AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR020133

Key Points:
- Statistical calibration of GEFS precipitation and temperature ensemble reforecasts
- Reconstruction of spatial, temporal, and intervariable dependence
- Improved meteorological forcings of a hydrological streamflow forecast model

Supporting Information:
- Supporting Information S1

Correspondence to:
M. Scheuerer,
michael.scheuerer@noaa.gov

Citation:
Scheuerer, M., T. M. Hamill, 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, doi:10.1002/2016WR020133.

Received 15 NOV 2016
Accepted 6 MAR 2017
Accepted article online 16 MAR 2017
Published online 13 APR 2017

A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation

Michael Scheuerer[1,2], Thomas M. Hamill[2], Brett Whitin[3], Minxue He[4], and Arthur Henkel[3]

[1]University of Colorado, Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado, USA, [2]NOAA/ESRL, Physical Sciences Division, Boulder, Colorado, USA, [3]NOAA/NWS, California Nevada River Forecast Center, Sacramento, California, USA, [4]California Department of Water Resources, Hydrology Branch, Sacramento, California, USA

Abstract Hydrological forecasts strongly rely on predictions of precipitation amounts and temperature as meteorological forcings for hydrological models. Ensemble weather predictions provide a number of different scenarios that reflect the uncertainty about these meteorological inputs, but these are often biased and under-dispersive, and therefore require statistical postprocessing. In addition to correcting the marginal distributions of the two weather variables, postprocessing methods must reconstruct their spatial, temporal, and intervariable dependence in order to generate physically realistic forecast trajectories that can be used as forcings of hydrological streamflow forecast models. For many years, a sample reordering method referred to as "Schaake shuffle" has been used successfully to address this multivariate aspect of forecast distributions by using historical observation trajectories as multivariate "dependence templates." This paper proposes a variant of the Schaake shuffle, in which the historical dates are selected such that the marginal distributions of the corresponding observation trajectories are similar to the forecast marginal distributions, thus making it more likely that spatial and temporal gradients are preserved during the reordering procedure. This new approach is demonstrated with temperature and precipitation forecasts over four river basins in California, and it is shown to improve upon the standard Schaake shuffle both with respect to verification metrics applied to the forcings, and verification metrics applied to the resulting streamflow predictions.

1. Introduction

Hydrological forecasts are valuable for a range of applications such as flood control, water supply, or environmental stream flow regulation. A probabilistic forecasting framework in which the associated uncertainty is represented and communicated can improve decision-making [*Roulin*, 2007; *Verkade and Werner*, 2011], and requires that both uncertainty about initial conditions and parameters of the hydrological model, and uncertainty about the meteorological inputs are quantified correctly. The latter is typically addressed by meteorological ensemble prediction systems (EPSs), which aim to approximate the probability distribution of the forcing data by a finite number of scenarios (an overview over ensemble use in flood forecasting is given by *Cloke and Pappenberger* [2009]). However, despite constant improvements to those meteorological EPSs over the last two decades, raw ensemble forecasts of surface weather variables are still unreliable [e.g., *Park et al.*, 2008], and statistical postprocessing is required to remove forecast biases and ensure adequate representation of forecast uncertainty. Since hydrological models are sensitive to the space-time covariability of the forcing forecasts, statistical postprocessing (throughout this article the term 'postprocessing' is used to denote the statistical adjustment of ensemble weather predictions based on prior discrepancies between forecasts and observations; in the lingo of hydrologists this process is called 'preprocessing') needs to model the joint distribution of meteorological inputs at all locations over the river basin of interest, all forecast lead times, and all relevant weather variables. In other words, statistical postprocessing should generate physically realistic, unbiased spatiotemporal forecast fields (we refer to such a spatiotemporal field as "trajectory") which adequately represent the forecast uncertainty.

Typically, this is done in two steps. First, a statistical model which yields reliable predictive marginal distribution for all weather variables of interest, separately for each location and forecast lead time is set up.

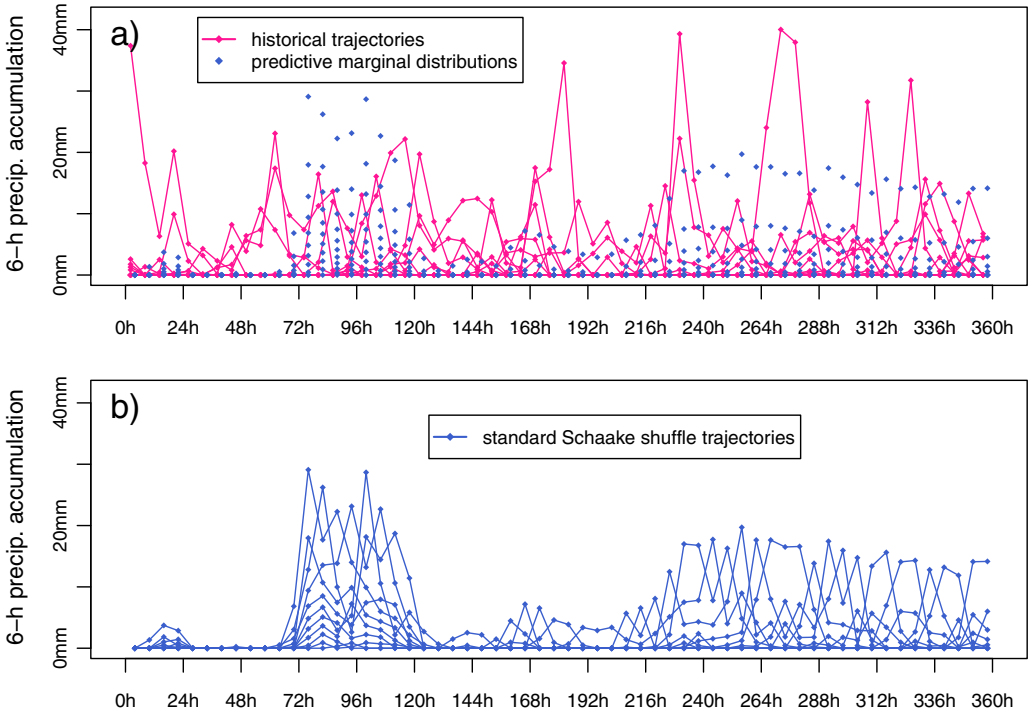

Figure 1. Illustration of the Schaake shuffle technique for constructing forecast trajectories of 6 h mean areal precipitation (MAP) amounts at Lake Mendocino subbasin. Predictive marginal distributions for each forecast lead time (initialization time: 9 January 2010, 00 UTC) are represented by quantile-based samples of size 11 (blue diamonds). (a) Historical observation trajectories from the same date in the previous 11 years. (b) The forecast trajectories obtained by reordering the marginal forecast samples at each lead time in the same way as the historical observation trajectories.

Subsequently, spatial, temporal, and intervariable dependence is addressed, with the predominant method in hydrological applications being the "Schaake shuffle" [*Clark et al*., 2004], which uses historical, observed trajectories as a multivariate "dependence template" onto which the univariate samples from the postprocessed predictive distributions are assembled in a way that duplicates the pairwise rank correlation structure in the historical ensemble [*Wilks*, 2015]. The Schaake shuffle technique can be used independently of how the marginal forecast distributions were obtained, and it has been applied successfully in numerous studies (and various time scales) in order to construct spatiotemporal forecast fields that adequately represent the multivariate forecast distribution [*Schaake et al*., 2007; *Wu et al*., 2011; *Verkade et al*., 2013; *Robertson et al*., 2013; *Vrac and Friederichs*, 2015]. Despite its success, the Schaake shuffle approach comes with a caveat. *Clark et al*. [2004] themselves note that "it assumes stationarity in the spatiotemporal correlation structure" and it "will not preserve the spatial gradients in precipitation and temperature fields for individual forecasts." The stationarity assumption implies in particular that spatiotemporal rank correlations are state-independent, i.e., for example, they are the same for low and high levels of precipitation. It can happen that historical observation trajectories associated with relatively low precipitation amounts are remapped to Schaake shuffle ensemble values with much higher values and vice versa, as illustrated in Figure 1. The postprocessed, marginal forecast distributions suggest very dry conditions during the first 66 h after forecast initialization, followed by a period of elevated levels of precipitation. The Schaake shuffle imposes the rank order of historical observations, which have values representative of climatology, on near-zero precipitation forecasts during the first 66 h, and on forecast values well above climatology during the 72–120 h forecast lead time period. It is not clear if the resulting forecast trajectories are still physically realistic, despite their foundation in observed trajectories. The idea of *Clark et al*. [2004] to address this by "preferentially select(ing) dates from the historical record that resemble forecasted atmospheric conditions and use the spatial correlation structure from this subset of dates to reconstruct the spatial variability for a specific forecast" has recently been implemented by *Schefzik* [2016], who defines a similarity-criterion for the raw ensemble forecasts and applies the Schaake shuffle technique to the subset of historical dates where the ensemble forecasts resemble those for the desired forecast date. His data example with surface

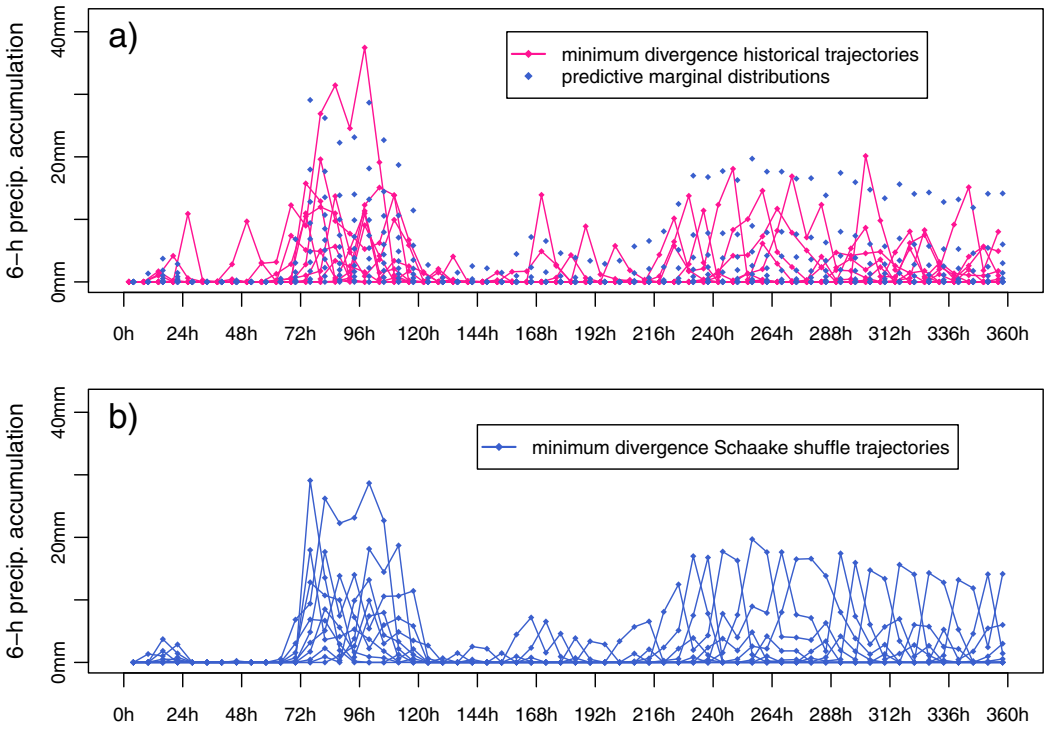

Figure 2. Same as Figure 1 but with historical dates selected as suggested in section 4.

temperature forecasts at several locations in Central Europe shows that this similarity-based implementation of the Schaake shuffle can improve the representation of the multivariate forecast distributions compared to the standard Schaake shuffle. The similarity-criterion suggested by *Schefzik* [2016] does not account for temporal correlation, but a corresponding adjustment is straightforward. In the typical setting encountered in hydrological applications, however, Schefzik's variant of the Schaake shuffle has two drawbacks:

1. Since the similarity criterion is based on the ensemble forecasts, it can only be defined for historical dates for which reforecasts are available. Even if reforecast datasets are available, they might not go back as far as the observation record. Or, they may be configured such as to skip dates, and those dates can then not be used in the Schaake shuffle.
2. Forecast skill typically decreases with increasing lead time, and forecast fields at longer lead times often have little resemblance with the observed fields, so that preferential selection of dates becomes less and less meaningful. If the similarity-criterion is extended to also address temporal correlation, longer lead times should be deemphasized, requiring some ad hoc weighting strategy to do so.

In this paper, we propose an alternative strategy for preferential selection of dates in the Schaake shuffle approach that can leverage the full observation record independent of reforecast availability. Our method chooses the historical dates in such a way that the marginal distributions of the observation trajectories resemble those of the postprocessed forecast distributions as illustrated in Figure 2. The 6 h precipitation amounts corresponding to the historical observations are now similar to the postprocessed, predictive samples, and imposing the rank order of the historical observations is therefore more likely to preserve spatial and temporal gradients, and yield physically realistic forecast trajectories. Since our approach chooses the historical dates based on the full, postprocessed predictive distributions, it automatically emphasizes locations and lead times where forecasts are skillful and forecast distributions differ significantly from the climatological distribution.

The forecast and observation data used in this study is presented in section 2, while sections 3 and 4 provide the details on the methods used for marginal calibration on the one hand and multivariate dependence modeling on the other hand. Section 5 compares the standard Schaake shuffle approach, the variant of it proposed in this paper, and the ensemble copula coupling technique [*Schefzik et al.*, 2013], by verifying

the forcing variables generated with both techniques. Subsequently, in section 6, results are presented which evaluate in how far the improvement of these forcings entails an improvement of predicted streamflow. Section 7 finally summarizes the results and discusses applications and limitations of the approach presented here.

2. Study Basins and Data Sets

The meteorological forcings considered here are 6 h surface temperature means and 6 h precipitation accumulations during the period from January 1985 to September 2010. Forecast data were obtained from the second-generation Global Ensemble Forecast System (GEFS) reforecast data set [*Hamill et al*., 2013], which consists of 11 ensemble member forecasts on a Gaussian grid at $\sim 1/2^{\circ}$ resolution. All forecasts were initialized at 00 UTC, and forecast lead times up to 15 days were considered for this study. Those forecasts are postprocessed with and verified against station-based observation data over four different basins in California: Russian River (7 subbasins), Eel River (11 subbasins), American River (2 subbasins), and Merced River (5 subbasins). While the four basins were processed separately, the mean areal temperature (MAT) and mean areal precipitation (MAP) values associated with the respective subbasins and the 15×4 forecast accumulation periods constitute the multivariate quantity for which a forecast distribution is sought. MAT and MAP observations are available for the period from October 1948 to September 2010, and while only the subperiod overlapping with the period of the reforecast data set can be used for model fitting and verification, the full observation record is used to create the Schaake shuffle ranks. Unlike the forecasts, observations are recorded in local time (PST/PDT) which lags behind UTC by 7 h during the warm season and 8 h during the cool season. The resulting challenge that the respective 6 h accumulation periods of forecasts and observation never fully overlap is dealt with by our marginal statistical postprocessing methods as discussed in the subsequent section. Finally, daily United States Geological Survey (USGS) archived streamflow data were obtained for Ukiah (Russian River Basin). Daily full natural flow data at Coyote Dam (Lake Mendocino, Russian River Basin) were derived from hourly inflow data obtained from the US Army Corps of Engineers, Sacramento District (USACE-SPK) and daily USGS Eel River diverted flows.

Figure 3 shows the locations of the four basins, and gives an idea of the topography in California. The American River and Merced River basins are located in the Sierra Nevada with area-averaged altitudes of 1250 m and 2800 m, respectively. Streamflow in the American River is about two-thirds wintertime rainfall and snowmelt runoff and less than one-third springtime snowmelt runoff, while the higher and cooler Merced River is dominated by springtime snowmelt runoff [*Dettinger et al*., 2014]. The Russian River and Eel River basins are coastal basins, and snowmelt runoff generally plays a less important role. In order to give an idea of the spatial variability of MAT and MAP within each of the four basins, Figure 3 further depicts the mean absolute difference (MD, a measure of dispersion) statistics of these quantities across the respective subbasins. Two different ways of calculating/aggregating these statistics are considered:

1. In order to measure the climatological, spatial variability in each basin, we calculate the MD of average MAT/MAP over all days of the respective month and all years from 1985 to 2010.
2. In order to measure the spatial variability on the 6 h time scale, we calculate the MD of 6 h MAT/MAP (for MAT, anomalies from the climatological average are considered) and average those 6 h MDs over all days of the respective month and all years from 1985 to 2010.

All MD statistics are calculated separately for each of the four 6 h periods of the day, but the plots in Figure 3 depict average statistics over those four 6 h periods. The American River and Merced River basins have a rather complex topography; the climatological spatial variability of the corresponding MATs (which is largely due to differences in elevation) is large compared to the spatial variability of the 6 h anomalies from the climatological average. The spatial variability of climatological MAT averages across the Russian River subbasins, on the contrary, is lower than the variability of 6 h anomalies, which entails that spatial MAT gradients are relatively more dependent on the particular atmospheric situation. The MAP mean absolute differences reflect the annual cycle of precipitation over California, which is characterized by very dry summers and precipitation mainly occurring during the cool season. The two Sierra basins, especially Merced River, are relatively small, and spatial differences of MAP between the different subbasins are moderate compared to those observed in the coastal basins. These basin characteristics may affect the effectiveness of

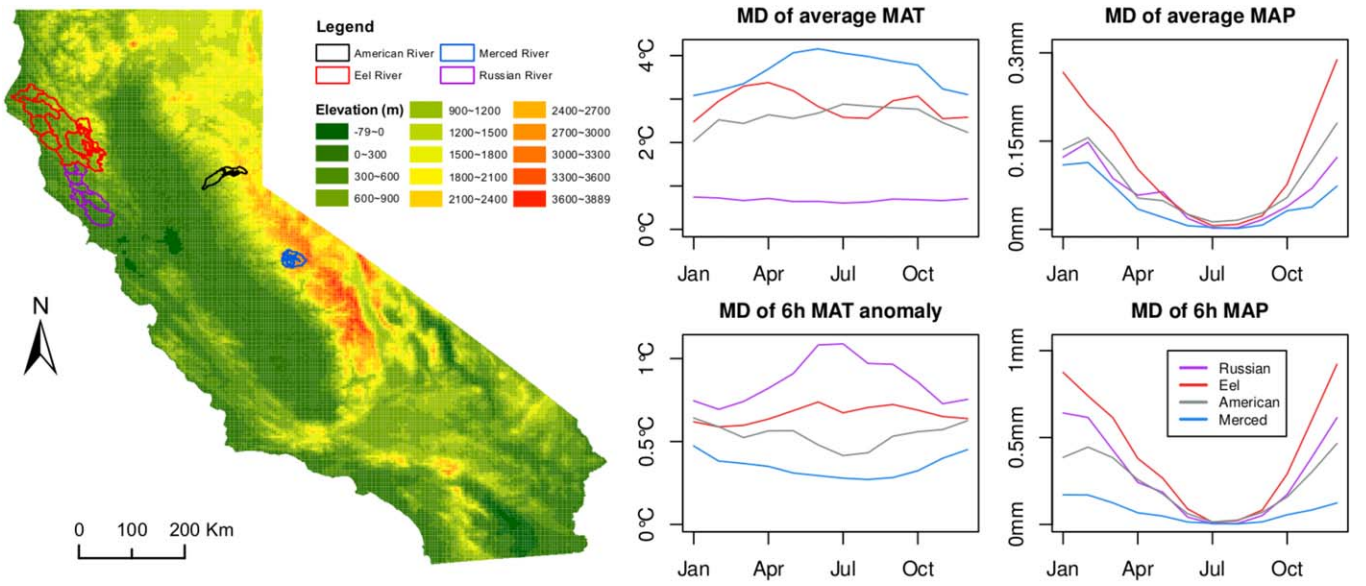

Figure 3. Location map showing the four study basins, and mean absolute difference (MD) statistics illustrating the spatial variability of MAT and MAP across the respective subbasins. The top row depicts the MD of climatological averages, the bottom row depicts the MD on the 6 h time scale.

the methods for modeling joint forecast distributions of MAT and MAP presented in the subsequent sections.

3. Methods for Marginal Calibration

3.1. Mean Areal Temperature (MAT)

For postprocessing the ensemble temperature forecasts, we use a variant of the nonhomogeneous Gaussian regression (NGR) [*Gneiting et al.*, 2005] approach proposed by *Scheuerer and Büermann* [2014], relating observed temperature anomalies to ensemble mean forecast anomalies, and using the ensemble variance as a predictor for the forecast uncertainty. Specifically, if y denotes the MAT to be predicted, $\bar{f}$ denotes the corresponding ensemble mean forecast, y_{cl} and $\bar{f}_{cl}$ denote the respective climatological MATs at the forecast time and date, and s^2 denotes the ensemble variance, then a predictive Gaussian distribution for y is defined via

$$y|\bar{f}, s^2 \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu = y_{cl} + a \cdot (\bar{f} - \bar{f}_{cl}), \quad \sigma^2 = b_0 + b_1 s^2. \tag{1}$$

The climatological MATs y_{cl} and $\bar{f}_{cl}$ are obtained as described by *Hemri et al.* [2014] by fitting a regression model with two harmonic terms (to represent the annual cycle) to the training forecasts and observations. The regression coefficients a, b_0, b_1 are obtained via CRPS minimization following *Gneiting et al.* [2005]. For temperature, we deal with the incomplete overlap of forecast and observation periods by calculating, for each 6 h observation period, a weighted average of both overlapping 6 h forecasts periods with weights proportional to the time overlap (4:2 for PST, 5:1 for PDT). Compared to an exact overlap this smoothes the diurnal cycle of forecast temperatures, but using different climatologies for forecasts and observations in equation (1) restores the correct amplitude of the diurnal cycle for y. The climatological MATs $y_{cl}, \bar{f}_{cl}$ are specific to each location, each time (6 h period) of the day, and each day of the year. The regression coefficients a, b_0, b_1 are estimated separately for each location, each lead time period, and based on training data that is composed specific to each month as described in section 5.

3.2. Mean Areal Precipitation Accumulations (MAP)

For postprocessing the ensemble precipitation forecasts, we largely follow the approach described by *Scheuerer and Hamill* [2015a] of fitting a nonhomogeneous, nonlinear regression model to training observations and statistics of the ensemble forecasts, using censored, shifted Gamma distributions (CSGDs). Starting from a CSGD representing the climatological distribution with mean, standard deviation, and shift

parameters μ_{cl}, σ_{cl}, and δ_{cl}, predictive CSGDs are obtained by fixing $\delta=\delta_{cl}$, and relating the predictive mean and standard deviation parameters μ and σ to the ensemble probability of precipitation POP_f, the ensemble mean $\bar{f}$, and the ensemble mean difference MD_f via

$$\mu=\frac{\mu_{cl}}{a_1}\,\mathrm{log1p}\left[\mathrm{expm1}(a_1)\left(a_2+a_3\mathrm{POP}_f+a_4\frac{\bar{f}}{\bar{f}_{cl}}\right)\right] \tag{2}$$

$$\sigma=b_1\sigma_{cl}\sqrt{\frac{\mu}{\mu_{cl}}}+b_2\mathrm{MD}_f, \tag{3}$$

where $\mathrm{log1p}(x)=\log(1+x)$, $\mathrm{expm1}(x)=\exp(x)-1$, and $\bar{f}_{cl}$ is the climatological mean of $\bar{f}$. For details about model fitting and a motivation of these equations we refer to *Scheuerer and Hamill* [2015a].

The method used here differs from *Scheuerer and Hamill* [2015a]—besides the slightly simpler equations for μ and σ—in the way we calculate the ensemble statistics $\mathrm{POP}_f, \bar{f}$, and MD_f. *Scheuerer and Hamill* [2015a] demonstrate that it is beneficial to augment the GEFS ensemble by adding forecasts from grid points within a certain neighborhood of the location s of interest, thus accounting for displacement errors. The optimal radius r of these neighborhoods increases with lead time [*Scheuerer and Hamill*, 2015a, their Figure 14], and trading off forecast performance and computational efficiency we choose $r=1^\circ\cdot\sqrt{1+t_e/24h}$, where t_e is the end of the lead time window in hours. In the present setup, we use a similar idea to address timing errors by interpreting the two 6 h forecasts periods overlapping each 6 h observation period as a "temporal neighborhood," and augmenting the ensemble by considering forecasts from both overlapping forecasts periods as predictors. Following *Scheuerer and Hamill* [2015a], we use an ad hoc spatial weighting scheme which deemphasizes forecast grid points further away from the observation location s, and combine it with the temporal weighting scheme (4:2 for PST, 5:1 for PDT) for the two overlapping forecast periods suggested above for MAT. As an alternative option, we study a data-driven weighting scheme:

1. let $\tilde{f}_{xtk}$ be the forecast of the kth member at grid point x and lead time t, adjusted via quantile mapping as described in *Scheuerer and Hamill* [2015a], section 4a,
2. compute the corresponding ensemble mean $\bar{f}_{xt}=\frac{1}{11}\sum_{k=1}^{11}\tilde{f}_{xtk}$ for each x, each t, and each training date,
3. compute the root mean squared forecast errors RMSE_{xt} of each of these ensemble means when verified against the training MAP observations at s,
4. assign weights w_{xt} proportional to the RMSE for x and t via

$$w_{xt}\sim\frac{\max_{\xi,\tau}\mathrm{RMSE}_{\xi\tau}-\mathrm{RMSE}_{xt}}{\max_{\xi,\tau}\mathrm{RMSE}_{\xi\tau}-\min_{\xi,\tau}\mathrm{RMSE}_{\xi\tau}} \tag{4}$$

That is, the pair (x, t) in the spatiotemporal neighborhood of s with the lowest RMSE is assigned the maximal weight, while the pair with the largest RMSE is assigned zero weight.

Our experiments show that this data-driven weighting scheme yields slightly better results than the fixed weighting scheme. It reflects the predominant westerly flow in that the earlier of the overlapping 6 h forecast periods has spatial weights more concentrated west of location s while the later period has weights concentrated east of s (see supporting information to this paper). The climatological parameters $\mu_{cl}, \sigma_{cl}, \delta_{cl}$ are specific to each location, time (6 h period) of the day, and each day of the year, and the regression coefficients $a_1, a_2, a_3, a_4, b_1, b_2$ and weights w_{xt} are derived separately for each location, each lead time period, and based on training data that is composed specific to each month.

4. Method for Modeling the Multivariate Dependence Structure

The methods described in section 3 yield reliable, predictive marginal distributions for each of the two weather variables, each location s, and each lead time t. These distributions can be sampled—either randomly, or systematically by choosing certain quantiles—and thus turned into an ensemble of any desired size K. In this study, we use the forecast quantiles with levels $\alpha_k=(k-0.5)/K$ for all $k=1,\ldots,K$, which is a CRPS-optimal sample of the predictive distribution [*Bröcker*, 2012]. An additional step is necessary, however, to link these individual samples in such a way that their joint probability distribution is represented adequately, and physically realistic, spatiotemporal forecast trajectories are obtained. To reconstruct the MAT and MAP

trajectories for a particular initialization date, the standard implementation of the Schaake shuffle proposed by *Clark et al*. [2004] chooses the dates in previous years within a short window surrounding that forecast initialization date, and uses the observed trajectories subsequent to these dates as "dependence templates" for the forecast trajectories to be constructed. In this process, the marginal forecast samples are reordered such that for each weather variable, location, and lead time their rank order is the same as that of the ensemble of historical trajectories (see *Clark et al*. [2004] or *Wilks* [2015] for a formal description and *Schefzik et al*. [2013] for general mathematical framework of this reordering process). In the following, we describe a variant of the Schaake shuffle that leaves the shuffling idea unchanged but chooses the historical dates in such a way that the marginal distributions of the sampled observation trajectories resemble those of the postprocessed forecast distributions. This way, the values of the historical trajectories that serve as "dependence templates" are closer to the forecast values to which they are mapped during the shuffling procedure, so that the assumption of state-independent spatiotemporal rank correlations that is somewhat implicit in the standard implementation of the Schaake shuffle (hereafter referred to as StSS) is substantially weakened.

The first step is to increase the pool of dates that are allowed as starting dates of the historical observation trajectories by widening the time window surrounding the forecast initialization date. Subsequently, this pool of candidate dates is thinned out to a subset of K historical starting dates. In order to preserve inter-variable correlations, that subset of dates must be the same for MAT and MAP, but for reasons of computational efficiency and due to the more complex nature of spatiotemporal correlations of precipitation fields, we let the MAT and MAP marginal distributions influence the thinning process in different ways. The MAT based selection criterion is applied first:

1. For each location s, and each lead time t compute the 99% prediction interval (0.005-quantile and 0.995-quantile of the predictive marginal distribution) for MAT.
2. Discard all candidate dates for which the corresponding MAT observation trajectory falls outside of more than m of those intervals.
3. That number m is chosen such that at least N_0 candidate dates are retained, where N_0 is a prespecified number larger than K.

In our examples we use this thinning criterion to bring the number of candidate dates down to approximately $N_0 = 500$. It is simple and computationally inexpensive, and additional experiments (not further discussed in this article) show that for MAT a more sophisticated approach yields only little further improvement. The highly asymmetric predictive distributions of MAP, on the contrary, require a different strategy to further reduce the remaining candidate dates from N_0 to K. Denote by F^f_{st} the MAP forecast cumulative distribution function (CDF) at location s and lead time t, and by $F^{\mathcal{H}}_{st}$ the empirical CDF calculated from observation trajectories corresponding to a set $\mathcal{H}$ of historical dates. Our goal is to choose $\mathcal{H}$ such that $F^{\mathcal{H}}_{st}$ and F^f_{st} are as similar as possible for all s and t, and we quantify similarity by studying the divergence [*Bröcker*, 2012; *Thorarinsdottir et al*., 2013] of the two distributions

$$\Delta^{\mathcal{H}}_{st} = \int (F^{\mathcal{H}}_{st}(x) - F^f_{st}(x))^2 dx. \tag{5}$$

Calculating the total divergence $\Delta^{\mathcal{H}}_{\text{tot}} = \sum_{s,t} \Delta^{\mathcal{H}}_{st}$ over all locations and lead times for any K-subset $\mathcal{H}$ is typically computationally infeasible since there are $\binom{N_0}{K}$ possible subsets. Instead, we propose the following backward elimination strategy:

1. Start with the set $\mathcal{H}_{N_0}$ of all observation trajectories retained after applying the MAT based criterion.
2. For each trajectory $j \in \mathcal{H}_{N_0}$, calculate the total divergence of the trajectories in the subset $\mathcal{H}_{N_0,-j}$ obtained by omitting j from $\mathcal{H}_{N_0}$.
3. Discard trajectories if their omission results in lower total divergence, keep the best N_1 trajectories and restart from step 2.
4. Iterate until only $N_{\text{final}} = K$ historical trajectories remain.

The sequence of the numbers $N_1 > N_2 > \cdots > N_{\text{final}}$ of trajectories retained after each iteration is a trade-off between statistical optimality and computational efficiency. Ideally, trajectories would be eliminated one by one, because the marginal CDFs $F^{\mathcal{H}}_{st}$ change each time a trajectory is removed. However, due to the computational cost for calculating the integrals in (5), reduction of the current set of trajectories by much more

than one may be required. In the data example discussed in sections 5 and 6, we use $K = 60$ and the sequence 420, 340, 270, 210, 160, 120, 90, 70, 60, which reduces the original N_0 to the desired K trajectories in nine iterations. Formulae for efficient computation of the integrals in (5) can be derived based on certain representations of the divergence given in *Bröcker* [2012], and are provided in Appendix A. Since the idea of minimizing the divergence between the marginal distributions of the historical observations and those of the postprocessed predictive distributions is the core of our algorithm, we refer to this approach as Minimum Divergence Schaake shuffle (MDSS).

5. Verification of the Postprocessed Meteorological Forcings

We now apply the marginal and multivariate calibration techniques presented in sections 3 and 4 to the GEFS reforecast data set and the MAT and MAP observations over California described in section 2. In addition, we apply the Ensemble Copula Coupling (ECC) technique to the same MAT and MAP marginal distributions. ECC is an another approach to reconstructing the space-time covariability of the postprocessed ensemble forecasts by reordering samples from the predictive marginal distributions; in contrast to StSS and MDSS, however, it uses the raw ensemble forecasts instead of historical observations to create multivariate "dependence templates" [*Schefzik et al*., 2013; *Wilks*, 2015]. The four river basins are processed separately. Results for the Russian River basin, which has been at the center of recent research efforts to assess the viability of forecast informed reservoir operations [*Jasperse et al*., 2015], are discussed in detail. A detailed analysis of the three other basins is provided in the supporting information to this paper, and a brief summary of some key performance metrics for all basins is given in Table 1.

To use the (almost) 26 years during which both reforecast and observation data are available in the most efficient way, we cross validate these data, leaving 1 year out for verification, fitting our statistical models with data from the remaining 25 years, and repeating that process for each year so that 26 years worth of forecasts and independent verifying observations are obtained. A different set of parameters for the univariate postprocessing models for MAT and MAP is fitted for each month, each location, and each lead time. Training data are composed of forecasts and observations during the ± 45 days around the 15th of this month, shifted by the respective lead time. With the cross-validation setting explained above, this amounts to 25×91 days worth of training data, which warrant stable parameter estimates for the CSGD regression equations (2), (3) even for longer lead times where the signal to noise ratio decreases and the risk of overfitting a complex model increases. A similar cross validation scheme is used for constructing the Schaake shuffle ranks, now using observation data from 1949 to 2009, and leaving out the data from the year that is currently being verified. This way, the StSS approach can generate $K = 60$ forecast trajectories by using just the forecast initialization date itself (i.e., window size is one). For the MDSS procedure proposed in section 4, we use again a 91 day time window to define the initial pool of candidate dates, yielding a initial set of 60×91 trajectories. This time window could be narrowed if, for example, the forcing ensemble is extended to lead times on a seasonal time scale, and too much deviation of the historical dates from the forecast date is undesirable. The ECC approach can generate a number of forecast trajectories that is equal to the number of members in the raw ensemble or multiples thereof. Two variants are considered here: ECC-Q11 generates $K = 11$ trajectories and uses quantiles (chosen as in section 4) to sample the predictive marginal distributions, while ECC-R66 generates $K = 66$ trajectories in six batches, each constructed by applying the rank reordering procedure to random samples of size 11. In order to gain insight into the role of K, we also report some results obtained with $K = 11$ StSS and MDSS trajectories.

Earlier studies [*Scheuerer and Büermann*, 2014; *Hemri et al*., 2014; *Scheuerer and Hamill*, 2015a] have already shown that the marginal postprocessing techniques employed here yield reliable and skillful forecasts, and this is confirmed by the reliability diagrams and Brier scores for CSGD postprocessed MAP forecasts provided as supporting information to this article. Here we focus on multivariate verification and compare the performance of StSS, MDSS, and ECC with regard to reconstructing the spatiotemporal dependence structure of the forcings, using multivariate proper scoring rules on the one hand, and univariate proper scoring rules applied to aggregate quantities that are sensitive to space-time covariability on the other hand. Scoring rules are a quantitative measure of the quality of a probabilistic forecast that takes both reliability and sharpness into account. Since scores are calculated case by case, they can be averaged over many cases without requiring any assumption about those cases having the same characteristics (e.g., same correlation

between locations A and B whatever the state of atmosphere). Specifically, we use the continuous ranked probability skill score (CRPSS) for univariate quantities and the energy skill score (ESS) and variogram skill score (VSS) for multivariate quantities. Definitions and details about these verification metrics are given in Appendix B. Climatological reference forecasts are obtained by building ensembles of climatological trajectories, separately for each month, using observation data from all days of this month (shifted by the respective forecast lead time) and all years from 1985 to 2010. We focus on results for winter, spring, and fall, because summers in California are usually very dry, numerical weather prediction is often unable to outperform a climatological, near-zero forecast, and representing the space-time structure of MAP is just not an issue of major importance.

5.1. Verification of Temporally and Spatially Upscaled MAT and MAP

One way of evaluating the statistical properties of a multivariate probability distribution represented by an ensemble is to aggregate the multivariate quantity to a set of univariate quantities whose distribution is sensitive to the spatial and/or temporal dependence structure. For MAT, we aggregate by considering the mean temperature across all subbasins and across blocks of 3 days. The uncertainty about the resulting spatially and temporally upscaled MATs strongly depends on the space-time covariability of the underlying 6 h MATs; e.g., averaging temperature trajectories that are either above or below normal for several consecutive 6 h periods yields upscaled MAT ensembles with much higher spread than averaging trajectories where the 6 h MATs alternate between below and above normal and the associated uncertainty partially averages out. For MAP, we aggregate by averaging over all subbasins and accumulating the original 6 h periods to 72 h periods. This kind of upscaling is especially relevant in the context of hydrological forecasting, where total runoff from the entire basin and cumulative precipitation amounts over an extended time period are of interest. Figure 4 depicts the CRPSSs obtained by verifying the upscaled StSS, MDSS, and ECC ensembles (in this and all subsequent figures, we show results for ECC-Q11 and the implementation of StSS/MDSS with 60 trajectories) against upscaled observations. For the MAT forecasts, the CRPSSs of StSS and MDSS are almost indistinguishable; a closer look (see also Table 1) suggests a very small but consistent improvement of MDSS over StSS. For MAP, on the contrary, that improvement is quite substantial during earlier forecast lead times; for lead times beyond day seven the CRPSSs of both methods tend to zero, and become more similar. The converse is true for the comparison between MDSS and ECC: their performance with respect to predicting aggregated MAT and MAP is very similar at earlier forecast lead times, but ECC skill decreases faster than MDSS skill as the forecast lead time increases. This could be expected since the space-time structure of MDSS trajectories converges toward the structure of observation trajectories representing climatology, while the ECC trajectories inherit the multivariate dependence structure from the raw ensemble which typically has little resemblance with the structure of observed precipitation fields at lead times beyond day 7. These conclusions are consistent over all months and all four basins (see supporting information to this paper for the corresponding figures). Table 1 summarizes these results, showing averages of CRPSSs obtained with the different implementations (StSS/MDSS with $K = 11$ and $K = 60$, ECC-Q11, and ECC-R66) of each multivariate postprocessing approach. It is interesting to note that the results for MAT confirm the conclusions of *Wilks* [2015] that increasing the number of trajectories that represent the multivariate forecast distribution leads to better forecast skill; in particular, ECC-R66 fares better than ECC-Q11, despite the additional sampling variability introduced by nonsystematically sampling the marginal forecast distributions. This is not true, however, for MAP forecasts, where ECC-Q11 often yields better results, especially during the dry summers where the sampling variability introduced by ECC-R66 entails forecasts that are significantly inferior to climatological forecasts. It is also worth noting that for MAP, the skill of MDSS-11 is comparable to the skill of StSS-60, suggesting that diligent construction of a small number of forecast trajectories can yield an equally good representation of the multivariate forecast distribution as a larger number of trajectories constructed without preferential selection of the dates that determine the "dependence template." Based on the results in this table, we chose to proceed with ECC-Q11 and the $K = 60$ implementation of StSS and MDSS.

5.2. Verification of the Spatial Structure of MAT and MAP Forecast Trajectories

We now take a closer look at the ability of StSS, MDSS, and ECC to generate forecast trajectories of MAT and MAP with appropriate spatial (i.e., across the different subbasins) structure. Figures 5 and 6 depict ESSs and VSSs of MAT forecast vectors corresponding to the seven subbasins of the Russian River basin for January, April, and October, separately for each lead time, but averaged over all cross-validation years and all days of

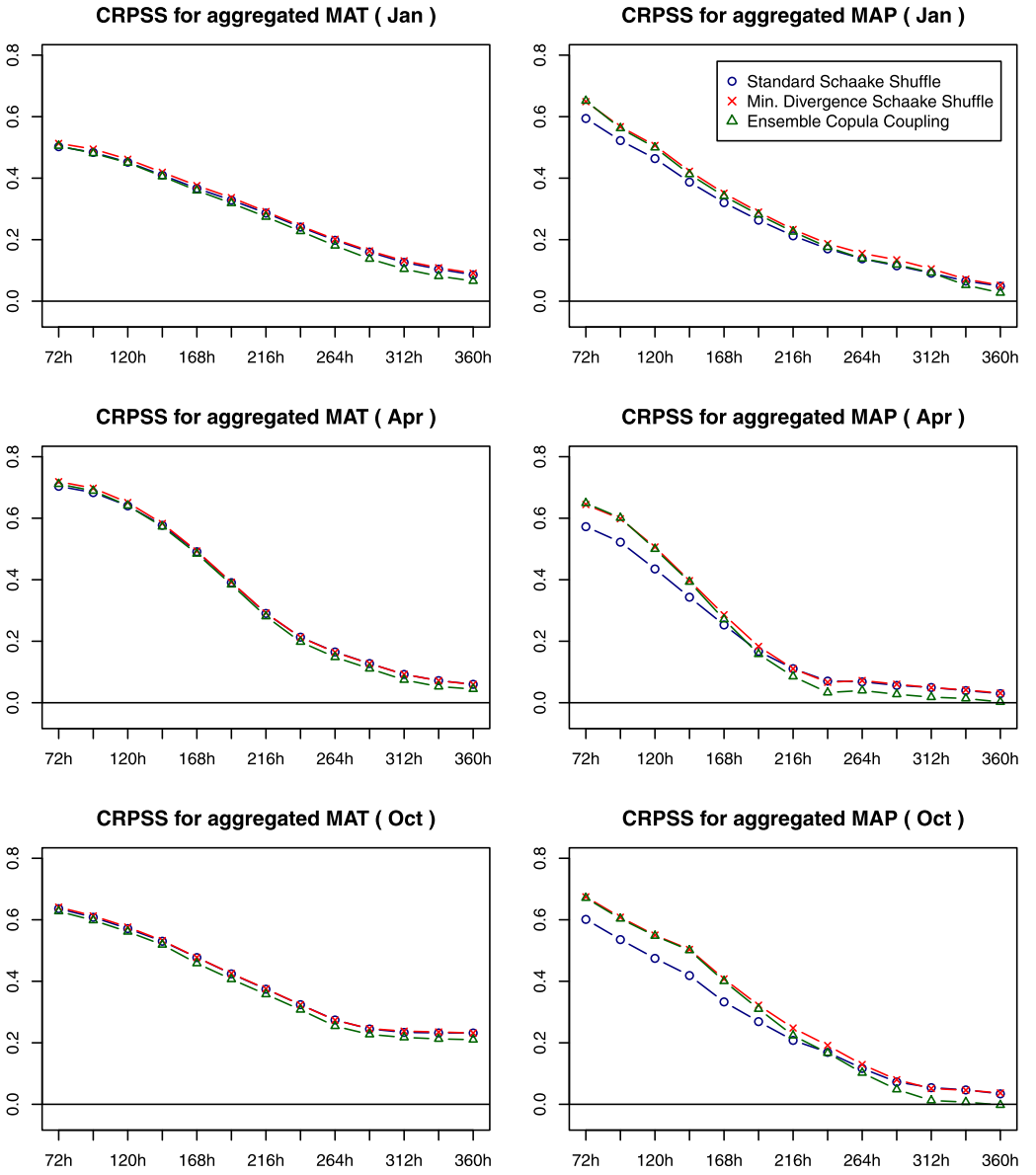

Figure 4. Continuous ranked probability skill scores of aggregated MAT (average over all subbasins and 72 h forecast periods) and MAP (average over all subbasins, accumulated to 72 h periods) forecasts over the Russian River basin.

the respective month. Forecast skill decreases with lead time, and can be very different across different seasons and different times of the day. ESS differences between the StSS, MDSS, and ECC technique are extremely small, but as explained in Appendix B this may be due to a lack of sensitivity of the ESS to genuinely multivariate properties of the forecasts. The VSS plots, on the contrary, suggest a clear improvement of MDSS over StSS with respect to the representation of the spatial structure of the forecasts. It may be surprising at first that the skill scores are close to zero or even negative for most lead times, but this is a consequence of the variogram score focusing on the evaluation of forecasts of *spatial differences* of MAT; unlike the energy score, it does not reward the forecast trajectories for predicting the correct magnitude of MAT, and it turns out that GEFS temperature forecasts provide little or no information on spatial MAT differences at the scale of the basins considered here. This is also likely the reason why ECC, which relies exclusively on the raw ensemble for information about the space-time structure, fares particularly bad. The MDSS approach yields forecast trajectories whose spatial structure is of comparable quality as that of climatological trajectories, while the StSS procedure degrades the spatial structure of the historical MAT trajectories on which it is based. For the three other basins (see Figures S15–S17 in the supporting information to this

Table 1. Average CRPSSs (Over All Lead Times) of Forecasts of Aggregated MAT and MAP, Obtained With Different Implementations of StSS, MDSS, and ECC

	MAT				MAP			
	January	April	July	October	January	April	July	October
Russian River								
StSS-11	0.273	0.332	0.237	0.382	0.245	0.196	−0.038	0.244
MDSS-11	0.274	0.331	0.234	0.378	0.268	0.211	−0.014	0.271
ECC-Q11	0.276	0.338	0.235	0.381	0.275	0.215	−0.017	0.276
StSS-60	0.288	0.347	0.252	0.397	0.261	0.209	0.000	0.257
MDSS-60	0.295	0.350	0.252	0.399	0.287	0.234	−0.001	0.296
ECC-R66	0.282	0.341	0.240	0.386	0.271	0.211	−0.198	0.278
Eel River								
StSS-11	0.234	0.337	0.272	0.353	0.249	0.200	0.030	0.266
MDSS-11	0.234	0.339	0.274	0.344	0.264	0.209	0.015	0.279
ECC-Q11	0.232	0.344	0.273	0.350	0.267	0.212	0.030	0.284
StSS-60	0.250	0.353	0.290	0.369	0.264	0.212	0.048	0.275
MDSS-60	0.253	0.358	0.291	0.369	0.284	0.232	0.048	0.304
ECC-R66	0.235	0.346	0.275	0.353	0.263	0.206	−0.128	0.282
American River								
StSS-11	0.320	0.353	0.313	0.362	0.219	0.198	−0.014	0.234
MDSS-11	0.318	0.354	0.312	0.360	0.233	0.206	−0.007	0.252
ECC-Q11	0.319	0.351	0.312	0.357	0.241	0.208	−0.017	0.257
StSS-60	0.332	0.365	0.327	0.373	0.232	0.209	0.002	0.250
MDSS-60	0.335	0.367	0.328	0.375	0.253	0.228	−0.018	0.272
ECC-R66	0.323	0.354	0.317	0.360	0.243	0.214	−0.096	0.264
Merced River								
StSS-11	0.331	0.360	0.277	0.328	0.198	0.164	0.033	0.193
MDSS-11	0.333	0.360	0.272	0.323	0.225	0.190	0.044	0.203
ECC-Q11	0.334	0.360	0.277	0.323	0.233	0.190	0.046	0.214
StSS-60	0.345	0.373	0.291	0.338	0.221	0.184	0.066	0.207
MDSS-60	0.347	0.374	0.289	0.340	0.248	0.214	0.085	0.234
ECC-R66	0.337	0.362	0.280	0.325	0.225	0.192	0.053	0.218

paper), the situation with respect to spatial MAT differences is even less favorable. A detailed analysis shows that GEFS forecasts of spatial MAT differences are uncorrelated or even negatively correlated with observed MAT differences between the respective subbasins, but since univariate postprocessing is based only on single-location forecast skill, the postprocessed marginal distributions still contain that erroneous information on spatial differences, and none of the multivariate methods are able to correct it.

For MAP, where spatial differences are more strongly interlinked with the magnitude of observed/predicted values, VSSs are much better than for MAT (ESSs are again very similar for all methods and therefore not shown here). Results shown in Figure 7 confirm previous conclusions that the preferential selection of dates by the MDSS procedure yields improved MAP trajectories compared to StSS for shorter lead times, with the degree of improvement varying across seasons and across the four basins (see Figures S18–S20 in the supporting information to this paper). For longer lead times the performance of both methods is comparable. ECC based trajectories are of poor quality for longer lead times, but are mostly also inferior to MDSS for shorter lead times, suggesting that the raw ensemble forecasts do not sufficiently represent the spatial structure of MAP at the scale of the basins considered here.

6. Verification of the Resulting Streamflow Forecasts

We finally study to what extent the improvement of the meteorological forcings translates into improved streamflow forecasts. To this end, the MAT and MAP ensembles studied in the previous section are now used as inputs to NOAA's Community Hydrologic Prediction System (CHPS) [*Roe et al.*, 2010] to generate streamflow hindcasts for the two headwater basins UKAC1 (Ukiah, Russian River basin) and LAMC1 (Coyote Dam, Lake Mendocino, Russian River Basin) during the period from 1 January 1985 to 15 September 2010. For these headwater basins, data of observed, unimpaired flow is available. For UKAC1 we study 1 day average flow forecasts, for LAMC1, 3 day average flow is considered since this is a reservoir and multiday volumes are more meaningful. Prior to generating streamflow hindcasts in CHPS, a historical set of watershed

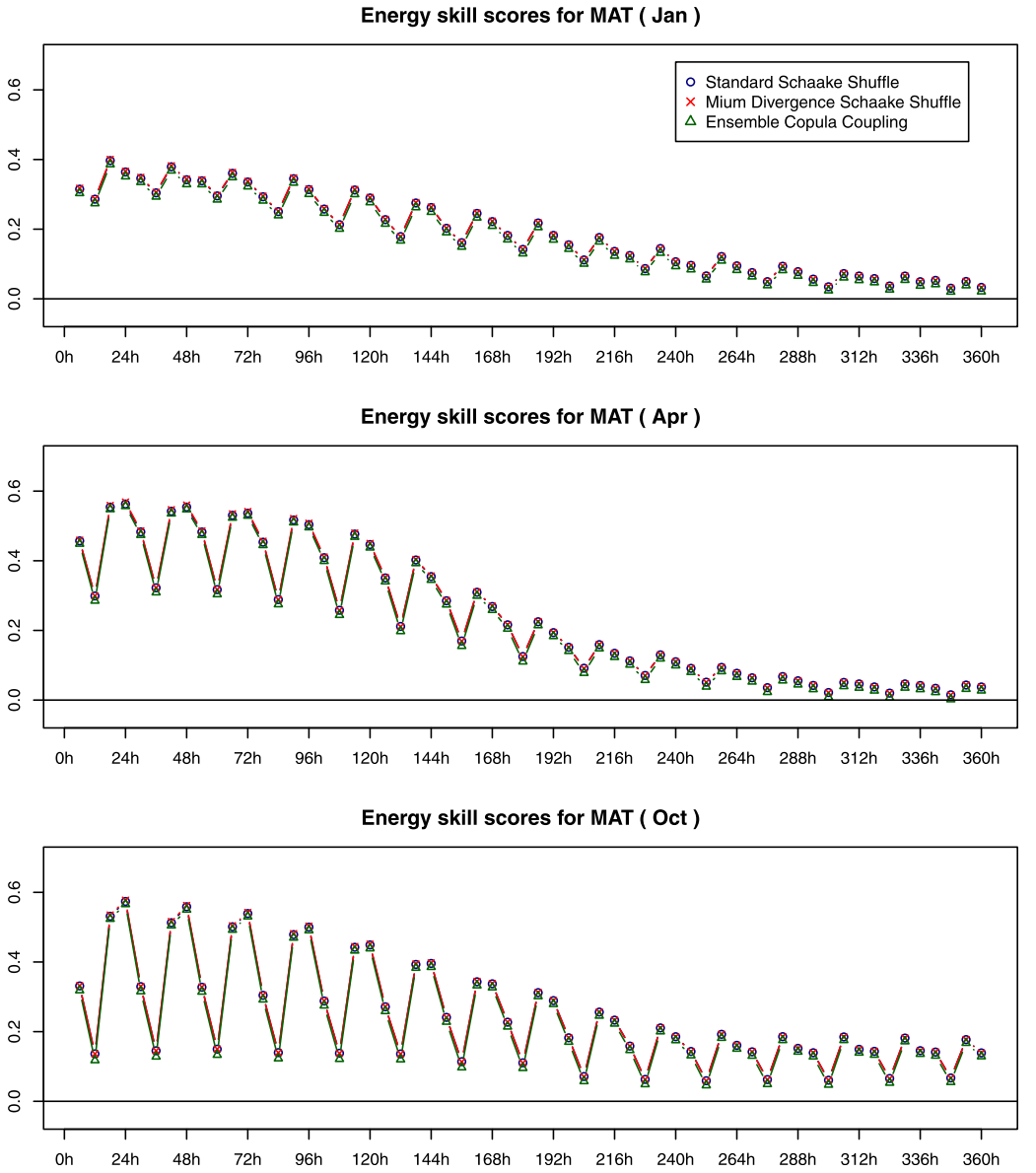

Figure 5. Energy skill scores of MAT forecasts over the Russian River basin.

states (warm states) was created by processing historical forcings through calibrated snow accumulation and ablation models (SNOW-17) [*Anderson*, 1973], and soil-moisture accounting models (SAC-SMA) [*Burnash et al.*, 1973] for the basins of interest. The streamflow hindcasts were then generated by looping through the hydrology forecast models 1 day at a time. For a given hindcast day, appropriate warm states were selected from the stored data set. The hydrology models were then forced with the meteorological inputs described previously resulting in a set of streamflow predictions equal to the number of meteorological forecast inputs. It is worth pointing out that both SNOW-17 and SAC-SMA models applied in routine forecasting operations are calibrated using data in the full record period (1985–2010). It is thus difficult to conduct a fully independent verification on streamflow forecasts (as is the case when verifying the meteorological inputs discussed in section 5). However, the StSS-derived, MDSS-derived, and ECC-Q11-derived forcing ensembles are run through the same hydrological model system (with parameters unchanged). As such, it is fair to compare the resulting streamflow hindcasts generated in all three scenarios with each other, though some care should be taken when interpreting the absolute skill of each set of streamflow hindcasts separately.

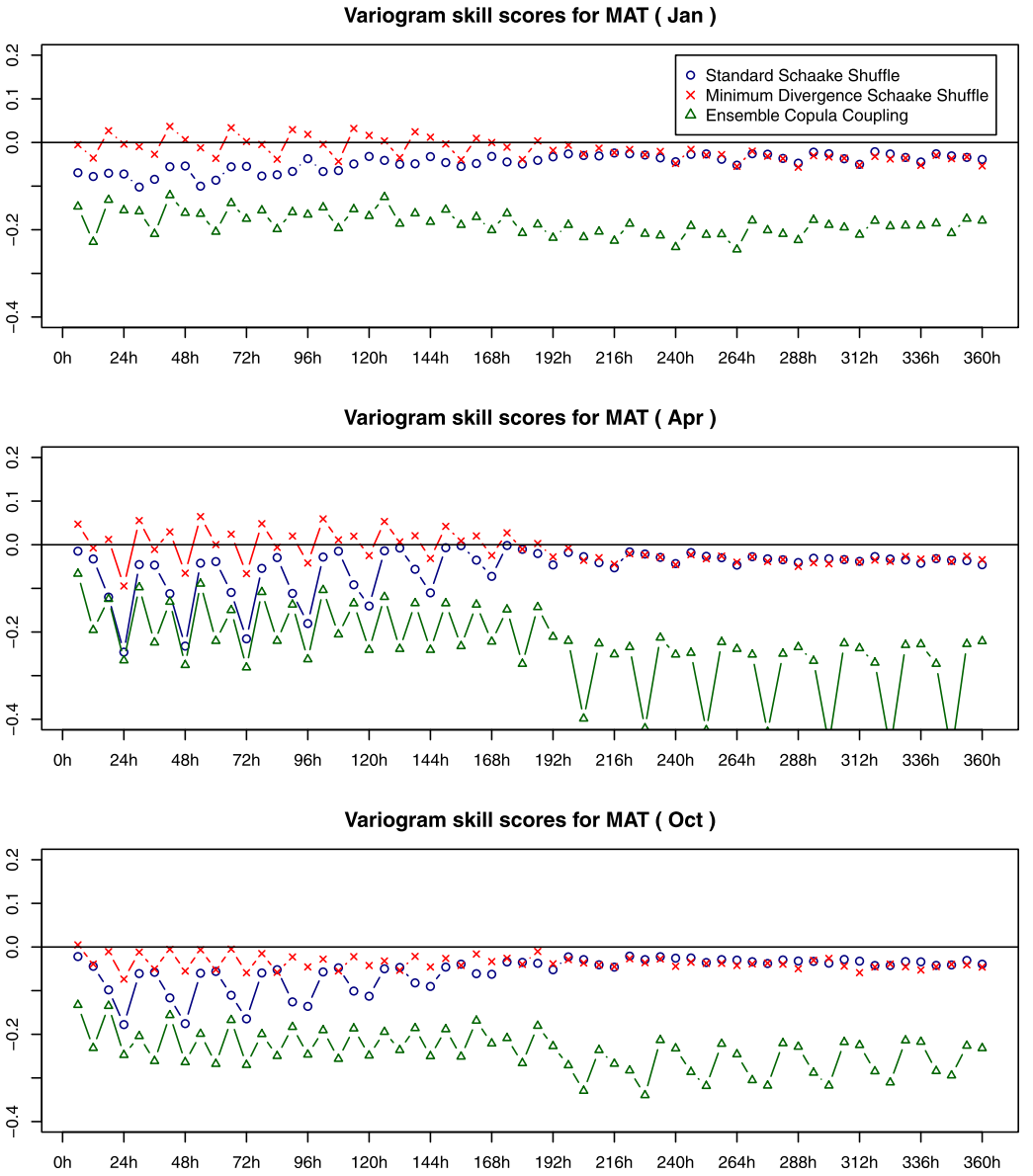

Figure 6. Variogram skill scores of MAT forecasts over the Russian River basin.

The CRPSS curves depicted in Figure 8 show that the improved temporal structure (note that spatial dependence is not a factor here as both UKAC1 and LAMC1 are headwater subbasins) of the MDSS forcing ensembles entails a clear improvement of the hydrological predictions. The ECC forcing ensembles fare better than the StSS forcing ensembles for shorter lead times but worse for longer lead times, and they are never better than the MDSS forcing ensembles, thus confirming the conclusion drawn before based on Figure 4. Clearly, the meteorological input variables are just one component of a hydrological forecast system, and for larger basins even substantially improved forcings might not have the same beneficial effect on streamflow forecasts. The results presented here demonstrate, however, that for some basins the additional complexity that comes with more sophisticated, multivariate calibration methods for the forcing variables can pay off and improve the overall accuracy of a hydrological forecast system.

7. Discussion

In this article, we have proposed a method for generating forecast trajectories of surface temperature and precipitation accumulations that have reliable marginal distributions and a physically realistic

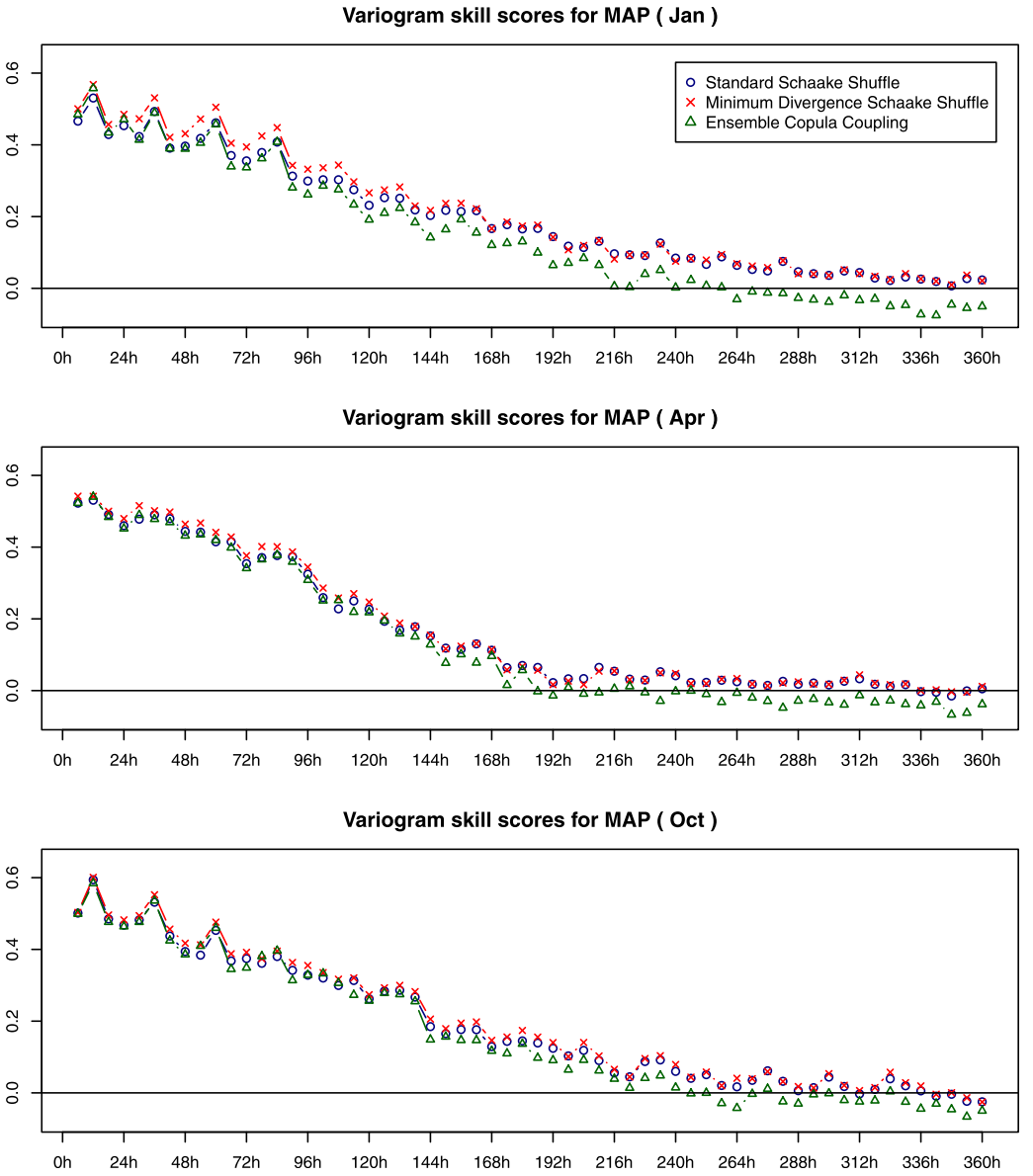

Figure 7. Variogram skill scores of MAP forecasts over the Russian River basin.

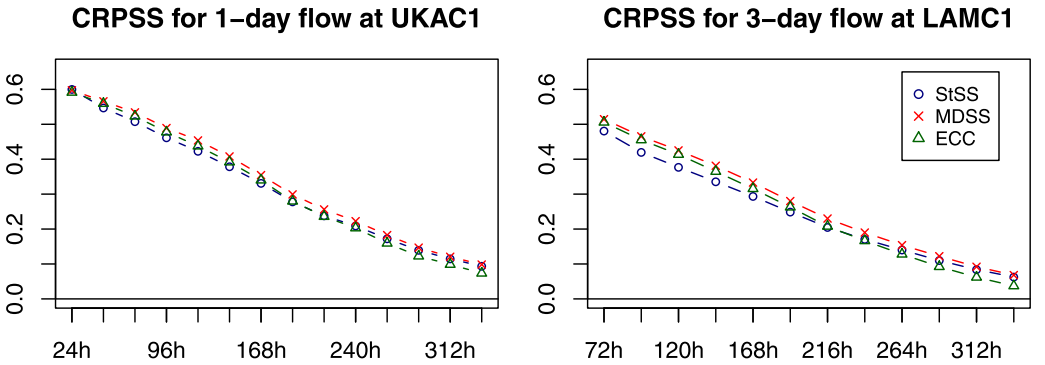

Figure 8. Continuous ranked probability skill scores of streamflow forecasts at Ukiah and Lake Mendocino.

spatiotemporal dependence structure. For univariate calibration of the GEFS ensemble predictions used here, a variant of the nonhomogeneous Gaussian regression approach [*Gneiting et al*., 2005] was employed while precipitation forecasts were postprocessed following the approach proposed by *Scheuerer and Hamill* [2015b]. These methods have been successfully used in the literature to generate reliable probabilistic forecasts based on the output of a global ensemble forecast system, but other approaches are available and could be used just as well for univariate calibration.

The main focus of this paper was to address the more difficult problem of modeling the joint distributions across the two variables, different subbasins, and a range of lead times. The Schaake shuffle [*Clark et al*., 2004] addresses this issue by using historical, observed trajectories as a "dependence template" and remapping the observed values to those derived from the marginal forecast distributions. To reduce the degradation of spatial and temporal gradients that can result from this remapping, we suggested a variant of the Schaake shuffle that aims at selecting the historical dates in such a way that the corresponding trajectories have marginal distributions that resemble the forecast distributions to which they are mapped. Our comparison of the standard Schaake shuffle (StSS) with this new variant (MDSS) focused on the spatial and temporal structure of the resulting forecast trajectories, and the results obtained for four river basins in California suggest that MDSS can indeed help generate trajectories with a physically more realistic structure. Covariability between the two forcing variables is addressed by simply using the same historical dates, but since the StSS and MDSS do not differ in that regard, verification of this intervariable dependence was omitted in our study. The forcing ensembles generated by StSS and MDSS, respectively, were used as input to a hydrological forecast system, and verification of the resulting streamflow ensemble forecasts confirmed the benefit of the improved multivariate structure obtained with MDSS.

Can the MDSS technique always improve upon the StSS? The biggest caveat about presumably any strategy for preferential selection of historical dates concerns the dimension of the multivariate distribution that is to be constructed. In the setup considered here, specifically, the Russian River basin, the dimension of that distribution was 840 (2 variables, 7 subbasins, 60 forecast lead times), which is pretty large already, but was feasible owing to a 61 year record of historical observations. For large basins with many more subbasins, this dimension would be much bigger which would not only increase the computational cost of the MDSS procedure dramatically, it would also make it harder and harder to find a set of historical trajectories with the desired marginal distributions. By construction, the MDSS should never perform significantly worse than the StSS, but for large basins its potential benefit is likely very small. Moreover, we have seen that the improvement of MDSS over StSS for constructing MAT trajectories is rather limited when verifying the quality of spatially and temporally upscaled forecasts. In some settings the representation of spatial differences by MDSS based trajectories was much better, while in other cases none of the multivariate techniques was able to correct misrepresentations of spatial properties inherited from the raw ensemble forecasts.

The Schaake shuffle is not the only approach for modeling joint forecast distributions, but there are good reasons why it has been the predominant approach in the context of hydrological forecasts. While Gaussian copulas [*Genest and Favre*, 2007; *Schölzel and Friederichs*, 2008; *Möller et al*., 2013] have been used successfully for modeling multivariate forecast distributions, the highly non-Gaussian, nonstationary, and spatially and temporally intermittent nature of precipitation makes them a suboptimal choice in hydrological forecast settings. Another discrete copula approach, ensemble copula coupling, has previously been used in the context of precipitation forecasting [*Roulin and Vannitsem*, 2012; *Flowerdew*, 2014] and was considered as a comparison method. Shortcomings of this approach found in the setup considered here were attributed to the fact that global ensemble forecast systems are typically unable to resolve the fine scale structure of precipitation and temperature fields, and for longer lead times (where significant displacement and timing errors are common) the structure of the raw ensemble forecast fields can be very different from the observed structure. The MDSS technique, on the contrary, is based on observations which have the desired spatial and temporal resolution. In contrast to the standard Schaake shuffle, it leverages real-time forecast information, and uses it to inform the choice of historical trajectories, thus combining the advantages of ECC with those of the standard Schaake shuffle and outperforming both of them. Our technique can be applied beyond the postprocessing of meteorological forcings, e.g., in a setting where streamflow forecasts themselves are postprocessed as in [*Hemri et al*., 2015], or in entirely different applications such as wind energy forecasting where forecast trajectories of wind power are required for decision-making [*Pinson*, 2013]. Different applications clearly entail different trade-offs concerning the advantages and disadvantages of the various multivariate modeling techniques, but we feel that the method proposed here is a valuable contribution to the modeling toolbox.

Appendix A: Computationally Efficient Calculation of the Divergence

The algorithm proposed in section 4 requires repeated evaluation of the integrals in equation (5), and it is crucial that this calculation is done efficiently to keep computational costs at a reasonable level. To simplify the notation, we consider a fixed location s and lead time t, and drop the corresponding subscripts from the notation. Denoting by S_{crp} the continuous ranked probability score (CRPS, see Appendix B) and defining

$$S_{\text{crp}}(F^{\mathcal{H}}, F^f) = \int S_{\text{crp}}(F^{\mathcal{H}}, x)\, dF^f(x)$$

the divergence integral can be expressed as [*Bröcker*, 2012]:

$$\Delta^{\mathcal{H}} = \int (F^{\mathcal{H}}(x) - F^f(x))^2 dx = S_{\text{crp}}(F^{\mathcal{H}}, F^f) - S_{\text{crp}}(F^f, F^f)$$

In our algorithm, $\Delta^{\mathcal{H}}$ needs to be evaluated for different choices of $\mathcal{H}$ while F^f is fixed, so the second term is a constant and can be omitted. Let $x_1, \ldots, x_K$ be a sample that represents the forecast distribution F^f and $y_1, \ldots, y_N$ the observation sample corresponding to the set $\mathcal{H}_N$. Assume that both samples are in increasing order. Using the CRPS representation in *Bröcker* [2012] and defining $\alpha_j := \frac{j-0.5}{N}, j \in \mathcal{H}_N$, the first term in the above equation can be expressed as

$$\begin{aligned} S_{\text{crp}}(F^{\mathcal{H}}, F^f) &= \frac{1}{K}\sum_{i=1}^{K} S_{\text{crp}}(F^{\mathcal{H}}, x_i) \\ &= \frac{2}{KN}\sum_{i=1}^{K}\sum_{j=1}^{N} \alpha_j (x_i - y_j)_+ + (1-\alpha_j)(y_j - x_i)_+ \\ &= \frac{2}{N}\sum_{j=1}^{N} \alpha_j \underbrace{\left(\frac{1}{K}\sum_{i=1}^{K}(x_i - y_j)_+\right)}_{=:\varepsilon_j^-} + (1-\alpha_j)\underbrace{\left(\frac{1}{K}\sum_{i=1}^{K}(y_j - x_i)_+\right)}_{=:\varepsilon_j^+} \end{aligned} \tag{A1}$$

The algorithm in section 4 requires, for a given set $\mathcal{H}_N$, the calculation of the divergence for all $(N-1)$-subsets of $\mathcal{H}_N$. The fastest way to do that is to precompute and store ε_j^- and ε_j^+ for all $j \in \mathcal{H}_N$. Then, for every $j \in \mathcal{H}_N$

1. define $\alpha_{j'} := \frac{j'-0.5}{N-1}, \quad j' = 1, \ldots, N-1,$
2. compute the outer sum in equation (A.1) over $j' \in \mathcal{H}_N \setminus \{j\}$, using these $\alpha_{j'}$
3. since $\varepsilon_{j'}^-$ and $\varepsilon_{j'}^+$ have been precomputed, the inner sums are only calculated once

Computation time can be further reduced by defining and storing $\varepsilon_j^\delta := \varepsilon_j^- - \varepsilon_j^+, j \in \mathcal{H}_N$, and rewriting equation (A1) as

$$S_{\text{crp}}(F^{\mathcal{H}}, F^f) = \frac{2}{N}\sum_{j=1}^{N} \varepsilon_j^+ + \frac{2}{N}\sum_{j=1}^{N} \alpha_j \varepsilon_j^\delta$$

This is an advantage when we consider the additional summation over s and t

$$\sum_{s,t} S_{\text{crp}}(F_{st}^{\mathcal{H}}, F_{st}^f) = \frac{2}{N}\sum_{s,t}\sum_{j=1}^{N} \varepsilon_j^+(s,t) + \frac{2}{N}\sum_{s,t}\sum_{j=1}^{N} \alpha_j \varepsilon_j^\delta(s,t)$$

The requirement that each subsample $\{y_{j'}(s,t) : j' \in \mathcal{H}_N \setminus \{j\}\}$ is in ascending order prevents any simplifications or precalculations of the second term, due to the multiplication with the coefficients α_j which rely on that ordering. In the first term, however, those coefficients are missing, rendering the ordering assumption unnecessary, and allowing one to interchange the two sums, precompute $\sum_{s,t} \varepsilon_j^+(s,t)$, and just omit the respective term that corresponds to the left-out $j \in \mathcal{H}_N$ from the summation over the elements in $\mathcal{H}_N \setminus \{j\}$.

Appendix B: Verification Metrics Used in This Paper

B.1. Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) is a popular measure for the overall skill of a probabilistic forecast of a univariate quantity. Given a forecast CDF F and a verifying observation y, the CRPS is defined as

$$S_{\mathrm{crp}}(F,y)=\int(F(x)-\mathbf{1}_{[y,\infty)}(x))^2dx$$

where $\mathbf{1}_{[y,\infty)}(x)$ is the indicator function that is 1 if $x \geq y$ and 0 otherwise. It is a negatively oriented (i.e., lower is better), strictly proper score which evaluates both reliability and sharpness of F as a forecast of y. Being "strictly proper" ensures that a forecaster who knows and issues the true conditional distribution of y attains a better score on average than any forecaster who issues a different distribution. *Gneiting and Raftery* [2007] provide a general discussion of (strictly) proper scoring rules and show that the CRPS can be represented as

$$S_{\mathrm{crp}}(F,y)=E_F|X-y|-\frac{1}{2}E_F|X-X'| \tag{B1}$$

where X and X' are independent random variables with CDF F, and E_F denotes the expectation under F.

B.2. Energy Score

Representation (B.1) of the CRPS is the starting point of its generalization to multivariate quantities. For an observation vector $\mathbf{y}=(y_1,\ldots,y_d)^t$ and independent random vectors $\mathbf{X}=(X_1,\ldots,X_d)^t$ and $\mathbf{X}'=(X_1',\ldots,X_d')^t$, representing a d-variate distribution F, the energy score is defined as

$$S_{\mathrm{en}}(F,\mathbf{y})=E_F\|\mathbf{X}-\mathbf{y}\|-\frac{1}{2}E_F\|\mathbf{X}-\mathbf{X}\prime\| \tag{B2}$$

where $||\cdot||$ denotes the Euclidean norm. The energy score (ES) is a strictly proper score for multivariate quantities [*Gneiting and Raftery*, 2007].

B.3. Variogram Score

In spite of being strictly proper, the ES has been found to be rather insensitive to misspecification of the correlation structure of the multivariate quantities. That is, if F is the true distribution and G is a forecast distribution with the same marginal distributions but different correlation structure than F, the systematic difference between $S_{\mathrm{en}}(F,\mathbf{y})$ and $S_{\mathrm{en}}(G,\mathbf{y})$ is often too small to detect the misspecification. *Scheuerer and Hamill* [2015b] therefore suggested an alternative score which compares forecast and observed discrete spatial/temporal differences of the multivariate quantity, and weighs these differences before combining them into a single score

$$S_{\gamma_p}(F,\mathbf{y})=\sum_{i,j=1}^{d}w_{ij}(|y_i-y_j|^p-E_F|X_i-X_j|^p)^2 \tag{B3}$$

which is referred to as variogram score of order p (VS-p). In this paper we set $p=0.5$, which *Scheuerer and Hamill* [2015b] and other studies found to be a choice with favorable sampling properties. We use the VS-0.5 in a setting where the components of $\mathbf{y}$ and $\mathbf{X}$ correspond to spatial locations, and we choose $w_{ij}\equiv 1$ since the basins considered here are relatively small and spatial dependence is typically strong among all locations. The VS-0.5 is usually better than the ES in distinguishing forecast distributions that differ only in their dependence structure. It is proper but not strictly proper, which implies that the "best" forecaster should never fare worse than any competitor, but it can happen that a different forecaster attains the exact same score despite issuing forecasts that are inferior in certain aspects of the forecast to which the VS-0.5 is simply not sensitive.

B.4. Skill Scores

Skill scores are a way to put scores into context and make them more interpretable by relating them to a reference score. Specifically, let $S_\star$ be any proper score (like those discussed above), S_{ref} a reference score to compare against, and denote by $\overline{S_\star}$ and $\overline{S_{\mathrm{ref}}}$ the respective average scores over a sufficiently large number of cases. Then the skill score corresponding to $S_\star$ is defined as

$$\mathcal{S}_\star=\frac{\overline{S_{\mathrm{ref}}}-\overline{S_\star}}{\overline{S_{\mathrm{ref}}}}$$

A perfect forecast has a skill score of 1, while a skill score of 0 indicates that the forecast is only as good as the reference forecast, and a negative skill score suggests that the forecast performance is inferior to the

reference. The standard reference forecast is a climatological forecast, which should be location specific and take seasonal variations into account, but does not depend on the particular state of the atmosphere.

Acknowledgments
Michael Scheuerer's research was supported by grants from the NOAA/NWS Sandy Supplemental (Disaster Relief Appropriations Act of 2013), award # NA14OAR4830123, the NOAA/NWS Research to Operations (R2O) initiative for the Next-Generation Global Prediction System (NGGPS), award # NA15OAR4320137, and additional funding from the NOAA/CIRES Cooperative Agreement to develop methods that help improve decision-making in reservoir management in California, award # NA15OAR4320137. Tom Hamill's work was supported by funding provided to NOAA/ESRL/PSD by NOAA/NWS/STI under the Next-Generation Global Prediction System, grant P8MWQNG-PTR. The reforecast data used in this study can be obtained from the website http://www.esrl.noaa.gov/psd/forecasts/reforecast2/download.html. Observed MAT/MAP and streamflow data for UKAC1 and LAMC1 are available at http://www.esrl.noaa.gov/psd/people/michael.scheuerer/forcings.zip and http://www.esrl.noaa.gov/psd/people/michael.scheuerer/streamflow.zip.

References

Anderson, E. A. (1973), National Weather Service River Forecast System—Snow accumulation and ablation model, *NOAA Tech. Memo. NWS HYDRO-17*, 217 pp., U.S. Department of Commerce, NOAA/NWS, Washington D. C.

Bröcker, J. (2012), Evaluating raw ensembles with the continuous ranked probability score, *Q. J. R. Meteorol. Soc.*, *138*, 1611–1617.

Burnash, R. J., R. L. Ferral, and R. A. McGuire (1973), *A generalized streamflow simulation system: Conceptual Modeling for Digital Computers*, U.S. Dep. of Comm., Natl. Weather Serv., Sacramento, Calif.

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. Hydrometeorol.*, *5*, 243–262.

Cloke, H., and F. Pappenberger (2009), Ensemble flood forecasting: A review, *J. Hydrol.*, *375*, 613–626.

Dettinger, M. D., D. R. Cayan, M. K. Meyer, and A. E. Jeton (2014), Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American river basins, Sierra Nevada, California, 1900–2099, *Clim. Change*, *62*, 283–317.

Flowerdew, J. (2014), Calibrating ensemble reliability whilst preserving spatial structure, *Tellus, Ser. A*, *66*, 22,662.

Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrol. Eng.*, *12*, 347–368.

Gneiting, T., and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *J. Am. Stat. Assoc.*, *102*, 359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, *133*, 1098–1118.

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 data set, *Bull. Am. Meteorol. Soc.*, *94*, 1553–1565.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden (2014), Trends in the predictive performance of raw ensemble weather forecasts, *Geophys. Res. Lett.*, *41*, 9197–9205, doi:10.1002/2014GL062472.

Hemri, S., D. Lisniak, and B. Klein (2015), Multivariate postprocessing techniques for probabilistic hydrological forecasting, *Water Resour. Res.*, *51*, 7436–7451, doi:10.1002/2014WR016473.

Jasperse, J., et al. (2015), A comprehensive plan to evaluate the viability of forecast informed reservoir operations for lake Mendocino, *Tech. Rep.*, Sonoma County Water Agency Rep. [Available at http://cw3e.ucsd.edu/FIRO/.]

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir (2013), Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas, *Q. J. R. Meteorol. Soc.*, *139*, 982–991.

Park, Y.-Y., R. Buizza, and M. Leutbecher (2008), Tigge: preliminary results on comparing and combining ensembles, *Q. J. R. Meteorol. Soc.*, *134*, 2029–2050.

Pinson, P. (2013), Wind energy: Forecasting challenges for its operational management, *Stat. Sci.*, *28*, 564–585.

Robertson, D. E., D. L. Shrestha, and Q. J. Wang (2013), Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, *Hydrol. Earth Syst. Sci.*, *17*, 3587– 3603.

Roe, J. M., et al. (2010), Introduction of NOAA's Community Hydrological Prediction System, in *26th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, vol. 7, Am. Meteorol. Soc. B, Atlanta, Ga. [Available at https://ams.confex.com/ams/pdfpapers/162687.pdf.]

Roulin, E. (2007), Skill and relative economic value of medium-range hydrological ensemble predictions, *Hydrol. Earth Syst. Sci.*, *11*, 725–737.

Roulin, E., and S. Vannitsem (2012), Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts, *Mon. Weather Rev.*, *140*, 874–888.

Schaake, J., J. Demargne, R. Hartman, M. Mullusky, E. Welles, L. Wu, H. Herr, X. Fan, and D.-J. Seo (2007), Precipitation and temperature ensemble forecasts from single-value forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, *4*, 655–717.

Schefzik, R. (2016), A similarity-based implementation of the Schaake shuffle, *Mon. Wea. Rev.*, *144*, 1909–1921.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting (2013), Uncertainty quantification in complex simulation models using ensemble copula coupling, *Stat. Sci.*, *28*, 616–640.

Scheuerer, M., and L. Büermann (2014), Spatially adaptive post-processing of ensemble forecasts for temperature, *J. R. Stat. Soc. C*, *63*, 405–422.

Scheuerer, M., and T. M. Hamill (2015a), Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions, *Mon. Weather Rev.*, *143*, 4578–4596.

Scheuerer, M., and T. M. Hamill (2015b), Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, *Mon. Weather Rev.*, *143*, 1321–1334.

Schölzel, C., and P. Friederichs (2008), Multivariate non-normally distributed random variables in climate research—Introduction to the copula approach, *Nonlinear Processes Geophys.*, *15*, 761–772.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl (2013), Using proper divergence functions to evaluate climate models, *SIAM/ASA J. Uncertainty Quantif.*, *1*, 522–534.

Verkade, J. S., and M. G. F. Werner (2011), Estimating the benefits of single value and probability forecasting for flood warning, *Hydrol. Earth Syst. Sci.*, *15*, 3751–3765.

Verkade, J. S., J. D. Brown, P. Reggiani, and A. H. Weerts (2013), Post-processing ecmwf precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, *J. Hydrol.*, *501*, 73–91.

Vrac, M., and P. Friederichs (2015), Multivariate-intervariable, spatial, and temporal-bias correction, *J. Clim.*, *28*, 218–237.

Wilks, D. S. (2015), Multivariate ensemble Model Output Statistics using empirical copulas, *Q. J. R. Meteorol. Soc.*, *141*, 945–952, doi:10.1002/qj.2414.

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.