# A Multi-Temporal-Scale Modulation Mechanism for the Postprocessing of Precipitation Ensemble Forecasts: Benefits for Streamflow Forecasting

JOSEPH BELLIER,[a,b] BRETT WHITIN,[c] MICHAEL SCHEUERER,[d] JAMES BROWN,[e] AND THOMAS M. HAMILL[b]

[a] *Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*
[b] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*
[c] *California–Nevada River Forecast Center, National Weather Service, Sacramento, California*
[d] *Norwegian Computing Center, Oslo, Norway*
[e] *Hydrologic Solutions Limited, Southampton, United Kingdom*

(Manuscript received 8 July 2022, accepted 27 November 2022)

ABSTRACT: In the postprocessing of ensemble forecasts of weather variables, it is standard practice to first calibrate the forecasts in a univariate setting, before reconstructing multivariate ensembles that have a correct covariability in space, time, and across variables, via so-called "reordering" methods. Within this framework though, postprocessors cannot fully extract the skill of the raw forecast that may exist at larger scales. A multi-temporal-scale modulation mechanism for precipitation is here presented, which aims at improving the forecasts over different accumulation periods, and which can be coupled with any univariate calibration and multivariate reordering techniques. The idea, originally known under the term "canonical events," has been implemented for more than a decade in the Meteorological Ensemble Forecast Processor (MEFP), a component of the U.S. National Weather Service's (NWS) Hydrologic Ensemble Forecast Service (HEFS), although users were left with material in the gray literature. This paper proposes a formal description of the mechanism and studies its intrinsic connection with the multivariate reordering process. The verification of modulated and unmodulated forecasts, when coupled with two popular methods for reordering, the Schaake shuffle and ensemble copula coupling (ECC), is performed on 11 Californian basins, on both precipitation and streamflow. Results demonstrate the clear benefit of the multi-temporal-scale modulation, in particular on multiday total streamflow. However, the relative gain depends on the method used for reordering, with more benefits expected when this latter method is not able to reconstruct an adequate temporal structure on the calibrated precipitation forecasts.

KEYWORDS: Statistical techniques; Ensembles; Probabilistic Quantitative Precipitation Forecasting (PQPF); Probability forecasts/models/distribution

## 1. Introduction

Probabilistic forecasting is becoming the standard practice in hydrologic prediction, where decision-makers can benefit from the knowledge of the forecast uncertainty to take more rational decisions. The ensemble approach is the most widely used approach to obtain probabilistic forecasts of streamflow, by ingesting meteorological ensembles issued by a numerical weather prediction (NWP) model into one (or more) hydrological model(s) (Cloke and Pappenberger 2009; Troin et al. 2021). Despite great progress over the last decades (Bauer et al. 2015), meteorological ensembles still tend to be underdispersed and sometimes biased, calling for a procedure known as statistical postprocessing [see Vannitsem et al. (2021) for a recent review of the field].

Meteorological postprocessors often follow the same two-step architecture. First, a *univariate calibration* is applied to the raw NWP ensembles at every combination of lead time, location, and weather variable (needed for hydrological modeling), producing a series of calibrated, univariate ensembles. Popular methods are, nonexhaustively, the ensemble model output statistics (EMOS; Gneiting et al. 2005), Bayesian model averaging (Raftery et al. 2005), or nonparametric techniques, such as quantile regression forest (Taillardat et al. 2016). Second, a *multivariate reordering* is performed to reconstruct multivariate ensembles that are coherent in space, time, and across weather variables, which is necessary for downstream applications such as hydrological modeling. Techniques generally used to this end are ensemble copula coupling (ECC; Schefzik et al. 2013), the Schaake shuffle (Clark et al. 2004), or variants thereof (e.g., Ben Bouallègue et al. 2016; Schefzik 2016; Scheuerer et al. 2017; Bellier et al. 2017). This two-step process must, in the perspective of streamflow forecasting, be performed at the spatiotemporal scale of the hydrological model, i.e., at the specific lead times and spatial units (e.g., basins, hydrological units, grid points) for which the hydrological model takes forcing data as input. When embedded in this framework though, most postprocessors are not able to fully extract the skill of the raw forecast that exists at larger scales. This is particularly true for precipitation, which NWP models can sometimes predict quite accurately in terms of multiday totals, while being slightly off in terms of spatial location and timing.

In this paper, we describe a postprocessing mechanism that aims to extract the skill of raw precipitation forecasts at multiple temporal scales. This mechanism, initially named "canonical events," traces back to the 2000s and an idea of Dr. J. Schaake at the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS).

*Corresponding author*: Joseph Bellier, joseph.bellier@noaa.gov

DOI: 10.1175/JHM-D-22-0119.1

It has been since then embedded in the Meteorological Ensemble Forecast Processor (MEFP), a component of the Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014) that is operationally used by the River Forecast Centers (RFCs) across the United States for probabilistic streamflow forecasting. It considers two types of temporal events (i.e., periods): the *base* events, which are the lead times that match the temporal scale of the hydrological model (e.g., 0–6, 6–12, …, 42–48 h), and the *modulation* events, which are temporal aggregations over multiple lead times (e.g., 0–24, 24–48, 0–48 h). Univariate calibration is performed not only on the base events, as in traditional postprocessors, but also on the modulation events. Then, the multivariate reordering step is performed in a sequential manner, by "modulating" the calibrated values of the base events according to the calibrated totals of the modulation events.

In the MEFP, the methods used for univariate calibration and multivariate reordering are the bivariate meta-Gaussian model (Schaake et al. 2007; Wu et al. 2011) and the Schaake shuffle (Clark et al. 2004), respectively. These two techniques have been thoroughly described in the above literature, while other studies (Brown et al. 2014; Demargne et al. 2014; Kim et al. 2018) that used the MEFP are limited to summarizing the main points. Meanwhile, the description of the canonical event approach is sparse. Articles limit to explaining the difference between base and modulation events, with the exception of Schaake et al. (2007), who touch upon how modulation events are incorporated into the multivariate reordering process. We found, though, that many aspects are overlooked or left out, and we would expect readers of this sole material to struggle to replicate the method. In addition, none of these studies have discussed the link that exists between the quality of the multivariate dependence structure and the gain brought by modulation. Despite the limited literature, the approach seems to yield a clear benefit, as stated, e.g., by Kim et al. (2018): "The use of the modulation events [ … ] significantly improves the predictive skill in ensemble precipitation and streamflow forecasts." In the same testbed as described later in section 2, we have also found modulation to greatly improve the MEFP streamflow forecasts, as Fig. 1 shows. We can therefore wonder whether the same principle can be used in connection with other postprocessing approaches, where the methods for univariate calibration and multivariate reordering are different from those currently used in the MEFP.

In this paper, we formally describe the basic principle of the canonical event approach, although from now on we will prefer the term *multi-temporal-scale modulation*, as we believe it is more self-explanatory. We follow the MEFP implementation, although without the many rules for dealing with specific cases that have been included in the original code. This modulation method is evaluated when it is coupled with the censored, shifted, gamma distribution (CSGD; Scheuerer and Hamill 2015) technique for the univariate calibration of the precipitation forecasts, as an alternative to the bivariate meta-Gaussian model that is used in the MEFP. For the multivariate reordering, we compare the use of the Schaake shuffle and ECC techniques, with the objective of studying how the multivariate dependence structure interacts with modulation. The postprocessed precipitation forecast are then propagated into a hydrological forecasting system, to verify that the gain in skill on precipitation brought by modulation does translate into better streamflow forecasts.

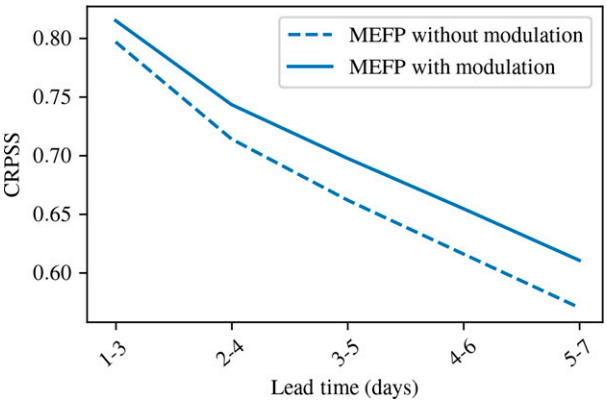


FIG. 1. Continuous ranked probability skill score (CRPSS) of 3-day total streamflow forecasts generated by the MEFP *with* and *without* modulation, in the same case study as described in section 2.

In summary, the objectives are 1) formally describe the modulation mechanism such that users outside of the MEFP community can replicate the method, 2) demonstrate that it can be successfully applied to other precipitation univariate calibration methods, 3) study the links between modulation and multivariate reordering, and 4) quantify the gains in forecast skill for precipitation, but also for streamflow. The paper is organized as follows. Section 2 presents the case study and data. Section 3 describes the methods, including the univariate calibration method used here, the two reordering techniques tested, and the multi-temporal-scale modulation method. Section 4 presents the results and discusses them, while section 5 concludes.

## 2. Case study and data

A set of 11 basins, which are monitored by the California–Nevada River Forecast Center (CNRFC), is considered. Table 1 provides some relevant hydrological characteristics, while Fig. 2 displays their location. As in operational forecasting, some of the larger basins are divided into subbasins based on elevation, leaving a total of 18 subbasins for which mean areal precipitation and temperature at 6-h intervals must be provided for hydrological modeling.

The meteorological forecasts come from the second-generation Global Ensemble Forecast System (GEFS) reforecast dataset (Hamill et al. 2013). They comprise 11 members, run each day from 0000 UTC initial conditions. Lead times up to +14 days, at 6-h intervals, are considered. The native resolution is approximately 0.5° grid spacing for week +1 and 0.75° for week +2. However, what will be referred in this paper to as the "raw forecasts" (of precipitation and temperature) are the gridded reforecasts that are spatially averaged to the 18 subbasins. A new

TABLE 1. Hydrological characteristics of the 11 basins of the case study.

| Basin | Subbasin(s) | Drainage area (km$^2$) | Mean elevation (m MSL) | Mean annual flow (m$^3$ s$^{-1}$) |
|---|---|---|---|---|
| BSRC1 | BSRC1HOF | 122 | 777 | 3 |
| CREC1 | CREC1HOF | 1588 | 766 | 104 |
| CWAC1 | CWAC1HLF, CWAC1HUF | 2391 | 676 | 25 |
| DOSC1 | DOSC1HLF, DOSC1HUF | 1930 | 1123 | 42 |
| FTJC1 | FTJC1HLF, FTJC1HUF | 1712 | 1320 | 17 |
| HPIC1 | HPIC1HMF, HPIC1HUF | 469 | 2744 | 10 |
| NFDC1 | NFDC1HLF, NFDC1HUF | 886 | 1331 | 23 |
| PFTC1 | PFTC1HLF, PFTC1HMF, PFTC1HUF | 3996 | 2328 | 67 |
| PRBC1 | PRBC1HOF | 1008 | 462 | 3 |
| UKAC1 | UKAC1HOF | 259 | 447 | 5 |
| WOOC1 | WOOC1HOF | 171 | 2458 | 3 |

version of the GEFS reforecasts is now available (Guan et al. 2022), but these were not used in this study. The observed data of mean areal precipitation and temperature are available from the CNRFC. All data are available over the 1989–2010 period (21 years). This will be our verification period, using a leave-one-year-out cross validation strategy. The month of September, however, is discarded because of occasional very low flow conditions that affected the hydrological simulations.

The hydrological models used in this study are the Sacramento soil moisture accounting (SAC-SMA) model coupled with the SNOW-17 snow model. They are embedded in the NOAA Community Hydrologic Prediction System (CHPS), a hydrological modeling platform that RFCs across the United States use for daily operations. The CHPS can also be used in hindcast mode for research purposes, as for this study. A hydrological simulation spanning the 1989–2010 period is then run, using the observed forcing. This provides a time series of model states that will be used as initial conditions for the hindcasts runs, which are processed one at a time and start each day at 1200 UTC. The hindcast procedure is fully automated with no interaction by the modeler, which differs from real-time conditions at RFCs where manual adjustments are made to improve both modeled watershed states and forecasted runoff response, resulting in reduced forecast error. In our study, to exclude hydrological modeling errors we perform the verification against the simulated streamflow.

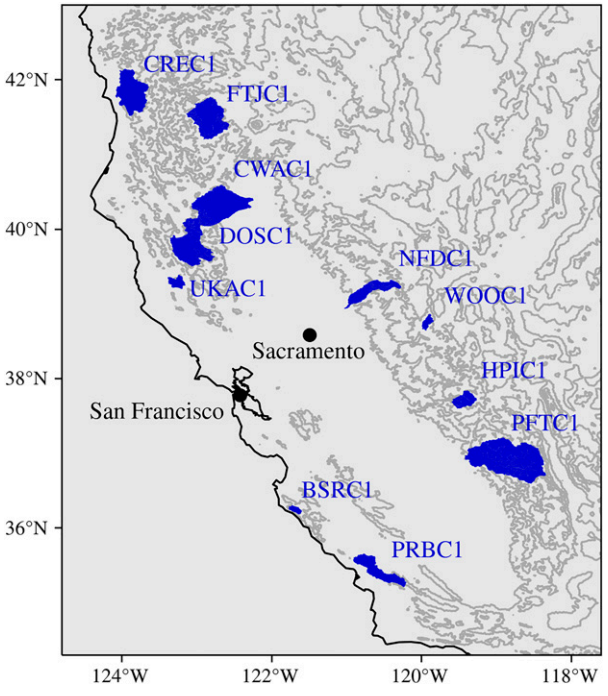


FIG. 2. Location of the 11 study basins in California.

## 3. Methods

Let us first present some aspects of notation that will be used throughout this section. The index $i \in \{1, \ldots, I\}$ refers to subbasins, $l \in \{1, \ldots, L\}$ to lead times, and $m \in \{1, \ldots, M\}$ to ensemble members. The raw ensemble forecast is denoted by $\mathbf{r}$. This quantity must be viewed as a three-dimensional array that writes $\mathbf{r} = \{\mathbf{r}^{1,1}, \ldots, \mathbf{r}^{I,L}\}$, where $\mathbf{r}^{i,l} = (r_1^{i,l}, \ldots, r_M^{i,l})$ denotes the $M$-member ensemble at subbasin $i = 1, \ldots, I$ and lead time $l = 1, \ldots, L$. Other multivariate ensembles, such as the forecast at different stages of the process, will be denoted by bold letters and interpreted similarly. Table 2 helps keep track of the notation.

We do not refer explicitly to weather variables, as this paper focuses on precipitation. The other variable required for hydrological modeling, the surface temperature, is here subject to univariate calibration with the standard EMOS technique (Gneiting et al. 2005). Its covariability with precipitation is reconstructed with either the Schaake shuffle or ECC as described in section 3b, as if it was an additional dimension to the subbasins and lead times. No modulation is applied to the

TABLE 2. Notation of the multivariate ensembles at different stages of the precipitation postprocessing process.

| Letter | Description |
|---|---|
| $\mathbf{r}$ | Raw ensemble |
| $\mathbf{x}$ | Calibrated ensemble |
| $\tilde{\mathbf{x}}$ | Reordered, calibrated ensemble |
| $\check{\mathbf{x}}$ | Modulated, reordered, calibrated ensemble (final output) |
| $\mathbf{z}$ | Template ensemble used for the reordering |
| $\mathbf{w}$ | Calibrated ensemble for the modulation events |

temperature forecasts. From now on, any quantity in the formal description is assumed to refer to precipitation.

### a. Univariate calibration

The goal of univariate calibration is to obtain, using $\mathbf{r}$ as input, a calibrated ensemble that we denote by $\mathbf{x} = (\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{I,L})$, where each univariate ensemble $\mathbf{x}^{i,l} = (x_1^{i,l}, \ldots, x_M^{i,l})$ is reliable. Because the procedure is univariate, we drop for this subsection the indices $i$ and $l$ from the equations. The method employed here, also described in Scheuerer and Hamill (2018), is a slightly simplified variant of the technique proposed by Scheuerer and Hamill (2015). It shares in common with the EMOS approach the assumption that predictive distributions take the form of a parametric distribution, which is here the CSGD, with parameters mean $\mu$, standard deviation $\sigma$, and shift $\delta$. We briefly describe the method in the next two paragraphs, although the modulation mechanism described in section 3c, which is the core of this paper, can be applied with any other univariate calibration technique.

The procedure starts by fitting a climatological CSGD (i.e., a distribution model for the observations), to obtain parameters $\mu_{\mathrm{cl}}$, $\sigma_{\mathrm{cl}}$, and $\delta_{\mathrm{cl}}$. The parameters $\mu$, $\sigma$, and $\delta$ of the predictive CSGDs are then related to three statistics [mean ($\mathrm{MEAN}_r$), mean absolute difference ($\mathrm{MD}_r$), and probability of precipitation ($\mathrm{POP}_r$)] of the augmented and homogenized (see next paragraph) raw ensemble, via

$$\mu = \frac{\mu_{\mathrm{cl}}}{a_1}\mathrm{log1p}[\mathrm{expm1}(a_1)(a_2 + a_3\mathrm{POP}_r + a_4\mathrm{MEAN}_r)], \quad (1)$$

$$\sigma = \sigma_{\mathrm{cl}}\left[b_1\sqrt{\frac{\mu}{\mu_{\mathrm{cl}}}} + b_2\mathrm{MD}_r\right], \quad (2)$$

$$\delta = \delta_{\mathrm{cl}}, \quad (3)$$

where $\mathrm{log1p}(x) = \log(1 + x)$ and $\mathrm{expm1}(x) = \exp(x) - 1$. These regression equations, which model the predictive CSGDs as deviations from the unconditional, climatological CSGD, have proven to be well suited for modeling the nonlinear, heteroscedastic behavior of precipitation (Scheuerer and Hamill 2015). A separate set of climatological parameters $\mu_{\mathrm{cl}}$, $\sigma_{\mathrm{cl}}$, and $\delta_{\mathrm{cl}}$ is determined for each subbasin, time of the day (6-h period), and day of the year, while the regression coefficients $a_1, a_2, a_3, a_4, b_1$, and $b_2$ are specific to each subbasin, lead time, and month of the year. These are determined by minimizing, over a training sample, the continuous ranked probability score (CRPS, see section a of the appendix) of the predictive CSGDs against the corresponding observations.

As suggested in Scheuerer et al. (2017), the ensemble statistics $\mathrm{MEAN}_r$, $\mathrm{MD}_r$, and $\mathrm{POP}_r$ are computed from forecasts within a spatial and temporal neighborhood of the respective subbasin, using a data-driven weighting scheme, with the objective to mitigate both displacement and timing errors. The size of the spatial neighborhood is defined in the same way as in Scheuerer et al. (2017) while the temporal neighborhood in our study consists of the current, the preceding, and the subsequent lead time. Here, we calculate the coefficient of predictive ability $\mathrm{CPA}_{s,t}$ (Gneiting and Walz 2022) to quantify the relative importance of each forecast grid point $(s, t)$ within the spatiotemporal neighborhood. The CPA can be viewed as a generalization of Spearman's rank correlation coefficient that is particularly suited for highly non-Gaussian quantities like precipitation amounts. The weight of grid point $(s, t)$ is then defined via

$$\omega_{s,t} \propto (\mathrm{CPA}_{s,t} - \min_{s',t'}\mathrm{CPA}_{s',t'})^2, \quad (4)$$

where the minimum is taken over the entire space–time neighborhood and weights are normalized such as to sum up to 1. Since $\mathrm{MEAN}_r$, $\mathrm{MD}_r$, and $\mathrm{POP}_r$ are computed from an augmented ensemble that contains forecasts at grid points with potentially different model climatologies, the raw ensemble forecasts are homogenized before calculating the weighted ensemble statistics. Following Scheuerer and Hamill (2018), we do this by dividing—separately for each grid point—each forecast by the climatological mean at this forecast grid point. The ensemble statistics are thus calculated from multiplicative forecast anomalies instead of the raw ensemble forecasts.

The last step of the univariate calibration is to obtain discrete ensembles from the continuous CSGDs. We here sample $M = 33$ members from the CSGDs, using an equidistant quantile sampling scheme with probabilities $1/(M + 1), \ldots, M/(M + 1)$. The choice of $M = 33$ aims at representing distributions better than with 11 members (as in the raw ensemble), while keeping $M$ as a multiple of 11, for a reason related to the ECC technique that is explained in the next subsection.

### b. Multivariate reordering

Univariate calibration has left us with the *calibrated* ensemble $\mathbf{x} = (\mathbf{x}^{1,1}, \ldots, \mathbf{x}^{I,L})$, but the member values within each univariate ensemble $\mathbf{x}^{i,l} = (x_1^{i,l}, \ldots, x_M^{i,l})$ are ordered based on the sampling scheme, not following any physical principle. The next step thus consists in "reordering" the member values within each combination $\{i, l\}$ such that it instills a specific multivariate dependence structure on the forecast. This will result in a *reordered*, *calibrated* ensemble that we denote by $\tilde{\mathbf{x}}$. The multivariate dependence structure in $\tilde{\mathbf{x}}$ must ideally be (i) *coherent*, which we define as representing in a plausible way (for the region and climate at hand) the covariability of the predictand across lead times and subbasins, but also (ii) *appropriate* to the forecast case, as spatiotemporal patterns of the predictand (here precipitation) may vary according to season, meteorological situations (stratiform versus convective), etc. Different reordering methods can be more or less effective in reproducing a dependence structure that satisfies these two attributes.

In this study, we compare two nonparametric reordering methods, the Schaake shuffle, in its standard version (Clark et al. 2004) and the ensemble copula coupling (ECC; Schefzik et al. 2013). Before looking at their difference, we describe their common mechanism. First, construct a *template* ensemble $\mathbf{z}$ (with the same dimensions as $\mathbf{x}$) with a desirable dependence structure. Second, set $\tilde{\mathbf{x}} = \mathbf{x}$, and then iterate over all combinations $\{i, l\}$ with $i = 1, \ldots, I$ and $l = 1, \ldots, L$, permuting the values $(\tilde{x}_1^{i,l}, \ldots, \tilde{x}_M^{i,l})$ such that

$$[\mathrm{rank}(\tilde{x}_1^{i,l}), \ldots, \mathrm{rank}(\tilde{x}_M^{i,l})] = [\mathrm{rank}(z_1^{i,l}), \ldots, \mathrm{rank}(z_M^{i,l})], \quad (5)$$

with ties resolved as random. In other words, reproduce in $\tilde{\mathbf{x}}$ the same *rank* dependence structure as in the template $\mathbf{z}$.

The Schaake shuffle and ECC techniques follow this same mechanism; however, they differ in the data used to populate $\mathbf{z}$. In the Schaake shuffle, $M$ historical (observed) trajectories are selected. We here replicate the Clark et al. (2004) implementation, where these trajectories are randomly selected from all available years, although they must start within a 14-day window centered around the calendar date of the forecast. The idea of the Schaake shuffle is therefore to impose on $\tilde{\mathbf{x}}$ the rank dependence structure of past observations selected based solely on the time of the year.

In the standard ECC, the template is the raw ensemble, i.e., $\mathbf{z} = \mathbf{r}$. The reordered, calibrated ensemble $\tilde{\mathbf{x}}$ therefore inherits the rank dependence structure of the NWP forecast, and for that reason ECC is sometimes referred to as a "flow-dependent" technique. In our case study, the raw ensemble from GEFS has 11 members, while $M = 33$ members have been sampled from the calibrated distributions. To circumvent that issue and populate $\mathbf{z}$ with 33 members, we consider a lagged ensemble, i.e., we "recycle" the raw forecasts issued one and two days prior to the forecast date, and shift their lead times by 24 and 48 h, respectively, such that the valid times match. The underlying assumption is that meteorological patterns are generally seen several days in advance by the NWP models, and so the spatiotemporal structure in three consecutive forecasts is likely to be similar [with the exception of occasional "jumpy" situations, as described in Zsoter et al. (2009)]. To verify this hypothesis, we have evaluated the forecast skill of three sets of 11-member postprocessed forecasts, all reordered with standard ECC, but using as template: (i) the current raw forecast, (ii) the 1-day lagged raw forecast, and (iii) the 2-day lagged raw forecast. Results (not shown) showed that forecast skills are nearly identical between the three experiments, which supports the hypothesis behind our lagged-ensemble ECC implementation.

In terms of coherence and appropriateness, the two desirable attributes of the dependence structure, the Schaake shuffle and ECC techniques provide different advantages. The Schaake shuffle is effective regarding the coherence, as the dependence structure is derived from observations that are measured at the same spatiotemporal scale of the hydrological model. However, in terms of appropriateness it limits to conditioning the dependence structure to the season, which can be a serious shortcoming in regions and/or seasons where meteorological patterns vary substantially from day to day. The ECC, as a flow-dependent method, is likely superior regarding the appropriateness of the dependence structure over the Schaake shuffle, at least for the lead times where the raw forecast has skill. However, its coherence will strongly depend on the spatiotemporal resolution of the NWP model, relative to the meteorological phenomenon. For instance, if multiple subbasins lie within a single NWP grid cell, ECC will not be able to reproduce spatial patterns that may occur on a subgrid scale, and it will therefore instill a disproportionately strong spatial dependence structure.

### c. *Multi-temporal-scale modulation*

The objective of multi-temporal-scale modulation is to adjust the postprocessed forecast such that it retains as much skill as possible from the lead-time-by-lead-time calibration, while being more skillful at larger temporal scales. In the MEFP terminology, individual lead times are called *base events*, while periods of temporal aggregations (over multiple lead times) are named *modulation events*. These modulation events are fixed and must be defined ahead of time, based on the typical precipitation patterns for the region and climate at hand. They may overlap one another, and their duration typically increases with lead time, following the idea that the timing of precipitation is harder to forecast with increasing lead times. Let $k = 1, \ldots, K$ denote the modulation events, and for any event $k$ let $s_k$ be the set containing the indices $l$ of the lead times (i.e., base events) that it encompasses. Table 3 presents the $K = 7$ modulation events that are used in this study, which are the same that the CNRFC uses for operational forecasting, although we have discarded those involving lead times beyond 14 days. We briefly discuss at the end of this subsection the rationale behind this choice of aggregation periods. However, the paper does not aim at testing the sensitivity of the forecast skill to the definition of the modulation events, but rather to demonstrate the potential benefit of the modulation mechanism with parameters that are already operationally used.

TABLE 3. Modulation events used in this study.

| Modulation event | Aggregation period (in hours) | Aggregation period (in days) |
|---|---|---|
| $k = 1$ | 0–72 | 0–3 |
| $k = 2$ | 36–108 | 1.5–4.5 |
| $k = 3$ | 72–144 | 3–6 |
| $k = 4$ | 108–180 | 4.5–7.5 |
| $k = 5$ | 144–216 | 6–9 |
| $k = 6$ | 216–288 | 9–12 |
| $k = 7$ | 180–336 | 7.5–14 |

Once the modulation events are defined, univariate calibration must be extended to precipitation accumulations over the modulation events. The simplest way is to use the same calibration method just changing the predictand, although one can use a different method too. In this study, we reuse the CSGD method described in section 3a, although we deactivate the temporal neighborhood scheme as it is less relevant for longer accumulation periods. While $\mathbf{x}^{i,l} = (x_1^{i,l}, \ldots, x_M^{i,l})$ for $l = 1, \ldots, L$ referred to the calibrated ensembles for the individual lead times, let $\mathbf{w}^{i,k} = (w_1^{i,k}, \ldots, w_M^{i,k})$ for $k = 1, \ldots, K$ denote the calibrated forecasts for the modulation events. In this subsection, subbasin indices $i$ are left in the equations for completeness, although the procedure applies independently for each subbasin, and therefore the reader can overlook them. Finally, let $\check{\mathbf{x}}$ denote the *modulated*, *reordered*, *calibrated* ensemble that will be output, using the method described as follows.

Because modulation and base events are defined such that they overlap in multiple ways, adjusting the values in $\check{\mathbf{x}}$ such that the distributions match the calibrated distributions for all the events (base and modulation) is an overconstrained problem. Therefore, the procedure proposes to sort all the events

(base and modulation) according to a metric that reflects their forecasting skill, and proceed to the adjustment of the forecast values one event at a time, in increasing order of skill, such that the calibrated values for more skillful events are given more importance. This sorting is specific to each subbasin and month of the year. In the MEFP original implementation, the metric used for sorting the events is the correlation parameter of the bivariate meta-Gaussian distribution that models the joint distribution of the forecast and the observation over the training period. Here, because a different method (the CSGD) is used for calibration, we do not have access to that exact same metric. As a simple alternative, we here use the Spearman's rank correlation coefficient between the calibrated ensemble means and the observations over the training period, leaving out of the computation the pairs of zeros. The impact of the choice of the metric used for sorting the events is discussed later in this subsection.

The modulation of the forecast values then works as follows:

Step 1: Using $\mathbf{x}$ and $\mathbf{z}$, perform a standard reordering (cf. section 3b), to obtain $\tilde{\mathbf{x}}$.

Step 2: Set $\check{\mathbf{x}} = \tilde{\mathbf{x}}$. At this point, $\check{\mathbf{x}}$ already has a dependence structure, so the computation of temporal precipitation accumulations over multiple lead times is meaningful.

Step 3: Iterate over the $L + K$ events in increasing order of skill, and proceed differently depending on whether it is a

- *base event* (indexed by $l$): Update the current ensemble $\check{\mathbf{x}}$ at lead time $l$ with its counterpart in the unmodulated ensemble $\tilde{\mathbf{x}}$:

$$\check{\mathbf{x}}^{i,l} \leftarrow \tilde{\mathbf{x}}^{i,l}, \tag{6}$$

- *modulation event* (indexed by $k$): For each member $m = 1, \ldots, M$, proceed as follows. First, compute the multiplicative factor

$$\alpha_m^{i,k} = \frac{w_{m'}^{i,k}}{\sum_{l \in s_k} \check{x}_m^{i,l}}, \tag{7}$$

where $m' \in \{1, \ldots, M\}$ is determined such that the rank of $w_{m'}^{i,k}$ among $(w_1^{i,k}, \ldots, w_M^{i,k})$ is equal to the rank of $\sum_{l \in s_k} \check{x}_m^{i,l}$ among $\left(\sum_{l \in s_k} \check{x}_1^{i,l}, \ldots, \sum_{l \in s_k} \check{x}_M^{i,l}\right)$. Then, update the current ensemble $\check{\mathbf{x}}$ for all the lead times covered by the modulation event $k$:

$$\check{x}_m^{i,l} \leftarrow \alpha_m^{i,k} \check{x}_m^{i,l} \tag{8}$$

for all $l \in s_k$.

As zero precipitation is frequent, it may happen that $w_{m'}^{i,k} > 0$ but $\sum_{l \in s_k} \check{x}_m^{i,l} = 0$, making a multiplicative modulation impossible. In such cases, as an ad hoc workaround we propose to randomly select one lead time $l$ within $s_k$, and assign to the corresponding element $\check{x}_m^{i,l}$ the entire value of $w_{m'}^{i,k}$. This generally concerns cases where $w_{m'}^{i,k}$ is low though, so the impact on forecast skill is negligible.

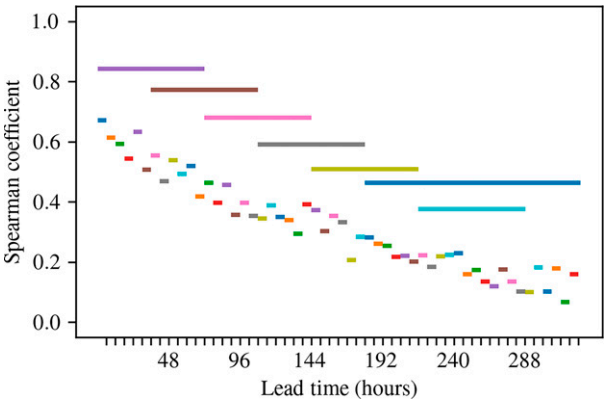


FIG. 3. Spearman's rank correlation coefficient of all events (base and modulation), for the month of January and the subbasin CWAC1HLF. Bars spanning multiple lead times correspond to modulation events. Colors are only used to facilitate distinguishing the events.

In the above algorithm, following J. Schaake's original idea, all events are pooled together and processed sequentially, leaving the possibility of a base event to be processed after the modulation event(s) that comprise it, in which cases the ensemble $\check{\mathbf{x}}^{i,l}$ takes back the unmodulated values from $\tilde{\mathbf{x}}^{i,l}$ [Eq. (6)]. However, this is rarely seen in practice, and modulation events generally happen to be processed on top of the base events. Figure 3 shows the Spearman's rank correlation coefficients plotted as a function of the events, for subbasin CWAC1HLF and the month of January. When sorting the events based on these criteria, it is indeed observed that every base event is ranked (and therefore processed) before the modulation event(s) that comprise it.

As an alternative to the Spearman's rank correlation coefficient for sorting the events, we have also tested the CRPS of the calibrated forecasts against the observations, a metric that takes into account the entire forecast ensemble and not only its mean. Interestingly, while the two metrics were sorting the base events among themselves sometimes slightly differently, they were systematically sorting the modulation events among themselves in the same order, but also the modulation events and their respective underlying bases events in the same order. This means that, in our case study, the base and modulation events have sufficiently different skill for the choice of the sorting metric to be impactless. Nonetheless, users wanting to replicate this modulation mechanism are encouraged to verify this hypothesis in their own forecast setup, in particular if significantly different aggregation periods to the ones described in Table 3 are considered.

Obviously, the sequentiality of the modulation process causes a strong dependence of the final ensemble $\check{\mathbf{x}}$ upon the highest skilled modulation events. Recall that processing the events sequentially is necessary to circumvent the issue of overconstraint that multiscale modulation necessarily raises, when the forecast must match the calibrated distribution for multiple temporal periods that overlap. In the context of streamflow postprocessing, Alizadeh et al. (2020) face a

similar issue, and propose a solution that shares some similarities with the one described here, although they do not allow ensemble values for a given lead time to be modulated more than once. Both rules can be considered as heuristic, and despite lacking a strong scientific basis they have the advantage of being very straightforward to implement. As potential improvements, one could try to solve the overconstraint modulation problem numerically, by finding a solution that minimizes the total amount of violation of the constraints, for instance. This avenue is left for future research.

The definition of the modulation events is another aspect that would benefit from improvements. At the CNRFC, the objective was to define the events based on local weather knowledge, including typical storm duration and lead time forecast skill. Multiple sets were developed using this knowledge, and had their resulting streamflow hindcasts evaluated. Ultimately, the final set was selected based on two criteria: (i) it could be explained in terms of important periods of precipitation forecast skill aggregation, and (ii) it provided stable results without noticeable discontinuities at modulation event boundaries (minimizing the amount of overlapping modulation events). No extensive research has been conducted to find a process that optimizes the definition of these events, and this represents a clear avenue for future studies.

Finally, it is important to understand the role played in this method by the template ensemble $\mathbf{z}$. In Schaake et al. (2007), modulation is described as embedded within the Schaake shuffle, but we show in this paper that it can be coupled with any reordering approach, just changing $\mathbf{z}$. This template $\mathbf{z}$ instills a specific temporal structure in $\tilde{\mathbf{x}}$ (step 1), which will directly affect the computation of the temporal accumulations $\sum_{l \in s_k} \check{x}_m^{i,l}$, and thereby of the multiplicative factors $\alpha_m^{i,k}$ (step 3). If $\mathbf{z}$ is "inadequate" with respect to the forecast temporal pattern, the multiplicative factors will be far from one, and as a result the univariate ensembles in the final forecast $\check{\mathbf{x}}$ may differ substantially from the univariate ensembles that were obtained via calibration. It is also worth mentioning a subtle difference that exists between the modulation algorithm described here and the original MEFP implementation. At step 2, $\check{\mathbf{x}}$ is here initialized with $\tilde{\mathbf{x}}$, while it is initialized with $\mathbf{z}$ in the MEFP. We believe our approach makes more sense, as the sequential processing of the events begins with an ensemble $\check{\mathbf{x}}$ which, unlike $\mathbf{z}$, is already calibrated in terms of both forecast distribution and dependence structure. Therefore, the temporal accumulations computed in Eq. (7) will more likely be closer to the calibrated values for the modulation events, resulting in multiplicative factors $\alpha_m^{i,k}$ that are closer to one, and therefore which modify the forecast values to a lesser extent. But again, in practice, modulation events are almost systematically processed after the underlying base events, so this subtle difference of implementation has virtually no impact on the final forecasts.

### d. Verification methodology

Verification of the modulated and unmodulated forecasts is conducted over the 21 years that the 1989–2010 period contains, using a cross-validation setting where the forecasts for every given year are postprocessed using the 20 remaining years as training. The quantitative evaluation is conducted for both precipitation and streamflow, using two verification metrics, the CRPS and the Brier score (BS). For ease of interpretation, these two scores are turned into skill scores, the CRPSS and the BSS, considering the climatological forecasts as reference. Forecasting schemes are compared two at a time, and the statistical significance of differences in skill is assessed using paired, stationary bootstrap (Politis and Romano 1994). In the charts, the 5th and 95th percentiles of the bootstrap distribution of the difference in skill score are plotted in addition to the nominal values, and whether or not this interval captures the zero line provides an indication of the statistical significance (at the 10% level) of that difference. Equations of the scores and details about the computation of the confidence intervals are given in the appendix.

Four *levels* of verification are conducted, by computing the CRPSS and BSS on forecasts of various quantities, with different objectives in mind:

- *Level 1*: CRPSS of the forecasts of precipitation at individual lead times and subbasins, with the objectives of (i) verifying that univariate calibration is effective, by comparing to the raw forecasts, and (ii) quantifying the impact of modulation on the univariate ensembles.
- *Level 2*: CRPSS of the forecasts of precipitation accumulated over 24 h and all subbasins of a given basin, with the objective of quantifying the effect of modulation on spatiotemporal rainfall totals, which are meaningful for hydrological modeling.
- *Level 3*: BSS of the forecasts of mean daily streamflow at individual lead times and basins, for two thresholds that correspond to the 90% and 99% quantiles of the simulated streamflow, with the objective of quantifying the benefits of modulation in a high-flow forecasting context.
- *Level 4*: CRPSS of forecasts of 3-day total streamflow at individual basins, with the objective of quantifying the benefits of modulation in a hydroelectricity or water supply forecasting context.

Skill scores are averaged over the 11 basins (except at level 1 where it is averaged over the 18 subbasins). At levels 3 and 4, the streamflow forecasts are verified against the simulated streamflow. All skill scores represent the forecast performance over the full year (except September, cf. section 2), as seasonal stratification (not shown) did not provide useful insights.

The reliability of the modulated and unmodulated forecasts is also assessed, using rank histograms. We here show rank histograms of univariate precipitation for base and modulation events, as well as 3-day total streamflow forecasts, for a few selected lead times.

## 4. Results and discussion

### a. Benefits of univariate calibration

Before turning to the other components of the postprocessing (reordering and modulation), we verify that the univariate calibration of the precipitation forecasts with the CSGD

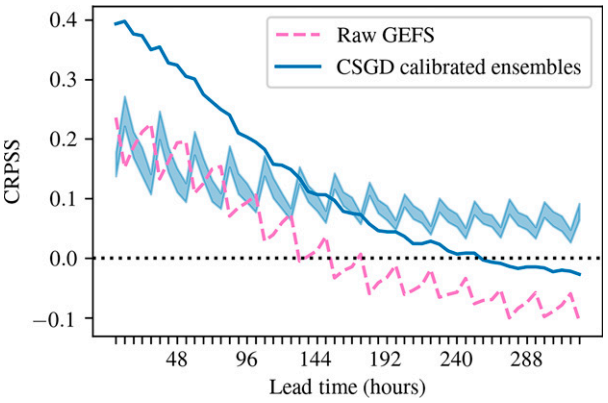


FIG. 4. Comparison of the raw GEFS forecasts vs the CSGD calibrated ensembles, at the level 1 of verification. The shaded area represents the uncertainty of the difference in skill score between the two forecasting schemes and must be interpreted in relation to the zero line, see text in section 3d and section d of the appendix.

technique is beneficial, by comparing the calibrated ensembles to the raw GEFS forecasts in a univariate framework (level 1), for the base events first. The results in Fig. 4 show that the forecast skill is, as expected, largely improved by calibration, and remains positive until approximately 10 days (240 h) in the forecast horizon, while the raw GEFS loses skill starting approximately 6 days (144 h). In terms of reliability, Figs. 5a–c show that a known shortcoming of the raw GEFS, the underdispersion of its members, is well corrected by calibration. Nonetheless, we observe in the calibrated ensembles a slight tendency to underestimation, increasing with lead time, which materializes with the very right bins of the histograms that remain overpopulated. This issue, which concerns the right tail of the distribution rather than the central tendency, is presumably due to a nonperfect fit, in our case study, of the CSGDs with the conditional distributions of the observation given the forecast statistics.

Finally, we look at the reliability of the raw versus calibrated forecasts for modulation events (Figs. 5d,e). Similar findings are observed, namely, a good calibration overall albeit a slight tendency to underestimation in the right tail. This demonstrates the versatility of the CSGD method for correcting most deficiencies in the raw ensemble at different temporal scales of aggregation.

### b. *Comparison of the Schaake shuffle versus ECC without modulation*

Figure 6 depicts the skill of the unmodulated forecasts reordered with either the Schaake shuffle or the ECC technique, at the level 2, 3, and 4 of verification (at level 1 the unmodulated forecasts are identical). When looking at the CRPSS of spatiotemporal precipitation aggregates (level 2), ECC outperforms the Schaake shuffle, with a gap that narrows as lead times increase and the raw forecasts lose skill. A gap is also visible on streamflow (mostly on 3-day total streamflow, i.e., level 4), although it takes a couple days to establish because of the strong autocorrelation of streamflow. Other studies (Bellier et al. 2017; Scheuerer et al. 2017) have also found the superiority of ECC over the Schaake shuffle on both precipitation and streamflow. It is important to remember though that the performance of ECC will depend on the spatiotemporal resolution of the NWP model with respect to the scale of the hydrological model. In our case study, the spatial resolution of the GEFS reforecasts is quite coarse with respect to the size of the basins, causing a lack of spatial coherence compared to the Schaake shuffle. However, ECC instills a more appropriate temporal dependence structure, which appears here to be sufficient to outperform the Schaake shuffle. Since upgrades in forecast systems generally come with finer spatial resolutions (as with the more recent GEFSv12 reforecasts), we hypothesize that the gap in skill between the Schaake shuffle and ECC in many case studies will increase in the future.

### c. *Effect of modulation on the univariate forecasts*

From now on we focus on the core of this paper, the effect of multi-temporal-scale modulation. To begin with, we look at how modulation impacts the univariate ensembles of precipitation

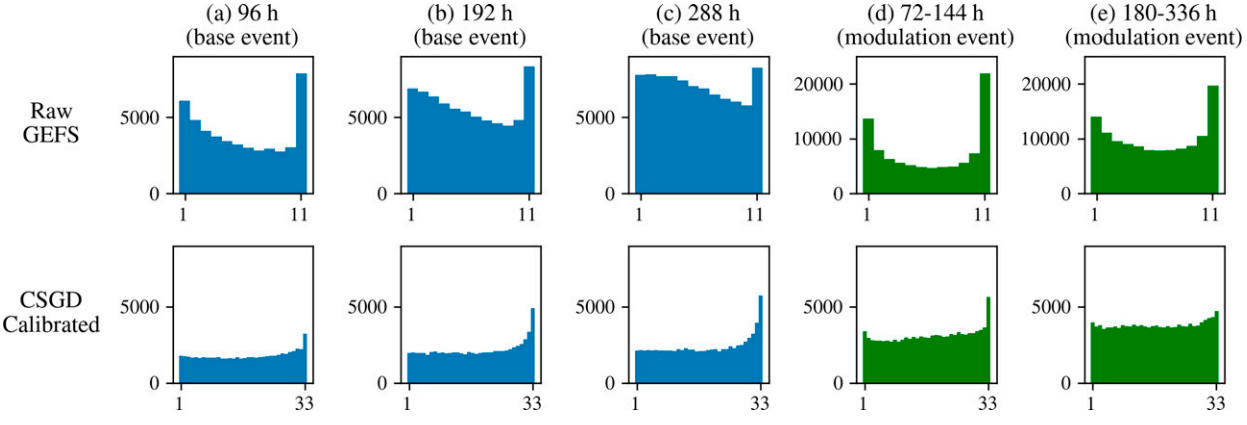


FIG. 5. Rank histograms of the (top) raw GEFS and (bottom) CSGD calibrated ensembles of precipitation, for all subbasins, and for (a)–(c) three selected base events and (e),(f) two selected modulation events. Note that the number of forecast cases on which the histograms are built may vary, as cases with all members plus the observation equaling zero are discarded (they do not carry any useful information about reliability).

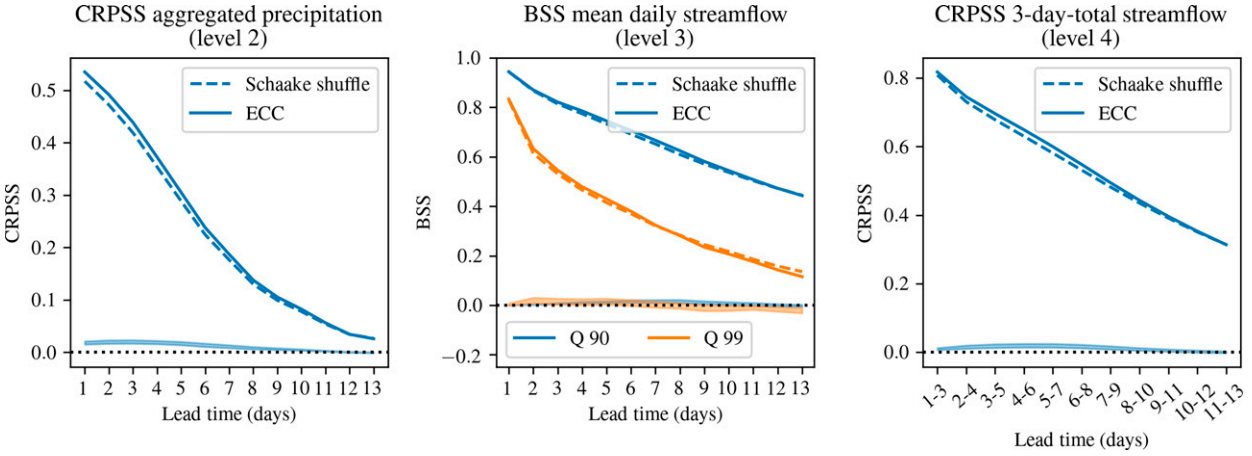


FIG. 6. Comparison of the reordering methods Schaake shuffle vs ECC *without* modulation, at the (left) level 2, (center) level 3, and (right) level 4 of verification. The shaded area represents the uncertainty of the difference in skill score between the two forecasting schemes, and must be interpreted in relation to the zero line, see text in section 3d and section d of the appendix.

(level 1). Figure 7 depicts the CRPSS of the modulated forecasts reordered with either the Schaake shuffle or ECC, compared to the unmodulated forecasts. It is seen that modulation deteriorates the univariate precipitation forecasts to a larger extent when it is coupled with the Schaake shuffle than with ECC. This finding can be explained as follows. The Schaake shuffle, unlike ECC, imposes on $\tilde{\mathbf{x}}$ a temporal structure that is not conditioned on the predicted weather situation. As a consequence, precipitation accumulations over multiple lead times in $\tilde{\mathbf{x}}$ are potentially quite different from the accumulation values that have been obtained via the calibration for the modulation events, values toward which they must be "pushed" via modulation [Eq. (8)]. Multiplicative factors will therefore take values further away from one, causing a more substantial modification of the original precipitation forecast values, and thus a loss of univariate CRPSS. This hypothesis can be verified on Fig. 8, which depicts the averaged multiplicative factors for each of the modulation events, and where it is indeed found that factors are further from one with the Schaake shuffle than with ECC.

One can also notice in Fig. 8 that with both reordering methods the multiplicative factors are greater than one, especially when lead times increase. By plotting in Fig. 9 the

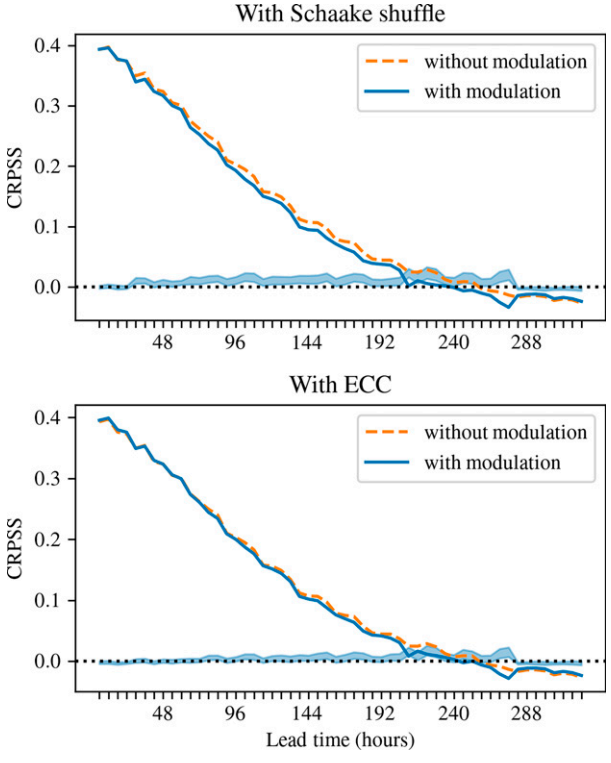


FIG. 7. Comparison of the unmodulated vs modulated forecasts, reordered with either (top) the Schaake shuffle or (bottom) the ECC, at the level 1 of verification. The shaded area represents the uncertainty of the difference in skill score between the two forecasting schemes, and must be interpreted in relation to the zero line, see text in section 3d and section d of the appendix.

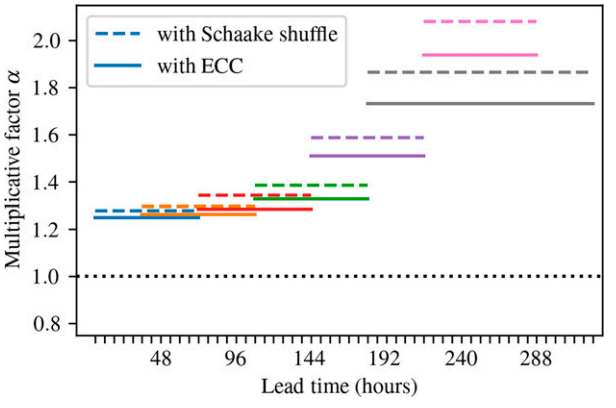


FIG. 8. Multiplicative factors averaged over all dates, subbasins, and members, for each of the modulation events (horizontal bars) and with the reordering methods Schaake shuffle and ECC. Colors are only used to facilitate distinguishing the events.

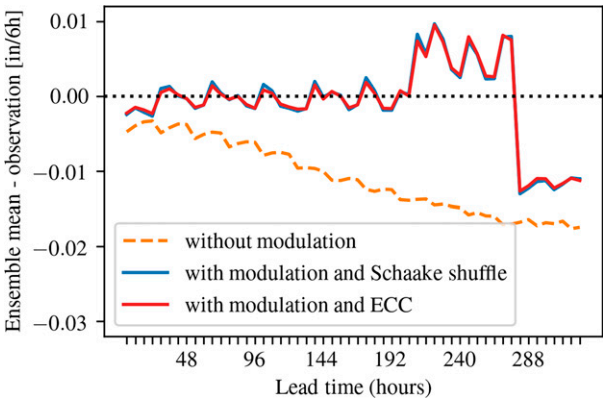


FIG. 9. Difference of the ensemble mean and the observed precipitation, averaged over the verification period and all subbasins, for the unmodulated forecasts and the modulated forecasts reordered with either the Schaake shuffle or the ECC.

averaged difference between the ensemble mean and observed precipitation (i.e., the bias), we observe that the calibrated ensemble mean tends to underestimate precipitation, but modulation mostly corrects for this bias, hence the multiplicative factors greater than one. However, modulation struggles to correct for the bias after day 9 (216 h), which happens to correspond to a modulation event boundary. This discontinuity is presumably due to a nonoptimal definition of the modulation events, as we have discarded for our study the events defined by CNRFC that extended beyond 14 days, but without refining the retained ones such that they "end" smoothly with the end of our forecast horizon. While this issue could have probably been reduced by an adjustment of the modulation events to our forecast setup, it illustrates an intrinsic limitation of the method, which is the discretization of the modulation events that does not allow for a seamless adjustment of the forecasts, as discussed in the conclusion.

Figure 10 finally depicts the rank histograms of the univariate ensembles without and with modulation (using either the Schaake shuffle or ECC), in order to assess the effect of modulation on the reliability of the univariate ensembles. Interestingly, we notice that, despite the correction of the ensemble central tendency as shown in Fig. 9 by the reduction in the bias, modulation has a limited effect on correcting the underestimation in the right tail of the distribution, an issue discussed in section 4a. This finding demonstrates that multiscale modulation is not a miracle add-on component that will compensate for any issue in the univariate calibration whatsoever. Rather, it suggests that the effects of modulation for streamflow forecasting, which will be quantified in the next subsection, should persist in the case where an alternative method to the CSGD would achieve a perfect calibration of the univariate ensembles. A further study that compares the relative gain of modulation with various univariate calibration methods would, however, be necessary to verify that hypothesis.

### d. *Benefits of modulation for streamflow forecasting*

After having studied the effect of modulation on univariate precipitation, we look at the benefits of modulation in the perspective of streamflow forecasting, by integrating the multivariate dependence structure into the evaluation (levels 2, 3, and 4 of verification). Figures 11 and 12 depict the skill scores of the unmodulated versus modulated forecasts when reordered with the Schaake shuffle and ECC, respectively. When coupled with the Schaake shuffle, modulation proves to be highly beneficial on the three levels of verification, although the gain is particularly noticeable for 3-day total streamflow.

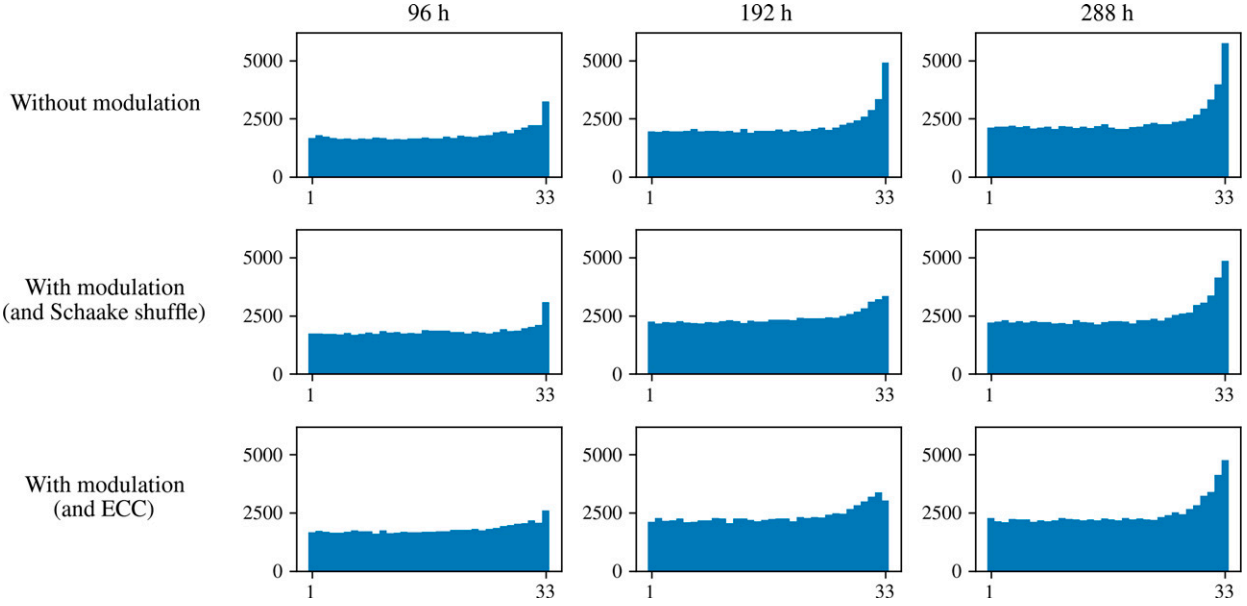


FIG. 10. Rank histograms of the univariate ensembles of precipitation, (top) unmodulated and modulated using either the (middle) Schaake shuffle or (bottom) ECC for reordering, for all subbasins, and three selected lead times.

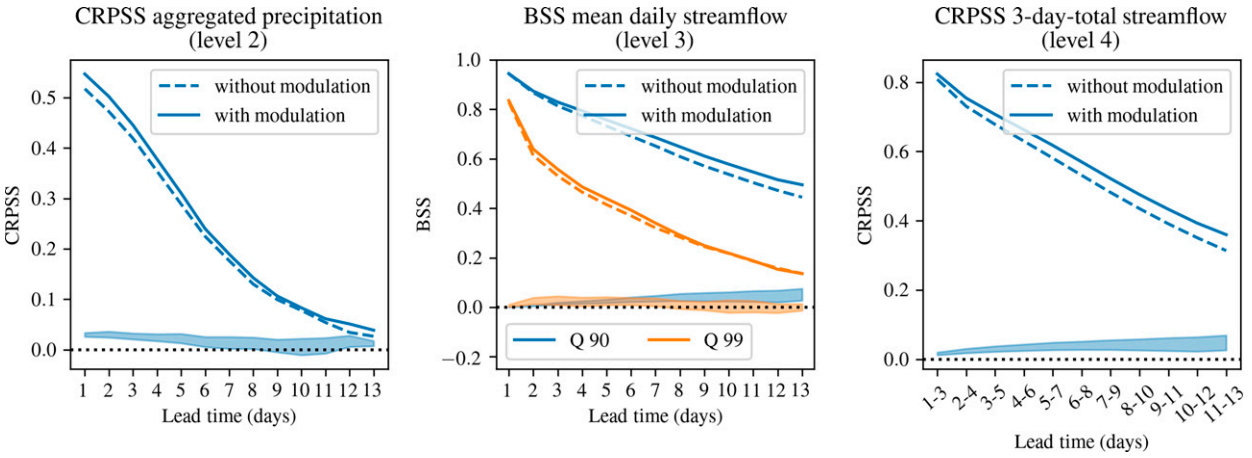


FIG. 11. Comparison of the unmodulated vs the modulated forecasts when reordered with the Schaake shuffle, at the (left) level 2, (center) level 3, and (right) level 4 of verification. The shaded area represents the uncertainty of the difference in skill score between the two forecasting schemes and must be interpreted in relation to the zero line; see text in section 3d and section d of the appendix.

One of the main shortcomings of the Schaake shuffle for streamflow forecasting, namely, to instill an unconditional temporal dependence structure in the forecast and thereby to misestimate the precipitation accumulations over multiple lead times, thus appears to be corrected, at least partially, by the modulation mechanism. With ECC (Fig. 12), the benefit of modulation is still visible, but the relative gain is smaller. As discussed in section 4b, in the setup of our study the main strength of ECC is a good appropriateness of the temporal structure, and therefore the method has less room for improvement by a modulation mechanism that concerns temporal scales only. The perspective of including spatial scales in the modulation mechanism will be discussed in the conclusion.

The rank histograms depicted in Fig. 13 show that modulation improves the reliability of 3-day total streamflow forecasts, when either of the two reordering methods is used. This improvement in reliability appears much greater than in the case of univariate precipitation (Fig. 10), and this again demonstrates that modulation acts primarily on multi-temporal-scale aggregates. Nonetheless, with both reordering methods the modulated 3-day total streamflow forecasts remain slightly underdispersive. The reason can presumably be found in the temporal dependence structure of the reordered precipitation forecasts, which in both cases slightly underestimates the temporal correlations as a result of the high proportion of zero precipitation values in the template **z**, as discussed in Bellier et al. (2017).

Finally, we compare in Fig. 14 the two reordering techniques when modulation is activated. Interestingly, the Schaake shuffle and the ECC now appear to perform almost equally well, while there was a clear gap when modulation was not used (cf. Fig. 6). However, if modulation allows here to close the gap between the two reordering methods, we cannot generalize to stating that all reordering methods will perform equally well after modulation is activated, as this is case-study specific. For instance, in a context where the spatial covariability of the precipitation processes across the basins is better resolved by the raw NWP forecast (with, e.g., a finer spatial resolution of the model), it is expected that ECC maintains an

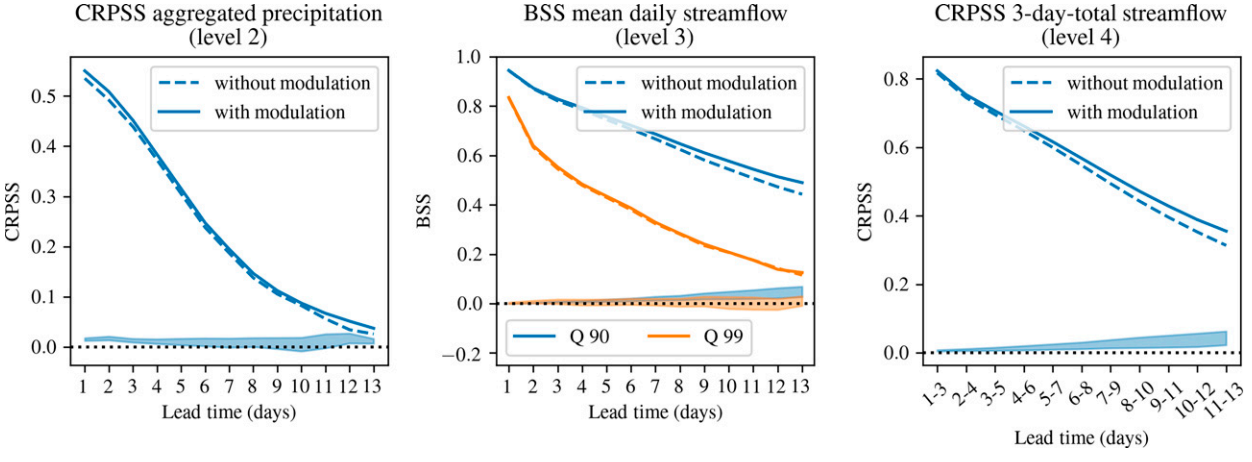


FIG. 12. As in Fig. 11, but when reordered with ECC.

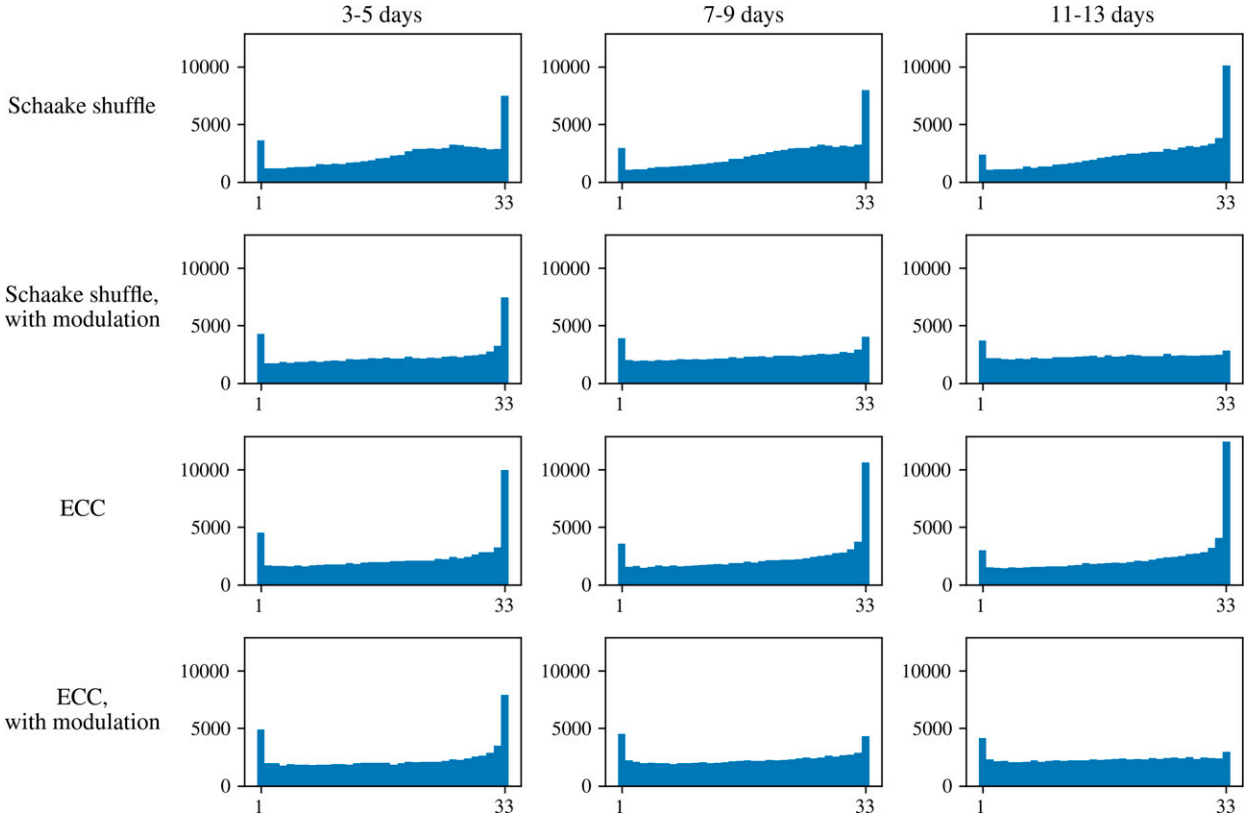


FIG. 13. Rank histograms of the 3-day total streamflow forecasts for all basins and three selected lead times (as 3-day windows) for (first row) the unmodulated and (second row) modulated Schaake shuffle and the (third row) unmodulated and (fourth row) modulated ECC.

advantage over the Schaake shuffle, even with modulation activated.

## 5. Summary and conclusions

This paper has focused on a simple multi-temporal-scale modulation mechanism that can be embedded in the postprocessing of precipitation ensemble forecasts, with the objective of improving the forecast skill for accumulations over multiple lead times, a highly relevant quantity for streamflow forecasting. This mechanism has been present for more than a decade in the MEFP, the meteorological postprocessing component of the HEFS, although it had never been the subject of a peer-reviewed publication, and users were left with

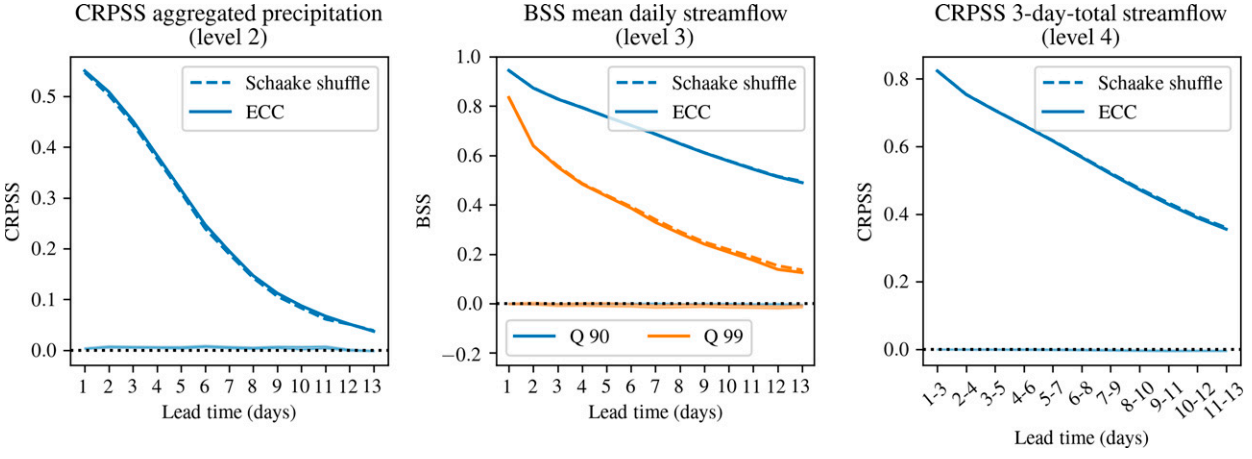


FIG. 14. Comparison of the reordering methods Schaake shuffle vs ECC, *with* modulation, at the (left) level 2, (center) level 3, and (right) level 4 of verification. The shaded area represents the uncertainty of the difference in skill score between the two forecasting schemes, and must be interpreted in relation to the zero line; see text in section 3d and section d of the appendix.

material in the gray literature. In this paper, we have formally described the method, not as an inner part of the MEFP but as a distinct component of precipitation postprocessing that can be coupled with any univariate calibration and multivariate reordering technique. We hope this can help users outside of the HEFS community to replicate the method or implement a similar mechanism.

Verification was conducted on 11 Californian basins, with modulation coupled with the CSGD method for univariate calibration, and with either the Schaake shuffle or the ECC technique for reordering, with the objective of studying the interactions between modulation and reordering. Benefits were quantified using four levels of verification, which not only focused on the precipitation forecasts but also on the resulting streamflow forecasts. Overall, modulation was found to greatly improve the streamflow forecasts, although with a relative gain that depends on the method used for reordering. Larger benefits were found when the Schaake shuffle is used, compared to when ECC is used. This was explained by the difficulty of the Schaake shuffle to correctly predict precipitation accumulations over multiple lead times, because of an unconditional temporal structure. The flow-dependent ECC technique had less room for improvement on that aspect, hence the smaller relative gain. From the results of our study, we can therefore infer that modulation is expected to be most beneficial in setups where (i) the dependence structure of the predictand is situation specific (i.e., it varies from day to day) and (ii) the method that is available for reordering is not able to effectively condition the dependence structure upon that forecast situation. Furthermore, the fact that the gains are noticeable on temporal aggregates but are negligible at the univariate level lets us hypothesize that modulation remains beneficial with most if not all univariate precipitation calibration techniques, as long as the conditions i and ii above are fulfilled. More work is needed to extend the analysis to other calibration methods, and also to a more diverse set of basins, in particular in other parts of the country where heavy precipitation is dominated by shorter time scales (e.g., convective precipitation).

While the mechanism described here concerns temporal scales only, there appears to be no technical constraints to an adaptation to spatial scales as well. What was here defined as "modulation events," the aggregation periods that comprise multiple lead times, could be extended to spatial aggregations over multiple subbasins, with a similar sequential modulation process that relies on the sorting of the events. However, keeping the same framework will still involve the arbitrary process of defining the events, as well as selecting a metric for their sorting. While this multiscale mechanism is simple yet skillful, we believe there is room for improvement on less heuristic approaches that capture in a seamless way the multiscale dependencies of forecasts, for instance using Fourier or wavelet transforms.

*Acknowledgments.* This research was supported by NOAA Physical Sciences Laboratory base funding and by the NOAA National Weather Service Office of Science and Technology Integration. We also thank the two anonymous reviewers who greatly helped improve the quality of this paper.

*Data availability statement.* Observed data of mean areal precipitation and temperature for the 11 Californian basins, as well as the simulated streamflow, are available on request (brett.whitin@noaa.gov). Because of the release of a new version of reforecasts, the second-generation GEFS reforecasts used in this study are no longer publicly available for download, but the chunk of data that has been used in this study can be supplied on request (psl.data@noaa.gov).

## APPENDIX

### Verification Scores

#### a. Continuous ranked probability score

Consider, for a given predictand, a verification sample containing $N$ pairs of forecast/observation. For each forecast case $n \in \{1, \ldots, N\}$, let $F_n$ be the univariate forecast distribution and $y_n$ the verifying observation. The continuous ranked probability score (CRPS) is defined as

$$\mathrm{CRPS}_n(F_n, y_n) = \int_{-\infty}^{+\infty} [F_n(u) - H(u - y_n)]^2 du, \tag{A1}$$

where $H$ is the Heaviside step function such that $H(u - y_n) = 1$ if $u \geq y_n$ and 0 otherwise. When forecasts are in the form of ensembles, Eq. (A1) is discretized for computation, using the empirical distribution function form of $F_n$ (Hersbach 2000). The average CRPS over all forecast cases is then

$$\overline{\mathrm{CRPS}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{CRPS}_n. \tag{A2}$$

#### b. Brier score

Consider a binary event that corresponds to the exceedance of a specific threshold, and for each forecast date $n \in \{1, \ldots, N\}$ let $p_n$ be the predicted probability assigned to this event (determined from $F_n$), while $o_n$ is the observed outcome (0 or 1). The Brier score (BS) is defined as

$$\mathrm{BS} = \frac{1}{N} \sum_{n=1}^{N} (p_n - o_n)^2. \tag{A3}$$

In this study, we compute the BS for two different thresholds that correspond to the 90% and 99% quantiles of the simulated streamflow.

#### c. Skill scores

For ease of interpretation, the average CRPS and the Brier scores are turned into skilled scores, noted CRPSS and BSS, respectively, and defined as

$$\mathrm{CRPSS} = 1 - \frac{\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{\mathrm{ref}}}, \tag{A4}$$

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{A5}$$

where $\overline{\text{CRPS}}_{\text{ref}}$ and $\text{BS}_{\text{ref}}$ are the average CRPS and the Brier score of a reference forecast dataset, respectively. A skill score of 1 corresponds to perfect forecasts; a skill score of 0 indicates that forecasts perform only as good as the reference forecasts, and negative values indicate forecast performances lower than the reference ones. In this study, the climatological forecasts systematically serve as reference. These are 50-quantile ensembles drawn from the 21-yr observation record, and specific to each subbasin, time of the day (6-h period) and day of the year (using a 30-day moving window).

In this paper, skill scores CRPSS and BSS are averaged over the 18 subbasins when the predictand is precipitation at individual lead times and subbasins (level 1), or over the 11 basins when it is space–time aggregate of precipitation (level 2) or streamflow (levels 3 and 4).

### *d. Confidence intervals*

Bootstrapping is used to quantify the uncertainty in the skill score computation that arises from the finite length of the verification sample. We here describe the method for the CRPSS, but the same applies for the BSS. Because we are interested in this paper in comparing competing forecast models (or schemes) two at a time (say $\mathcal{M}1$ versus $\mathcal{M}2$), and because the performance of both models on a given day are related to the meteorological situation on that day, the verification sample is bootstrapped in a *paired* setting, with a common resampling structure between the two models. Let $b = 1, \ldots, B$ denote the bootstrap replications, with $B = 1000$ in this study, and for any $b$ let $s_b$ be the set containing $N$ dates randomly selected from the full verification period $(1, \ldots, N)$, as per the chosen bootstrap method (see next paragraph). For each bootstrap replication $b$, we first compute $\overline{\text{CRPS}}^*_{\mathcal{M}1,b}$, $\overline{\text{CRPS}}^*_{\mathcal{M}2,b}$, and $\overline{\text{CRPS}}^*_{\text{ref},b}$, the CRPS of $\mathcal{M}1$, $\mathcal{M}2$, and of the climatology, respectively, averaged over the dates in $s_b$. Then, we calculate the skill scores $\text{CRPSS}^*_{\mathcal{M}1,b}$ and $\text{CRPSS}^*_{\mathcal{M}2,b}$, and average them over the locations (subbasins or basins, depending on the predictand at hand). Finally, we compute the difference in skill between $\mathcal{M}1$ and $\mathcal{M}2$:

$$\Delta\text{CRPSS}^*_b = \text{CRPSS}^*_{\mathcal{M}1,b} - \text{CRPSS}^*_{\mathcal{M}2,b}. \tag{A6}$$

Repeating this process for $b = (1, \ldots, B)$ we obtain $(\Delta\text{CRPSS}^*_1, \ldots, \Delta\text{CRPSS}^*_B)$, a bootstrap distribution of the difference in skill between $\mathcal{M}1$ and $\mathcal{M}2$, from which we can compute the 5th and 95th percentiles. In the charts (Figs. 4, 6, 7, 11, 12, and 14), the interval in between these two quantities is plotted as shaded area, and must be interpreted in relation to the zero line. If it excludes zero, the difference in skill score between the two competing model $\mathcal{M}1$ and $\mathcal{M}2$ can be considered as statistically significant.

Because the data to resample exhibit temporal correlation, we here used the stationary bootstrap (Politis and Romano 1994) which, rather than randomly selecting individual dates to form $s_b$, picks blocks of consecutive dates and concatenate them to form pseudo time series of size $N$. The stationary bootstrap differs from other block bootstrap methods in that the lengths of the blocks are random samples from a geometric distribution with prescribed mean $\overline{L}$, while the positions are random samples from a uniform distribution over the timeline. This has desirable properties in terms of stationarity of the pseudo time series (Politis and Romano 1994). In case of spatial correlation, the concept of blocks must also be extended to the space domain. Here, given the relative proximity between the locations, we assume a perfect spatial dependence, which is equivalent to considering a single spatial block. In other words, the selected (blocks of) dates are, in every bootstrap replication $b$, identical for all locations. Finally, to define the average block size $\overline{L}$ we use the formulation proposed Politis and White (2004) (and corrected in Patton et al. 2009), which determines the optimal value based on the correlogram of the data to bootstrap. In our case though, the temporal time series to bootstrap is multivariate (two models $\mathcal{M}1$ and $\mathcal{M}2$, plus multiple locations), while a unique $\overline{L}$ is required. We therefore apply Politis and White's (2004) formulation to each univariate time series and take $\overline{L}$ as the average of all values.

## REFERENCES


Alizadeh, B., R. A. Limon, D.-J. Seo, H. Lee, and J. Brown, 2020: Multiscale postprocessor for ensemble streamflow prediction for short to long ranges. *J. Hydrometeor.*, **21**, 265–285, https://doi.org/10.1175/JHM-D-19-0164.1.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Bellier, J., G. Bontron, and I. Zin, 2017: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resour. Res.*, **53**, 10 085–10 107, https://doi.org/10.1002/2017WR021245.

Ben Bouallègue, Z., T. Heppelmann, S. E. Theis, and P. Pinson, 2016: Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Mon. Wea. Rev.*, **144**, 4737–4750, https://doi.org/10.1175/MWR-D-15-0403.1.

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

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Gneiting, T., and E.-M. Walz, 2022: Receiver operating characteristic (ROC) movies, universal roc (UROC) curves, and coefficient

of predictive ability (CPA). *Mach. Learn.*, **111**, 2769–2797, https://doi.org/10.1007/s10994-021-06114-3.

——, A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Kim, S., and Coauthors, 2018: Assessing the skill of medium-range ensemble precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) for the upper trinity river basin in North Texas. *J. Hydrometeor.*, **19**, 1467–1483, https://doi.org/10.1175/JHM-D-18-0027.1.

Patton, A., D. N. Politis, and H. White, 2009: Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Rev.*, **28**, 372–375, https://doi.org/10.1080/07474930802459016.

Politis, D. N., and J. P. Romano, 1994: The stationary bootstrap. *J. Amer. Stat. Assoc.*, **89**, 1303–1313, https://doi.org/10.1080/01621459.1994.10476870.

——, and H. White, 2004: Automatic block-length selection for the dependent bootstrap. *Econometric Rev.*, **23**, 53–70, https://doi.org/10.1081/ETC-120028836.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci.*, **4**, 655–717, https://doi.org/10.5194/hessd-4-655-2007.

Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, https://doi.org/10.1175/MWR-D-15-0227.1.

——, T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, https://doi.org/10.1214/13-STS443.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

——, and ——, 2018: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *J. Hydrometeor.*, **19**, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1.

——, ——, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Troin, M., R. Arsenault, A. W. Wood, F. Brissette, and J.-L. Martel, 2021: Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years. *Water Resour. Res.*, **57**, e2020WR028392, https://doi.org/10.1029/2020WR028392.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wu, L., D.-J. Seo, J. Demargne, J. D. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, https://doi.org/10.1016/j.jhydrol.2011.01.013.

Zsoter, E., R. Buizza, and D. Richardson, 2009: "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Mon. Wea. Rev.*, **137**, 3823–3836, https://doi.org/10.1175/2009MWR2960.1.