A hybrid, non-stationary Stochastic Watershed Model (SWM) for uncertain hydrologic projections under climate change

Zach Brodeur[1], Sungwook Wi[2], Ghazal Shabestanipour[3], Jon Lamontagne[4], Scott Steinschneider[5]
Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY

1. Graduate Research Assistant, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: zpb4@cornell.edu, Phone: 607-255-2155 (Corresponding Author).

2. Research Scientist, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: sw2275@cornell.edu, Phone: 607-255-2155.

3. Graduate Research Assistant, 200 College Avenue, Department of Civil and Environmental Engineering, Tufts University, Medford, MA, 02155. Email: ghazal.shabestanipour@tufts.edu, Phone: 617-627-3211.

4. Assistant Professor, 200 College Avenue, Department of Civil and Environmental Engineering, Tufts University, Medford, MA, 02155. Email: jonathan.lamontagne@tufts.edu, Phone: 617-627-3211.

5. Assistant Professor, 111 Wing Drive, Riley-Robb Hall, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853. Email: ss3378@cornell.edu, Phone: 607-255-2155.

Key Points:

- We document non-stationarity of hydrologic model errors under plausible climate change in an idealized experimental design
- We leverage state variable – model error relationships to develop a hybrid machine learning based error model
- The hybrid model exhibits a robust capability to predict out-of-sample and non-stationary error properties

Abstract

Stochastic Watershed Models (SWMs) are emerging tools in hydrologic modeling used to propagate uncertainty into model predictions by adding samples of model error to deterministic simulations. One of the most promising uses of SWMs is uncertainty propagation for hydrologic simulations under climate change. However, a core challenge with this approach is that the predictive uncertainty inferred from hydrologic model errors in the historical record may not correctly characterize the error distribution under future climate. For example, the frequency of physical processes (e.g., snow accumulation and melt, droughts and hydrologic recessions) may change under climate change, and so too may the frequency of errors associated with those processes. In this work, we explore for the first time non-stationarity in hydrologic model errors under climate change in an idealized experimental design. We fit one hydrologic model to historical observations, and then fit a second model to the simulations of the first, treating the first model as the true hydrologic system. We then force both models with climate change impacted meteorology and investigate changes to the error distribution between the models in historical and future periods. We develop a hybrid machine learning method that maps model input and state variables to predictive errors, allowing for non-stationary error distributions based on changes in the frequency of internal state variables. We find that this procedure provides an internally consistent methodology to overcome stationarity assumptions in error modeling and offers an important path forward in developing stochastic hydrologic simulations under climate change.

1. Introduction

Climate change and its uncertain impacts on the hydrologic system pose major challenges to the adaptation of existing water resources infrastructure and the design and construction of new infrastructure (Stakhiv & Hiroki, 2021). This challenge is particularly notable in the developing world, where the infrastructure needed for robust and resilient water resources systems is either inadequate or non-existent (Stakhiv & Hiroki, 2021; Boland & Loucks, 2021), and data to support precise hydrologic modeling are limited. Considering this challenge, methods that quantify uncertainty in future hydrology play an increasingly critical role in the modern practice of water resources planning and management (Milly et al., 2008; Brown et al., 2015; Read & Vogel, 2015; Hui et al., 2018; Sterle et al., 2019).

In the past, historical hydrologic variability was deemed an adequate representation of future hydrologic uncertainty, motivating the use of stationary, stochastic streamflow models in engineering design and planning (Thomas & Fiering, 1962; Loucks & Van-Beek, 2017; Teegavarapu et al., 2019). As the impacts of climate change (and other land use change) have become increasingly apparent, many have questioned the suitability of such stationary statistical models for infrastructure planning (Milly et al., 2008, Galloway, 2011, Montanari & Koutsoyiannis, 2014). While the parameters of these models can be modified to enable the simulation of new hydrologic behavior (e.g., Hadjimichael et al., 2020; Bracken et al., 2014), the range of plausible change is difficult to infer without a modeling framework that can predict emergent patterns of hydrologic response to climate change and other biological, biophysical, and human feedbacks on the hydrologic system.

Stochastic watershed models (SWM) have been forwarded to address this challenge (Vogel, 2017). SWMs combine deterministic predictions from process-based hydrologic models with a stochastic element that captures model uncertainty (Steinschneider et al., 2015; Sikorska et al., 2015, Farmer & Vogel, 2016; Vogel, 2017). The use of process-based models enables hydrologic projections that explicitly represent changes to meteorological forcings and landscape characteristics (e.g., vegetation or land use) and their non-linear impacts on hydrologic response. The stochastic component of a SWM represents hydrologic uncertainty that the deterministic model cannot capture. In the most straightforward case, this uncertainty is approximated by the predictive uncertainty of the model (i.e., based on errors between model predictions and the observations). The predictive uncertainty reflects the integration of input, parametric, and model uncertainty (Montanari & Koutsoyiannis, 2012) and can be represented by a variety of error modeling approaches (Vogel, 2017; McInerney et al., 2017; Koutsoyiannis & Montanari, 2022; Shabestanipour et al., 2023). The addition of simulated model errors and deterministic hydrologic model simulations creates a SWM simulation, and repetition of this process using multiple random samples of error yields a SWM ensemble that can be used for both short term probabilistic prediction (e.g., flood forecasting; Sikorska et al., 2015; McInerney et al., 2018; Koutsoyiannis & Montanari, 2022) and long-term planning (e.g., design event estimation; Farmer & Vogel, 2016; Shabestanipour et al., 2023).

To date, one important issue in stochastic watershed modeling that remains unresolved relates to non-stationarity in the stochastic process for predictive uncertainty. When used to develop hydrologic projections under climate change, past studies have made the implicit assumption that predictive uncertainty inferred from historical errors is sufficient to characterize future

uncertainty (Sikorska et al., 2015; Vogel, 2017; Shabestanipour et al., 2023). Some have argued this approach is sufficient if the deterministic component of the model can account for non-stationarity (Montanari & Koutsoyiannis, 2014). However, there are reasons to doubt this assumption in the context of hydrologic model prediction. The stochastic component of a SWM fit to historical prediction errors relies on historical relationships between model states and observed hydrology (Liu & Gupta, 2007; Renard et al., 2011; Vogel, 2017). The effects of climate change go deeper than simply amplifying or attenuating hydrologic response, instead affecting fundamental process relationships within catchments, including the timing and rate of snow accumulation and melt (Musselman, 2017; Mote et al., 2018), timing of peak soil moisture (Xu et al., 2021), and changes to runoff efficiency through both physical (Lehner et al., 2017; Overpeck & Udall, 2020) and biophysical (Mankin et al., 2019) effects. These climate change induced effects will alter the frequency, timing, and intensity of model states, activate model components in configurations not seen in the historical record, and change the way meteorological forcing is converted to streamflow. In turn, the model predictive errors would be expected to exhibit fundamental departures from the distributional properties observed in the historical period. For instance, if within the historical record a hydrologic model exhibits different error distributions during periods of snow accumulation and melt versus periods of direct rainfall-runoff response (e.g., because of different, incorrect process representations under those two different hydrologic regimes), and under climate change the former process becomes less frequent and the latter more frequent, the distribution of model errors under climate change would almost certainly change compared to the historical period. To the authors' knowledge, this issue in SWM has not yet been documented in the literature.

The potential for non-stationary predictive errors complicates an already difficult problem in stochastic watershed modeling (Beven, 2016). Hydrologic prediction errors exhibit a number of challenging characteristics including autocorrelation, heteroscedasticity, and non-normality, even in the stationary case (Schoups & Vrugt, 2010; Mcinerny et al., 2017; Mcinerny et al., 2018; Hunter et al., 2021). Efforts to understand and quantify these errors (Liu & Gupta, 2007) have progressed from simple autoregressive techniques (Toth et al., 1999) to more complex statistical methods using either decomposition (Kuczera et al., 2006; Renard et al., 2011) or aggregate approaches to predictive error modeling (Montanari & Koutsoyiannis, 2012; Sikorska el., 2015; McInerny et al., 2018; Shabestanipour et al., 2023). One recent approach that has gained significant traction is the use of machine learning (ML) to correct prediction errors of process-based hydrologic models (Konapala et al., 2020; Shen et al., 2022; Hah et al., 2022; Quilty et al., 2022). These approaches (often termed 'hybrid' or 'physics informed data driven' models) range from simpler ML-based error correction models (Shamseldin & O'Connor, 2001; Konapala et al., 2020; Shen et al., 2022) to more complex stochastic formulations that utilize ensembles of hydrologic model simulations, each with different parameter sets and ML-based error correction models (Quilty et al., 2021), and possibly including contributions from additional uncertainties (e.g., input, parameter; Quilty et al., 2022; Hah et al., 2022).

Hybrid approaches capitalize on the capability of ML models to better capture non-linear hydrologic responses as compared to process models (Kratzert et al., 2018; Nearing et al., 2019; Nearing et al., 2021). However, they do so by mapping endogenous physical model states or exogenous information (e.g., meteorological variables) to process-model errors, enabling more accurate and reliable predictions while still being constrained by first order physical relationships

in the process-based model (Beven, 2020; Shen et al., 2021; Hah et al., 2022; Quilty et al., 2022). While hybrid methods do not consistently improve hydrologic predictive performance over more direct ML methods (Frame et al., 2021), they can help to address the issue of uncertainty representation in these methods (Klotz et al., 2022). In addition, some initial work suggests hybrid models may be more appropriate for long-term projections that extrapolate hydrologic responses under unprecedented climate change (Wi and Steinschneider, 2022). Hybrid methods that map process model states to predictive errors may also be able to exploit these relationships to capture non-stationarity in error structure based on changes in the frequency of hydrologic regimes (i.e., changing frequency of projected model state variables). This approach decouples the error models from static empirical relationships that may change fundamentally in a future climate, such as seasonality in the error distribution. To date, the potential of hybrid models to support non-stationary SWMs remains unexplored.

In this work, we demonstrate for the first time the challenge of non-stationary prediction errors in stochastic watershed modeling under climate change, and we advance a novel, hybrid modeling framework to address this challenge. We demonstrate this work in a case study of the Feather River basin upstream of Oroville Dam in northern California, where climate change is expected to significantly impact hydrologic response through reduced snowpack, earlier snowmelt, and changing precipitation characteristics (Hanak et al., 2011; Huang et al., 2012; Sterle et al., 2019). We first forward an idealized experimental design where one hydrologic model is calibrated to observed streamflow and treated as the true hydrologic system (hereafter the "truth model"), while a second model (hereafter the "process model") is then calibrated to simulations from the truth model. We force both models with the same set of non-stationary meteorological inputs and

183 document non-stationarity in the error distribution between them. This approach is similar to so-
184 called 'model-as-truth' or 'perfect model' experiments that are relatively common in the
185 assessment of climate model ensembles under non-stationary climate change scenarios
186 (Abramowitz & Bishop, 2015; Knutti et al., 2017; Herger et al., 2018).
187
188 We then develop a hybrid error model composed of an ML-based error correction model and a
189 dynamic residual noise model, both of which use process model state variables to infer error
190 properties. The ML correction model, based on the Random Forest (RF) approach in Shen et al.
191 (2022), captures conditional bias in the process model. The dynamic residual model is based on a
192 modified version of the generalized likelihood (GL) approach of Schoups and Vrugt (2010) and
193 maps process model state variables into time-varying autocorrelation, variance, skew, and
194 kurtosis of the ML-based error correction residuals. We assess the ability of this hybrid error
195 model, coupled with simulations from the process model, to preserve the statistical properties of
196 the truth model in out-of-sample cases with and without the impacts of climate change and
197 compare results to a static SWM approach as a benchmark. We conclude the study by
198 demonstrating the same technique for a process model fit to actual streamflow observations in
199 the Feather River basin.
200
201 **2. Data**
202 The Feather River basin upstream of Lake Oroville drains an area of 9338 km^2 on the west facing
203 slopes of the northern Sierra Nevada mountain range (Figure 1). This portion of the Sierra
204 Nevada reaches altitudes of nearly 3000 m, making the Feather River a snow-dominated
205 catchment. The precipitation regime is driven by large, infrequent atmospheric rivers (ARs) that

206 exhibit significant inter-annual variability and occur primarily in the cold season (November –
207 April). Accordingly, streamflow varies considerably across years and also across seasons, as
208 snowmelt drives higher flows in the spring and early summer months and high
209 evapotranspiration drives lower flows in late summer and fall. Winter flows can vary
210 considerably in response to winter storms, particularly when associated with AR-induced
211 warming or rain-on-snow events (Hanak et al., 2011; Huang et al., 2012).
212
213 Observed daily streamflow data for this watershed were taken from the California Data
214 Exchange Center (CDEC) Full Natural Flow (FNF) database for water years (WY) 1988-2013 at
215 Oroville Dam on the Feather River (CDEC ID: ORO). These data account for human
216 modifications to the natural hydrology, which occur at upstream reservoirs through diversions
217 and export/import of water between watersheds (see Figure 1). We used daily precipitation and
218 mean temperature from the 1/16-degree climate product of Livneh et al. (2015) as input forcings
219 to all hydrologic models used in this work.

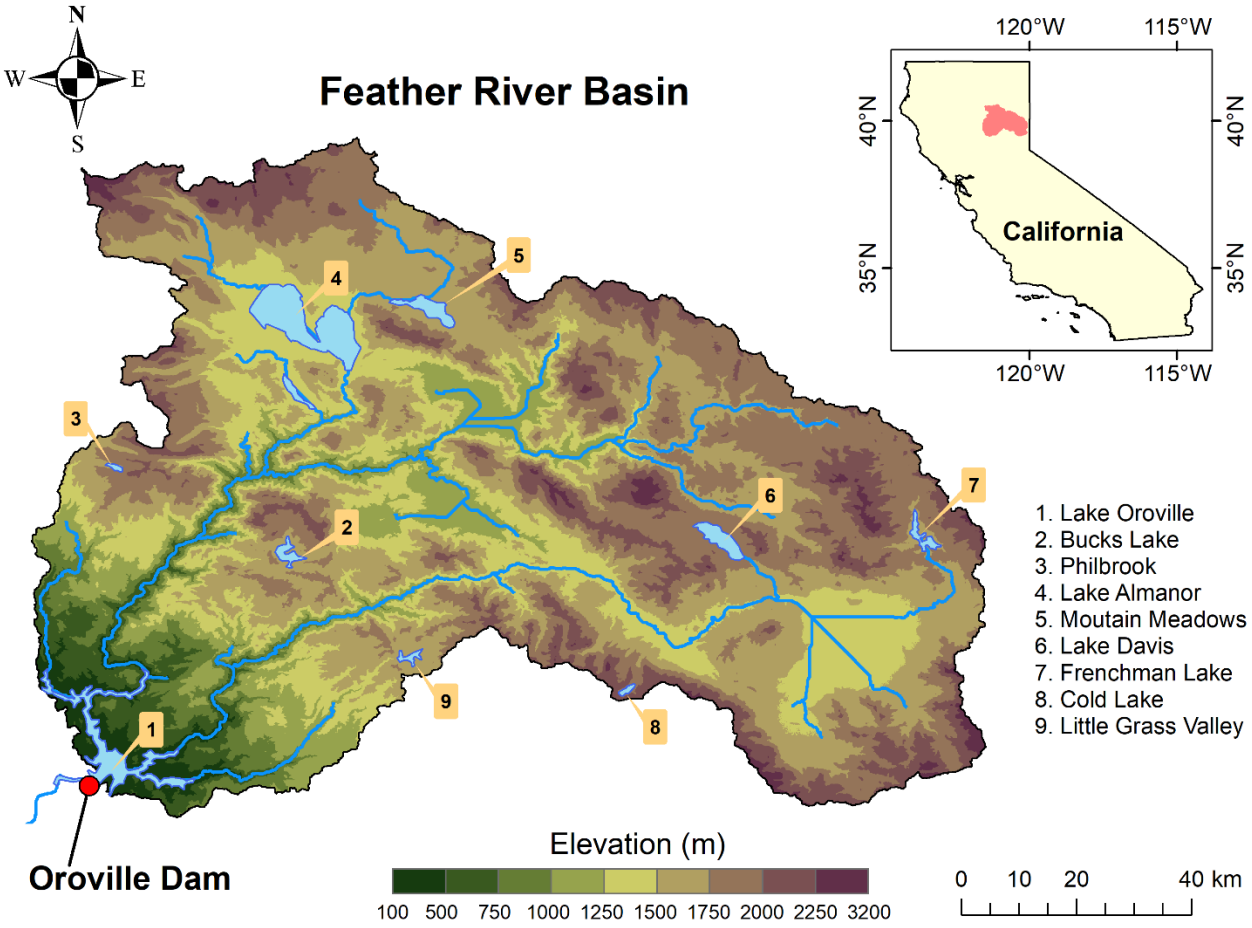


Figure 1. *Geographical area of study depicting Feather River inflow to Oroville Dam (1) as well as significant upstream diversions (2-9).*

3. Methods

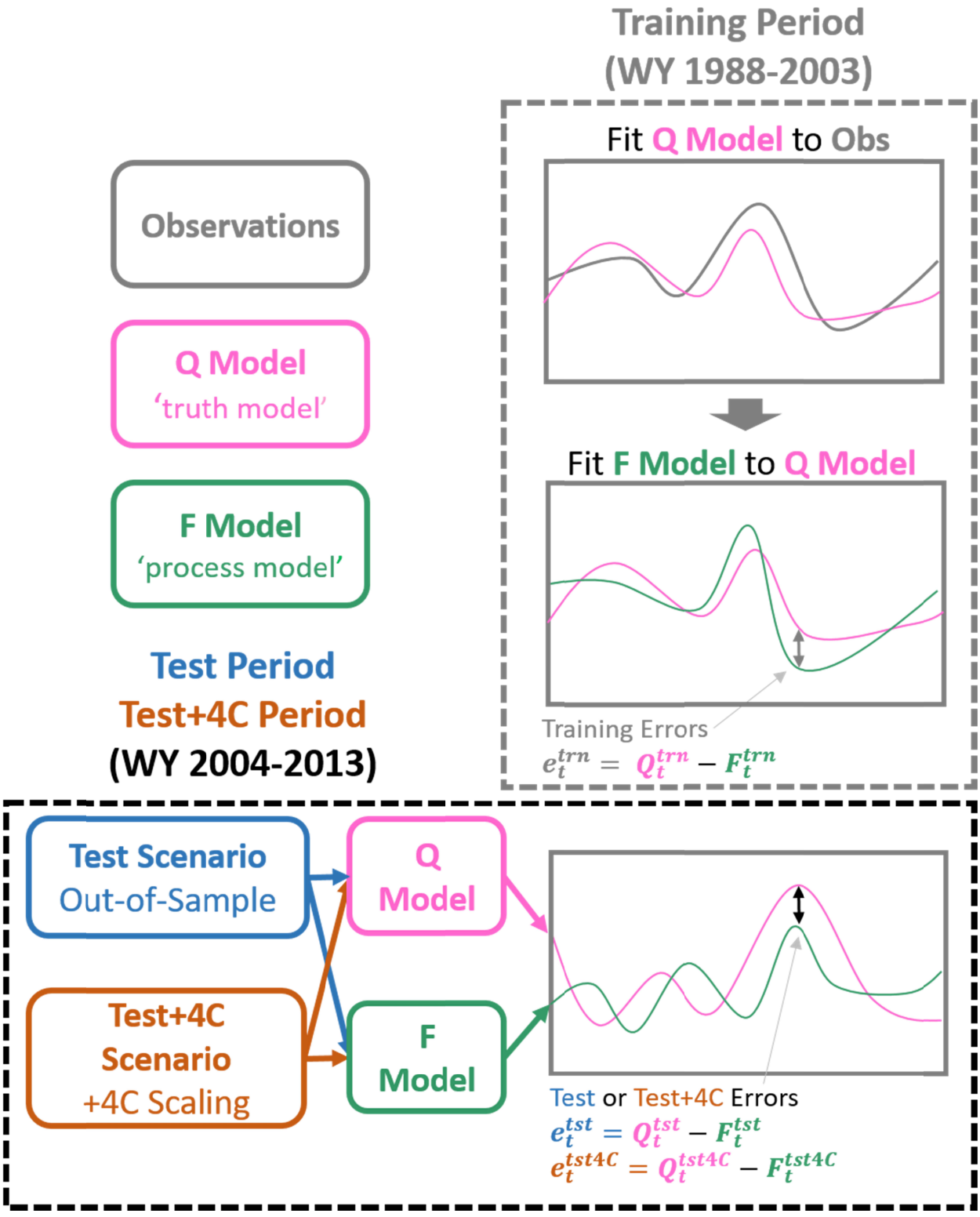
3.1. Experimental Design

This study employs a stylized experimental design to demonstrate the challenge of non-stationary prediction errors in SWMs under climate change and to evaluate whether a novel, hybrid modeling framework can address this challenge (Figure 2). We first select two hydrologic models, designating the 'truth model' and the 'process model'. The truth model is taken to be the true hydrologic system, and simulations from this model under alternative meteorological forcing are taken to be the true hydrologic response to that forcing. The process model represents an

233 (imperfect) model of the true hydrologic system that can only approximate new hydrologic
234 responses under alternative meteorological forcing. We use this 'model-as-truth' approach to
235 overcome the practical conundrum of having no future observations with which to compare our
236 process model, but fully acknowledge the limitations of using another model to represent the true
237 system (discussed more in section 3.5). We describe the hydrologic models used for the truth and
238 process models in Section 3.2 below.
239
240 In the experiment, we split the available record into a training period for calibration and
241 validation and a test period for out-of-sample model evaluation (WY 1988-2003 and 2004-2013
242 in our case study, respectively). During a portion of the training period, we calibrate the truth
243 model to observed streamflow data and then calibrate the process model to the truth model in
244 that same period (gray dashed box in Figure 2). Errors between the truth and process model in
245 the training period, along with state variables from the process model, are used to fit a hybrid
246 SWM (described in Section 3.3). We then examine the distribution of hydrologic prediction
247 errors between the truth and process models in the test period, calculated based on historical
248 precipitation and temperature data from this period (Test), as well as historical precipitation and
249 temperature warmed by 4^{o} C (Test+4C). Hereafter, we occasionally use 'historical' and
250 'warmed' when referring to the Test and Test+4C scenarios. Our primary focus is to 1) document
251 changes to the error distribution between the truth and process models in the historical test period
252 scenario versus the warmed test period scenario; and 2) evaluate whether our proposed hybrid
253 SWM can capture potential changes in the error distribution between these two scenarios. We
254 also use interpretability methods to understand how the hybrid SWM uses model state variables
255 to estimate changes to the error distribution (see Section 3.4). Finally, we repeat the experiment

256 in a real-world (non-stylized) setting, training and testing the performance of a hybrid SWM
257 against actual streamflow observations in the Feather River basin over the historical record
258 (Section 3.5).

259


Training Period
(WY 1988-2003)
Observations
Q Model
'truth model'
F Model
'process model'
Fit Q Model to Obs
Fit F Model to Q Model
Training Errors
$e_t^{trn} = Q_t^{trn} - F_t^{trn}$
Test Period
Test+4C Period
(WY 2004-2013)
Test Scenario
Out-of-Sample
Test+4C
Scenario
+4C Scaling
Q
Model
F
Model
Test or Test+4C Errors
$e_t^{tst} = Q_t^{tst} - F_t^{tst}$
$e_t^{tst4C} = Q_t^{tst4C} - F_t^{tst4C}$

260

Figure 2. *Conceptual diagram showing the stylized experimental design, where a 'truth' (Q) and 'process' (F) model are designated to test the effect of alternate hydrologic model forcing on predictive errors between the two models. In the Test scenario, both models are forced with out-of-sample but stationary forcings, whereas in the Test+4C scenario, both models are forced with non-stationary and out-of-sample forcings incorporating 4^{o}C of applied warming.*

3.2. Hydrologic Model Setup

We calibrate two hydrologic models for the Feather River basin, SAC-SMA (Burnash, 1995) and HYMOD (Boyle, 2001), that are used as the truth and process models, respectively. These two models are built using 828 hydrologic response units (HRUs) defined for the basin by segregating each 1/16° Livneh climate grid cell into different soil classes from the 1-km-resolution State Soil Geographic dataset (Miller & White, 1998). The Livneh temperature forcings are adjusted for each HRU using the monthly lapse rates derived by Wi & Steinschneider (2022) for the area. The Lohmann routing model (Lohmann et al., 1998) traces the runoff from HRUs through the river channel to simulate streamflow at the basin outlet (i.e., daily inflows into Oroville Dam).

We use a genetic algorithm (Wang et al., 1991) to calibrate the hydrologic models and use Nash Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) as a performance metric. We first calibrate SAC-SMA (truth model) to the full natural flows for the period of WY1988-2003. The flows simulated by SAC-SMA were then used to calibrate HYMOD (process model) for the same period. For the training (test) periods of WY1988-2003 (WY2004-2013), SAC-SMA simulations achieved training (test) NSE of 0.92 (0.89), whereas HYMOD NSE was 0.95 (0.92).

Internal state variables simulated by HYMOD (the process model) are used to inform our error model. These include simulated streamflow (sim), runoff, baseflow, snow water equivalent

(swe), and upper and lower soil moisture (upr_sm, lwr_sm), and represent basin-wide average states (i.e., the sum of HRU state variables weighted by percent area). We also include meteorological input variables (e.g., temperature, precipitation) in this list, and use the term 'state variables' hereafter to refer to both meteorological and internal hydrologic model state variables, as in Shen et al. (2022). A summary table of state variables and their detailed descriptions is provided in supporting information (S1).

3.3. Hybrid SWM

We develop a novel, hybrid SWM that is composed of deterministic and stochastic components:

$$Q_t = F(X_t, \pi) + e_t \quad \text{(Eq. 1)}$$

Here, Q_t is the true streamflow at time t, $F(X_t,\pi)$ is a deterministic streamflow estimate from a process model F conditioned on meteorological and other inputs X_t and parameters π, and e_t is the stochastic prediction error. To estimate the SWM, we follow the approach of Montanari & Brath (2004) and first train the model F to observed flows Q_t (i.e., estimate the model parameters π), and then afterwards we develop an error model to represent the stochastic behavior of e_t. While more sophisticated approaches are possible that estimate the error model jointly with F and quantify parameter uncertainty in π (Kuczera et al., 2006; Renard et al., 2011), we opt for a simpler, staged approach that is easier to implement and helps avoid complex interactions between process and error model estimation.

309 The primary methodological contribution of this work is an adaptive, state-variable-dependent
310 hybrid model for e_t, illustrated in Figure 3. There are two main components of this model. The
311 first is an initial model updating step referred to as 'error correction' (Shen et al. 2022). The error
312 correction model *f* creates a mapping between the process model state variables ($\theta_{t,SV}$) and the
313 raw errors (e_t), including autocorrelation in the errors through lagged error terms ($e_{t-1:t-p}$) out
314 to lag *p:*
315
316 $$e_t = f\left(\theta_{t,SV}, e_{t-1:t-p}\right) + \varepsilon_t \quad \text{(Eq. 2)}$$
317
318 This model corrects for conditional bias, i.e., biases in the process model predictions that are
319 dependent on the internal states of the model and recent prediction errors. The second component
320 is a dynamic residual model to capture the remaining stochasticity in the error correction model
321 residual, ε_t, also as a function of process model state variables. Each of these components is
322 described in more detail in Sections 3.3.1 and 3.3.2 below.
323
324 Importantly, we use a split-sample calibration/validation approach to fit these two components in
325 a similar fashion to Hah et al. (2022). That is, we first fit the error correction model *f* to one
326 subset of the training data (termed the calibration set), and then we fit the dynamic residual
327 model to a separate subset of the training data (termed the validation set), after the error
328 correction model has been applied to that validation set (see Figure 3). This strategy helps ensure
329 that the dynamic residual model will represent the true variability of out-of-sample residuals
330 from the error correction model. In this work we employ an approximate 70%/30% split of the
331 training data between calibration (WY 1988-1997) and validation (WY 1998-2003) periods,

following common practice in the ML literature (Shalev-Shwartz & Ben-David, 2013; Hastie et al., 2017).

The hybrid model can then be used to simulate errors (e_t^*) in a new time period using the state variables associated with the process model simulated in that new period. We hypothesize that the model-based hydrologic states will vary considerably in periods with very different climates (e.g., Test vs. Test+4C; see Figure 3), and this will propagate into new error distributions for the SWM. Simulated errors can be subtracted from the process model simulation to yield a single SWM trace of streamflow; a SWM ensemble is generated by repeating this process for many independent simulations of error.

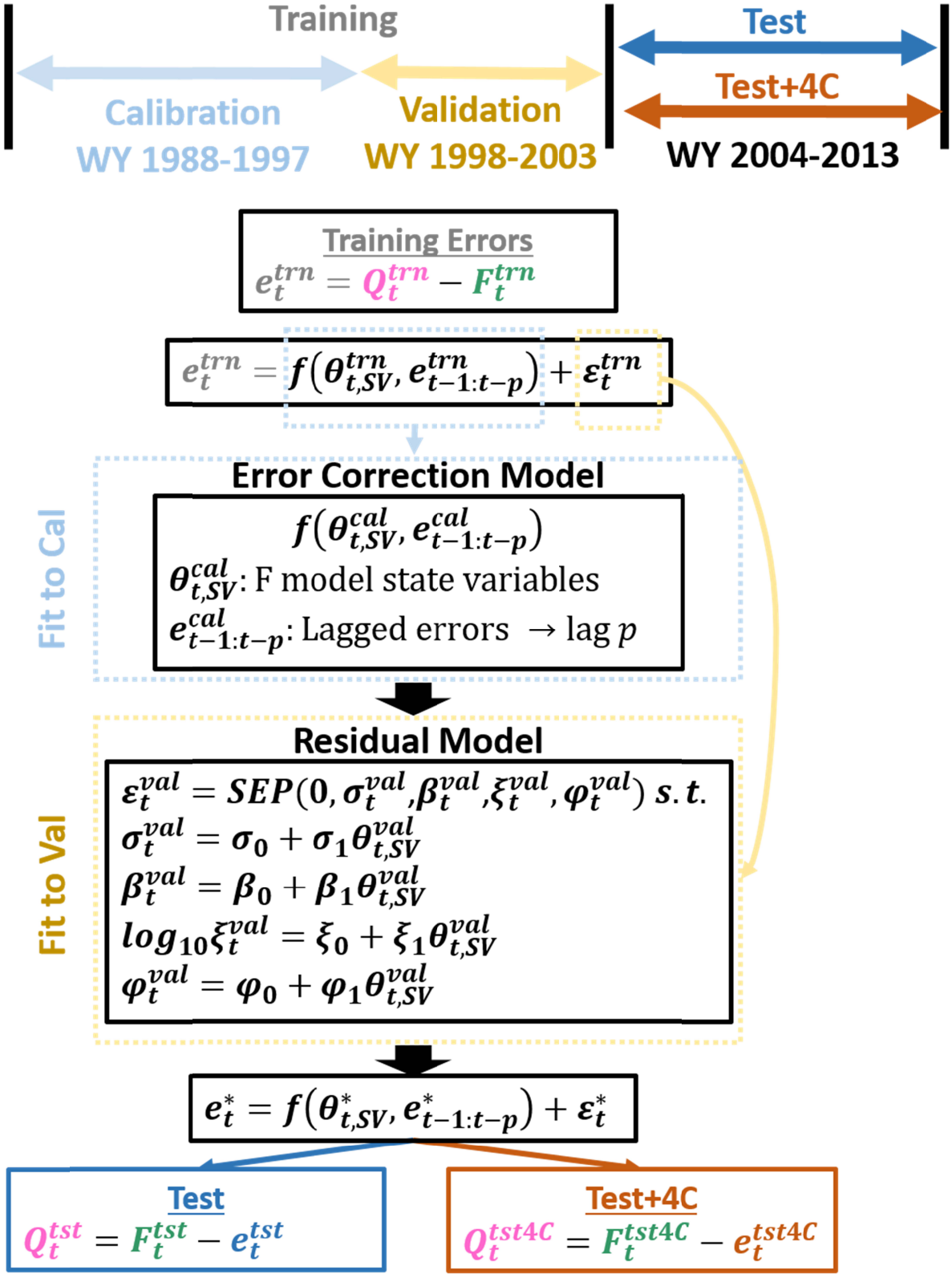


Figure 3. *Diagram of hybrid SWM structure including setup of calibration, validation, and testing periods (Test and Test+4C) within the data. This diagram highlights the staged nature of*

the hybrid SWM, where an error correction model is first fitted to calibration data, which is then used to fit the residual model on the validation data. The resultant model can be used to generate new sequences of predictive uncertainty in out-of-sample scenarios (e.g. Test & Test+4C) using state variable timeseries associated with the 'process' model.

3.3.1. Error Correction Model

As described in Eq. 2, the error correction model *f* uses process model state variables ($\theta_{t,SV}$) and lagged errors ($e_{t-1:t-p}$) to estimate current process model errors for time *t*. Many different error correction models could be selected for *f* (Konapala et al., 2020; Frame et al., 2021; Shen et al., 2022). In this work, we select *f* to be a random forest (RF) model, leveraging its parsimony, demonstrated hydrologic performance, and out-of-sample robustness (Tyralis et al., 2019). The primary hyper-parameters of an RF model are the number of trees in the forest ('ntree') and the number of features to randomly select at each split ('mtry').

The RF model was implemented using the 'ranger' package in R (Wright & Ziegler, 2017) and the default hyper-parameter settings of 'ntree' = 500 and 'mtry' = $\sqrt[2]{k}$, where k is the number of variables. While we found that some improvement in error correction was possible through hyper-parameter selection on the 'out-of-bag' prediction error, these improvements were modest and had the negative effect of apportioning more variable importance to the lagged errors, which degraded simulation performance (see supporting information S2).

The RF model is fit to calibration period data and then used to predict the errors in the validation set ($\hat{e}_t^{val}$). These predicted errors are subtracted from the raw errors e_t^{val} to yield residuals ε_t^{val} in the validation period, which are used to train the dynamic residual model, described next.

3.3.2. Dynamic Residual Model

The residual model captures the stochastic properties of ε_t, and is 'dynamic' in the sense that it allows the stochastic properties to vary over time based on hydrologic model state. This model is fit to the validation set of error correction model residuals (ε_t^{val}), which ensures it does not underestimate the variability of out-of-sample residuals from the error correction model (see supporting information S3).

To construct this model, we leverage the generalized likelihood (GL) approach of Schoups and Vrugt (2010) that utilizes the flexible skew exponential power (SEP) distribution (also known as the skew generalized error distribution; Wurtz et al., 2020). The original GL approach includes an autoregressive model and a linear model for heteroscedasticity which results in a set of random deviates (a_t) that are modeled via the SEP with a mean μ of 0, a standard deviation σ of 1 (i.e., after standardization by the heteroscedastic model), kurtosis β, and skew ξ. We modify this formulation to allow all free parameters of the GL model (standard deviation σ, kurtosis β, skew ξ, and lag-1 autoregressive coefficient φ) to vary over time:

$$\mathcal{L}(\eta|\varepsilon) = \sum_{t=1}^{n} log\frac{2\sigma_{\xi_t}\omega_{\beta_t}}{\xi_t+\xi_t^{-1}} - log\sigma_t - c_{\beta_t}|a_{\xi,t}|^{2/(1+\beta_t)} \quad \text{Eq. (3)}$$

$$\sigma_t = \sigma_0 + \sigma_1\theta_{SV,t} \quad \text{Eq. (3a)}$$

$$\beta_t = \beta_0 + \beta_1\theta_{SV,t} \quad \text{Eq. (3b)}$$

$$log_{10}\xi_t = \xi_0 + \xi_1\theta_{SV,t} \quad \text{Eq. (3c)}$$

$$\varphi_t = \varphi_0 + \varphi_1\theta_{SV,t} \quad \text{Eq. (3d)}$$

The log-likelihood function for the SEP distribution of ε (Eq. 3) is a function of the parameter vector $\eta = \{\sigma_0, \sigma_1, \beta_0, \beta_1, \xi_0, \xi_1, \varphi_0, \varphi_1\}$, which determines how the standard deviation, kurtosis, skew, and lag-1 autocorrelation change based on model state variables ($\theta_{SV,t}$) (Eq. 3a-d). Maximization of the log-likelihood function simultaneously estimates all $4(m+1)$ parameters in η, where m is the number of state variables. In the Appendix, we define other intermediate terms ($\sigma_{\xi_t}, \omega_{\beta_t}, c_{\beta_t}, a_{\xi,t}$) required in the likelihood function, following Schoups and Vrugt (2010). Prior to maximum likelihood estimation, we scale all state variables to prevent discrepancies in magnitude from impacting the inferred parameters, and we preserve this scaling when simulating residuals from the SEP distribution in new time periods. When maximizing the likelihood function, we ensure the free parameters remain within valid ranges ($\sigma_t > 0;\ \beta_t > (-1);\ 0.1 < \xi_t < 10; 0 \leq \varphi_t \leq 1$) by penalizing parameter selections that result in parameter values outside of these ranges. As part of this constraint for σ_t, we require σ_0 to be no lower than the mean of the absolute value of the lowest decile of the residuals, and we require all elements of the vector σ_1 to be non-negative. Finally, ξ_t is log-transformed to linearize its relationship with $\theta_{SV,t}$ (see supporting information S4 for more detail).

This modified GL approach allows the residual model to capture state dependent, time varying properties of variance, autocorrelation, and distributional form. Moreover, the dynamic model allows for adaptive prediction of residual error distributions even if the model state variables extend beyond their historical range, which is a challenge for other recently developed local uncertainty estimation procedures (e.g., BLUECAT, Montanari & Koutsoyiannis, 2022).

3.3.3. SWM Ensemble Generation and Benchmark Static SWM

To generate SWM simulations, we first generate new sets of random deviates $\tilde{a}_t$ from the SEP distribution with $\mu = 0, \sigma = 1$ and the time-varying estimates $\widehat{\beta_t}$ and $\hat{\xi}_t$, which are determined by the model state variables via Eq. 3b-c. These $\tilde{a}_t$ are then converted to new $\tilde{\varepsilon}_t$ timeseries via Eq. 4, where estimates of the lag-1 autoregressive parameter $\hat{\varphi}_t$ and the heteroscedastic parameter $\hat{\sigma}_t$ are inferred from Eq. 3a and 3d:

$$\tilde{\varepsilon}_t = \hat{\varphi}_t \tilde{\varepsilon}_{t-1} + \hat{\sigma}_t \tilde{a}_t \;\; where \;\; \tilde{a}_t \sim SEP(0,1,\hat{\beta}_t, \hat{\xi}_t) \qquad \text{Eq. (4)}$$

We then combine the simulated residuals $\tilde{\varepsilon}_t$ with the error correction model to simulate new errors $\tilde{e}_t$ (Eq. 5). These errors are generated as the sum of the predicted error from the error correction model, $f\left(\theta_{SV,t}, \tilde{e}_{t-3}, \tilde{e}_{t-2}, \tilde{e}_{t-1}\right)$, which depends on the state variables at time t $(\theta_{SV,t})$ and the generated errors from the previous 3 timesteps $(\tilde{e}_{t-3}, \tilde{e}_{t-2}, \tilde{e}_{t-1})$, and the generated residual error $(\tilde{\varepsilon}_t)$.

$$\tilde{e}_t = f\left(\theta_{SV,t}, \tilde{e}_{t-3}, \tilde{e}_{t-2}, \tilde{e}_{t-1}\right) + \tilde{\varepsilon}_t \qquad \text{Eq. (5)}$$

Since this model includes lag-1 to lag-3 errors, it is initialized with 3 randomly generated deviates from the residual error model. Importantly, this simulation procedure includes contributions from model state variables directly (via the error correction model), as well as through the stochastic distribution of $\tilde{\varepsilon}_t$ (via the dynamic residual model). This novel integration of dynamic, state variable dependent components enables generalizable error simulation with intrinsic adaptability for out-of-sample and non-stationary error distributions.

To benchmark the hybrid error model, we also introduce a static SWM designed similar to the hybrid model but without dependence on hydrologic state variables. The static SWM has an error model fit to historical errors e_t from the calibration period (1987-1997) on a monthly basis to capture seasonality. First, monthly mean biases are estimated and removed from e_t, producing residuals similar to ε_t in Eq. 2. Then, an autocorrelative model (AR(3)) and a heteroscedastic transform are fit to ε_t to remove autocorrelation and capture variance that changes with simulated flow, and an SEP distribution is fit to the decorrelated and scaled residuals. This is the basic approach proposed in Schoups and Vrugt (2010) and is very similar to the dynamic residual model in Eq. 3, although these fits are conducted separately by month without dependence on state variables. Simulation from the static SWM follows a similar procedure to the dynamic residual model, with monthly biases added back in during simulation.

3.4. Local Interpretable Model-Agnostic Explanation (LIME)

To understand the time dependent importance of state variables in the RF error correction process (section 3.3.1), we use an explainable artificial intelligence (xAI, Holzinger et al., 2022) technique referred to as LIME (Ribeiro et al., 2016). LIME randomly perturbs the inputs around each model prediction to develop a local, sparse linear approximation to the more complex ML model's predictive logic. This linear model provides a representation of the relative importance of each input to the ML model's final prediction at each time step (Ribeiro et al., 2016; Hvitfeldt et al., 2022). In this context, LIME has similarities to time-varying sensitivity analyses used in hydrologic model diagnoses (Herman et al., 2013). Importantly, LIME offers a uniquely different perspective than the aggregate variable importance metrics generated internally by the RF

algorithm, since the explanations can be analyzed at the precision of individual events or aggregated over subsets of interest.

3.5. Real-World Application

As a final experiment, we apply the hybrid SWM framework developed in section 3.3 to a non-stylized, real-world setting, where the hybrid error model is constructed on the errors between the actual observed streamflow and a process based hydrologic model (SAC-SMA) calibrated to those observations. In this case, the 'truth model' is now the more complex real world streamflow generating process against which hydrologic models are a simplified representation, and we assess the ability of the hybrid SWM framework to learn an error distribution in the training period that generalizes to the test period. We note that the actual observations are themselves modeled via a full natural flow estimation procedure that accounts for upstream diversions and reservoir operations (Zimmerman et al., 2018), but these human influences can change in nature over time in ways not totally captured by the natural flow estimation algorithm. Any non-stationarity in human influences over observed inflows not corrected by the natural flow estimation algorithm will lead to nonstationarity in errors between the observations and process model predictions unrelated to hydrologic model state variables, making the state variable – error relationships much more challenging to learn and simulate.

4. Results

4.1. Non-stationarity in raw errors

To first highlight the potential for non-stationarity in predictive errors, we examine the raw error distribution (e_t, Eq. 1) by month between the process (HYMOD) and truth (SAC-SMA) models

482 in the out-of-sample test period with and without 4^{o}C warming applied to the meteorological data
483 (Figure 4a). Note that predictive errors are defined as truth minus process model output. There is
484 a substantial divergence in error distributions between the historical and warmed scenarios across
485 seasons, with the most notable differences occurring in the late winter and early spring
486 (February-May). In February and March, when mean daily flows are rising towards their annual
487 peak (Figure 4b), the errors in the two cases are biased in opposite directions, with process model
488 outflows systematically overpredicting truth model flows in the Test case but underpredicting
489 them in the Test+4C case. In April and May, when the two cases exhibit the greatest disparities
490 in mean daily flow (Figure 4b), error biases are of the same sign but are more severe in the
491 Test+4C case. In addition, there are several months when the error dispersion (i.e., interquartile
492 range) differs substantially between the two cases, with January, March, and April being the
493 most prominent examples. Overall, the error distribution between the truth and process models
494 are significantly different under the Test and Test+4C scenarios, highlighting the potential for
495 nonstationary predictive errors.

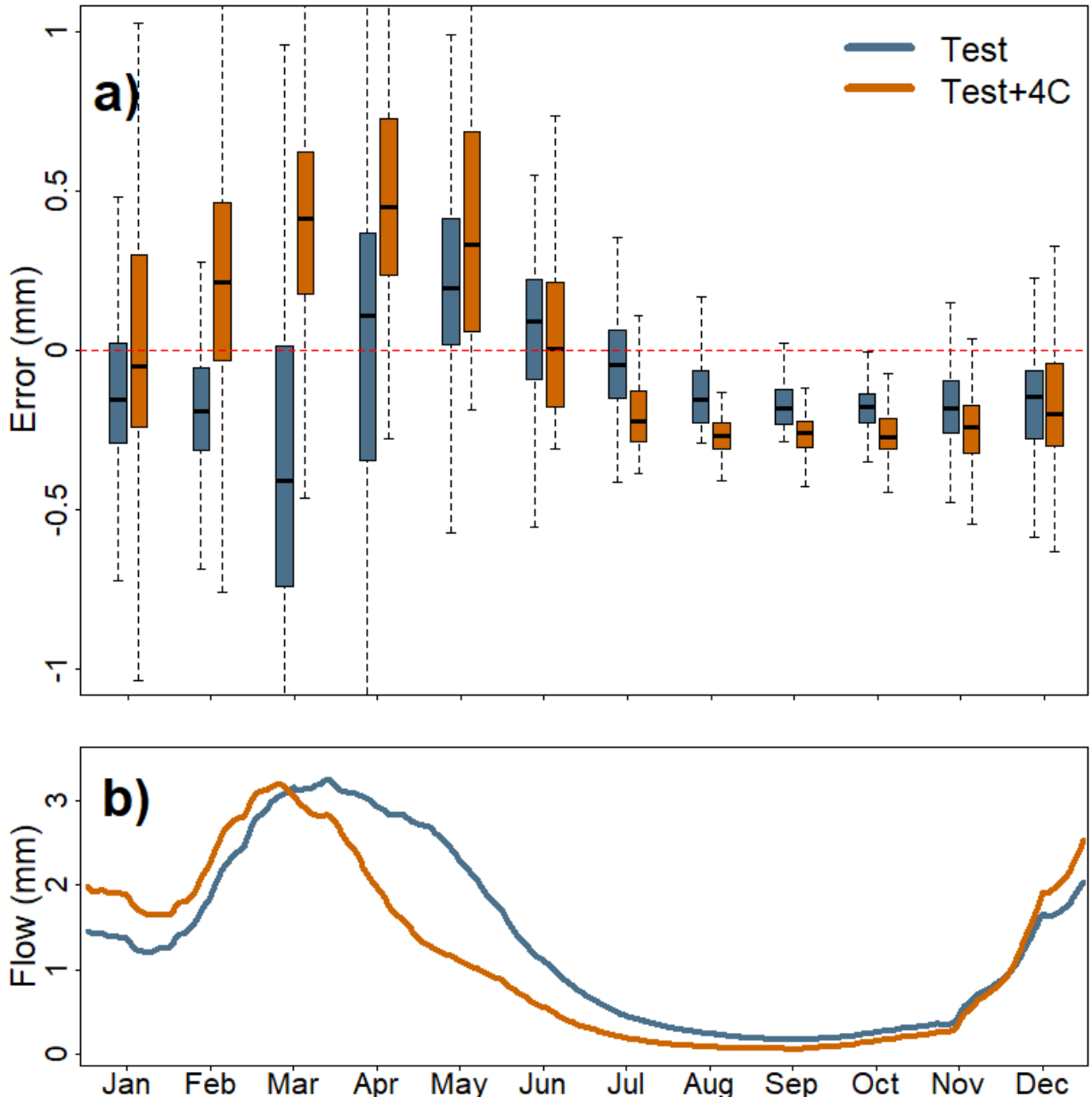
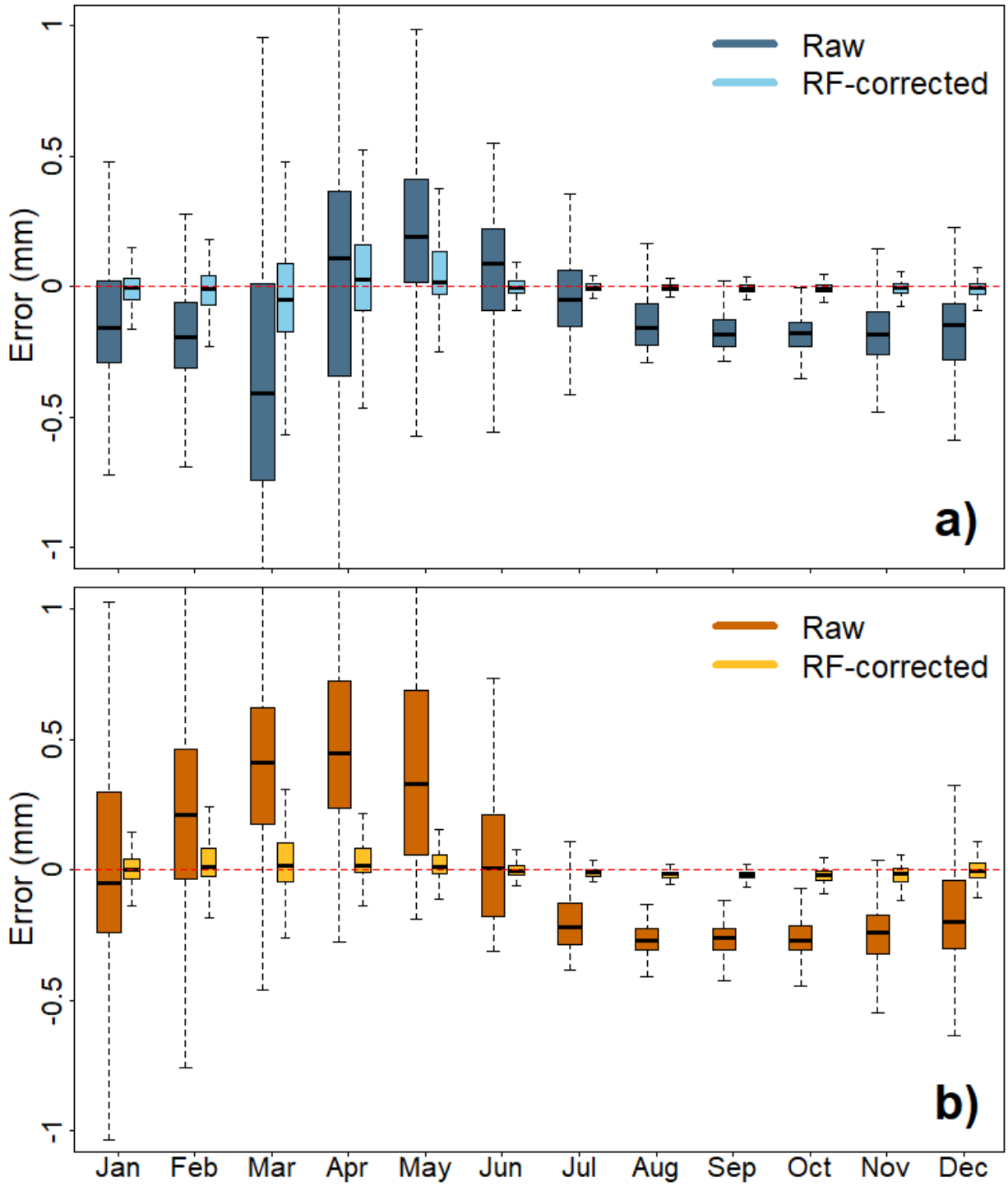


Figure 4. *a) Monthly comparison of out-of-sample (WY2004-2013) raw error distributions (e_t) between truth (SAC-SMA) and process (HYMOD) models without (Test) and with (Test + 4C) warming. B) Mean daily flow of the SAC-SMA truth model across WY2004-2013 for the two scenarios, smoothed with a 30-day moving average.*

4.2. RF error correction model performance

In the first step of our modeling approach, we apply RF error correction to remove systematic biases that can vary through time conditional on hydrologic state. Figure 5 shows the residual distributions (ε_t, Eq. 2) for both Test and Test + 4C cases after fitting this error correction procedure to the training set and applying it to the test set. Figure 5 also shows the raw error distributions (e_t, Eq. 1) for comparison. There is a clear reduction in conditional bias across months, with residuals now consistently centered around zero. Notably, the raw errors were successfully debiased in both Test and Test + 4C cases in the late winter and spring (February-May), when biases in the two cases were of different sign or substantially different magnitude. In addition, the error correction model reduces the dispersion of the raw predictive errors considerably across months. These results showcase three important properties of the error correction process: a) the model's ability to learn state variable-error relationships that enable debiasing across varying seasonal behavior; b) the model's resistance to overfitting (i.e., the RF model provides effective error correction on unseen data in both the Test and Test+4C case); and c) stability of the learned relationships even with prominent shifts to the raw error distributions under non-stationary forcing.

519



520

521 ***Figure 5.*** *a) Monthly comparison of out-of-sample Test period (2003-2013) raw error (e_t)*
522 *distributions versus residual distributions (ε_t) after correction by the Random Forest (RF)*
523 *model. b) As in a) but for the Test+4C period (2003-2013).*
524

The RF model calculates variable importance as the fractional contribution of each variable to reducing prediction variance across the entire dataset. Figure 6 shows that lag-1 and -2 autocorrelation in the raw errors are the most influential predictor variables. The most important state variables are the runoff component of simulated flow and the simulated flow itself, implying that conditional biases in the error are related to differences in how rainfall is apportioned to overland flow between the truth and process models. The remaining state variables show similar, lower values of importance, but we note that some variables that would be important only in specific times of year (e.g., snow water equivalent, SWE) will likely be less important in the aggregate.

The dominance of autocorrelation in variable importance suggests that a simpler autoregressive (AR) model could be sufficient as an error correction procedure. However, an AR model cannot simulate conditional bias that changes in nature under non-stationary conditions (Shabestanipour et al., 2023). A RF error correction model based solely on state variables (no lag terms) can infer conditional bias in both out-of-sample and non-stationary out-of-sample cases, but underpredicts the magnitude of the bias (see supporting information S5). This supports the integration of autoregressive and state variables in the RF error correction model.

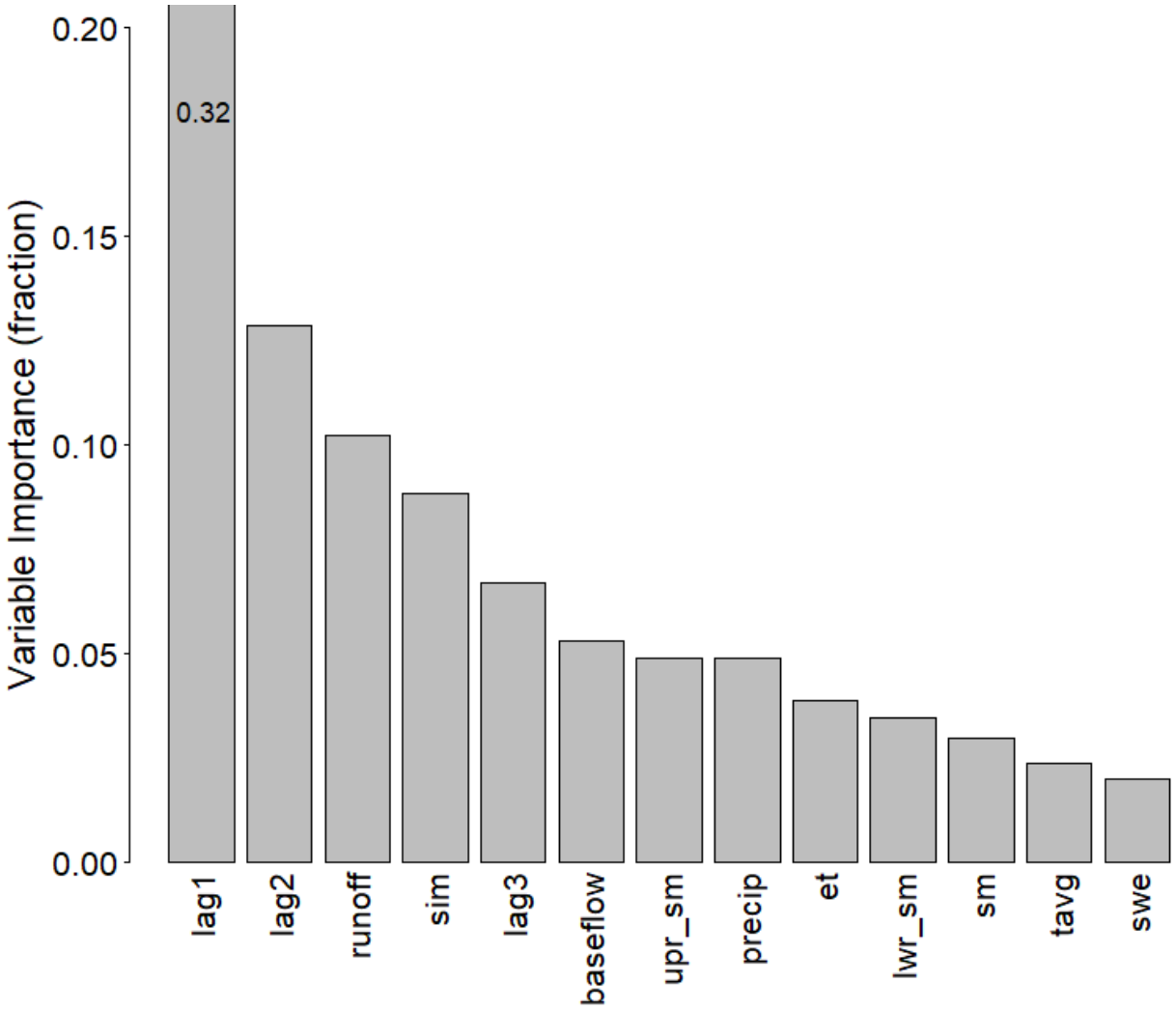


Figure 6. *Variable importance from RF error correction model fit to calibration period (1987-1997). Note: 'lag1' variable importance equals 0.32, which extends outside of plot bounds.*

While the RF model calculates variable importance across the entire dataset, we use LIME to explore the time varying importance of state variables to the error correction model. This is shown in Figure 7 for the Test and Test+4C cases, using results in March for illustration. Here, we confine our analysis to daily empirical errors that bias in opposite directions for the Test and Test+4C cases, in order to better emphasize how state variable impacts on error correction change based on background climate state. That is, we take the daily feature weights from LIME in March only when the predictive errors are negative (positive) for the Test (Test+4C) cases,

and show the median feature weight for these days in Figure 7. We do not include the lag-1 to lag-3 features in Figure 7 to concentrate focus on the state variable effects (supporting information S6 shows all features).

Figure 7 shows that the process model simulated flow (sim) and runoff exhibit the largest absolute feature weights, but these feature weights are of opposite sign between the Test and Test+4C cases and the Test+4C weights are of lower magnitude. The difference in sign between the two cases suggests that the RF model uses simulated flow and runoff in fundamentally different ways to correct bias in March depending on the background climate state. The lower magnitude weights in the Test+4C case likely reflect the smaller bias for the Test+4C errors in March (see Figure 5). We also note a change in sign for the feature weight on SWE, though the absolute weights are relatively small. There is also a noticeable reduction in the magnitude of the baseflow feature weight from the Test to the Test+4C case.

Overall, the results from Figures 6 and 7 show that the RF error correction model is able to infer biases of changing sign and magnitude based on changes to the process model simulated flow itself, particularly the runoff component, with some lesser contributions from SWE.

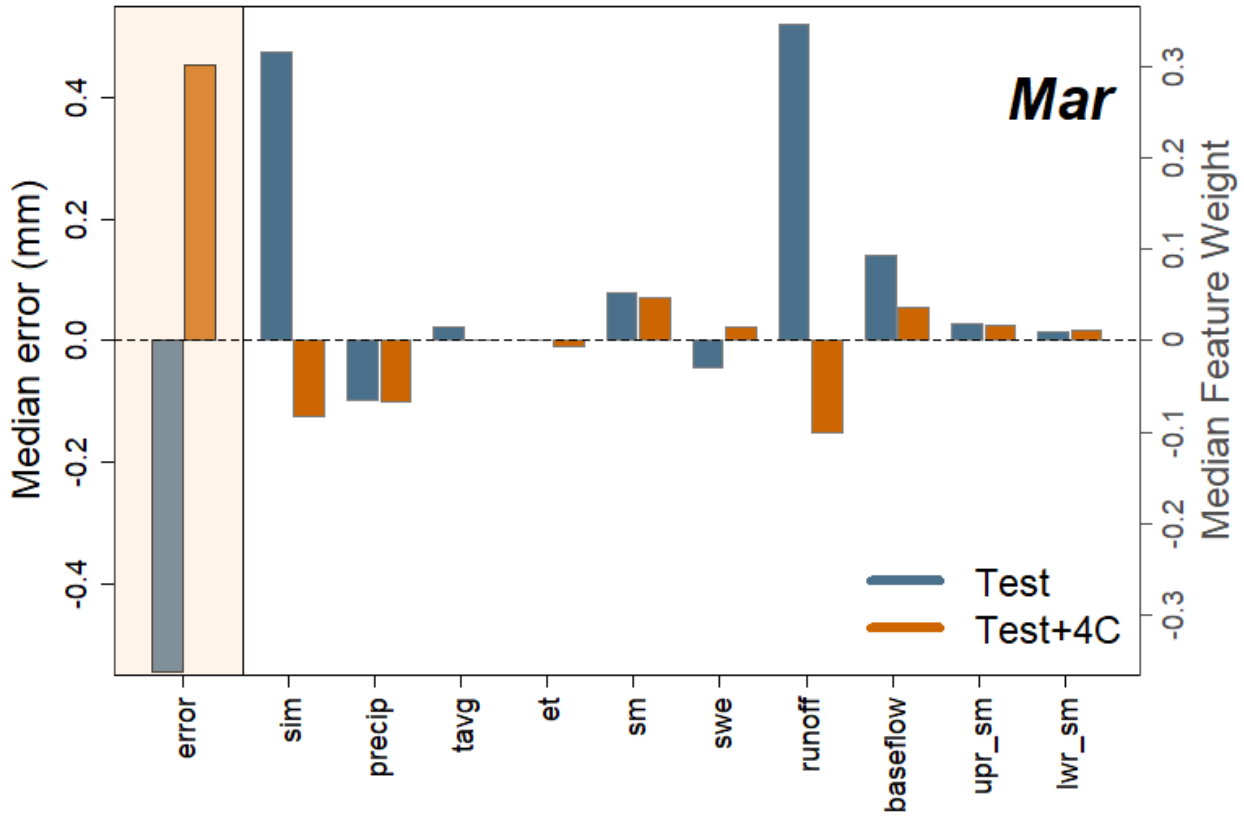


Figure 7. *Locally Interpretable Model-agnostic Explanation (LIME) median feature weight (white background, gray outline) comparison against median error (light orange background, black outline) for a selected month, where errors and associated feature weights are aggregated for errors less than (greater than) zero in the Test (Test+4C) cases.*

4.3. Dynamic residual model performance

Overall, the error correction process yields residuals (ε_t) in both Test and Test+4C cases that are unbiased, but that still exhibit time dependent properties (e.g., variance that changes by month; see Figure 5). This suggests that important dependencies between the model states and the residuals may still exist after error correction. We assess the ability of the dynamic residual model to capture these dependencies by comparing the empirical residual distribution (i.e., the distribution of ε_t calculated from Eq. 2) to the residual distribution simulated by the dynamic residual model ($\tilde{\varepsilon}_t$ in Eq. 4), all for the out-of-sample Test and Test+4C cases. Figure 8 shows this comparison for selected months (March-May) that exhibited the most notable differences

between residual distributions in the Test and Test+4C cases (see Figure 5), but a comparison across all months is presented in Figure S7. Results from Figure 8 show that in the Test case (top row), the dynamic residual model captures seasonal changes to the residual distribution's shape, variance, and skew. In the Test+4C case (bottom row), the empirical residual distributions become more peaked compared to the Test case, and the dynamic residual model is able to infer these changes. The close agreement between empirical and simulated residuals in Figure 8 confirms that the dynamic residual model is able to use state variable information to capture changes in higher moments of the residuals ε_t across months and very different climate conditions.

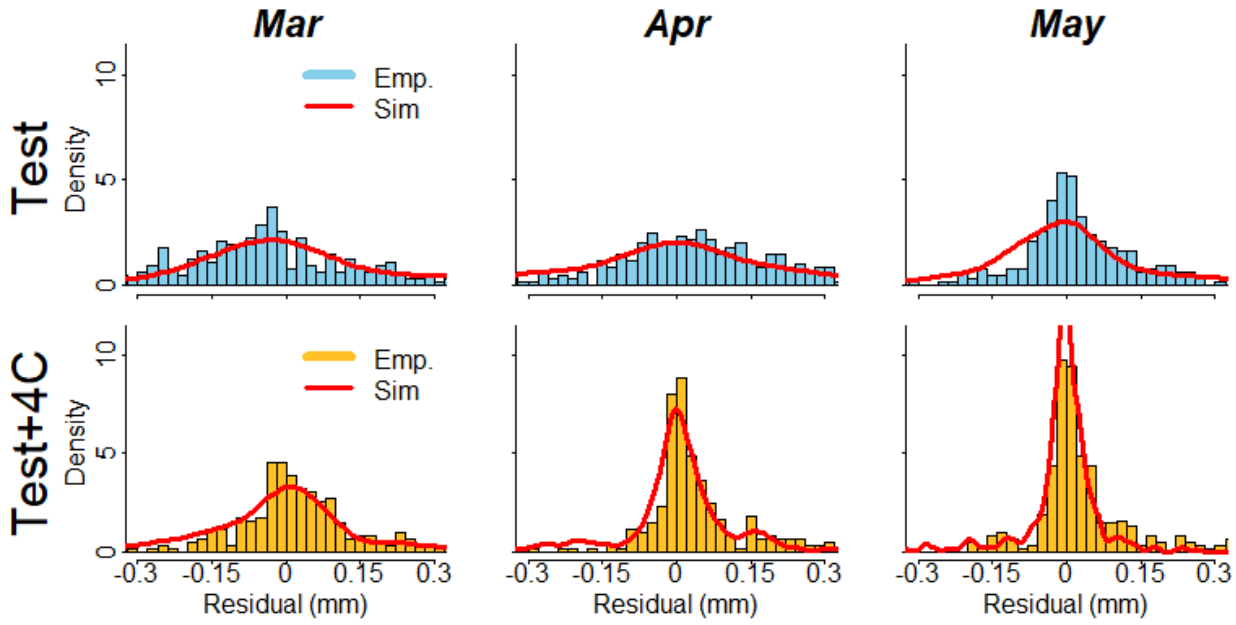


Figure 8. *Top row: Empirical distribution of RF-corrected residuals ε_t (histogram) versus the kernel density estimate of a simulated sample of residuals $\tilde{\varepsilon}_t$ from the dynamic residual model (red line) for selected months from the Test case. Bottom row: As in top row, but for the Test+4C case.*

603 Table 1 shows the state variable effects for the different parameters in the SEP model (see Eqs.
604 3a-3d), while Figure 9 shows the seasonality in SEP parameters in Test and Test+4C cases. For
605 the parameter σ_t (heteroscedasticity), the runoff state variable is the most influential, followed
606 closely by the simulated flow and then precipitation (Table 1). This result reflects the strong
607 relationship between error variance and flow magnitude, as noted in previous literature (Schoups
608 & Vrugt, 2010). This is also seen in the strong seasonal signal in both the mean and variability in
609 σ_t (Figure 9, top left). Further, the greatest divergence in σ_t between the Test and Test+4C cases
610 occurs in the late winter and spring months where mean flow magnitudes diverge most
611 substantially (see Figure 4b).
612
613 Skewness (ξ_t) shows a relatively weak seasonal signal centered around 1 (i.e., $log_{10}\xi_t = 0$; no
614 skew) across months, with only weak relationships with evapotranspiration (et), snow water
615 equivalent (swe), runoff, and soil moisture (sm). In contrast, the kurtosis parameter β_t exhibits a
616 strong seasonal signal that is primarily tied to temperature (tavg). The residual distributions
617 exhibit values of β_t close to 1 (i.e., a Laplace distribution) in the cold season that become
618 progressively more peaked and fat-tailed ($\beta_t > 1$) in the summer months. This reflects a
619 concentration of probability mass around small residuals ε_t in low flow months with high
620 probability of large (scaled) residuals (see Figure S4). Both the Test and Test+4C cases show
621 similar seasonal characteristics with slightly higher β_t values in the late winter to spring for the
622 Test case.
623
624 Finally, lag-1 autocorrelation (φ_t) exhibits a seasonal peak in summer, similar to β_t but with a
625 decrease in variability during those months. The lowest values of φ_t, but highest variability,

occur in the cold season. It is important to note that the autocorrelation captured in the residual model is the leftover autocorrelation after error correction (which included lag-1 to lag-3 terms). Thus, the results highlight that the dynamic residual model infers relatively high degrees of additional autocorrelation in the summer months with little variability and moderate autocorrelation with more variability in other seasons. The greatest positive contributors to φ_t are temperature and lower-zone soil moisture (lwr_sm), while precipitation is the most important negative contributor. Across the seasons, there is very little difference between the Test and Test+4C cases, indicating that the autocorrelation structure of the residuals ε_t is not highly influenced by warming.

Table 1. *State variable coefficients for multiple linear regression models of the 4 parameters in the dynamic residual model. The 'intcpt' row corresponds to* $[\sigma_0, \beta_0, \xi_0, \varphi_0]$ *from Eq. 3a-d, while the remaining rows correspond to the coefficient vectors* $[\sigma_1, \beta_1, \xi_1, \varphi_1]$ *from Eq. 3a-d from left to right. All state variables are scaled, so the magnitude of the coefficient is proportional to its effect on the parameter.*

Residual Model Coefficients				
	σ_t	β_t	$\log_{10}(\xi_t)$	φ_t
intcpt	0.002	0.083	0.016	0.503
sim	0.103	0.043	−0.014	0.003
runoff	0.149	0.022	−0.035	0.008
baseflow	0.000	0.055	0.006	0.000
sm	0.000	0.074	−0.030	0.010
upr_sm	0.000	0.073	−0.011	−0.015
lwr_sm	0.000	0.078	0.015	0.036
swe	0.000	0.060	0.029	0.001
et	0.000	0.021	0.039	0.015
tavg	0.000	0.877	−0.015	0.040
precip	0.054	0.027	−0.009	−0.077

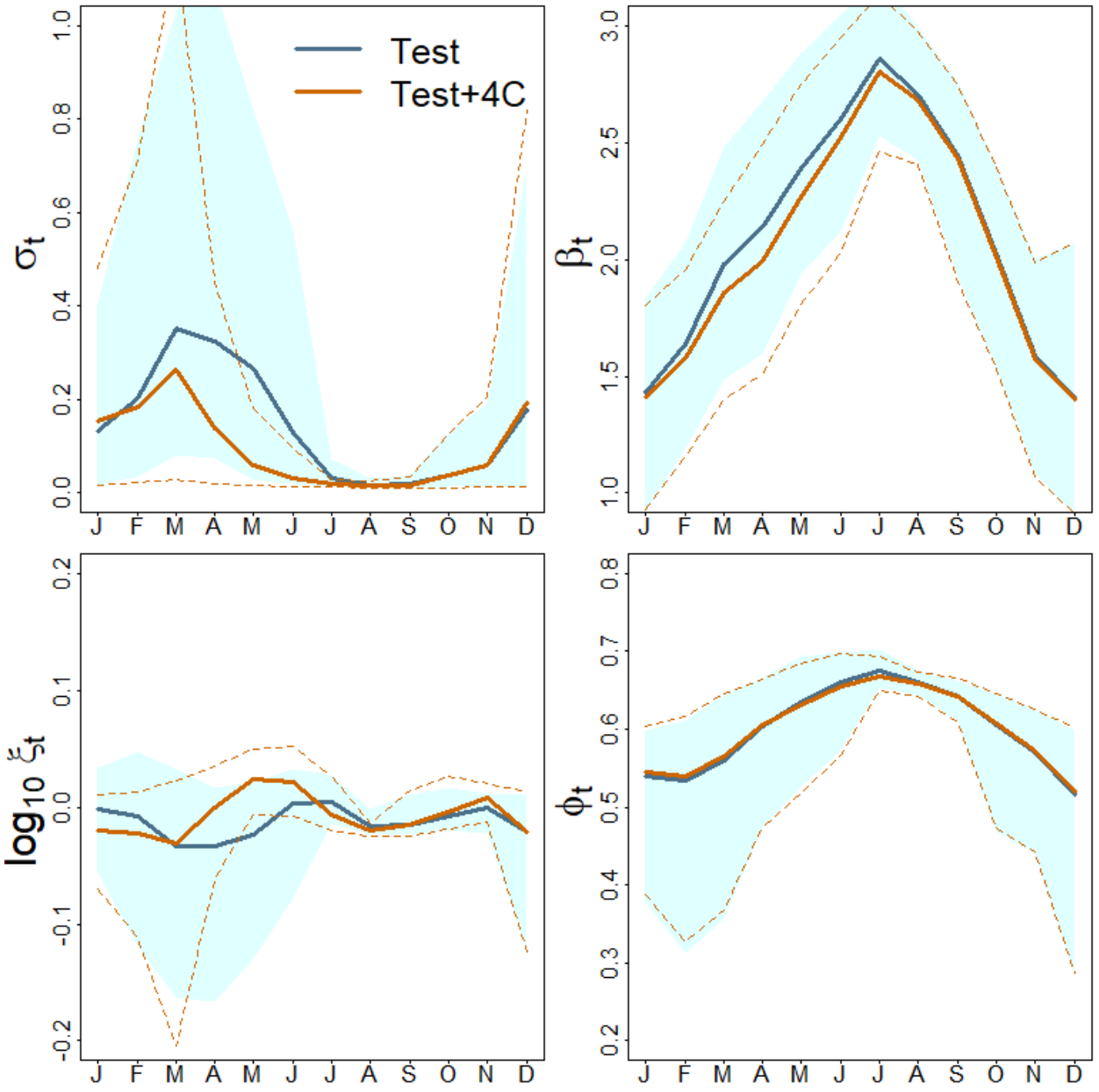


Figure 9. *Values for the four free parameters in the dynamic residual model aggregated by month for the Test and Test+4C cases. The bold lines are the mean parameter values while the blue shading is the 90% confidence interval for the Test case and the orange dashed line is the 90% confidence interval for the Test+4C case.*

4.4. Hybrid model simulation performance

After fitting both components of the hybrid error model, we simulate new errors ($\tilde{e}_t$, Eq. 5) via the generation procedure detailed in section 3.3.3 and evaluate how well their distribution

matches that of the raw empirical errors (e_t, Eq. 1). In Figure 10a,b, we show this comparison separately by month for both Test and Test+4C cases . The hybrid model is able to reproduce the direction of bias and general patterns of variance across both the Test and Test+4C cases. For instance, in February the Test empirical errors are negatively biased and have less variance in comparison to the Test+4C errors, which are positively biased and have greater variance. The model is able to simulate both of these shifts. A similar result is seen in March, although the variance is smaller in the Test+4C period. In April, the Test empirical errors have relatively low positive bias and large variance as compared to the Test+4C errors, which have a large positive bias and less variance. The simulated errors capture this general pattern.

There are some deficiencies in the simulated errors, including biases that are often smaller than the observed bias (e.g., see March – May in Test+4C) and some overestimation of variance (particularly in the Test case). However, the hybrid model significantly outperforms a static SWM based purely on seasonality in the error distribution (Figure 10c,d). For the static model, although error distributions are well simulated by month during the Test case, changes in the direction of bias during the winter and spring are completely missed because monthly biases are fixed (further information on static model in supporting information S8). The state variable dependencies built into the hybrid model allow a more faithful emulation of these shifts, even if imperfect. We also note that there are some (albeit very minor) improvements over the static SWM simulations in terms of coverage probabilities in the Test+4C case, although both models tend to be overdispersed (see supporting information S9).

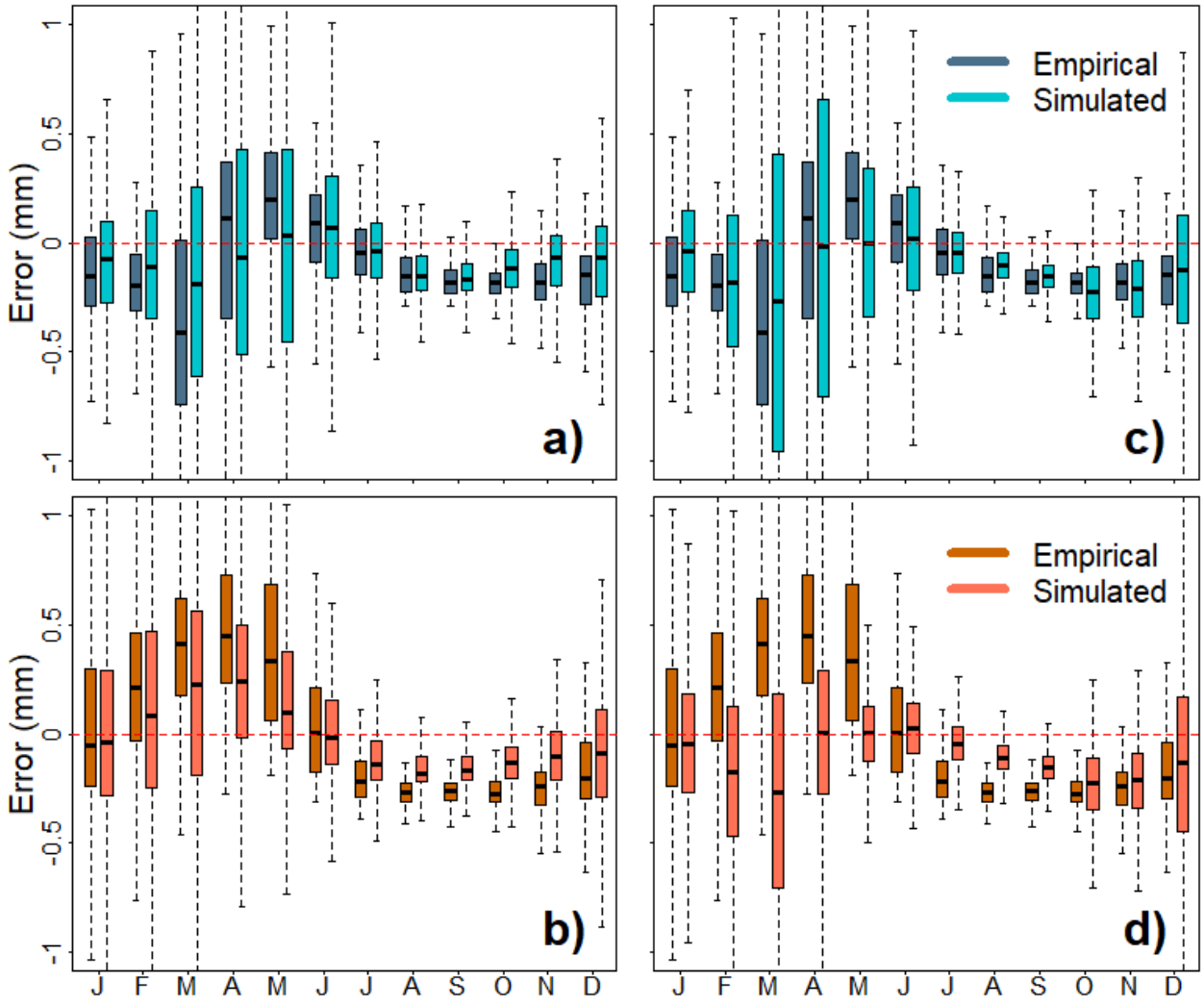


Figure 10. *a) Monthly empirical distribution of errors in the Test subset (dark blue) versus 1000 aggregated samples of hybrid model simulated errors (light blue). b) Same as (a) but for Test+4C errors, where empirical (simulated) errors are dark orange (coral). c-d) As in (a) and (b), but simulated errors are from the static SWM model.*

To further illustrate the performance of the hybrid SWM, Figure 11 shows the simulated timeseries of flow for a 6-month subperiod (February-July 2011) in the Test and Test+4C cases that spans both wet and dry seasons. We first highlight the markedly different truth model flows for the Test and Test+4C cases, where again the only difference is the applied +4^{o}C temperature adjustment to the Test+4C forcings. The peak flow event in March in the Test case is weaker and

of shorter length compared to the larger, sustained multi-peak event in the Test+4C case. In contrast, the snowmelt recession is longer and of higher magnitude in the Test case versus the Test+4C case. The hybrid SWM results show that the largest uncertainty is inferred around peak events as well as flow recessions, with particularly large uncertainty in the Test April-June snowmelt season. The results also illustrate the method's adaptive bias correction, where the hybrid SWM corrects much of the process model's overprediction bias in the Test case, particularly in April-June, whereas in the Test+4C case for the same months, the model helps correct for process model underprediction bias. Around the larger peak flows in February-March for both Test and Test+4C, there is little bias correction for this particular year. Overall, the hybrid SWM simulations improve the process model simulation based on the ensemble median and capture many of the observations within the ensemble spread.

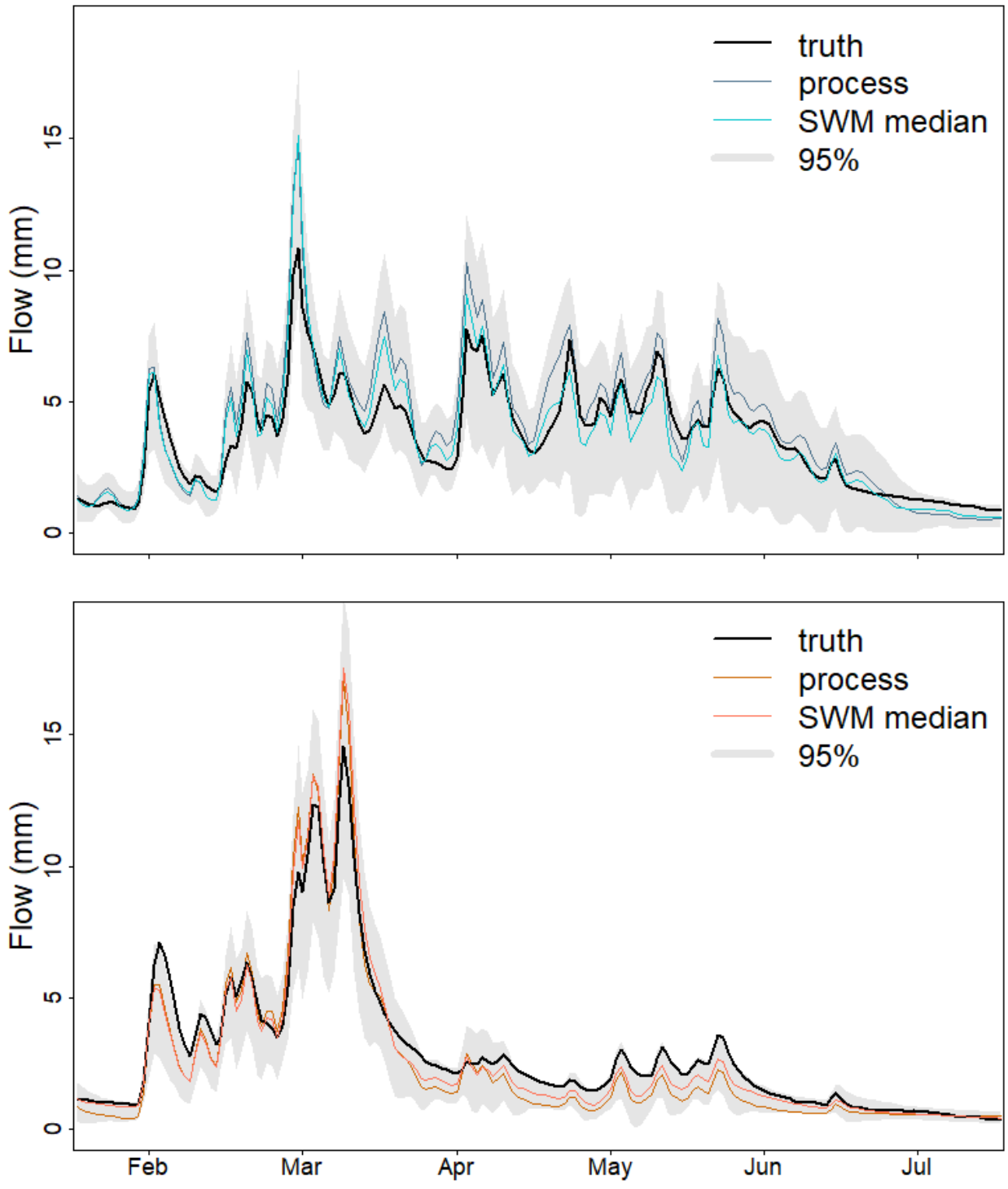


Figure 11. *a) Truth model flow (black) compared against process model flow (dark blue) and the median flow of 1000 samples (light blue) from the hybrid SWM for the February-July period in 2011 from the Test scenario. 95% coverage interval for the 1000 samples are shown in light gray. b) As in (a) but for the Test+4C scenario, where process model (SWM median) flow is dark orange (light orange).*

717 **4.5. Real-world application**

718 We conclude by employing the hybrid model in a real-world setting to assess whether the model
719 is effectively inferring conditional error distributions when the truth model is the actual
720 streamflow observations and the process model is SAC-SMA. In this application we utilize two
721 sampling schemes to determine calibration, validation, and test periods: 1) a 'split-sample'
722 approach identical to the idealized example described in Figure 3; and 2) a 'skip-sample'
723 approach, where calibration, validation, and test periods are sampled evenly over the historical
724 record in three-year increments (i.e., calibration water-years: 1988, 1991,…, 2012; validation
725 water-years: 1989, 1992,…, 2013; testing water-years 1990, 1993,…, 2011). The skip-sample
726 approach addresses the challenges associated with any non-stationarity in the observed data (and
727 errors) over the period of record (WY1988-2013) that is not represented by modeled flows
728 forced with historical climate (e.g., changes in diversions over time not captured by the natural
729 flow estimation procedure; see Section 3.5). The hybrid model under the split-sample approach
730 would not be able to capture this type of non-stationarity, since hydrologic state variables would
731 be uncorrelated with these shifts in model error. We note that there is no appreciable trend in
732 temperature or precipitation over the 1988-2013 period, so any non-stationarity in errors across
733 the 1988-2013 period is not likely driven by climatic factors (see supporting information S10).
734
735 Figure 12 shows the results of applying the hybrid model to errors between the process model
736 (SAC-SMA) and observed streamflow. There is slightly more bias in the raw, empirical errors by
737 month in the split-sample case as compared to the skip-sample case, although these biases are
738 significantly smaller than those that emerged under the idealized Test+4C scenario. In addition,
739 the test period empirical biases are relatively similar across the split-sample and skip-sample

approaches, despite the fact that there is little overlap in the test data under these two sampling methods. Under the split-sample approach, the hybrid model does capture the direction of biases in the summer season but not in the winter and spring months. This deficiency appears to stem from learned biases in the calibration period that do not transfer to the test period (see Figure S11). In addition, the model overestimates variance considerably across months. In contrast, the hybrid model does well in estimating both the error bias and variance in the 'skip-sample' case (also see Figure S12). Without more knowledge about what is driving changes in the error distribution across training and testing periods (e.g., climate, human activity), it is hard to draw strong conclusions about the hybrid model in this real-world setting. However, these results do suggest that the hybrid model's performance may be sensitive to non-stationarity in errors associated with factors uncorrelated with hydrologic state variables.

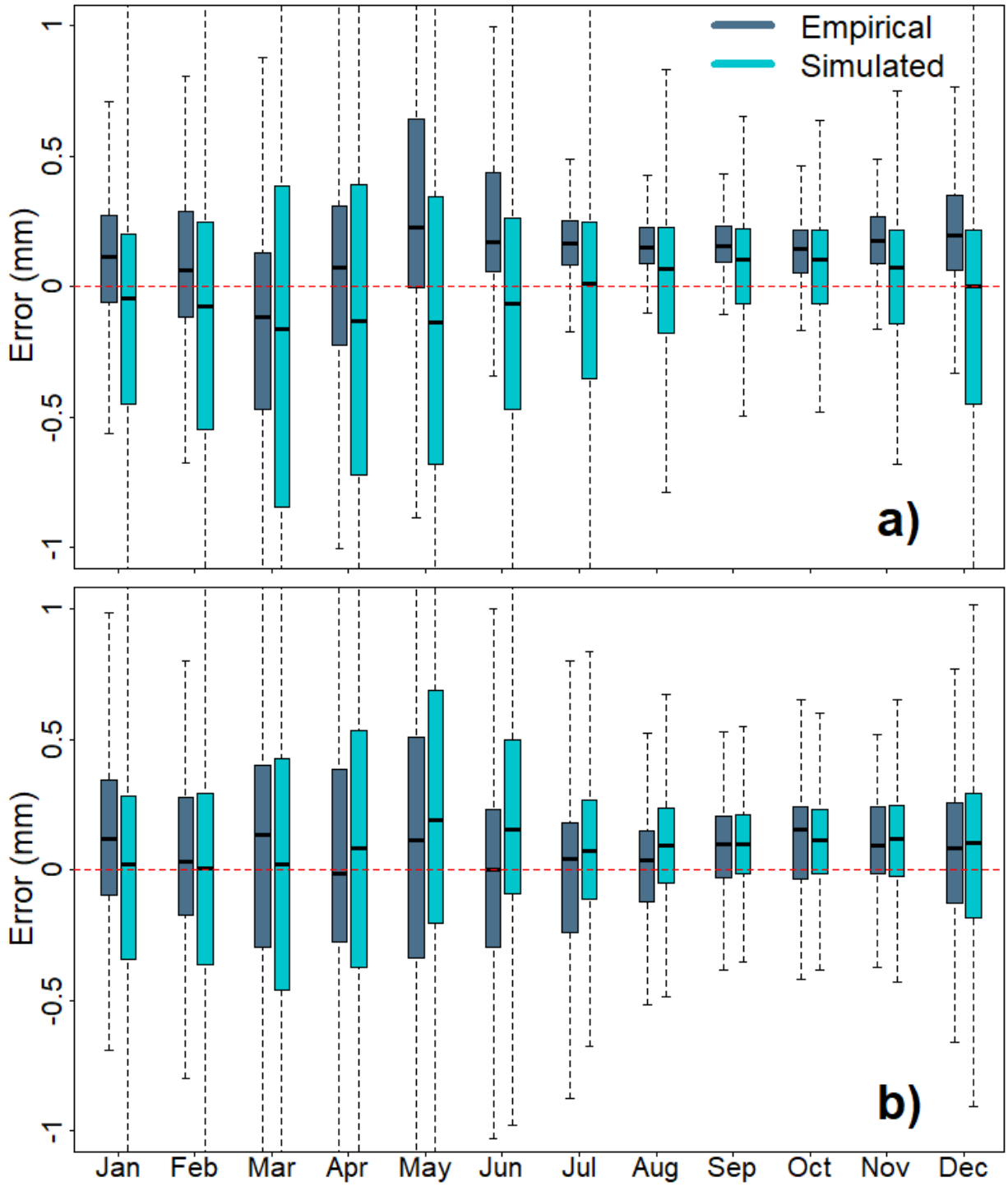


Figure 12. *a) Monthly empirical raw error distributions (dark blue) between observed streamflow and the process model (SAC-SMA) against simulated errors from the hybrid model (light blue) for the out-of-sample Test period (WY 2004-2013) under the split-sample approach. b) Same as (a), but for the 'skip-sample' approach.*

5. Discussion and Conclusion

In this work, we examined the assumption that historical predictive uncertainty of hydrologic models is sufficient to characterize future predictive uncertainty under non-stationary climate. We developed an idealized 'model as truth' experimental design to test this assumption, where we designated one hydrologic model as 'truth' and another as the 'process' model. This design allowed us to analyze predictive uncertainty under both the historical meteorological conditions to which the models were fit and also under significant warming. We found that there were substantial shifts in the predictive error distribution under climate change, which were manifest in changes to bias, variance, and (to a lesser extent) higher moments of error. These results suggest that SWMs fit to historical data may not perform well when used to simulate future, climate change impacted hydrology.

One of the most important contributions of SWMs is the reduction of simulation bias in hydrological model predictions at the upper and lower flow quantiles (Farmer & Vogel, 2016; Vogel, 2017). This simulation bias often leads to systematic errors in the estimation of extreme low flow (e.g., 7Q10), high flow (e.g., 100-year flood), and other design events (Shabestanipour et al., 2023). Although not explicitly shown here, the differences in error distributions between the Test and Test+4C cases imply that the current generation of SWMs trained to a historical period may not improve the estimation of these design criteria under future climate conditions.

To address these issues, we developed a novel, hybrid SWM to leverage information in hydrologic model state variables to predict changes in predictive uncertainty. The model used

781 ML error correction to remove biases conditional on hydrologic state, and then used dynamic
782 residual modeling to capture the dependencies between hydrologic state and higher order
783 moments of the error distribution. To better emulate out-of-sample predictive uncertainty, we
784 introduced a training approach whereby we fit the error correction model to a calibration set and
785 then subsequently fit the dynamic residual model to a separate validation set, before evaluating
786 the approach on an independent test set.
787
788 We found that the hybrid model was able to capture prominent shifts in predictive uncertainty in
789 the test set, both for historical climate (Test) and under warming (Test+4C). This included
790 significant changes in bias during the winter and spring months, when snow accumulation and
791 melt dynamics differed significantly between the truth and process models in the Test and
792 Test+4C cases. Notably, a static benchmark SWM was unable to emulate these shifting biases.
793 The hybrid modeling framework was also able to predict changes in error variance and kurtosis
794 in the spring months under warming, and autocorrelation that varied across the year. Overall,
795 predictive uncertainty estimated using the hybrid error model matched that observed between the
796 truth and process model reasonably well, even though some attributes of predictive uncertainty
797 (e.g., magnitude of bias; coverage probabilities) were not captured. While improvements in some
798 of these attributes should be the focus of future work, our methodology provides an important
799 step towards addressing a gap in the hydrologic ML literature of how to adequately assess
800 uncertainty under plausible but unprecedented future conditions (Klotz et al., 2022).
801
802 Using different approaches for model interpretability (e.g., feature importance, LIME), we
803 showed that lagged error terms and components of simulated streamflow were the most

important features when correcting for bias, while a variety of meteorological and internal state variables helped model changes in higher order moments and autocorrelation of the residuals. Importantly, the effects of certain features in the error correction model changed in sign depending on the background climate and month of interest, suggesting that changes to predictive uncertainty under non-stationarity are more complex than just shifts in timing (e.g., Xu et al., 2021) or simple scaling relationships (e.g., Read & Vogel, 2015). This work demonstrates an approach to leverage relationships between model state and model error to infer these complex changes, paving the way for future work in this area.

We also tested the hybrid model in a more challenging real-world setting, where the hybrid error model had to predict changes in predictive uncertainty between a process model and actual streamflow observations in the context of (potentially unobserved) changes to the real hydrologic system that would be uncorrelated with hydrologic model state. We found that the hybrid error model worked reasonably well in some months but struggled in others when the data were separated into sequential training and testing periods. These issues were resolved if data used for model training and testing were evenly spread out across the record, suggesting that the hybrid error model may be sensitive to nonstationarity in the true hydrologic system, especially if the source of nonstationarity is unrelated to features simulated by the process model. Future work should examine this problem in more depth, since understanding how hydrologic predictive uncertainty might change against a future, unobserved real world system under non-stationarity is the ultimate goal.

In constructing the hybrid error model used in this work, we emphasized generalizability, interpretability, and parsimony over complexity. Future work could explore more complex error correction procedures (boosted trees, convolutional neural networks, or long short-term memory networks), more complex optimization schemes for the dynamic residual model, or non-linear relationships to state variables in the dynamic residual model. Furthermore, this study only evaluated the hybrid error model in one location, and so future work should assess the spatial generalizability of the approach and whether performance varies by region.

Finally, the experimental design forwarded in this study offers a unique way to explore the tradeoff between traditional 'predictive' performance objectives (e.g., NSE) of hydrologic modeling versus 'functional' performance (structural adequacy) objectives (Ruddel et al., 2019) in the context of SWMs. Complexity in hydrologic model predictive uncertainty arises from the predictive errors 'doing the work the model should be doing' (Vogel, 2017). Thus, a structurally deficient hydrologic model coupled with a sufficiently complex error model may perform well in scenarios that are similar to its training data, but will likely perform poorly under nonstationarity, where the hydrologic model's structural faults will be amplified. In contrast, a more structurally adequate hydrologic model may exhibit more (but less complex) predictive uncertainty in stationary conditions, but show more consistency in error distributions under nonstationarity. Using the information theoretic framework developed in Ruddel et al. (2019), one could explore this tradeoff explicitly across different hydrologic models (or hydrologic model parameterizations) in relation to its effect on the complexity of the predictive error distributions as well as their homogeneity between the stationary and non-stationary cases.

Appendix

We provide the intermediate equations derived in Schoups and Vrugt (2010) to define the conditional generalized likelihood (GL) function with modifications to account for time varying kurtosis (β_t), skew (ξ_t), and lag-1 autocorrelation (φ_t). The reader is referred to this manuscript for further details on the derivations.

$$\omega_{\beta,t} = \frac{\Gamma^{1/2}[3(1+\beta_t)/2]}{(1+\beta_t)\Gamma^{3/2}[(1+\beta_t)/2]} \quad \text{Eq. (A1)}$$

$$c_{\beta,t} = \frac{\Gamma[3(1+\beta_t)/2]^{1/(1+\beta_t)}}{\Gamma[(1+\beta_t)/2]} \quad \text{Eq. (A2)}$$

$$M_{1,t} = \frac{\Gamma[1+\beta_t]}{\Gamma^{1/2}[3(1+\beta_t)]\Gamma^{1/2}[(1+\beta_t)/2]} \quad \text{Eq. (A3)}$$

$$M_2 = 1 \quad \text{Eq. (A4)}$$

$$\mu_{\xi,t} = M_{1,t}(\xi_t + \xi_t^{-1}) \quad \text{Eq. (A5)}$$

$$\sigma_{\xi,t} = \sqrt{\left(M_2 - M_{1,t}^2\right)(\xi_t^2 + \xi_t^{-2}) + 2M_{1.t}^2 - M_2} \quad \text{Eq. (A6)}$$

$$a_t = \frac{\varepsilon_t - \varphi_t \varepsilon_{t-1}}{\sigma_t} \quad \text{Eq. (A7)}$$

$$a_{\xi,t} = \xi_t^{-sign(\mu_{\xi,t}+\sigma_{\xi,t}a_t)}(\mu_{\xi,t} + \sigma_{\xi,t}a_t) \quad \text{Eq. (A8)}$$

Data Availability Statement

All code and data are available in public GitHub (https://github.com/zpb4/Nonstationary-SWM) and Zenodo repositories (https://doi.org/10.5281/zenodo.7689054) respectively and cited in the references.

Acknowledgements

This study was supported by the U.S. National Science Foundation grant EnvS-1803563.

Reference

Abramowitz, G., & Bishop, C. H. (2015). Climate model dependence and the ensemble dependence transformation of CMIP projections. *Journal of Climate*, *28*(6), 2332–2348. https://doi.org/10.1175/JCLI-D-14-00364.1

Baecher, G. B., & Galloway, G. E. (2021). US Flood risk management in changing times. *Water Policy*, *23*, 202–215. https://doi.org/10.2166/wp.2021.269

Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, *61*(9), 1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A, et al. (2019). Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., et al. (2019). Changing climate both increases and decreases European river floods. *Nature*, *573*(7772), 108–111. https://doi.org/10.1038/s41586-019-1495-6

Boland, J. J., & Loucks, D. P. (2021). Infrastructure capacity planning for reducing risks of future hydrologic extremes. *Water Policy*, *23*, 188–201. https://doi.org/10.2166/wp.2021.242

Boyle, D. P. (2001). Multicriteria calibration of hydrologic models, (Doctoral dissertation). Retrieved from UA Campus Repository (http://hdl.handle.net/10150/290657), Tucson, AZ: The University of Arizona.

Bracken, C., Rajagopalan, B., & Zagona, E. (2014). A hidden Markov model combined with climate indices for multidecadal streamflow simulation. *Water Resources Research*, *50*, 7836–7846. https://doi.org/10.1002/2013WR014979

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1007/9781441993267_5

Brodeur, Z. P. (2023). Data from: Nonstationary-SWM. GitHub repository. https://github.com/zpb4/Nonstationary-SWM. Accessed 1 March 2023.

Brodeur, Z. P. (2023). Data from: Non-stationary SWM dataset. Zenodo repository. https://doi.org/10.5281/zenodo.7689054. Accessed 1 March 2023.

Brown, C. M., Lund, J. R., Cai, X., Reed, P. M., Zagona, E. A., Ostfeld, A., et al. (2015). Scientific Framework for Sustainable Water Management. *Water Resources Research*, 6110–6124. https://doi.org/10.1002/2015WR017114

Burnash, R. J. (1995). The NWS river forecast system - catchment modeling. In Singh, V. (Ed.), *Computer Models of Watershed Hydrology* (pp. 311-366). Littleton, CO: Water Resources Publication.

Farmer, W., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*, 5619–5633. https://doi.org/:10.1002/2016WR019129

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., & Nearing, G. S. (2021). Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. *Journal of the American Water Resources Association*, *57*(6), 885–905. https://doi.org/10.1111/1752-1688.12964

Galloway, G. E. (2011). If stationarity is dead, what do we do now? *JAWRA Journal of the American Water Resources Association, 47*(3), 563–570.

Hadjimichael, A., Quinn, J., Wilson, E., Reed, P., Basdekas, L., Yates, D., & Garrison, M. (2020). Defining Robustness, Vulnerabilities, and Consequential Scenarios for Diverse Stakeholder Interests in Institutionally Complex River Basins. *Earth's Future*, *8*(7), 1–22. https://doi.org/10.1029/2020EF001503

Hah, D., Quilty, J. M., & Sikorska-Senoner, A. E. (2022). Ensemble and stochastic conceptual data-driven approaches for improving streamflow simulations: Exploring different hydrological and data-driven models and a diagnostic tool. *Environmental Modelling and Software*, *157*(August), 105474. https://doi.org/10.1016/j.envsoft.2022.105474

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

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The Elements of Statistical Learning* (Second). Berlin: Springer.

Herger, N., Angélil, O., Abramowitz, G., Donat, M., Stone, D., & Lehmann, K. (2018). Calibrating Climate Model Ensembles for Assessing Extremes in a Changing Climate. *Journal of Geophysical Research: Atmospheres*, *123*(11), 5988–6004. https://doi.org/10.1029/2018JD028549

Herman, J. D., Reed, P. M., & Wagener, T. (2013). Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research*, *49*(3), 1400–1414. https://doi.org/10.1002/wrcr.20124

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, Klaus-Robert, & Samek, W. (Eds.)(2022). *xxAI-Beyond Explainable AI*. Springer, Switzerland. https://doi.org/10.1007/978-3-031-04083-2. Retrieved from https://link.springer.com/bookseries/1244

Huang, G., Kadir, T., & Chung, F. (2012). Hydrological response to climate warming: The Upper Feather River Watershed. *Journal of Hydrology*, *426–427*, 138–150. https://doi.org/10.1016/j.jhydrol.2012.01.034

Hui, R., Herman, J., Lund, J., & Madani, K. (2018). Adaptive water infrastructure planning for nonstationary hydrology. *Advances in Water Resources*, *118*(October 2017), 83–94. https://doi.org/10.1016/j.advwatres.2018.05.009

Hunter, J., Thyer, M., McInerney, D., & Kavetski, D. (2021). Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. *Journal of Hydrology*, *603*(PA), 126578. https://doi.org/10.1016/j.jhydrol.2021.126578

Hvitfeldt, E., Pedersen, T., Benesty, M. (2022). *lime: Local Interpretable Model-Agnostic Explanations*. https://lime.data-imaginist.com, https://github.com/thomasp85/lime.

Inter-American Development Bank (IDB) (2017). Inter-American Development Bank Sustainability Report 2017. p. 68. http:// dx.doi.org/10.18235/0001034. Available at: https://publications.iadb.org/publications/english/document/Inter-American- Development-Bank-Sustainability-Report-2017.pdf.

Inter-government Panel on Climate Change (IPCC) (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/9781009157896.002.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., … Nearing, G. (2022). Uncertainty estimation with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences*, *26*(6), 1673–1693. https://doi.org/10.5194/hess-26-1673-2022

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. https://doi.org/10.1002/2016GL072012

Konapala, G., Kao, S. C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10). https://doi.org/10.1088/1748-9326/aba927

Koutsoyiannis, D., & Montanari, A. (2015). Negligent killing of scientific concepts: the stationarity case. *Hydrological Sciences Journal*, *60*(7–8), 1174–1183. https://doi.org/10.1080/02626667.2014.959959

Koutsoyiannis, D., & Montanari, A. (2022). Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions. *Water Resources Research*, *58*(1). https://doi.org/10.1029/2021WR031215

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1–2), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010

Lehner, F., Wahl, E. R., Wood, A. W., Blatchford, D. B., & Llewellyn, D. (2017). Assessing recent declines in Upper Rio Grande runoff efficiency from a paleoclimate perspective. *Geophysical Research Letters*, *44*(9), 4124–4133. https://doi.org/10.1002/2017GL073253

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., & Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions. *Hydrology and Earth System Sciences*, *16*(4), 1239–1254. https://doi.org/10.5194/hess-16-1239-2012

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*(7), 1–18. https://doi.org/10.1029/2006WR005756

Livneh, B., Bohn, T., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., et al. (2015). A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950-2013. *Scientific Data*, *2*, 150042. https://doi.org/10.1038/sdata.2015.42

Lohmann, D., Raschke, E., Nijssen, G., & Lettenmaier, D. (1998). Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model. *Hydrological Sciences Journal*, *43:1*, 131-141. https://doi.org/10.1080/02626669809492107

Loucks, D. P., & van Beek, E. (2017). *Water Resource Systems Planning and* 873 *Management*. Springer International Publishing. https://doi.org/10.1007/978-3-319-874 44234-1

Mankin, J. S., Seager, R., Smerdon, J. E., Cook, B. I., & Williams, A. P. (2019). Mid-latitude freshwater availability reduced by projected vegetation responses to climate change. *Nature Geoscience*, *12*(12), 983–988. https://doi.org/10.1038/s41561-019-0480-x

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*, 2199–2239. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., & Kuczera, G. (2018). A simplified approach to produce probabilistic hydrological model predictions. *Environmental Modelling and Software*, *109*(July), 306–314. https://doi.org/10.1016/j.envsoft.2018.07.001

Miller, D., & White, R. A. (1998). A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling, *Earth Interactions*, *2*(2), 1-26. https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Zbigniew, W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity Is Dead : Whither Water Management ? *Science*, *319*(February), 573–575.

Montanari, A., & Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, *40*(1), 1–11. https://doi.org/10.1029/2003WR002540

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of 895 uncertain hydrological systems. *Water Resources Research*, *48*(9), 896 2011WR011412. https://doi.org/10.1029/2011WR011412

Montanari, A., & Koutsoyiannis, D. (2014). Modeling and mitigating natural hazards: Stationarity is immortal! *Water Resources Research*, *50*, 9748–9756. https://doi.org/10.1002/ 2014WR016092

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water (Switzerland)*, *10*(11), 1–40. https://doi.org/10.3390/w10111536

Mote, P. W., Li, S., Lettenmaier, D. P., Xiao, M., & Engel, R. (2018). Dramatic declines in snowpack in the western US. *Npj Climate and Atmospheric Science*, *1*(1). https://doi.org/10.1038/s41612-018-0012-1

Musselman, K. N., Clark, M. P., Liu, C., Ikeda, K., & Rasmussen, R. (2017). Slower snowmelt in a warmer world. *Nature Climate Change*, *7*(3), 214–219. https://doi.org/10.1038/nclimate3225

Nash, J. E., & Sutcliff, J. V. (1970). River flow forecasting through conceptual models part I–A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nearing, G. S., Pelissier, C. S., Kratzert, F., Klotz, D., Gupta, H. V, Frame, J. M., & Sampson, A. K. (2019). Physically Informed Machine Learning for Hydrological Modeling Under Climate Nonstationarity. *Science and Technology Infusion Climate Bulletin; NOAA's National Weather Service 44th NOAA Annual Climate Diagnostics and Prediction Workshop Durham, NC, 22-24 October 2019*, (October), 22–24. Retrieved from https://www.nws.noaa.gov/ost/climate/STIP/44CDPW/44cdpw-GNearing.pdf

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, *57*(3). https://doi.org/10.1029/2020WR028091

Overpeck, J. T., & Udall, B. (2020). Climate change and the aridification of North America. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(22), 11856–11858. https://doi.org/10.1073/pnas.2006323117

Quilty, J. M., Sikorska-Senoner, A. E., & Hah, D. (2022). A stochastic conceptual-data-driven approach for improved hydrological simulations. *Environmental Modelling and Software*, *149*(January), 105326. https://doi.org/10.1016/j.envsoft.2022.105326

Read, L. K., & Vogel, R. M. (2015). Reliability, return periods, and risk under nonstationarity. *Water Resources Research*, *51*, 6381–6398. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011). Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resources Research*, *47*(11). https://doi.org/10.1029/2011WR010643

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. https://doi.org/10.18653/v1/n16-3020

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information Theory for Model Diagnostics: Structural Error is Indicated by Trade-Off Between Functional and Predictive Performance. *Water Resources Research*, *55*(8), 6534–6554. https://doi.org/10.1029/2018WR023692

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

Seo, D. J., Herr, H. D., & Schaake, J. C. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.*, *3*, 1987–2035. Retrieved from www.hydrol-earth-syst-sci-discuss.net/3/1987/2006/

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic Watershed Model Ensembles for Long-Range Planning : Verification and Validation. *Water Resources Research*, *59*. https://doi.org/10.1029/2022WR032201

Shalev-Shwartz, S., & Ben-David, S. (2013). *Understanding machine learning: From theory to algorithms*. *Understanding Machine Learning: From Theory to Algorithms* (Vol. 9781107057). https://doi.org/10.1017/CBO9781107298019

Shamseldin, A.Y., O'Connor, K.M. (2001). A non-linear neural network technique for updating of river flow forecasts. *Hydrol. Earth Syst. Sci. 5*, 577–598. https://doi.org/10.5194/hess-5-577- 2001

Sharma, S., Ghimire, G. R., & Siddique, R. (2021). Machine learning for postprocessing ensemble streamflow forecasts. arXiv preprint arXiv:2106.09547.

Shen, C., Chen, X., & Laloy, E. (2021). Editorial: Broadening the Use of Machine Learning in Hydrology. *Frontiers in Water*, *3*(May), 1–4. https://doi.org/10.3389/frwa.2021.681023

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H., & Karssenberg, D. (2022). Random forests based error correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. *Computers and Geosciences*. https://doi.org/10.1016/j.cageo.2021.105019

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to Update the Split-Sample Approach in Hydrological Model Calibration. *Water Resources Research*, *58*(3), 1–26. https://doi.org/10.1029/2021WR031523

Sikorska, A. E., Montanari, A., & Koutsoyiannis, D. (2015). Estimating the Uncertainty of Hydrological Predictions through Data-Driven Resampling Techniques. *Journal of Hydrologic Engineering*, *20*(1), 1–10. https://doi.org/10.1061/(asce)he.1943-5584.0000926

Stakhiv, E. Z., & Hiroki, K. (2021). Special Issue for UN HELP: "Water infrastructure planning, management and design under climate uncertainty." *Water Policy*, *23*, 1–9. https://doi.org/10.2166/wp.2021.268

Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*(12), 2823–2839. https://doi.org/10.1002/hyp.10409

Sterle, K., Hatchett, B. J., Singletary, L., & Pohll, G. (2019). Hydroclimate Variability in Snow-fed River Systems: Local Water Managers' Perspectives on Adapting to the New Normal. *Bulletin of the American Meteorological Society*, (June), BAMS-D-18-0031.1. https://doi.org/10.1175/BAMS-D-18-0031.1

Teegavarapu, R. S. V., Salas, J. D., & Stedinger, J. R. (Eds.). (2019). *Statistical Analysis of Hydrologic Variables: Methods and Applications*. American Society of Civil Engineers. https://doi.org/10.1061/9780784415177

Thomas, H., & Fiering, M. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. In *Design of water resources systems*, edited by A. Mass, et al., 459–493. Cambridge, MA: Harvard University Press.

Toth, E., Montanari, A., Brath, A., 1999. Real-time flood forecasting via combined use of conceptual and stochastic models. Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos. 24, 793– 798. https://doi.org/10.1016/S1464-1909(99)00082-9

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of Random Forests for water scientists and practitioners and their recent history in water resources. *Water*, *11*(910), 1–37. https://doi.org/10.3390/w11050910

Xu, D., Ivanov, V. Y., Li, X., & Troy, T. J. (2021). Peak Runoff Timing Is Linked to Global Warming Trajectories. *Earth's Future*, *9*(8). https://doi.org/10.1029/2021EF002083

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H., & Anh, D. T. (2020). Deep learning convolutional neural network in rainfall-runoff modelling. *Journal of Hydroinformatics*, *22*(3), 541–561. https://doi.org/10.2166/hydro.2020.095

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

Wang, Q. J. (1991). The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resources Research*, *27*(9), 2467-2471. https://doi.org/10.1029/91WR01305

Wilks, D. S., (2019). *Statistical Methods in the Atmospheric Sciences, 4th ed.* Cambridge, MA: Elsevier.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R., *Journal of Statistical Software, 77*(1), 1-17. http://doi.org/10.18637/jss.v077.i01

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. https://cran.r-project.org/web/packages/fGarch

Zha, X., Xiong, L., Guo, S., Kim, J.-S., & Liu, D. (2020). AR-GARCH with Exogenous Variables as a Postprocessing Model for Improving Streamflow Forecasts. *Journal of Hydrologic Engineering*, *25*(8), 1-16. https://doi.org/10.1061/(asce)he.1943-5584.0001955

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