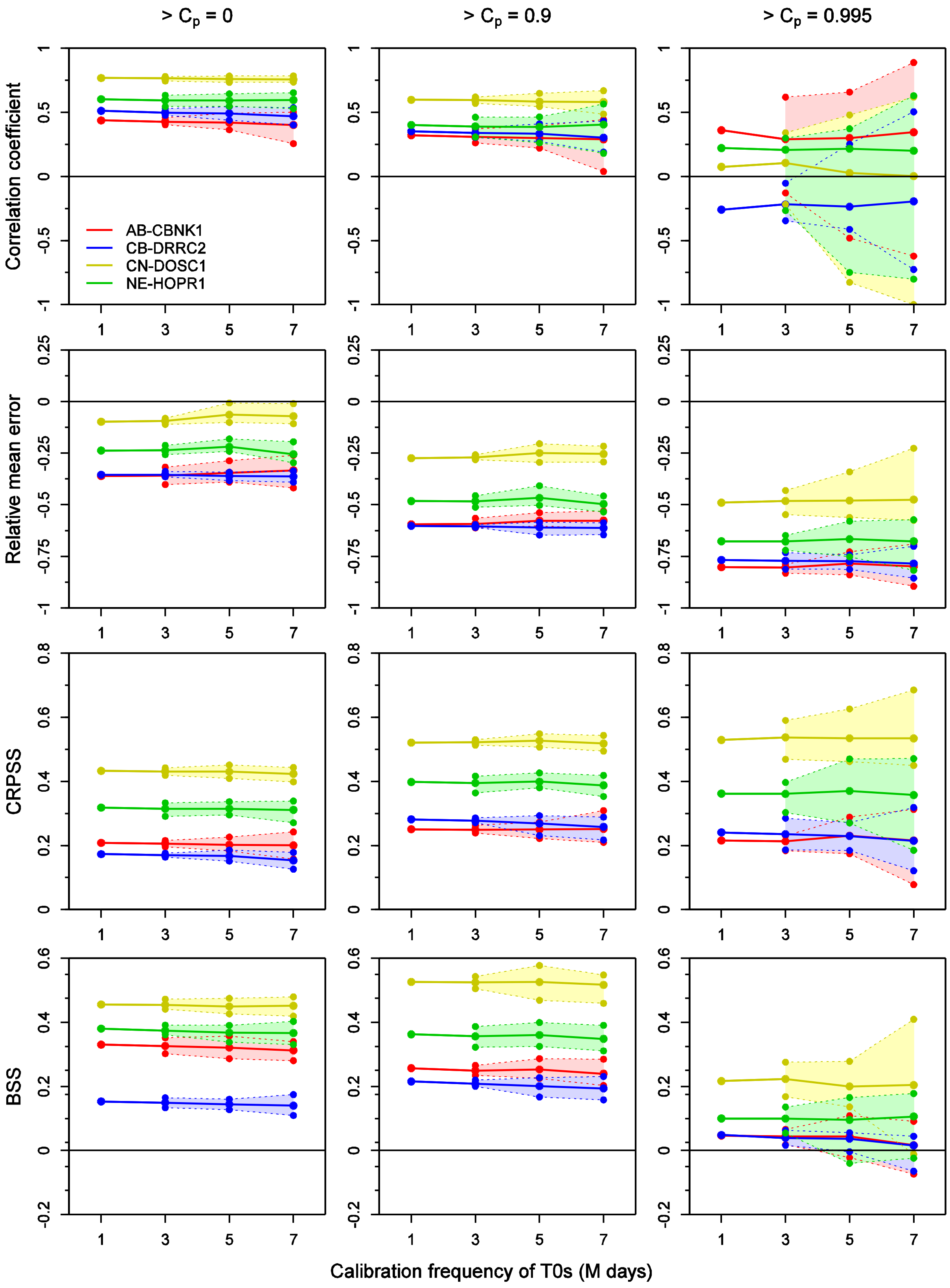

Figure 42: Selected verification scores for the MEFP-GEFS precipitation forecasts. The nominal scores are shown for each scenario of M (solid lines), together with the range of scores across the subcases of each scenario. The results include several non-exceedence climatological probabilities (C_p). The reference forecasts for the CRPSS and the BSS comprise the MEFP-CLIM forecasts.

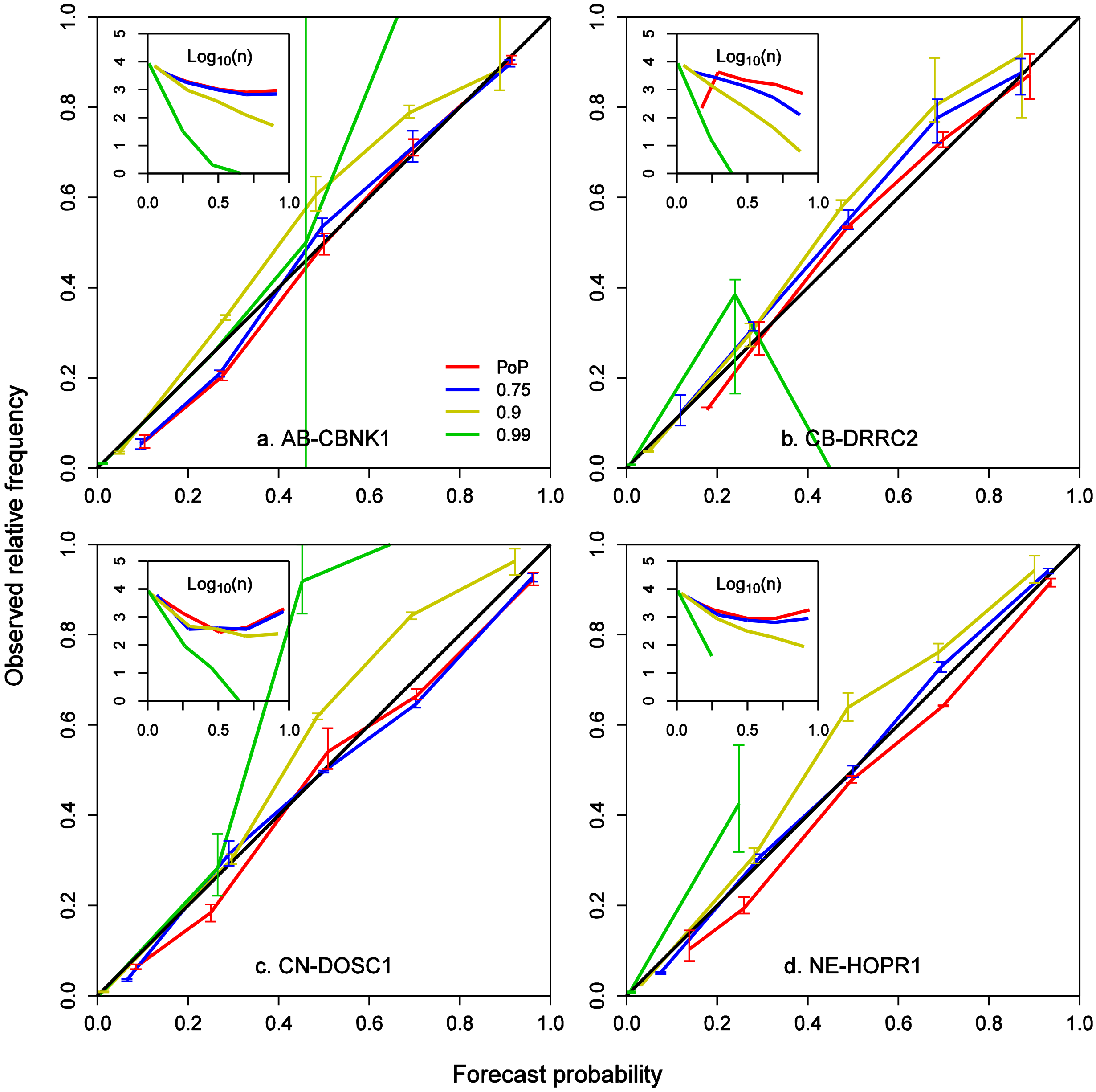

Figure 43: Reliability diagrams and corresponding sharpness plots (base 10 logarithm of the sample size, n) for the MEFP-GEFS precipitation forecasts at N=12. The results are shown for selected climatological non-exceedence probabilities (C_p), including the Probability of Precipitation (PoP; C_p=0.0), and comprise a daily aggregation between 0-24 hours. Alongside the nominal values (bold lines), the range of scores is shown for the two sub-periods of N=12.

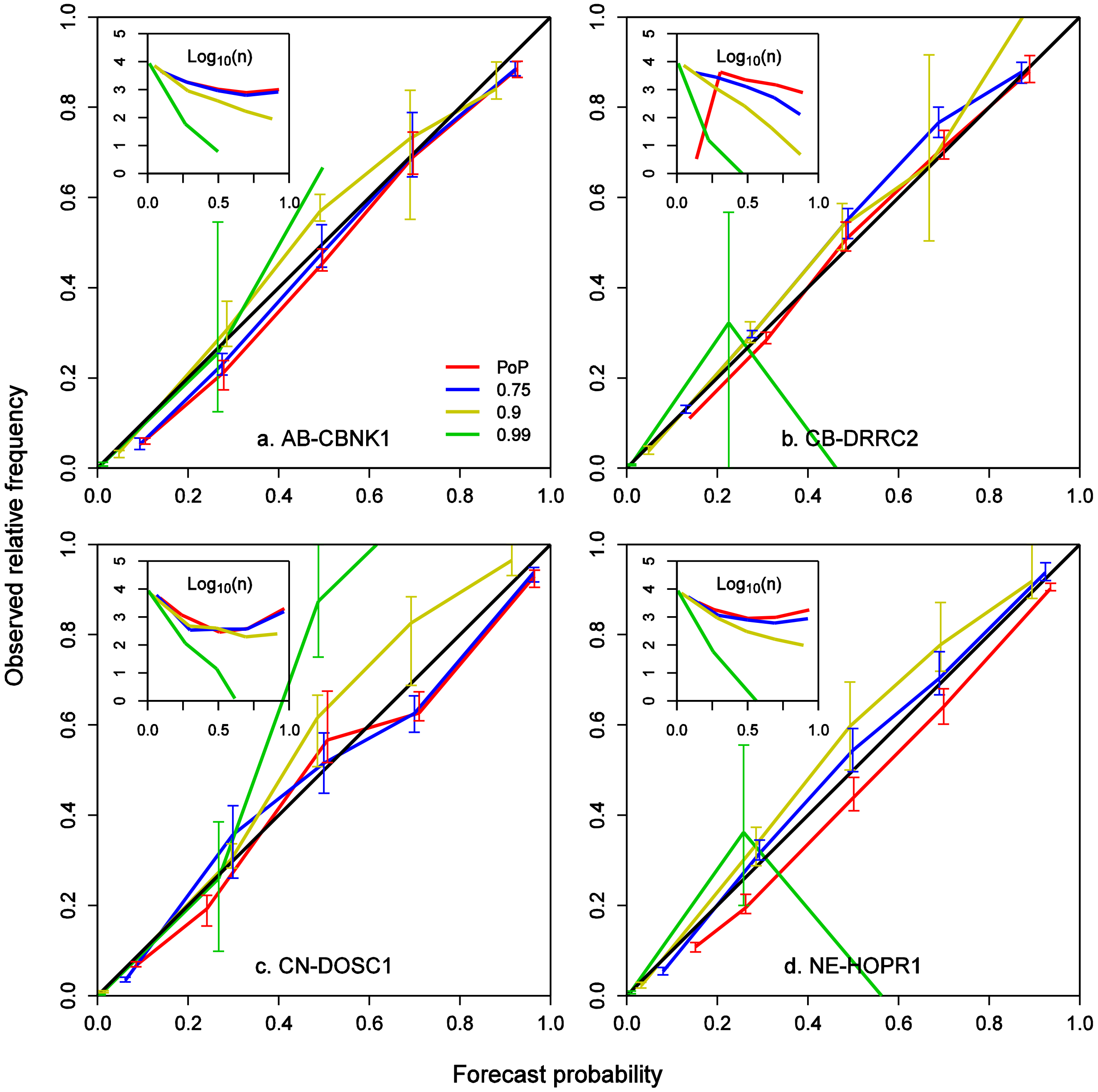

Figure 44: Reliability diagrams and corresponding sharpness plots (base 10 logarithm of the sample size, n) for the MEFP-GEFS precipitation forecasts at M=5. The results are shown for selected climatological non-exceedence probabilities (C_p), including the Probability of Precipitation (PoP; C_p=0.0), and comprise a daily aggregation between 0-24 hours. Alongside the nominal values (bold lines), the range of scores is shown for the five sub-periods of M=5.

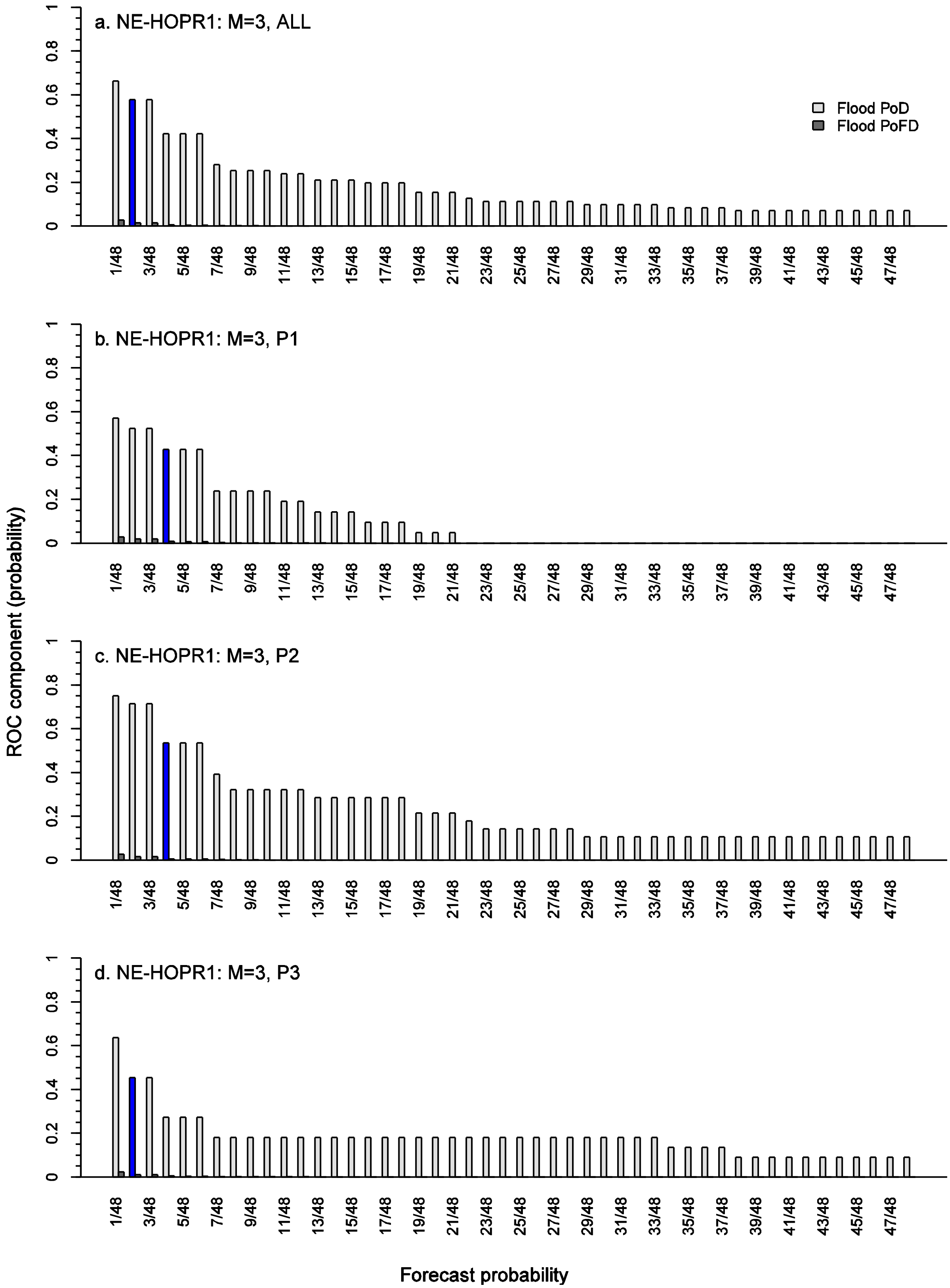

Figure 45: Probability of Detection (PoD) and Probability of False Detection (PoFD) for flooding at NE-HOPR1. The results are shown for each ensemble member (48 in total) and for three validation scenarios at a reforecast interval of M=3, namely the full period of record (daily reforecasts) and the three sub-periods (reforecasts every 3 days, offset by 1 day). The PoD is highlighted at PoFD≤0.015.

APPENDIX A: The Hydrologic Ensemble Forecast Service (HEFS)

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

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

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

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

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

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

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

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

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

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

The total uncertainty in Eqn. A1 is approximated, numerically, by integrating a finite number of "equally likely" ensemble members through the operational forecasting system. The HEFS is embedded within the Community Hydrologic Prediction System (CHPS), which provides the operational forecasting environment. A phased implementation of the HEFS is currently underway, with the first version (HEFSv1) due to be implemented across all RFCs by 2014. In support of this phased implementation, hindcasting and verification has been conducted at selected river basins in five RFCs (Brown, 2013, 2014; Brown et al., 2014a/b). The hindcasts are also being used by the NYCDEP in their Operational Support Tool (OST) for managing water supply to NYC.

APPENDIX B: Verification measures

a. Relative mean error

The relative mean error (RME), or fractional 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, $\overline{\mathrm{y}}$, and the corresponding observation, x, over n pairs of forecasts and observations

$$RME = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\bar{y}_{\mathrm{i}} - x_{\mathrm{i}}) \Big/ \sum_{\mathrm{i}=1}^{\mathrm{n}} x_{\mathrm{i}} . \tag{B1}$$

The RME provides a measure of relative bias in the ensemble mean forecast, and may be positive, zero, or negative. A positive RME denotes over-forecasting and a negative RME denotes under-forecasting (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 the variable, Q, exceeds a discrete threshold, 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. Normalizing by the BS of a reference forecast, BS_{REF}, leads to the Brier Skill Score (BSS),

$$BSS = 1 - \frac{BS}{BS_{REF}}. \tag{B3}$$

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 (Hersbach, 2000),

$$CRPS = \int \left\{ F_X(q) - F_Y(q) \right\}^2 dq. \tag{B4}$$

The mean CRPS comprises the CRPS averaged across n pairs of forecasts and observations. The Continuous Ranked Probability Skill Score (CRPSS) measures the difference in CRPS of the main prediction system, $\overline{CRPS}$, relative to a reference system, $\overline{CRPS}_{REF}$, as a fraction of the $\overline{CRPS}_{REF}$,

$$CRPSS = \frac{\overline{CRPS}_{REF} - \overline{CRPS}}{\overline{CRPS}_{REF}}. \tag{B5}$$

d. Reliability diagram

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

$$\bar{F}_{X_k}(q) = \frac{1}{|I_k|} \sum_{I_k} F_{X_i}(q), \; where \; I_k = \{i : i \in B_k\}. \tag{B6}$$

The corresponding fraction of observations is

$$\bar{F}_{Y_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{Y_i}(q), \textit{where } F_{Y_i}(q) = \begin{cases} 1, \mathrm{Y_i} > q; \\ 0, \textit{ otherwise} \end{cases}. \quad \text{(B7)}$$

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

e. Relative Operating Characteristic

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

$$PoD = \sum_{i=0}^{n} I_{X_i}\left(F_{X_i}(q) > d \mid Y_i > q\right) \Big/ \sum_{i=0}^{n} I_{Y_i}(Y_i > q). \quad \text{(B8)}$$

where I denotes the indicator function. The empirical PoFD is

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

f. Cumulative rank histogram

The rank histogram measures the reliability of an ensemble forecasting system. It involves counting the fraction of observations that fall between any two ranked ensemble members in the forecast distribution. For an ensemble forecast that comprises m ensemble members ranked in ascending order, $X=\{x_1,\ldots,x_m\}$, there are $m+1$ “gaps” between any two ensemble members into which the corresponding observation, y, could fall. The cumulative rank histogram measures the fraction of observations that fall below the upper bound of each gap

$$h_i = {}^{1}\!/_{n} \sum_{j=1}^{n} I\left(y_j < x_{ij}\right), \tag{B10}$$

where h_i is the fraction of observations that fall below the ith ranked ensemble member of the jth forecast, x_{ij}, and I is a step function that assumes value 1 if the condition is met and 0 otherwise.

If the forecasting system is reliable in terms of the rank histogram, the probability that an observation falls between any two ranked ensemble members is approximately uniform. Indeed, the actual reliability can be tested for goodness-of-fit of the sample fractions to a uniform probability distribution (e.g. using the one-sided Cramer von Mises test; Anderson, 1962).