AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR029453

A Multivariate Approach to Generate Synthetic Short-To-Medium Range Hydro-Meteorological Forecasts Across Locations, Variables, and Lead Times

Zachary P. Brodeur[1] and **Scott Steinschneider**[1]

[1]Department of Biological and Environmental Engineering, Cornell University, NY, USA

Key Points:

- We develop a multivariate synthetic forecast methodology to support design, validation, and testing of forecast informed water management
- We validate our method in two case studies that illustrate the model's ability to capture complex forecast behavior
- We find good model performance in developing synthetic forecasts of both streamflow and meteorology (temperature and precipitation)

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Z. P. Brodeur,
zpb4@cornell.edu

Citation:

Brodeur, Z. P., & Steinschneider, S. (2021). A multivariate approach to generate synthetic short-to-medium range hydro-meteorological forecasts across locations, variables, and lead times. *Water Resources Research*, *57*, e2020WR029453. https://doi.org/10.1029/2020WR029453

Received 14 DEC 2020
Accepted 18 MAY 2021

Abstract The use of hydro-meteorological forecasts in water resources management holds great promise as a soft pathway to improve system performance. Methods for generating synthetic forecasts of hydro-meteorological variables are crucial for robust validation of forecast use, as numerical weather prediction hindcasts are only available for a relatively short period (10–40 years) that is insufficient for assessing risk related to forecast-informed decision-making during extreme events. We develop a generalized error model for synthetic forecast generation that is applicable to a range of forecasted variables used in water resources management. The approach samples from the distribution of forecast errors over the available hindcast period and adds them to long records of observed data to generate synthetic forecasts. The approach utilizes the Skew Generalized Error Distribution (SGED) to model marginal distributions of forecast errors that can exhibit heteroskedastic, auto-correlated, and non-Gaussian behavior. An empirical copula is used to capture covariance between variables, forecast lead times, and across space. We demonstrate the method for medium-range forecasts across Northern California in two case studies for (1) streamflow and (2) temperature and precipitation, which are based on hindcasts from the NOAA/NWS Hydrologic Ensemble Forecast System (HEFS) and the NCEP GEFS/R V2 climate model, respectively. The case studies highlight the flexibility of the model and its ability to emulate space-time structures in forecasts at scales critical for water resources management. The proposed method is generalizable to other locations and computationally efficient, enabling fast generation of long synthetic forecast ensembles that are appropriate for risk analysis.

1. Introduction

The use of forecast information in water resources decision-making has become an increasingly important soft path toward water sustainability (Brown et al., 2015; Gleick, 2002). Forecasts are used to inform several management actions (Miller et al., 2018), including emergency flood response (Valeriano et al., 2010; Wang et al., 2012), water use restrictions (Zeff et al., 2016), and forecast informed reservoir operations (Delaney et al., 2020; Jasperse et al., 2020). These actions are often organized through policies that specify how forecast information should trigger different decisions. To ensure they are robust, forecast informed policies should be developed and tested using climate and hydrologic hindcasts or retrospective forecasts based on models that are initialized using initial conditions that were present over a historical period. Hindcasts can be split into calibration and testing periods, such that policies are designed and/or optimized for the calibration period and then tested out of sample using testing period data. A major challenge with this approach is that it is limited by the timeframe over which hindcasts are available. Advanced meteorological hindcasts are generally bounded to the period when satellite data are available for climate model initialization (i.e., from 1979 onward; Hartmann, 2016). This limitation extends to hydrologic forecasts forced by those climate hindcasts (Demargne et al., 2014). Thus, a fairly short time period of available hindcasts (at most ~40 years) must be parsed into even smaller periods to enable calibration and testing of policies, creating the potential for overfitting and poor out-of-sample performance (Brodeur et al., 2020; Herman et al., 2020; Nayak et al., 2018).

Synthetic forecasts offer a solution to overcome this challenge. Synthetic forecasts are generated by adding random error to observational records, such that the resulting series is statistically indistinguishable from forecasts developed using a physically based model. Many water resources projects in the United States have instrumental records of streamflow and climate that extend back to the early twentieth century (Loucks &

Van-Beek, 2017). Synthetic forecasts based on these extended observational records, which often contain multiple floods and droughts, can provide a rich source of information for understanding how forecasts would perform during past extreme events outside of the available hindcast period. Moreover, synthetic forecasts can address the aforementioned overfitting problem in forecast informed policy design and selection. For instance, previous work (Brodeur, 2021) has shown that selecting an optimal policy by its calibration performance against a single historical sequence can lead to poor out of sample performance, whereas policies selected by validation across many synthetic sequences are more robust. Synthetic forecasts can enable this type of procedure in the forecast informed setting, aiding in robust policy selection frameworks that emulate model selection procedures in machine learning (Shalev-Shwartz & Ben-David, 2013).

Synthetic forecasts can be beneficial at all timescales, including seasonal forecasts that are used to inform water supply based decisions (Ahn et al., 2017; Anghileri et al., 2016; Denaro et al., 2017; Giuliani et al., 2019; Turner et al., 2017; Yuan et al., 2015), and shorter range forecasts that are used for hazard management (Valeriano et al., 2010; You & Cai, 2008). However, uncertainty in seasonal forecasts surpasses that of short-to-medium range forecasts (Giuliani et al., 2019). In addition, regions that receive a significant portion of their annual water supply from a small number of events (e.g., California; Dettinger et al., 2011; Hanak et al., 2011) benefit more from short time scale forecasts associated with those events (Jasperse et al., 2020; Nayak et al., 2018; Raso et al., 2014). Thus, shorter lead forecasts can have more value for decision-making in many regions and further increase the likelihood that water managers would incorporate them into decision-making (Rayner et al., 2005). Accordingly, the present study concentrates on short-to-medium range synthetic forecast generation.

The two primary methods for synthetic streamflow forecasting are (1) the "direct statistical error model" approach and (2) the "conceptual hydrologic model" approach (Lamontagne & Stedinger, 2018). The direct approach derives synthetic forecasts of streamflow based on a statistical model of the errors between streamflow forecasts and observations (Grygier et al., 1989; Lettenmaier, 1984; Maurer & Lettenmaier, 2004; Sankarasubramanian et al., 2009; Turner et al., 2017). In contrast, the conceptual approach uses hindcasts of meteorological variables (generally temperature and precipitation) to drive hydrologic forecast model outputs (Alemu et al., 2011). Where long observed streamflow records exist, the direct approach is straightforward to apply and captures both hydrologic and meteorological sources of forecast uncertainty. In many regions, extensive streamflow records are not available (Teegavarapu et al., 2019), making the conceptual approach an attractive alternative. Synthetic meteorological forecasts are common for energy system applications that are dependent on forecasts of wind and solar power generation (Barth et al., 2011; Mello et al., 2011; Hodge et al., 2012; Olauson et al., 2016; Pelland et al., 2013; Sun et al., 2020), but are relatively rare in water resources management applications (Nayak et al., 2018).

This study forwards a novel method for synthetic forecast generation that can be applied to either streamflow or meteorological data, supporting both the direct and conceptual approach to synthetic streamflow forecasting. The proposed methodology addresses two challenges in synthetic forecast development that are currently unresolved. First, daily hydro-meteorological forecast errors often exhibit highly non-normal distributions, similar to the hydrologic and climate variables being modeled (Teegavarapu et al., 2019; Wilks, 2019). However, previous synthetic forecast generation methodologies (Alemu et al., 2011; Grygier et al., 1989; Lettenmaier,1984; Maurer & Lettenmaier, 2004; Sankarasubramanian et al., 2009; Turner et al., 2017) have utilized simple error distribution models for sub-seasonal to seasonal forecasts that would not be appropriate for daily resolution forecasts. We forward the generalized error distribution (GED) to address this challenge. The GED has a long history in the statistical literature (Subbotin, 1923) and has been referenced in alternate forms as the "exponential power (EP)" distribution (Box & Tiao, 1992), the "generalized Laplace" (Ernst, 1998), and the "generalized normal" distribution (Nadarajah, 2005). A uniting feature of all variants is that they can fit data with varying degrees of kurtosis (Cerqueti et al., 2019). This flexibility allows modeling of common Gaussian distributions and fat-tailed distributions with higher probabilities assigned to large errors. Methods to add skew to the distribution (Bottazzi & Secchi, 2011; Buckle, 1995; Fernández & Steel, 1998) further increase the flexibility of the model to account for asymmetries in probability mass around the mode. Use of the GED is widespread in the econometrics literature (e.g., Cerqueti et al., 2019; Nelson, 1991; So et al., 2008), where skewed and fat-tailed errors are common, but its use in the hydrologic literature is less established (with some exceptions, e.g., Schoups & Vrugt, 2010).

Another challenge is that synthetic hydro-meteorological forecasts must preserve correlations across space, variables, and lead times (Demargne et al., 2009; Wilks, 2019), especially for short-to-medium range timescales over which transient storms move across the landscape (Hartmann, 2016; Wilks, 2019). In addition, Lamontagne and Stedinger (2018) recently noted that forecasts errors can be correlated with the observed data, and these correlations have not been appropriately replicated in past synthetic forecast methods (Maurer & Lettenmaier, 2004; Sankarasubramanian et al., 2009). They forward a corrected stochastic method and an updated generalized maintenance of variance extension (GMOVE) procedure to address these issues. However, these proposed approaches either lack a multivariate extension or are not designed for ensemble synthetic generation to capture forecast uncertainty. As an alternative, copulas have emerged in hydrology as an effective tool to capture dependence across variables and spatio-temporal domains (Chen & Guo, 2019; Teegavarapu et al., 2019 and references therein), and they have been used for synthetic flood forecasting (e.g., the Martingale Model of Forecast Error; Heath & Jackson, 1994; Zhao et al., 2013). However, we were unable to find examples of copula-based synthetic meteorological forecasts, despite their previous use for climate forecast bias correction (Piani & Haerter, 2012) and ensemble post-processing (Wilks, 2015).

This work develops an adaptable synthetic forecast generation methodology that can (a) model both hydrological and meteorological forecast errors exhibiting auto-correlation, heteroscedasticity, and a variety of distributional forms, (b) link those errors across space, time and/or variables with empirical copulas, and (c) preserve critical relationships between the observed data and forecast errors. This approach mirrors multivariate Generalized Autoregressive Conditional Heteroskedastic (GARCH) models in the econometrics literature (Rao & Vinod, 2019; Wei, 2019) and employs the Skew Generalized Error Distribution (SGED) (Wurtz et al., 2020) as the underlying model for marginal errors. We demonstrate this approach in two separate applications for hydrological and meteorological synthetic forecast generation. In the first, we generate synthetic streamflow forecasts for Folsom Reservoir and Lake Mendocino, CA, emulating those of the Hydrologic Ensemble Forecast System (HEFS) currently in operational use in NOAA/NWS River Forecast Centers (RFC) (Demargne et al., 2014). Here, we focus on model performance in the temporal domain across 1–10 days forecast lead times. In the second application, we synthetically generate forecasts of temperature and precipitation based on the NCEP GEFS/R numerical weather prediction model (Hamill et al., 2013) at 5 lead times and across 30 grid cells that span Northern California. This case study highlights the ability of the approach to capture key features of conditional dependence between forecasted variables across space and time. These two applications demonstrate the ability of the generalized approach to support the direct statistical model or conceptual hydrologic model approach for synthetic forecast generation (Lamontagne & Stedinger, 2018).

2. Data

All hydro-meteorological data were collected from a region within Northern California (36.5°–42.5°N and 119.5°–124.5°W; Figure 1). Within this region, our first synthetic forecast application focuses on hydrologic forecasts in the American River and Russian River watersheds (insets of Figure 1). We obtained hindcast daily streamflow data directly upstream of Folsom Reservoir and Lake Mendocino from the NOAA/NWS California/Nevada River Forecast Center (CA/NV RFC, 2020) for the period of January 1, 1985 to September 15, 2010 and observed streamflow data from the California Department of Water Resources Data Exchange Center (CDEC) for the period of October 1, 1948 to September 15, 2010 (CA/DWR, 2020). As observed data, we use full natural flow (FNF), which is an estimated time series of natural streamflow that has been adjusted from the gauged record to remove the impacts of upstream regulation and diversions (Zimmerman et al., 2018). This process is imperfect, and negative flows are produced across ~8.5% of the record, which we corrected to zero flow in this study.

The hindcasted streamflow data are medium-range ensemble mean output from the HEFS. These forecasts are driven by 6-hourly meteorological forcing from the NCEP Global Ensemble Reforecast System (GEFS) version two ensemble mean hindcast of precipitation (PRECIP), maximum temperature (TMAX), and minimum temperature (TMIN). Streamflow hindcasts are provided at an hourly timescale and initialized at 12:00 GMT daily. The HEFS model converts a single raw meteorological forecast trace into an ensemble of hydrologic forecasts using an internal Meteorological Ensemble Forecast Processor (MEFP) coupled with a hydrologic processor that incorporates observed and forecast information from hydrologic, snowmelt, and

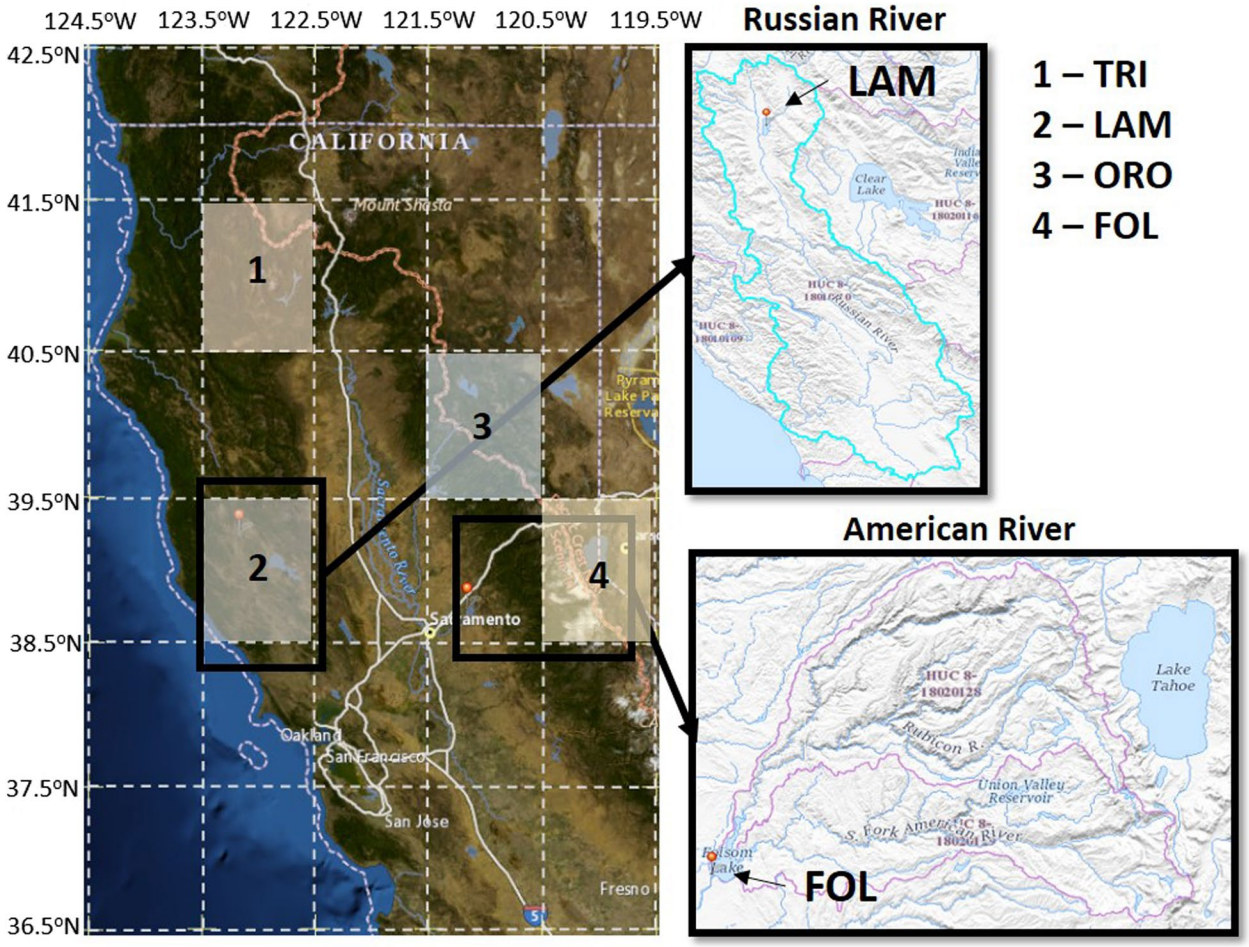

Figure 1. Geographical area of study (northern CA). White dashed lines indicate the 30 grid cells used in the meteorological analysis, where the four highlighted grid cells overlay substantial portions of the reservoir watersheds indicated (TRI–Trinity Reservoir, LAM–Lake Mendocino, ORO–Oroville Reservoir, and FOL–Folsom Reservoir). The regions outlined in black detail the HUC-8 Russian River watershed boundary (cyan) with inflow to Lake Mendocino, and the HUC-8 sub-basin boundaries (pink) of the north and south forks of the American River with inflow into Folsom Reservoir (USGS, 2020).

reservoir models among others (Demargne et al., 2014). For the purposes of this study, we aggregate the hourly HEFS ensemble mean model output to a daily scale between the time period from 08:00 – 07:59 GMT to match the observed FNF data, which is recorded at 00:00 local time. There were some missing data in the HEFS output, most notably in the period from September 16–30 for all years. We estimated these missing values using linear interpolation, since they often occur during times of low flow and little natural variability or for a small number of individual days scattered throughout the record.

Our second synthetic forecast application focuses on meteorological forecasts across 30 grid cells within Northern California (see Figure 1), including four grid cells that overlay the watersheds for Trinity Reservoir (TRI – 40.5° – 41.5°N, 122.5° – 123.5°N), Lake Mendocino (LAM – 38.5° – 39.5°N, 122.5° – 123.5°N), Oroville Reservoir (ORO – 39.5° – 40.5°N, 120.5° – 121.5°N), and Folsom Reservoir (FOL – 38.5° – 39.5°N, 119.5° – 120.5°N). These four locations, which we will focus on in the results, span the eastern and western slopes of the coastal ranges (TRI and LAM, respectively) and the middle and high elevations of the western slope of the Sierra Nevada range (ORO and FOL, respectively).

We obtained data for PRECIP, TMAX, and TMIN from the NOAA-CIRES-DOE twentieth Century Version 3 (20CRV3) historical reanalysis dataset between October 1, 1948 and December 31, 2015 (NOAA PSL, 2020). We use reanalysis meteorology, instead of gauge-based meteorology, for its parity with the NCEP GEFS/R version 2 reforecast model (described below). The reanalysis data are cataloged at the same spatio-temporal scale as the reforecasts (1° × 1°, 6 hourly) and are also produced using the NCEP GFS as the underlying

model, albeit a somewhat newer version (Slivinski et al., 2019). This ensures that underlying physical processes are emulated consistently between the observational and reforecast datasets. We acknowledge that reanalysis data are not true observations, but hereafter use the term "observations" when referring to reanalysis data to ensure consistent terminology between the streamflow and meteorological case studies.

We obtained hindcast meteorological data for forecast lead times of 1–5 days from the NCEP GEFS/R V2 data repository with the same variables and temporal resolution as the observational data, but starting at December 1, 1984 (first available hindcast date). These data come from a single "frozen" version of the GEFS reforecast model across an 11-member ensemble and we used the ensemble mean for all variables (Hamill et al., 2013; NOAA/NCEP, 2013).

3. Generalized Synthetic Forecast Generator

The synthetic forecast generator begins with a matrix (O) of $t = 1,...,n$ observations across $j = 1,...K$ dimensions, where K can denote different lead times (e.g., 1-day ahead, 2-day ahead, etc.), locations (e.g., grid cells or sites at a single lead), or both (e.g., multiple forecast leads and locations). For example, the first row ($t = 1$) of O, associated with a particular date (e.g., January 1, 1985), could contain observations across $K = 10$ lead times (January 1, 1985 through January 10, 1985), $K = 4$ locations (observed data for January 1, 1985 at sites 1 through 4), or $K = 40$ site-lead time combinations. The basic structure of the procedure is a linear additive error model where each element of this multivariate $n\,x\,K$ set of observed data ($O_{t,j}$) is modeled as the sum of the corresponding forecast value ($F_{t,j}$) and an error component ($\varepsilon_{t,j}$), with the allowance that errors may be auto-correlated, heteroskedastic, and non-Gaussian:

$$O_{t,j} = F_{t,j} + \varepsilon_{t,j} \; where \; t \in \left(1,2,\ldots,n\right) and \; j \in \left(1,2,\ldots,K\right) \tag{1}$$

We model and generate errors in three primary steps (see Figure 2). First, we use a vector auto-regressive (VAR) model to account for temporal auto-correlation within each of the K time-series (Section 3.1). We then fit a generalized likelihood (GL) model (Schoups & Vrugt, 2010) to each of the K residual series (ω_t) of the VAR model, accounting for heteroscedasticity and transforming the result to K series of random deviates (a_t) that we model with standardized SGED distributions (Section 3.2). Finally, we model the correlation of each of the K a_t series via an empirical copula and simulate new series of a_t using a K-nearest neighbor (KNN) and Schaake Shuffle approach. We generate synthetic forecast errors using these simulated a_t series by inverting the process of Sections 3.1 and 3.2 (Section 3.3). The remainder of the methods (Section 3.4) describes aspects of the procedure that are specific to given variables analyzed in this study (streamflow vs. meteorology).

3.1. Vector Auto-Regressive (VAR) Model

We use a VAR model with K dimensions and lag order p to model auto-correlation in the forecast errors (Wilks, 2019):

$$\varepsilon_t = \sum_{i=1}^{p}\left(\left[\varphi^i\right]\varepsilon_{t-i}\right) + \omega_t \tag{2}$$

Here, the vector of original forecast errors ($\boldsymbol{\varepsilon}_t$) at time-step t are equal to a linear function of lagged errors ($\boldsymbol{\varepsilon}_{t-i}$), with $K\,x\,K$ matrices of coefficients $\left(\left[\varphi^i\right]\right)$ out to lag order p. Equation 2 reduces to a set of K equations that are solved via linear regression, creating both large parameter dimensions (the number of VAR coefficients φ grow proportional to K^2p) and potential problems arising from high multi-collinearity between lagged terms, including inflated variance and inaccurate significance testing (Elith et al., 2012; Nicholson et al., 2020; Wilks, 2019). A number of regularized methods have been proposed to account for these issues in VAR models (Wei, 2019 and references therein); we utilize the approach of Nicholson et al., (2020) employing a group LASSO penalized model to estimate the regression coefficients while driving redundant parameters to zero. This approach selects the LASSO penalty parameter (λ) based on a rolling cross-validation

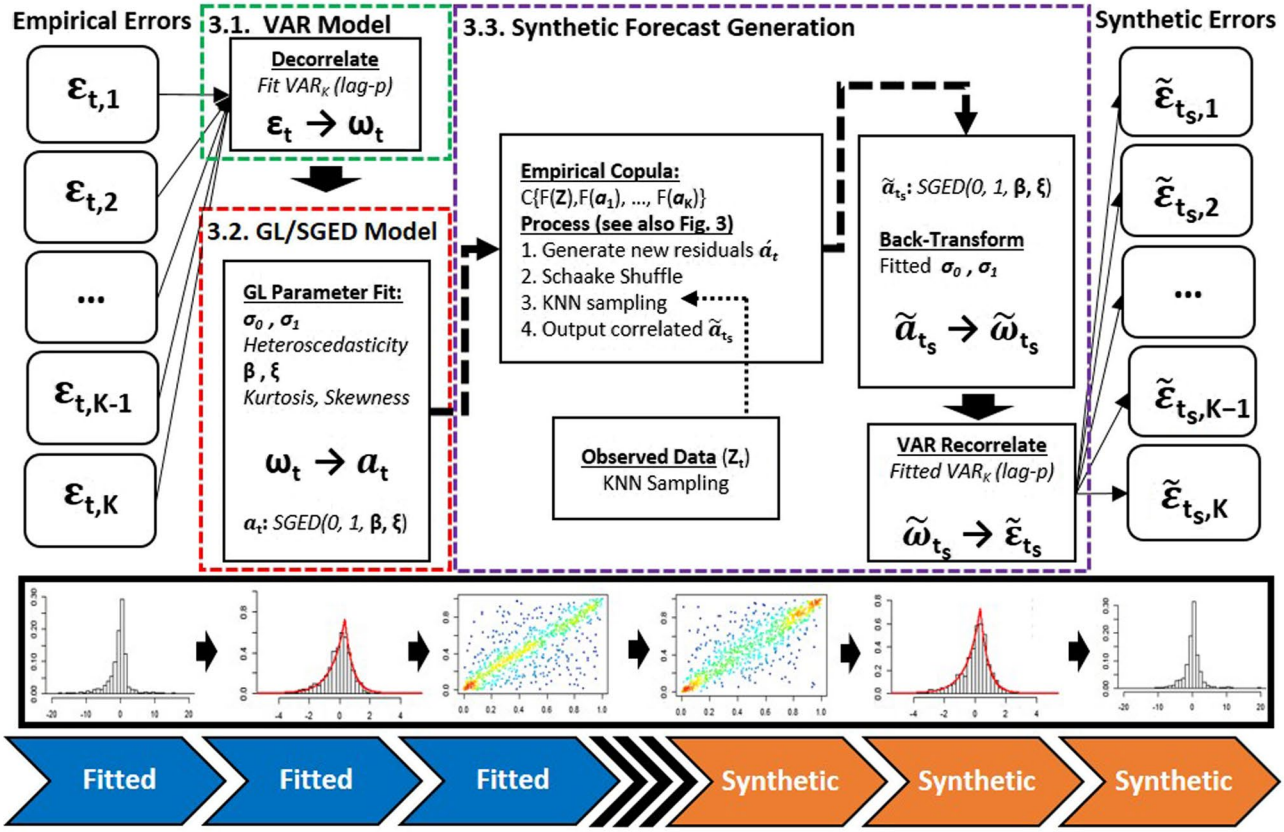

Figure 2. Conceptual model of synthetic forecast generation process, with tilde (~) notation for synthetic outputs. Section 3.1 describes the VAR model decorrelation process. Section 3.2 defines the fitting of the GL/SGED model and output of standardized deviates (a_t). Section 3.3 illustrates the synthetic generation model, with the center section showing the empirical copula approach (described in more detail in Figure 3) to model correlations across the time-series of empirical deviates (a_t) and output new time-series of synthetic deviates ($\tilde{a}_{ts}$). The right-hand section of Section 3.3 then depicts the back-transformation of these deviates to raw errors ($\tilde{\varepsilon}_{ts}$) that are used to create the synthetic forecasts. The bottom section of the figure shows graphical representations of each step in the process and whether those steps are associated with the fitted or synthetic period of the historical data (blue and orange chevrons; see Section 3.3 for more detail).

and out-of-sample mean standard forecast error, helping to stabilize estimation and mitigate overfitting. We fit this model using the R-package "BigVAR" (Nicholson et al., 2019). We used a maximal lag order (p) of 3, which we found was sufficient to reduce autocorrelation while maintaining model parsimony (see Figure S1).

The residuals of the VAR model ($\omega_{t,j}$) are assumed to be the product of a standard random deviate ($a_{t,j}$) and an associated term to capture any temporal variations in the standard deviation ($\sigma_{t,j}$):

$$\omega_{t,j} = \sigma_{t,j} a_{t,j} \tag{3}$$

Models for these terms are discussed next in Section 3.2.

3.2. Generalized Likelihood and Skew Generalized Error Distribution (GL/SGED) Model

We use the generalized likelihood (GL) method of Schoups and Vrugt (2010) to fit each of the K univariate time-series of ω_t to a standardized SGED distribution with corrections for heteroscedasticity. To model heteroscedasticity, the standard deviation at time t ($\sigma_{t,j}$) is modeled as a linear function of the observed value ($O_{t,j}$):

$$\sigma_{t,j} = \sigma_{0,j} + \sigma_{1,j} O_{t,j} \tag{4}$$

The estimated standard deviation $\sigma_{t,j}$ is paired with the VAR model residual $\omega_{t,j}$ to estimate a standardized random deviate $a_{t,j}$:

$$a_{t,j} = \omega_{t,j} / \sigma_{t,j} \tag{5}$$

These deviates $a_{t,j}$ are assumed to follow a SGED distribution with zero mean, unit variance, and parameterized with skew (ξ_j) and kurtosis (β_j) parameters. This is equivalent to the parameterization of the skew exponential power (SEP) distribution in Schoups and Vrugt (2010), although other parameterizations of the SGED are available (e.g., Wurtz et al., 2020). The kurtosis parameter (β_j) in this model can vary continuously between −1 and 1, where values of 1, 0, and approaching (in the limit) −1 correspond to Laplacian, Gaussian, and uniform distributions, respectively. Skewness (ξ_j) can vary continuously between 0.1 and 10 with 1 being a centered distribution and values less than 1 (greater than 1) corresponding to negative skewed (positive skewed) distributions. We use maximum likelihood estimation via numerical optimization to estimate the four parameters ($\sigma_{0,j}, \sigma_{1,j}, \beta_j, \xi_j$) simultaneously. Note that different parameters are estimated for each of the $j = 1,...,K$ series.

3.3. Synthetic Forecast Generation

To generate consistent forecast errors across variables, space, and/or lead time, we must preserve the correlations among the K series of a_t. These correlations reflect the tendency of meteorological phenomena (e.g., frontal systems and atmospheric rivers) and forecasts thereof to organize in space and time, which in turn forces space-time organization in forecast errors. In addition, it is also important to preserve the correlation between the observations and each of the $a_{t,j}$ series (Lamontagne & Stedinger, 2018). This is because actual forecasts tend to underestimate the variance of the observations, particularly if post-processed via model output statistics (MOS, Wilks, 2019). For instance, a post-processed forecast will often under-predict large magnitude events and over-predict small magnitude events to perform well in statistical (e.g., least squares) calibration. However, if forecast errors are assumed to be independent of the observations, then synthetic forecasts generated by adding synthetic forecast errors to observations will have a variance greater than the actual forecasts.

The multivariate relationships between the K series of a_t and the observations may be difficult to model using a parametric approach (e.g., Gaussian or student-t copulas). Empirical copulas, on the other hand, preserve the observed correlation structure exactly, but random samples from an empirical copula will be limited to the range of values observed in the historic record. To address this limitation, we employ a version of the Schaake Shuffle (Clark et al., 2004), which can be interpreted as a type of empirical copula method. The Schaake Shuffle is a procedure where a new matrix of randomly sampled data with the same dimensions as the empirical data is rearranged (shuffled) to replicate the ordering of the empirical data. We use the Schaake Shuffle, coupled with a K-nearest neighbor (KNN) sampling technique, to synthesize series of $a_{t,j}$ outside of their historic range but that exhibit the same rank structure as the original data.

The steps to generate synthetic forecasts are summarized in Figures 2 and 3, the latter of which provides an overview of the procedure for sampling synthetic forecast errors. We use the terminology "fitted period" to refer to the time period (length n) in which the model parameters are estimated and "synthetic period" to refer to the time period (length n_s) over which synthetic forecasts are generated. The fitted period aligns with the period containing available hindcast data, while the synthetic period aligns with the available observational period excluding the fitted period. We use the notation tilde (∼) to refer to data in the synthetic period, an accent to refer to randomly generated data (e.g., $a'_{t,j}$), and the generic variable Z_t to refer to observational data used for KNN sampling. The steps to generate a synthetic forecast are as follows:

1. Rank empirical $a_{t,j}$ values in each of K dimensions for the fitted period ($t = 1,..., n$)
2. Generate a new set of standardized random deviates ($a'_{t,j}$) for each dimension over the fitted period using the fitted SGED distributions
3. Rank $a'_{t,j}$ values and "Schaake shuffle" to match original rank structure from step 1 (i.e., reorder $a'_{1:n,j}$ values so $a'_{t,j}$ will have the same rank as $a_{t,j}$ for all $t = 1,...,n$, separately for each $j = 1,...,K$)
4. For each time step $t_s = 1,..., n_s$ in the synthetic period

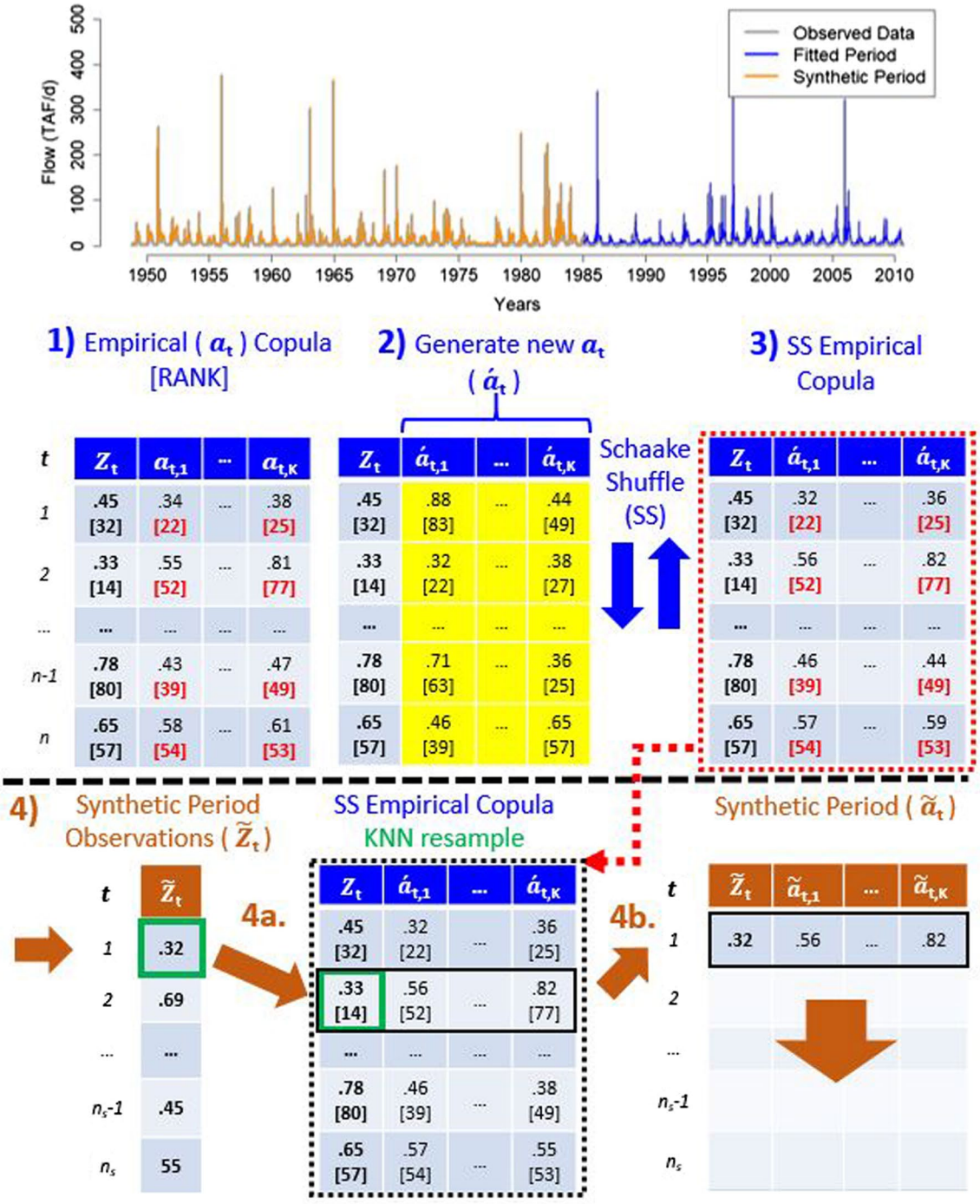

Figure 3. Graphical depiction of empirical copula/Schaake shuffle/KNN resampling procedure from Section 3.3, with large bolded numbers (1,..,4b) corresponding to steps in the text. Top Row–Illustration of partitioning of observed record into "fitted" periods (blue) and "synthetic" periods (orange); color scheme is maintained through remainder of figure. Middle Row–Depiction of steps 1–3 of Empirical Copula/Schaake Shuffle procedure after generation of new random deviates for each of K time series ($a'_{t,j}$). Bracketed values indicate the rank of each deviate, with red bracketed text showing rank structures that are preserved in the Schaake shuffle process. Bottom Row–Illustration of step 4 for each time step t in the synthetic period. Green outline depicts KNN sampling procedure (step 4a) while population of the synthetic period standardized deviate matrix to length n_s is shown in step 4b.

a. Sample via KNN an observation Z_t from the fitted period based on the synthetic period observation $\tilde{Z}_{t_s}$
b. Populate synthetic forecast error matrix with (reordered) $a'_{t,j}$ values associated with Z_t
Result: An n_s *x* K dimension matrix composed of K series of (rank correlated) synthetic random deviates $\tilde{a}_{t_s,j}$ over the synthetic period

5. Back-transform all $\tilde{a}_{t_s,j}$ to $\tilde{\omega}_{t_s,j}$ using Equations 4 and 5, fitted $\sigma_{0,j}$ and $\sigma_{1,j}$, and $O_{t_s,j}$
6. Convert $\tilde{\omega}_{t_s,j}$ to raw forecast errors $\tilde{\varepsilon}_{t_s,j}$ using Equation 2 and fitted VAR coefficients $\left[\varphi^i\right]$
7. Convert $\tilde{\varepsilon}_{t_s,j}$ to synthetic forecasts $\tilde{F}_{t_s,j}$ using Equation 1 and $O_{t_s,j}$
8. Repeat steps 2–7 M times to create M separate synthetic forecasts

In the procedure above, each synthetic forecast (composed of an n_s *x* K matrix of deviates over the synthetic period) is populated with randomly generated standardized deviates ($a'_{t,j}$) that are "Schaake shuffled" to match the original rank structure over the fitted period. A row of these standardized deviates is then sampled and used for time step t_s in the synthetic period. This sampling is based on a KNN approach, whereby the synthetic period observation ($\tilde{Z}_{t_s}$) for time step t_s is used to select a value of Z_t from the fitted period (along with the randomly generated standardized deviates associated with it). The KNN procedure uses a k value of $\sim\sqrt[2]{n}$ (Lall & Sharma, 1996), a discrete kernel function (Steinschneider et al., 2015) to weight the k neighbors for each sample, and a Euclidean distance metric. This approach ensures that the deviate values in the synthetic forecast error matrix ($\tilde{a}_{t_s,j}$) retain the correlation structure of the original empirical copula in the fitted period. The synthetic deviates $\tilde{a}_{t_s,j}$ are then converted back to raw errors ($\tilde{\varepsilon}_{t_s,j}$) and synthetic forecasts ($\tilde{F}_{t_s,j}$) by reversing the procedures in Sections 3.1 and 3.2.

It is important to note that Z_t may refer to the observations directly, or it may refer to transformations of the observed data. In particular, Z_t could be a scalar observation (e.g., observed flow) or a vector of transformed observations (e.g., principal components of precipitation occurrence). This is discussed in Section 3.4. However, when calculating the heteroscedastic components in Equation 4 or manipulating Equation 1, the direct observational data $O_{t,j}$ is used.

3.4. Application to Hydrology and Meteorology

The general methodology for synthetic forecasts above is applied in two case studies: (1) synthetic streamflow forecasts for Folsom Reservoir, CA that emulate the RFC HEFS forecasting system; and (2) synthetic temperature and precipitation forecasts across Northern California. Certain details of the generalized approach differ across these two applications.

For streamflow forecasts, we fit all model parameters separately by month to capture seasonal behavior in forecast errors and the observed FNF data are used directly for KNN sampling (Z_t and $\tilde{Z}_{t_s}$). In the KNN resampling, nearest neighbors are ambiguous when observed FNF values are zero, which occurs often in certain seasons. KNN samples in these instances are chosen randomly from all samples associated with zero observed flow values.

In the meteorological case, initial experiments (not shown) suggested that it was sufficient to fit parameters separately for the cold season (October–March) and warm season (April–September) to account for seasonality in error behavior. These experiments also suggested that the use of observed precipitation occurrence (0/1 for non-occurrence/occurrence) for the given day (t) and day prior ($t-1$) as the basis for selecting nearest neighbors led to good synthetic model performance. For this study, we consider 30 grid cells across the Northern California region, and create a sampling matrix of dimension n *x* 60 (n *x* {30 [day t] observations + 30 [day $t-1$] observations}). We reduce this matrix to 10 dimensions (91.8% proportion of variance explained) using logistic principal component analysis (logistic PCA; Landgraf & Lee, 2015), which we implement with the "logisticPCA" R-package (Landgraf, 2016). Then, for each precipitation occurrence observation from the synthetic period (i.e., $\tilde{Z}_{t_s}$, a vector of 10 PC values), the KNN algorithm is used to select a precipitation occurrence observation from the fitted period (Z_t, also a vector of 10 PC values), along with the associated standardized deviates ($a'_{t,j}$). As in the streamflow example, precipitation occurrence observations when there is no precipitation in any grid cell for day t or $t-1$ are randomly sampled from

all days in the fitted period where the same condition is observed. Also importantly, we impose the occurrence-based structure from each KNN sample on our synthetic forecast. That is, whatever grid cells had forecasted non-zero precipitation from the resampled day in the fitted period (Z_t) are also assumed to have forecasted non-zero precipitation for the associated synthetic day ($\tilde{Z}_{t_s}$). Then, for those grid cells with forecasted non-zero precipitation, we develop synthetic forecasts of precipitation magnitude based on synthesized deviates and the procedures in Section 3.3. This approach enables a straightforward way to capture realistic proportions of true positives/negatives and false positives/negatives from the associated forecasts. In cases where the synthetically generated errors produce a negative precipitation forecast, we resample until a non-negative result is produced.

We note that NCEP GEFS/R V2 temperature forecasts exhibited consistently biased behavior, particularly TMIN. To improve modeling, we subtract these biases based on a monthly mean, model the resultant unbiased forecast errors, and then add the biases back in when creating the synthetic forecasts. Finally, computation time to fit the model and generate 1,000 synthetic samples was approximately an hour for the synthetic streamflow forecast example and ~48 h for the higher dimension synthetic meteorological forecast example on a Dell OptiPlex Desktop with an 8-core, 3.00 GHz i7 processor in a non-parallel configuration.

4. Results

4.1. Synthetic Streamflow Forecasts

We first analyzed model performance against HEFS inflow forecasts across 1–10 days lead times ($K = 10$) for Folsom Reservoir; results for Lake Mendocino are shown in the supporting information and discussed later. Synthetic streamflow forecast models were developed separately by month, but in the results below, we selectively highlight model behavior in representative months across the year. Figure 4 shows error behavior and some components of model fit in January and July for 1-day and 5-day lead times. Similar results for other months are presented in the supporting information (Figures S2–S5). The top two rows of Figure 4 show the distribution and autocorrelation function of the raw errors (ε_t). For both months and lead times, the raw errors are skewed and leptokurtic. The errors in January exhibit a larger range, less autocorrelation, and a clear left skew, indicating a tendency toward over-prediction. The error range is consistent with greater and more frequent precipitation in January that increases the chances for large streamflow errors. Conversely, in July, errors are less skewed and much more persistent, reflecting the base-flow dominated hydrology typical of the warm season.

The standardized deviates (a_t) for both months and lead-times exhibit a similar distribution as the raw errors (Figure 4, row 3), with a fat-tailed Laplacian distribution ($\beta \approx 1$) and some amount of negative skew ($\xi < 1$). Both the heteroscedastic intercept (σ_0) and scaling coefficient (σ_1) terms are substantially higher in January than July, suggesting greater baseline variability and conditional heteroscedasticity in that month. Though not shown, autocorrelation in the a_t series has mostly been removed via the VAR model. The most notable difference between the a_t distributions across the two months is the more prominent negative skew in the January deviates, which follows the clear negative skew in the raw errors. In both cases, the fitted SGED pdfs appear to fit the data well, but we also confirm goodness-of-fit (GoF) visually through P–P plots (Figure S6), albeit with some disparities in January at 5- and 10-days forecast leads.

The final two rows of Figure 4 illustrate the complicated relationship between the observed streamflow values and the a_t series (row 4), as well as the correlation between the a_t series at different lead times (row 5). These relationships are shown in terms of the empirical non-exceedance probabilities (NEPs) for all variables. The NEP values for the a_t series at one- and 5-day lead times are clearly and strongly correlated for both months, although the dependence is weaker in January but with more upper and lower tail dependence (row 5). In contrast, the relationships between the NEPs of observed flow and the a_t series at different leads times exhibit clustered and asymmetric behavior (row 4). These error clusters are unique to the calibration of the HEFS model and should be captured in synthetic forecast generation, which motivates our use of an empirical copula in the generation process (see Section 3.3).

Figure 5 more clearly shows the seasonality of parameters in the GL/SGED model (and hence error structure) at lead times of 1, 3, 5, and 10 days. Seasonality in the heteroscedastic intercept (σ_0) is consistent across lead times throughout the warm season (AMJJAS), but diverges substantially in the cold season (ONDJFM),

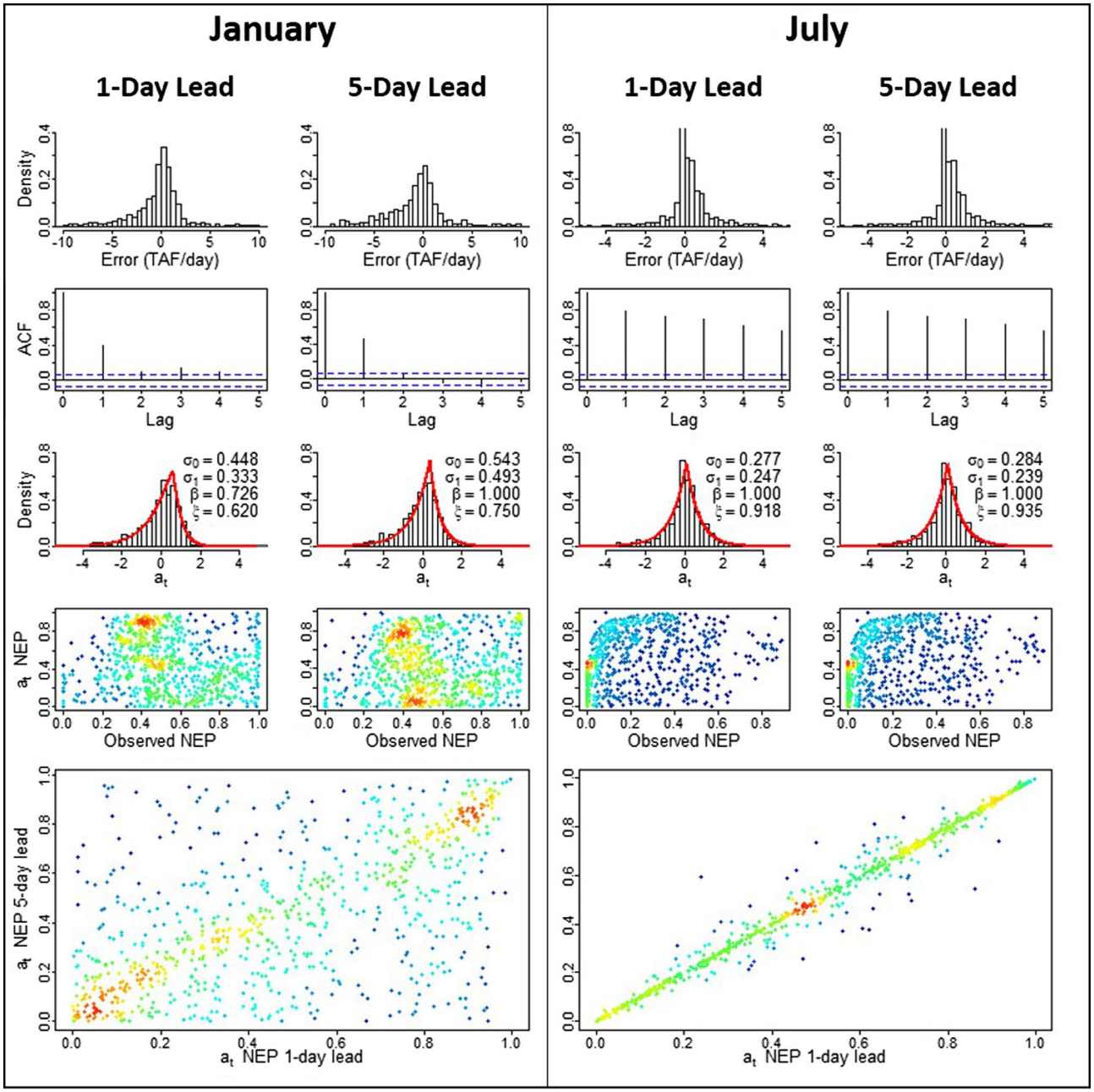

Figure 4. Graphical depiction of forecast error behavior for Folsom Reservoir hydrologic forecasts for January (July) in the left (right) halves of the figure. The top four rows show error behavior at 1-day (5-day) leads in the left (right) halves of the monthly sub-sections. The top row shows the empirical distribution of raw forecast errors, and the second row shows the autocorrelation function (ACF). The third row shows the empirical distribution of standardized deviates (a_t), with the red line showing the fitted SGED ($0, 1, \beta, \xi$) pdf and fitted parameters indicated in black text. The fourth row shows scatter density plots of observed flow non-exceedance probabilities (NEP) versus a_t NEPs, where red (blue) coloration indicates high (low) density. The bottom row shows a scatter density plot of a_t NEPs between 1- and 5-day leads.

with larger values more common at longer lead times. In general, there is higher static error variance in the cold season when storms are frequent. The heteroscedastic scaling term (σ_1) is very similar across lead times, and generally is lower during months when flows are higher due to snowmelt.

The SGED kurtosis parameter (β) remains at or near 1 (Laplacian distribution) for most months and lead times, with the most noticeable exception being the 1-day lead forecasts that decrease below 1 (i.e., become more Gaussian) during the cold season. This reflects a higher probability for small to moderate forecast

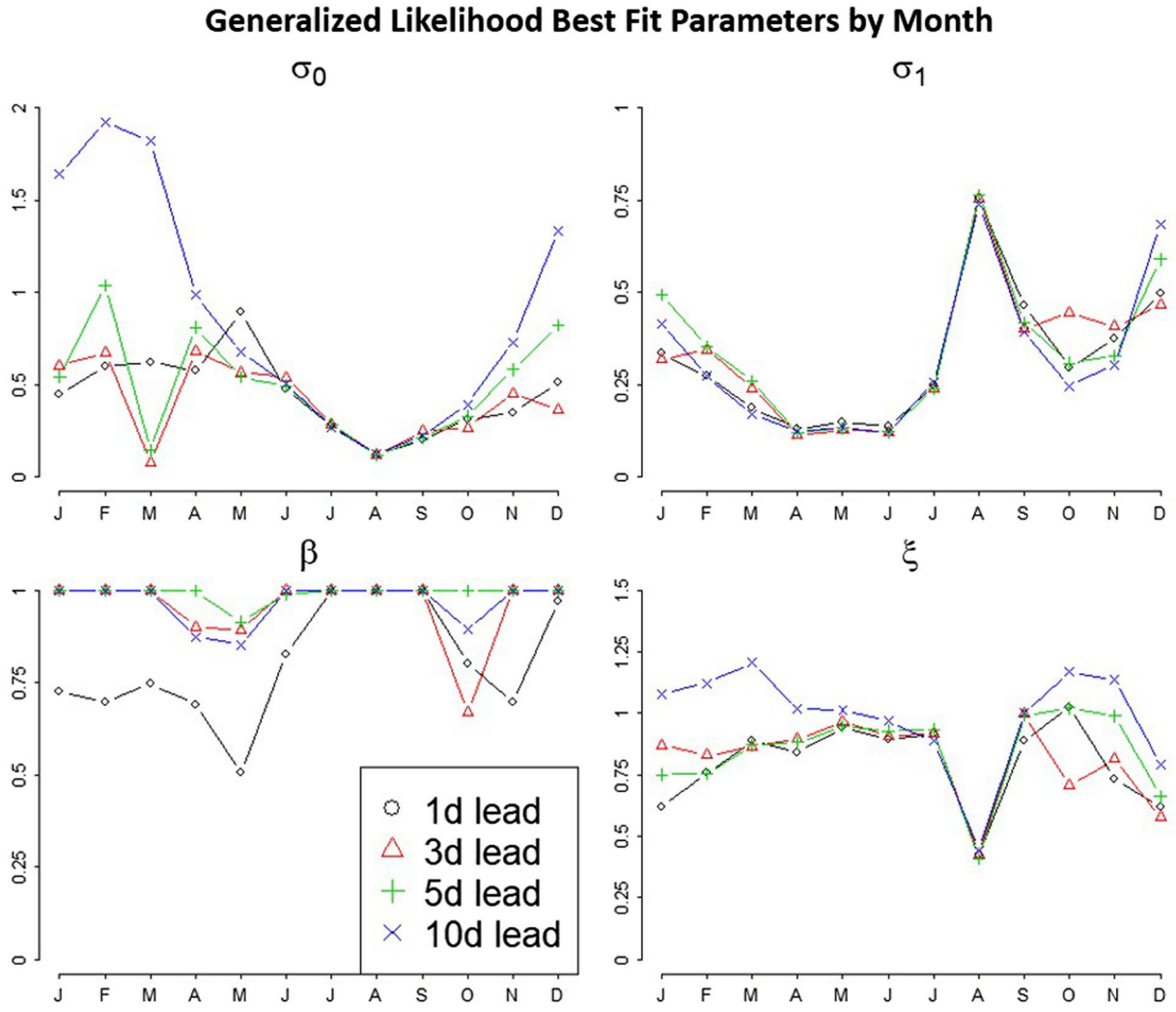

Figure 5. Fitted values (*y*-axis) by month (*x*-axis) for the four parameters of the GL/SGED model across 1, 3, 5, and 10 days of forecast leads. σ_0 and σ_1 are the intercept and slope parameters for heteroscedasticity, respectively, while β and ξ are the kurtosis and skewness parameters of the standardized SGED distribution.

errors at a 1-day lead, especially during the more variable cold season months. At 1, 3, and 5 days of lead times, the SGED skewness is generally negative ($\xi < 1$). At a 10-day lead, the skewness disappears or becomes slightly positive ($\xi \geq 1$). Finally, we note that the skewness parameter and heteroscedastic scaling term (σ_1) are negatively correlated, which likely explains some of the aberrant spikes in these parameters during certain months (August, December).

To assess performance of the model for streamflow forecasts, we create 1,000 synthetic forecasts over the fitted period [when hindcasts are available, 1985–2010 (25 years)] and compare them to the actual HEFS forecasts. Figure 6 shows the distribution of these synthetic forecasts (expressed as 50% and 95% prediction intervals) for four months in 1986, as well as a single synthetic forecast trace. Similar results for a year (1955) in the synthetic period [when hindcasts are unavailable, 1948–1985 (37 years)] are shown in supporting information (Figure S7). Several results emerge from Figure 6. First, we note that across all months and lead times, the model preserves cross-correlation in forecast error structure. For example, on January 17, 1986, a large spike in the observed inflow is systematically under-predicted by the 1, 3, 5, and 10 days actual HEFS forecast (HEFS-sim). This behavior is reflected by the synthetic forecast trace (HEFS-syn) that also under-predicts across all lead times. In addition, the model captures important relationships between the observed flow and forecasts, including the general tendency for forecasts to under-predict large, infrequent

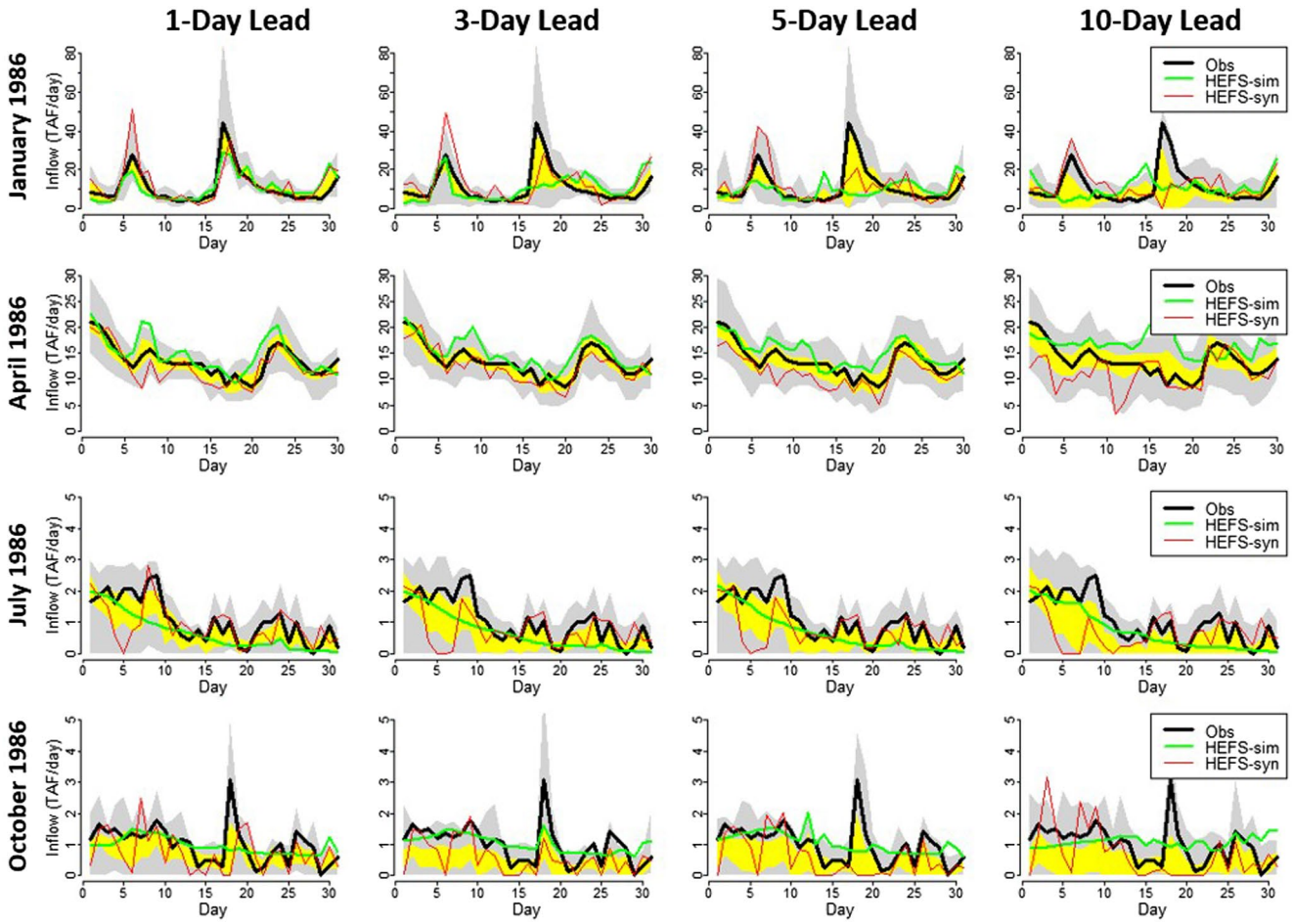

Figure 6. Folsom Reservoir (FOL) hydrographs from a selected year (1986) for four different months spaced evenly across the year showing synthetic forecast performance at 1-, 3-, 5-, and 10-day forecast leads. Observed full natural flow is indicated by the black solid line, the HEFS ensemble mean forecast by the green solid line, and a randomly sampled synthetic forecast by the red solid line. Light gray (yellow) shading indicate the 95th (50th) percentile bounds from 1,000 synthetic forecast samples.

events, especially at long lead-times. This is seen most clearly by the 50th percentile bounds that are depressed below the highest observed flow values, and is also confirmed through direct comparisons between observed flow and forecast residuals for both the empirical and synthetically generated data (see Figure S8).

Figure 6 also highlights how the synthetic model preserves auto-correlation in the forecasts. For instance, at a 5-day lead in April, the actual HEFS forecast (HEFS-sim) shows persistent over-predictions across the April 5–25 interval, while the synthetic forecast trace (HEFS-syn) shows a similar degree of persistence but for under-predictions. Autocorrelation appears to increase with lead time, especially in months driven by snowmelt.

The 10-day lead-time example for January 1986 displays a noteworthy limitation of the modeling structure. The empirical copula maintains the correlation between observed flows and forecast errors, which generally leads to forecast under-predictions for large inflow events. When coupled with the autocorrelation structure imposed by the VAR model, this tends to drive even greater under-predictions in the following time steps. For very large inflow events, this can cause the synthetic forecasts to be unrealistically low, as shown by the January 10-day lead synthetic forecast trace (HEFS-syn) reaching zero flow.

For the remaining two months (July and October, row 3–4), we note many of the same characteristics as for January and April. However, the actual HEFS forecasts in these months display smoothed behavior and are increasingly detached from the variability in the observed inflow. The uncertainty bounds capture the actual forecast traces reasonably well, but the physical behavior of the synthetic forecast trace, which is tied

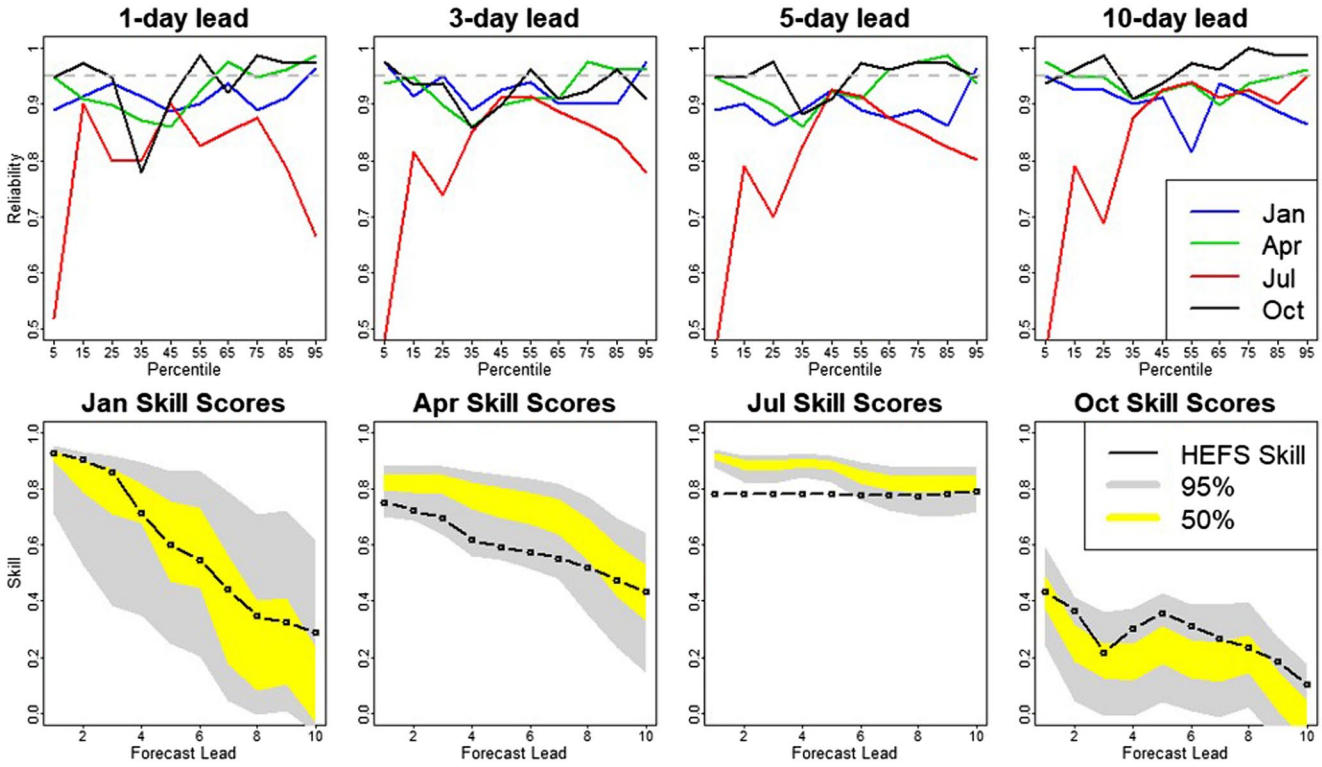

Figure 7. Top row–Reliability plots for 95% uncertainty bounds and 1, 3, 5, and 10-days forecast leads across four selected months. The dashed gray line indicates the target reliability of 95% while the *x*-axis labels show the center of each of the 10 percentile bins (e.g., "5 percentile" indicates 0–10 percentile values of the observed flow). Bottom row–Skill score based on mean squared error. The black line shows the HEFS ensemble mean forecast skill across the 10 forecast leads while the gray (yellow) shading indicate 95th (50th) percentile bounds across 1,000 synthetic forecast samples.

to the observations, is substantially different than that of the smooth HEFS forecasts. We note, though, that the flow magnitudes in these months are rather low, so the practical implications of these differences in forecast behavior are likely small.

To more systematically validate synthetic forecast performance, Figure 7 shows synthetic forecast reliability and skill. Reliability refers to the frequency that the HEFS forecasts lie within the 95% synthetic forecast bounds (which should be 95% if the synthetic forecasts were generated correctly). We assess reliability for different observed flow ranges, discretized into percentile bins (0–10th percentile, 10–20th percentile, etc.). Across lead times, January, April, and October reliability is generally near the 95% target across flow percentiles, albeit at times slightly below. The 1-day lead forecasts are slightly less reliable than the other lead times, particularly at low to moderate flows, which may be tied to the slightly more Gaussian (i.e., less fat-tailed) fits noted above for this lead-time that would lead to tighter uncertainty bounds. The July reliability diverges substantially from the other months and is usually well below 95%. Most notably, reliability is low at both the lowest and highest flows. The HEFS forecasts exhibit smoothed and sometimes biased behavior in these low-flow months (see Figure 6). This could lead to extended periods where observed flows are at or near zero and the actual HEFS forecasts are biased above the synthetic forecast uncertainty bounds, explaining the low reliability at low flows. During rare high flows in July the HEFS forecasts often substantially under-predict, and when coupled with the low error variance for this month, synthetic forecasts tend to be less able to capture these under-predicted HEFS forecasts.

Next, we assess skill in the synthetic forecasts compared to the actual forecasts using a common mean squared error (MSE) climatological skill score ($SS_{clim} = 1 - \frac{MSE}{MSE_{clim}}$; Wilks, 2019). This score captures the ability of the forecasts to outperform a climatological forecast, which in this case is a 7-day rolling average for each day of the year across the observational record. A value of one is a perfect forecast, a value of 0 is equivalent to climatology, and a negative value is worse than climatology. Figure 7 compares whether the

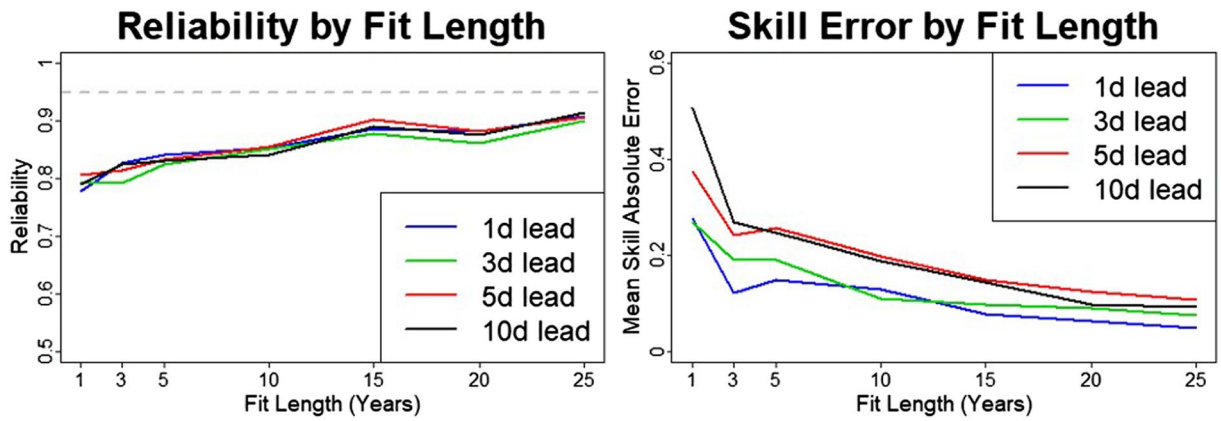

Figure 8. Left–Comparison of total reliability across the 1985–2010 period for selected lead times based on differing fit lengths. The dashed gray line indicates the target reliability of 95%. Right–The absolute difference between the ensemble mean skill score of the synthetic forecasts and the actual HEFS skill score for the same period, lead times, and fit lengths as left panel.

synthetic forecasts match the skill of the actual HEFS forecasts for different lead times and months. January skill for the synthetic forecasts shows good correspondence to the actual forecasts, with the actual forecast skill remaining well within the 50th percentile bounds. April and July synthetic forecasts have too much skill compared to the actual forecasts at shorter lead times but capture the actual forecast skill more accurately at longer leads. The overestimation of skill is particularly prevalent in July, when synthetic forecast reliability is also lowest. October underestimates skill relative to the actual forecasts, although the actual forecast skill still lies within the 95% bounds of the synthetic forecasts. Overall, the skill in the synthetic forecasts generally reflects that in the actual forecasts across months and lead times, but with some deviations that are specific to different times of year.

We performed two final streamflow forecast analyses to assess model generalizability. First, we fit and validated the model on HEFS streamflow forecasts for the Russian River inflow to Lake Mendocino (see Figures S9–S13). The results were largely similar to that of Folsom Reservoir, even though the Lake Mendocino HEFS forecast skill was substantially lower than that of Folsom Reservoir for all months tested. Second, we assessed the sensitivity of our model to varying lengths of available hindcast data by fitting our model to 1, 3, 5, 10, 15, and 20 years subsets of the data (all starting in 1985) and validating on the full dataset for reliability and skill score performance (Figure 8). For reliability, we note a near monotonic increase in total reliability (i.e., across all months) for each of the lead times shown. In the right panel, we show the difference between the ensemble mean estimate of the skill score for the synthetic forecast output and the actual HEFS skill score. Similar to reliability, the difference between the synthetic and HEFS skill scores decreases near-monotonically with fit length. For both metrics, only modest gains in performance occur beyond a 15 years fit length, indicating that this is likely a sufficiently long training period to attain reasonable performance. Additionally, parameter estimates are unstable for fit lengths less than 5 years (see Figures S14–S16).

4.2. Synthetic Meteorological Forecasts

The meteorological case study illustrates a much higher dimensional problem, as we model 3 variables (PRECIP, TMAX, TMIN) across 30 grid cells at 5 lead times, resulting in $K = 450$ dimensions. We split the meteorological data into cold-season (ONDJFM) and warm-season subsets (AMJJAS) for model fitting, and focus our results on synthetic forecast performance during the cold season at four grid cells that overlay key watersheds (TRI, LAM, ORO, and FOL; see Figure 1). Figure 9 shows error behavior and some components of model fit in the cold season for two of the three variables (PRECIP, TMAX) at Lake Mendocino (LAM), since TMIN behaves in a similar manner to TMAX. Plots for additional variables, sites, and lead times are shown in Figures S17–S22, where we note that inter-site correlations among the same variables appears to be much stronger than inter-variable correlations at each site. In Figure 9, PRECIP standardized deviates (a_t) for the LAM site show similar distributional qualities to those for streamflow, while TMAX

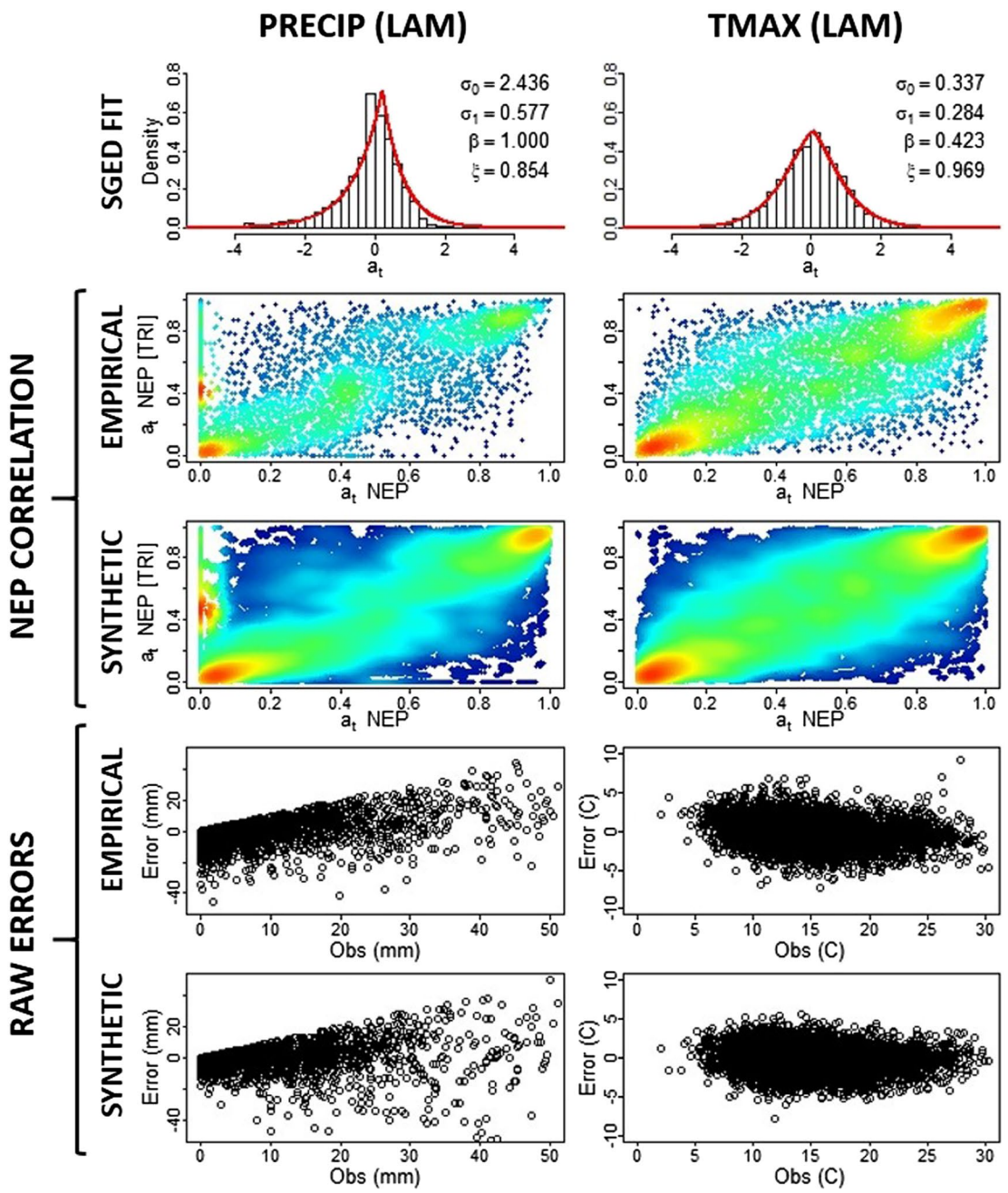

Figure 9. Forecast error behavior for precipitation (PRECIP–left column) and maximum temperature (TMAX–right column) at a 5-days forecast lead for the LAM watershed grid cell. Top row - The empirical distribution of standardized deviates (a_t), with the red line showing the fitted SGED ($0, 1, \beta, \xi$) pdf and fitted parameters indicated in black text. Second and third row–Scatter density plots between a_t NEPs at the LAM and TRI grid cells for empirical (synthetic) data in the second (third) row, where the third row shows the density of 100 synthetic samples. Fourth and fifth row–Empirical (synthetic) forecast errors plotted against observed values for in the fourth (fifth) row. Synthetic forecast errors are shown for a single synthetic forecast sample.

standardized deviates are more Gaussian (top row). The SGED model is able to capture this behavior well, and goodness-of-fit is consistent across sites and variables (see Figure S23). Both forecasts show some level of conditional heteroscedasticity ($\sigma_1 > 0$), though TMIN and TMAX are sometimes fit with no conditional heteroscedasticity at other locations/lead times. The relationship between empirical a_t NEP values at LAM and TRI (row 2) show symmetric tail-dependence typical of a t-copula (Chen & Guo, 2019) for TMAX, while PRECIP shows more pronounced lower tail dependence and other asymmetric behavior. The synthetic forecasts capture this behavior well (row 3). The final two rows of Figure 8 show raw errors versus observed

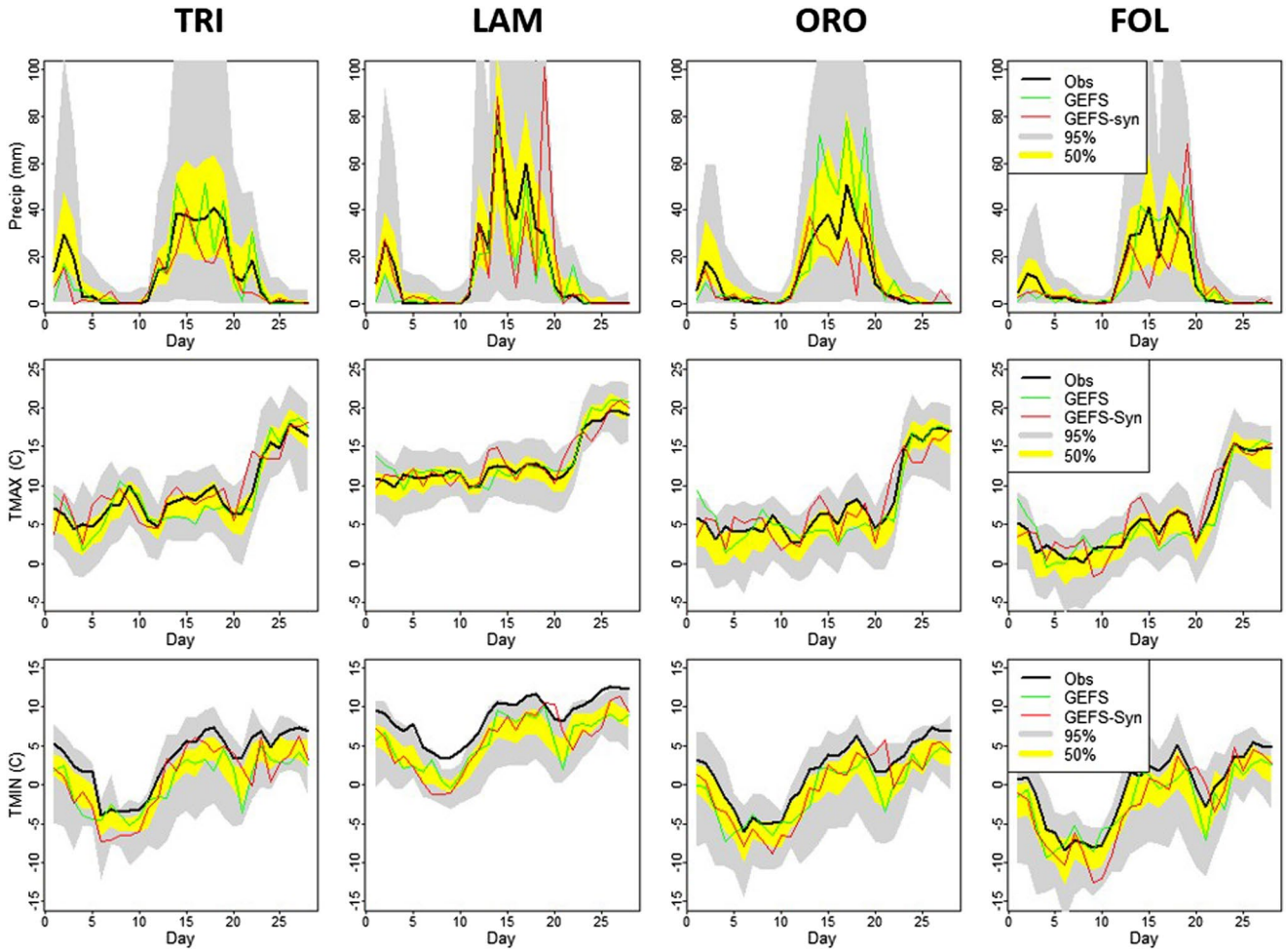

Figure 10. Synthetic forecast performance for February 1986 across four selected watersheds at a 5-day lead for precipitation (PRECIP), maximum temperature (TMAX), and minimum temperature (TMIN). Observed values are indicated by the solid black line, the GEFS ensemble mean forecast by the solid green line, and the synthetic forecast (single random sample) by the solid red line. Gray (yellow) shading show 95th (50th) percentile bounds across 1,000 synthetic forecast traces.

precipitation and temperature for the original hindcasts (row 4) and the synthetic forecasts (row 5). For both variables, the model accurately preserves much of the relationship between the raw errors and the observations. We do note though that the synthetic errors for PRECIP show slightly less variability at low observed values than the empirical errors and vice versa for moderate to high observations.

Similar to our approach for streamflow, we assess synthetic meteorological forecast performance using 1,000 synthetic forecast traces over the fitted period [1985–2015 (30 years)]. Figure 10 shows the distribution of these synthetic forecasts for PRECIP, TMAX, and TMIN in February 1986 at a 5-days lead. Similar results for another month and year (December 1955) from the synthetic period [1948–1984 (36 years)] are shown in the supporting information (Figure S24). In Figure 10, we note similar cross-correlated behavior in the forecasts to those of streamflow, except in this instance the correlations are spatial. For example, at all sites, the sampled synthetic PRECIP forecast trace (GEFS-syn) primarily underestimates the observed event from February 13–17, and then at three sites (LAM, ORO, and FOL), it overestimates the observations between February 18 and 19. This shows how the synthetic forecast trace captures the synchronized error in event timing across locations. This is also shown spatially for this event in Figure S25.

The TMAX and TMIN GEFS forecasts exhibit less variable behavior, which is well captured by the synthetic forecast model. Cross-correlations still exist (note the over-prediction of TMAX near February 14 and 22 across sites). There is some moderate negative bias in the TMIN forecasts that is especially evident at

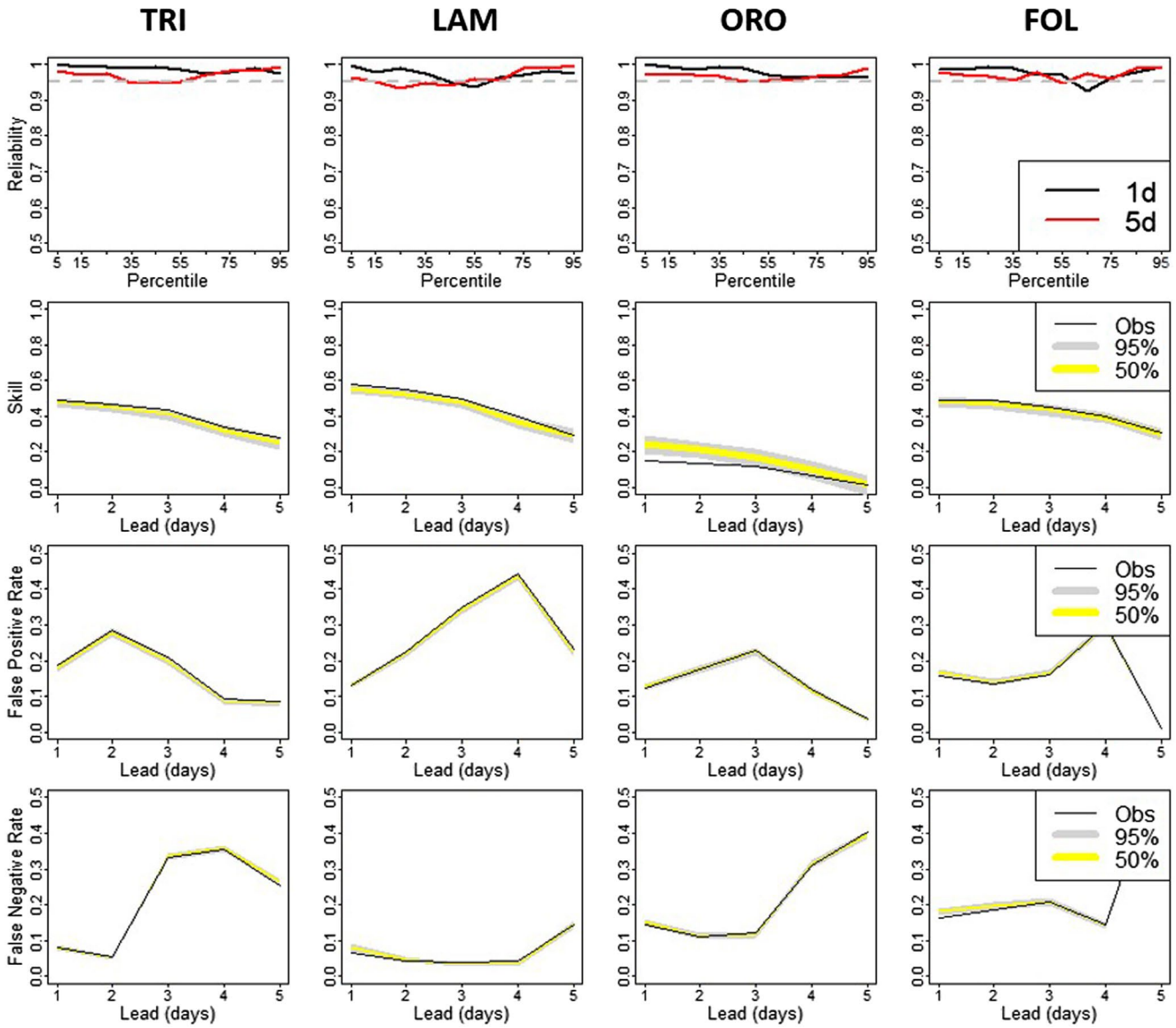

Figure 11. Precipitation (PRECIP) forecast metrics across 4 selected watershed grid cells. Top row–Reliability plots for 95% uncertainty bounds and 1- and 5-days forecast leads. The dashed gray line indicates the target reliability of 95% while the *x*-axis labels show the center of each of the 10 percentile bins (e.g., "5 percentile" indicates 0–10 percentile values of the observed flow). Second row–Climatological skill score by mean absolute error (SS_{MAE}) across 5 forecast leads, with the solid black line showing observed skill and gray (yellow) shading indicating 95th (50th) percentile bounds of 1,000 synthetic forecast traces. Third and fourth row: As in row 2, but showing false positive (FP) and false negative (FN) rates, respectively.

the LAM location, which the synthetic forecasts are able to capture through a simple bias correction (see Section 3.4).

Figure 11 shows forecast validation metrics across sites for PRECIP, where reliability and skill are calculated only for nonzero precipitation days. Reliability for precipitation (top row) is generally above the 95% target for all sites, especially for the lower percentile bins, suggesting the synthetic forecasts are somewhat over-dispersed. Some slight dips in reliability occur across sites and lead times near the middle percentile bins, but the only noticeable excursions below the 95% line occur at LAM (1 and 5-day lead) and FOL (1-day lead). PRECIP forecast skill (second row) is summarized using a mean absolute error climatological skill score, which is similar to the metric used for streamflow but with absolute instead of squared errors (Wilks, 2019). The synthetic bounds for precipitation skill closely match that of the actual forecast, with only one exception of slightly over-estimated skill at shorter lead times for the ORO site. Uncertainty

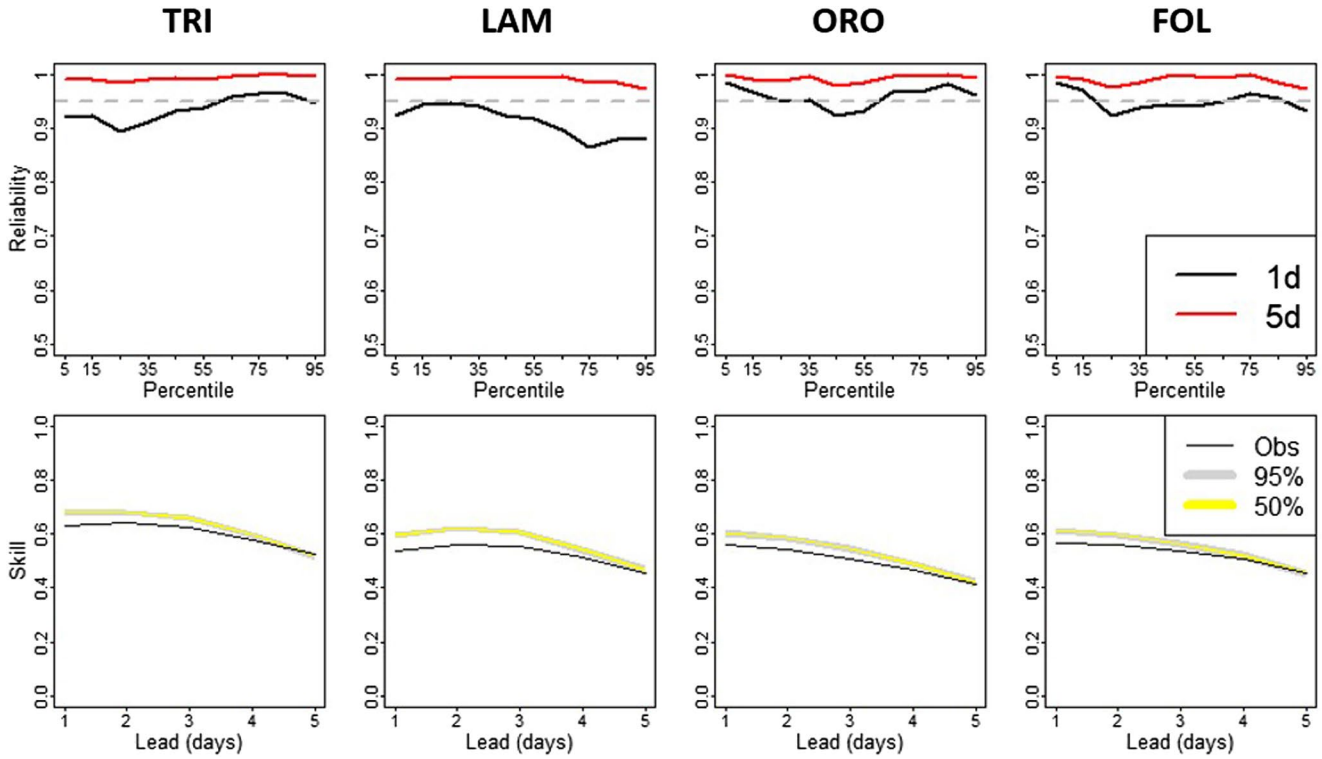

Figure 12. Maximum temperature (TMAX) forecast metrics across 4 selected watershed grid cells. Top row–TMAX reliability, as in top row of Figure 11. Bottom row–SS_{MAE} for TMAX, as in row two of Figure 11.

bounds for the skill metric are much tighter than those for the streamflow case study (Figure 7), largely because the sample size of observations is much larger for the meteorological case study (all data within the cold season, rather than just one month). Finally, we assess the ability of synthetic precipitation forecasts to replicate false positive and false negative PRECIP rates in the last two rows of Figure 11. As noted in Section 3.4, these occurrence-based attributes of precipitation are sampled along with the synthetic a_t values, so we are primarily validating the meteorological KNN sampling procedure. Both false positive (row 3) and false negative (row 4) behavior is maintained accurately.

Figure 12 shows reliability and skill for TMAX forecasts. Reliability is somewhat higher than the 95% target at a 5-day lead at all sites (i.e., over-dispersed) but is near or sometimes slightly below the target at a 1-day lead. The pattern of forecast skill across lead times is also closely matched by the synthetic TMAX forecasts, although biased slightly positive. Again, the uncertainty bounds are very tight in these skill figures due to a large sample size. Overall, Figures 11 and 12 suggest that the synthetic meteorological forecasts accurately preserve many of the properties of the empirical hindcasts at multiple variables, sites, and lead times.

5. Discussion

In the streamflow case study, we found that synthetic forecasts performed better (i.e., more accurately captured hindcast behavior) in the cold season months (ONDJFM) when observed streamflow was greater and more variable. The correspondence between the actual HEFS forecasts and observations deteriorated in the warm season (AMJJAS), causing a corresponding decrease in synthetic forecast performance. In general, our empirical copula and KNN sampling approach preserved key statistical relationships between the observations and the forecasts and between forecasts at different lead times. However, some challenges did remain; for instance when autocorrelation in forecast errors during large, infrequent flow events led to some unrealistically low synthetic forecasts during those events.

In our meteorological case study, we produced correlated forecasts of PRECIP, TMAX, and TMIN across locations and lead times. We found that our methodology readily adapted to the different distributional

forms found in these variables and that synthetic output performed well against the actual forecasts. We also found that our sampling procedure, tailored specifically to capture the occurrence-based statistics of PRECIP, enabled an accurate representation of false positives and false negatives in the synthetic forecasts.

There are a number of important assumptions in our approach that require discussion. First, we assumed that the statistical properties of the forecast errors are stationary, in part because we fit our model to reforecasts from a single "frozen" version of the GEFS and HEFS models. However, even with a single version of the physical forecast models, the distribution of errors may vary as a function of hydroclimate regime in ways not currently captured by our approach, let alone if those physical models are updated over time. Another assumption in our model is that the VAR component of the modeling framework primarily captures auto-correlation in forecast errors, while the empirical copula primarily captures cross-correlations across lead times, locations, and variables. In actuality, the VAR model can capture both auto-correlation and cross-correlation in the errors and we do not distinguish in our results between cross-correlations captured by this portion of the model and those that are captured by the empirical copula. Furthermore, we used an empirical copula, Schaake shuffle, and KNN sampling approach to preserve complex correlation structures in the data. However, a parametric copula or multivariate kernel density estimate (Scott, 1992) of the empirical copula could also be employed, and this may allow for a richer characterization of forecast uncertainty across space, variables, and lead times. There are also other methods for bias correction, heteroscedastic modeling, and observational scaling that could be considered in future applications (see Schoups & Vrugt, 2010).

One important choice we made in our approach was to model ensemble mean forecasts from the HEFS and GEFS, rather than modeling the entire forecast ensemble. This choice was motivated by past work (Christiansen, 2018; Rougier, 2016; Thompson, 1977) showing that the ensemble mean is more skillful than any single ensemble member. Additionally, we model the uncertainty in the ensemble mean forecast based on its errors in the hindcast period, reducing the need to recreate the entire forecast ensemble. Still, certain applications may benefit from synthetic generation of the entire ensemble, and this should be the focus of future work. These synthetic approaches will have to contend with additional challenges, e.g., dynamic forecast traces that exhibit state dependent variations in predictability (Palmer, 1993; Wilks, 2019) and regime structures that can cause clustering among ensemble members. The ensemble mean forecast tends to smooth out this behavior (Wilks, 2019), providing an easier target for synthetic forecast generation.

Additionally, we chose to fit our synthetic forecast models to subsets of the data (monthly for streamflow and cold/warm season for meteorology). This partitioning was satisfactory for our purposes given the specific behavior of the reforecasts in our application, but these choices are subject to discretion and are not prescriptive. In addition, much shorter lengths of available hindcast data will affect how this choice impacts synthetic forecast accuracy and precision, as shown in Figure 8. More generally, our approach does not consider the uncertainty in parameter estimates of the VAR or SGED components, which if propagated, could improve synthetic forecast reliability (Stedinger & Taylor, 1982). The relatively long records (25 years of daily data) used for model fitting in our application likely reduced the impact of parameter uncertainty, but this impact will grow with shorter hindcast periods or if the model is fit to smaller subsets of the data. In these cases, formal methods to estimate and propagate parameter uncertainty should be considered (Kuczera, 1983; Schoups & Vrugt, 2010; Sorooshian and Dracup, 1980).

6. Conclusions

This study contributes a generalized methodology to create synthetic forecasts from observed data that preserve empirical space-time and inter-variable relationships in forecast errors. The methodology is adaptable to a wide variety of distributional forms and explicitly accounts for auto-correlation and heteroscedasticity in the forecast errors. We demonstrated short-to-medium range synthetic forecast generation in two case studies that highlighted the ability of the model to simulate accurately streamflow and meteorological forecasts across multiple lead times and locations from modern operational forecast systems. The two applications highlighted the model's potential for developing long records of streamflow forecasts via either the direct statistical approach or the conceptual hydrologic model approach commonly used in designing, testing, and validating forecast informed water management policies (Lamontagne & Stedinger, 2018).

Future work should explore how the design of forecast informed policies depends on the forecast data used for training and testing, including data from the original hindcasts as well as different synthetic forecast methods. In addition, it will be important to test how sensitive policy designs are to deficiencies in the synthetic forecast algorithm (e.g., those seen in our hydrologic case study for low streamflows). We hypothesize that the impact of statistical artifacts in synthetic forecasts will depend on the application, e.g., poor synthetic forecasts at low flows would be inconsequential for flood risk operations but important for ecological flow management. However, these consequences should be tested, as no statistical model can capture all the nuanced behavior of a deterministic model and the consequences of these deviations should be quantified.

The flexibility of our modeling framework directly addresses critiques that research efforts in water resources often suffer from a lack of generalizability across spatio-temporal scales (Brown et al., 2015; Lall, 2014). Our model captures correlations across space and time explicitly, although there may be a limit to the number of dimensions that can be modeled, especially when available reforecast data is limited (Wilks, 2019). For multiple sites within regional or basin scale settings with ~20–40 years of reforecast data typically available in CONUS, this is unlikely to be a limitation. However, the number of dimensions that can be modeled may be constrained with substantially shorter data sets, a very large number of sites or lead times, or larger spatial domains. The use of moving spatial windows offers one possible approach to address these types of dimensionality constraints (Voisin et al., 2011).

In addition to the case studies presented here that focused on using synthetic streamflow and meteorological forecasts for hindcast extension, the model could be modified to quantify forward-looking forecast uncertainty in support of forecast informed operations. Furthermore, the model could be extended to other applications that require space-time scalability. For instance, recent work has stressed the importance of stochastic watershed modeling (Steinschneider et al., 2015; Vogel, 2017) for long-term hydrologic risk assessment (rather than short-term forecasting). The model presented in this work could readily be extended to help develop correlated stochastic watershed models for river basins across a region, allowing for better characterization of risk in complex, multi-basin water systems. Furthermore, the adaptability of the methodology makes it well suited to exploratory efforts in forecast informed design like high-dimensional input/indicator variable selection (IVS) (Fernando et al., 2009; Giuliani et al., 2015; Herman et al., 2020).

Lastly, we note that the proposed approach to synthetic forecast generation is applicable anywhere there is sufficient overlap in hindcast and observational data to fit the model. This confers the advantage of forecast record extension in areas with long observational records, but limited hindcasts. In cases where the observational record is also limited, there is the possibility of producing synthetic forecasts from traces of stochastically generated weather (Baxevani & Lennartsson, 2015; Steinschneider et al., 2019), which could significantly expand the data available to calibrate and test forecast-informed policies. This effort is left for future work.

Data Availability Statement

The data used in this manuscript are in the following repository and cited in the references: HydroShare, https://doi.org/10.4211/hs.4382404b935f4fde99c7ff4ada264867.

Acknowledgments
This study was supported by the U.S. National Science Foundation Grant EnvS-1803563. We thank the members of the FIRO Steering Committee and the California/Nevada River Forecast Center for their assistance in provisioning the streamflow forecast data and also thank three reviewers for their constructive comments that helped to significantly improve the article.

References

Ahn, K.-H., Palmer, R., & Steinschneider, S. (2017). A hierarchical Bayesian model for regionalized seasonal forecasts: Application to low flows in the northeastern United States. *Water Resource Research*, *53*, 503–521. https://doi.org/10.1002/2016WR019605

Alemu, E. T., Palmer, R. N., Polebitski, A., & Meaker, B. (2011). Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *ASCE Journal of Water Resources Planning and Management*, *137*(1), 72–82. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000088

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., & Lettenmaier, D. P. (2016). Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research*, *52*, 4209–4225. https://doi.org/10.1002/2015WR017864

Barth, R., Meibom, P., & Weber, C. (2011). *Simulation of short-term forecasts of wind and load for a stochastic scheduling model*. In Proceedings of the IEEE Power and Energy Society General Meeting (pp. 24–28). Detroit, MI, USA: IEEE .

Baxevani, A., & Lennartsson, J. (2015). A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resources Research*, *51*, 4338–4358. https://doi.org/10.1002/2015WR017200.A10.1002/2014wr016455

Bottazzi, G., & Secchi, A. (2011). A new class of asymmetric exponential power densities with applications to economics and finance. *Industrial and Corporate Change*, *20*(4), 991–1030. https://doi.org/10.1093/icc/dtr036

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian Inference in statistical analysis*. New York: Wiley. https://doi.org/10.1002/9781118033197
Brodeur, Z. P. (2021). Data repository for: A multivariate approach to generate synthetic short-to-medium range hydro-meteorological forecasts across locations, variables, and lead times. *HydroShare*. https://doi.org/10.4211/hs.4382404b935f4fde99c7ff4ada264867
Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap aggregation and cross-validation methods to reduce overfitting in reservoir control policy search. *Water Resources Research*, *56*, e2020WR027184. https://doi.org/10.1029/2020WR027184
Brown, C. M., Lund, J. R., Cai, X., Reed, P. M., Zagona, E. A., Ostfeld, A., et al. (2015). The future of water resources systems analysis: Toward a scientific framework for sustainable water management. *Water Resources Research*, *51*, 6110–6124. https://doi.org/10.1002/2015WR017114
Buckle, D. J. (1995). Bayesian Inference for Stable Distributions. *Journal of the American Statistical Association*, *90*(430), 605–613. https://doi.org/10.1063/1.357361010.1080/01621459.1995.10476553
California Department of Water Resources (CA/DWR). (2020). *California data Exchange center (CDEC) observed full natural flow streamflow output, Folsom reservoir, CA (FOLC1)*. Retrieved from https://cdec.water.ca.gov
Cerqueti, R., Giacalone, M., & Panarello, D. (2019). A generalized error distribution copula-based method for portfolios risk assessment. *Physica A: Statistical Mechanics and its Applications*, *524*, 687–695. https://doi.org/10.1016/j.physa.2019.04.077
Chen, L., & Guo, S. (2019). *Copulas and its application in hydrology and water resources*. Singapore: Springer. https://doi.org/10.1007/978-981-13-0574-0
Christiansen, B. (2018). Ensemble averaging and the curse of dimensionality. *Journal of Climate*, *31*(4), 1587–1596. https://doi.org/10.1175/JCLI-D-17-0197.1
Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., & Wilby, R. (2004). The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, *5*(1), 243–262. https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2
Delaney, C. J., Hartman, R. K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., et al. (2020). Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in Northern California. *Water Resources Research*, *56*(9). https://doi.org/10.1029/2019WR026604
Demargne, J., Mullusky, M., Werner, K., Adams, T., Lindsey, S., Schwein, N., et al. (2009). Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society*, 778–784. https://doi.org/10.1175/2008BAMS26I9.I
Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., et al. (2014). The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society*, *95*(1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1
Denaro, S., Anghileri, D., Giuliani, M., & Castelletti, A. (2017). Informing the operations of water reservoirs over multiple temporal scales by direct use of hydro-meteorological data. *Advances in Water Resources*, *103*, 51–63. https://doi.org/10.1016/j.advwatres.2017.02.012
Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., & Cayan, D. R. (2011). Atmospheric rivers, floods and the water resources of California. *Water*, *3*(4), 445–478. https://doi.org/10.3390/w3020445
Elith, J., Lautenbach, S., McClean, C., Bacher, S., Dormann, C. F., Skidmore, A. K., et al. (2012). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x
Ernst, M. D. (1998). A multivariate generalized Laplace distribution. *Computational Statistics*, *13*, 227–232. https://doi.org/10.1007/PL00022717
Fernández, C., & Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359–371. https://doi.org/10.1080/01621459.1998.10474117
Fernando, T. M. K. G., Maier, H. R., & Dandy, G. C. (2009). Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology*, 367(3–4), 165–176. https://doi.org/10.1016/j.jhydrol.2008.10.019
Giuliani, M., Pianosi, F., & Castelletti, A. (2015). Making the most of data: An information selection and assessment framework to improve water systems operations. *Water Resources Research*, *51*, 9073–9093. https://doi.org/10.1002/2015WR017200.A10.1002/2015wr017044
Giuliani, M., Zaniolo, M., Castelletti, A., Davoli, G., & Block, P. (2019). Detecting the State of the Climate System via Artificial Intelligence to Improve Seasonal Forecasts and Inform Reservoir Operations. *Water Resources Research*, *55*(11), 9133–9147. https://doi.org/10.1029/2019WR025035
Gleick, P. H. (2002). Soft water paths. *Nature*, *418*(July), 373. https://doi.org/10.1038/418373a
Grygier, J. C., Stedinger, J. R., & Yin, H.-B. (1989). A generalized maintenance of variance extension procedure for extending correlated series. *Water Resources Research*, *25*(3), 345–349. https://doi.org/10.1029/wr025i003p00345
Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, *94*(10), 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1
Hanak, E., Lund, J., Dinar, A., Gray, B., Howitt, R., Mount, J., et al. (2011). *Managing California's water*. Retrieved from, http://www.ppic.org/content/pubs/report/R_211EHR.pdf
Hartmann, D. L. (2016). *Global physical climatology* (2nd ed.). Waltham, MA: Elsevier.
Heath, D. C., & Jackson, P. L. (1994). Modeling the evolution of demand forecasts Ith application to safety stock analysis in production/distribution systems. *IIE Transactions*, *26*(3), 17–30. https://doi.org/10.1080/07408179408966604
Herman, J. D., Quinn, J. D., Steinschneider, S., Giuliani, M., & Fletcher, S. (2020). Climate Adaptation as a Control Problem: Review and Perspectives on Dynamic Water Resources Planning Under Uncertainty. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR025502
Hodge, B. M., Lew, D., Milligan, M., Holttinen, H., Sillanpää, S., Gómez-Lázaro, E., et al. (2012). *Wind Power Forecasting Error Distributions: An International Comparison*. In Proceedings of the International Workshop on large-scale Integration of Wind Power into Power Systems as Well as on Transmission Networks for Offshore Wind Power Plants (pp. 13–15), Lisbon, Portugal
Jasperse, J., Ralph, F. M., Anderson, M., Brekke, L. D., Malasavage, N., Dettinger, M. D., et al. (2020). *Lake Mendocino forecast informed reservoir operations final Viability assessment. Lake Mendocino FIRO Steering Committee* (p. 28). San Diego, CA: Scripps Institute Center for Western Weather and Water Extremes. Retrieved from https://escholarship.org/uc/item/3b63q04n
Kuczera, G. (1983). Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. *Water Resources Research*, *19*(5), 1151–1162. https://doi.org/10.1029/WR019i005p01151
Lall, U. (2014). Debates-The future of hydrological sciences: A (common) path forward? One water. One world. Many climes. Many souls. *Water Resources Research*, *50*, 5335–5341. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x10.1002/2014wr015402
Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research*, *32*(3), 679–693. https://doi.org/10.1029/95wr02966

Lamontagne, J. R., & Stedinger, J. R. (2018). Generating synthetic streamflow forecasts with specified precision. *Journal of Water Resources Planning and Management*, *144*(4), 04018007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000915

Landgraf, A. J., & Lee, Y. (2015). *Dimensionality reduction for binary data through the projection of natural parameters* (Technical Report 890URL). Department of Statistics, Ohio State University. Retrieved from http://arxiv.org/abs/1510.06112

Landgraf, M. A. J. (2016). *logisticPCA: Binary dimensionality reduction*. R package version 0.2. Retrieved from. https://doi.org/10.5194/amt-2016-114-ac3

Lettenmaier, D. P. (1984). Synthetic streamflow forecast generation. *Journal of Hydraulic Engineering*, *110*(3), 277–289. https://doi.org/10.1061/(ASCE)0733-9429(1984)110:3(277)

Loucks, D. P., & Van-Beek, E. (2017). *Water resources systems planning and management: An introduction to methods, models, and applications*. Switzerland: Springer. https://doi.org/10.1007/978-3-319-44234-1

Maurer, E. P., & Lettenmaier, D. P. (2004). Potential effects of long-lead hydrologic predictability on Missouri River main-stem reservoirs*. *Journal of Climate*, *17*(1), 174–186. https://doi.org/10.1175/1520-0442(2004)017<0174:PEOLHP>2.0.CO;2

Mello, P. E., Lu, N., & Makarov, Y. (2011). An optimized autoregressive forecast error generator for wind and load uncertainty study. *Wind Energy*, *14*, 967–976. https://doi.org/10.1002/we.460

Miller, A., Blum, A. G., & Higgins, P. A. T. (2018). *Opportunities for forecast-informed water resources management: An AMS Policy Program Study*. Washington, DC: The American Meteorological Society.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*(7), 685–694. https://doi.org/10.1080/02664760500079464

Nayak, M. A., Herman, J. D., & Steinschneider, S. (2018). Balancing flood risk and water supply in California: Policy search integrating short-term forecast ensembles with conjunctive use. *Water Resources Research*, *54*, 7557–7576. https://doi.org/10.1029/2018WR023177

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, *59*(2), 347–370. https://doi.org/10.2307/2938260

Nicholson, W. B., Matteson, D. S., & Bien, J. (2019). *BigVAR: Dimension reduction methods for multivariate time series*. Package version 1.0.6. Retrieved from https://cran.r-project.org/web/packages/BigVAR

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. S. (2020). High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, *21*, 1–52.Retrieved from https://www.jmlr.org/papers/volume21/19-777/19-777.pdf

NOAA/NWS California/Nevada River Forecast Center. (2020). *(CA/NV RFC). Hydrologic ensemble forecast system (HEFS) streamflow forecast output, Folsom reservoir, CA (FOLC1)*. Retrieved from Brett Whitin, P.E., CA/NV RFC, Accessed date March 1.

NOAA Physical Sciences Laboratory. (NOAA PSL). *NCEP global ensemble forecast system (GEFS) data archive*. Retrieved from https://psl.noaa.gov/forecasts/reforecast2/download.html

NOAA Physical Sciences Laboratory *(NOAA PSL). NOAA-CIRES-DOE 20th century reanalysis version 3 (20CR V3)*. Retrieved from https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html

NWS Office of Hydrologic Development (NWS OHD) (2016). *HEFS overview and getting started, version OHD-CORE-CHPS-4.4.a*. Release May 2016.,

Olauson, J., Bladh, J., Lönnberg, J., & Bergkvist, M. (2016). A new approach to obtain synthetic wind power forecasts for integration studies. *Energies*, *9*(10), 800–816. https://doi.org/10.3390/en9100800

Palmer, T. N. (1993). Extended-range atmospheric prediction and the Lorenz model. *Bulletin of the American Meteorological Society*, *74*(1), 49–65. https://doi.org/10.1175/1520-0477(1993)074<0049:ERAPAT>2.0.CO;2

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, *21*, 284–296. https://doi.org/10.1002/pip10.1002/pip.1180

Piani, C., & Haerter, J. O. (2012). Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophysical Research Letters*, *39*(20), 1–6. https://doi.org/10.1029/2012GL053839

Rao, C. R., & Vinod, H. D. (Eds.), (2019). *Handbook of statistics: Conceptual econometrics using R* (Vol. 41). Waltham, MA: Elsevier.

Raso, L., Schwanenberg, D., van de Giesen, N. C., & van Overloop, P. J. (2014). Short-term optimal operation of water systems using ensemble forecasts. *Advances in Water Resources*, *71*, 200–208. https://doi.org/10.1016/j.advwatres.2014.06.009

Rayner, S., Lach, D., & Ingram, H. (2005). Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. *Climatic Change*, *69*(2–3), 197–227. https://doi.org/10.1007/s10584-005-3148-z

Rougier, J. (2016). Ensemble averaging and mean squared error. *Journal of Climate*, *29*(24), 8865–8870. https://doi.org/10.1175/JCLI-D-16-0012.1

Sankarasubramanian, A., Lall, U., Souza Filho, F. A., & Sharma, A. (2009). Improved water allocation utilizing probabilistic climate forecasts: Short-term water contracts in a risk management framework. *Water Resources Research*, *45*(11), 1–18. https://doi.org/10.1029/2009WR007821

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), 1–17. https://doi.org/10.1029/2009WR008933

Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley. https://doi.org/10.1002/9780470316849

Shalev-Shwartz, S., & Ben-David, S. (2013). *Understanding machine learning: From theory to algorithms*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., et al. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. https://doi.org/10.1002/qj.3598

So, M. K. P., Chen, C. W. S., Lee, J.-Y., & Chang, Y.-P. (2008). An empirical evaluation of fat-tailed distributions in modeling financial time series. *Mathematics and Computers in Simulation*, *77*(1), 96–108. https://doi.org/10.1016/j.matcom.2007.02.008

Sorooshian, S., & Dracup, J. A. (1980). Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resources Research*, *16*(2), 430–442. https://doi.org/10.1029/wr016i002p00430

Stedinger, J. R., & Taylor, M. R. (1982). Synthetic streamflow generation: 2. Effect of parameter uncertainty. *Water Resources Research*, *18*(4), 919–924. https://doi.org/10.1029/WR018i004p00919

Steinschneider, S., Ray, P., Rahat, S. H., & Kucharski, J. (2019). A weather-regime-based stochastic weather generator for climate vulnerability assessments of water systems in the Western United States. *Water Resources Research*, *55*(8), 6923–6945. https://doi.org/10.1029/2018WR024446

Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*, 2823–2839. https://doi.org/10.1002/hyp.10409

Subbotin, M. T. (1923). On the law of frequency of error. *Matematicheskii Sbornik*, *31*(2), 296–301.
Sun, M., Feng, C., & Zhang, J. (2020). Probabilistic solar power forecasting based on weather scenario generation. *Applied Energy*, *266*, 114823. https://doi.org/10.1016/j.apenergy.2020.114823
Teegavarapu, R. S. V., Salas, J. D., & Stedinger, J. R. (Eds.), (2019). *Statistical analysis of hydrologic variables: Methods and applications*. Reston, VA: American Society of Civil Engineers.
Thompson, P. D. (1977). How to improve accuracy by combining independent forecasts. *Monthly Weather Review*, *105*, 228–229. https://doi.org/10.1175/1520-0493(1977)105<0228:htiabc>2.0.co;2
Turner, S. W. D., Bennett, J. C., Robertson, D. E., & Galelli, S. (2017). Complex relationship between seasonal streamflow forecast skill and value in reservoir operations. *Hydrology and Earth System Sciences*, *21*(9), 4841–4859. https://doi.org/10.5194/hess-21-4841-2017
USGS (2020) *U.S. Geological Survey in cooperation with U.S. Environmental Protection Agency, USDA Forest Service, and other Federal, State and local partners*. Watershed Boundaries, HUC 8, for California, Retrieved from: https://www.sciencebase.gov/catalog/item/5696a727e4b039675d00a4ef
Valeriano, O. C. S., Koike, T., Yang, K., Graf, T., Li, X., Wang, L., & Han, X. (2010). Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts. *Water Resources Research*, *46*(10). https://doi.org/10.1029/2010WR009502
Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, *1*, 28–35. https://doi.org/10.1016/j.wasec.2017.06.001
Voisin, N., Pappenberger, F., Lettenmaier, D. P., Buizza, R., & Schaake, J. C. (2011). Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. *Weather and Forecasting*, *26*(4), 425–446. https://doi.org/10.1175/waf-d-10-05032.1
Wang, F., Wang, L., Zhou, H., Saavedra Valeriano, O. C., Koike, T., & Li, W. (2012). Ensemble hydrological prediction-based real-time optimization of a multiobjective reservoir during flood season in a semiarid basin with global numerical weather predictions. *Water Resources Research*, *48*(7), 1–21. https://doi.org/10.1029/2011WR011366
Wei, W. W. S. (2019). *Multivariate time series analysis and its applications*. New York: Wiley. https://doi.org/10.1002/9781119502951
Wilks, D. S. (2015). Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, *141*(688), 945–952. https://doi.org/10.1002/qj.2414
Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Cambridge, MA: Elsevier.
Wurtz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2020). *fGarch: Rmetrics – Autoregressive conditional heteroskedastic modeling*. R package version 3042.83.2. Retrieved from https://cran.r-project.org/web/packages/fGarch
You, J.-Y., & Cai, X. (2008). Determining forecast and decision horizons for reservoir operations under hedging policies. *Water Resources Research*, *44*(11), 1–14. https://doi.org/10.1029/2008WR006978
Yuan, X., Wood, E. F., & Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: Physical understanding and system development. *WIREs Water*, *2*(5), 523–536. https://doi.org/10.1002/wat2.1088
Zeff, H. B., Herman, J. D., Reed, P. M., & Characklis, G. W. (2016). Cooperative drought adaptation: Integrating infrastructure development, conservation, and water transfers into adaptive policy pathways. *Water Resources Research*, *52*, 7327–7346. https://doi.org/10.1002/2016WR018771
Zhao, T., Zhao, J., Yang, D., & Wang, H. (2013). Generalized martingale model of the uncertainty evolution of streamflow forecasts. *Advances in Water Resources*, *57*, 41–51. https://doi.org/10.1016/j.advwatres.2013.03.008
Zimmerman, J. K. H., Carlisle, D. M., May, J. T., Klausmeyer, K. R., Grantham, T. E., Brown, L. R., & Howard, J. K. (2018). Patterns and magnitude of flow alteration in California, USA. *Freshwater Biology*, *63*(8), 859–873. https://doi.org/10.1111/fwb.13058