A hybrid, non-stationary Stochastic Watershed Model (SWM) for uncertain hydrologic simulations 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 promise in predicting 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 is that the historical predictive uncertainty may not correctly characterize the error distribution under future climate. For example, the frequency of physical processes (e.g., snow accumulation and melt)) 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.. We develop a hybrid machine learning method that maps model state variables to predictive errors, allowing for non-stationary error distributions based on changes in the frequency of model states. We find that this procedure provides an internally consistent methodology to overcome stationarity assumptions in error modeling and offers an important advance for implementing SWMs under climate change. We test this method on three hydrologically distinct watersheds in California (Feather River, Sacramento River, Calaveras River), finding that the hybrid model performs best in large, snowmelt dominated basins.

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 regions where 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 (SSMs) 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 SSMs 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.

Stochastic watershed models (SWM) were developed 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 short term probabilistic prediction (e.g., flood forecasting; Sikorska et al., 2015; McInerney et al., 2018; Koutsoyiannis & Montanari, 2022), subseasonal-to-seasonal forecasting (McInerney et al., 2020), 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 distribution of hydrologic model prediction errors in the historical period implicitly reflects the historical distribution of 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 SWMs 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). 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 add realistic physical constraints to model predictions and 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 than direct ML prediction models for long-term projections that extrapolate hydrologic responses under unprecedented climate change (Wi and Steinschneider, 2022, 2023; Feng et al., 2023; Reichart et al., 2023). A similar logic may support non-stationary models of hydrologic model error for future uncertainty quantification. That is, 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 Lake in northern California. We first introduce a stylized 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 document non-stationarity in the error distribution between them. We then develop a hybrid error model composed of an ML-based error correction model and a dynamic residual noise model, both of which use process model state variables to infer error properties. We assess the ability of this hybrid error model, coupled with simulations from the process model, to preserve the statistical properties of the truth model in out-of-sample cases with and without the impacts of climate change and compare results to a static SWM approach as a benchmark. We conclude the study by demonstrating the same technique for a process model fit to actual streamflow observations in the Feather River basin, as well as two other California basins (Sacramento River above Shasta Lake, Calaveras River above New Hogan Lake) that differ in size and hydrologic regime.

2. Study Area and Data

The Feather River basin upstream of Lake Oroville drains an area of 9338 km^2 on the west facing slopes of the northern Sierra Nevada Range (Figure 1), and serves as the largest water source directly contributing to the State Water Project of California. This portion of the Sierra Nevada reaches altitudes of nearly 3000 m, making the Feather River a snow-dominated catchment. The precipitation regime is driven by large, infrequent atmospheric rivers (ARs) that exhibit significant inter-annual variability and occur primarily in the cold season (November – April) (Hanak et al., 2011). Accordingly, streamflow varies considerably across years and also across seasons, as snowmelt drives higher flows in the spring and early summer months and high evapotranspiration drives lower flows in late summer and fall. Winter flows can vary considerably in response to winter storms, particularly when associated with AR-induced warming or rain-on-snow events (Hanak et al., 2011; Huang et al., 2012). Climate change is expected to significantly impact hydrologic response in this region through reduced snowpack, earlier snowmelt, and changing precipitation characteristics (Hanak et al., 2011; Huang et al., 2012; Sterle et al., 2019).

We also consider flows in the Sacramento River upstream of Shasta Lake (SHA) and the Calaveras River upstream of New Hogan Lake (NHG) for part of our analysis described in Section 3.5. The Sacramento River basin is approximately two times larger than the Feather River basin (12262 km^2) and flows out of the southern Cascade Range, which is lower in elevation than the Sierra Nevada and less snowmelt dominated. The Calaveras River basin is approximately one tenth the size of the Feather River basin (940 km^2) and originates in the foothills of the Sierra Nevada, making it primarily rainfall dominated and its hydrology flashier than the Feather or Sacramento Rivers.

Daily streamflow data were taken from the California Data Exchange Center (CDEC) Full Natural Flow (FNF) database for water years (WY) 1989-2018 (October 1, 1989 – September 30, 2018) at Oroville Dam on the Feather River (CDEC ID: ORO), Shasta Dam on the Sacramento River (CDEC ID: SHA), and New Hogan Dam on the Calaveras River (CDEC ID: NHG) (CA DWR, 2024). These data represent unimpaired flows, or the natural water production of a river basin unaltered by upstream diversions, storage, and imports of water (see Figure 1). FNFs are calculated from observed flows and estimates of water diversion and imports, reservoir operations, and reservoir evaporation. We note that even though FNFs on the Feather River are calculated rather than observed natural flows, this will have little impact on our stylized experiment described in Section 3, which fits and compares one hydrologic model to another. In addition, the two other study locations are relative unimpaired, with only a few, very small reservoirs upstream of Shasta Lake with insignificant storage capacity (~ 1% of mean annual flow), and no reservoirs on the Calaveras River upstream of New Hogan Lake.

We used daily precipitation from the 6 km climate product of Pierce et al. (2021) as input forcings to all hydrologic models used in this work. Daily minimum and maximum temperature were taken from the 6 km dataset in Livneh et al. (2015) up through December 2015 and then extended to September 2018 using the PRISM daily dataset (PRISM Climate Group, 2014) to match the timeframe of the precipitation data.

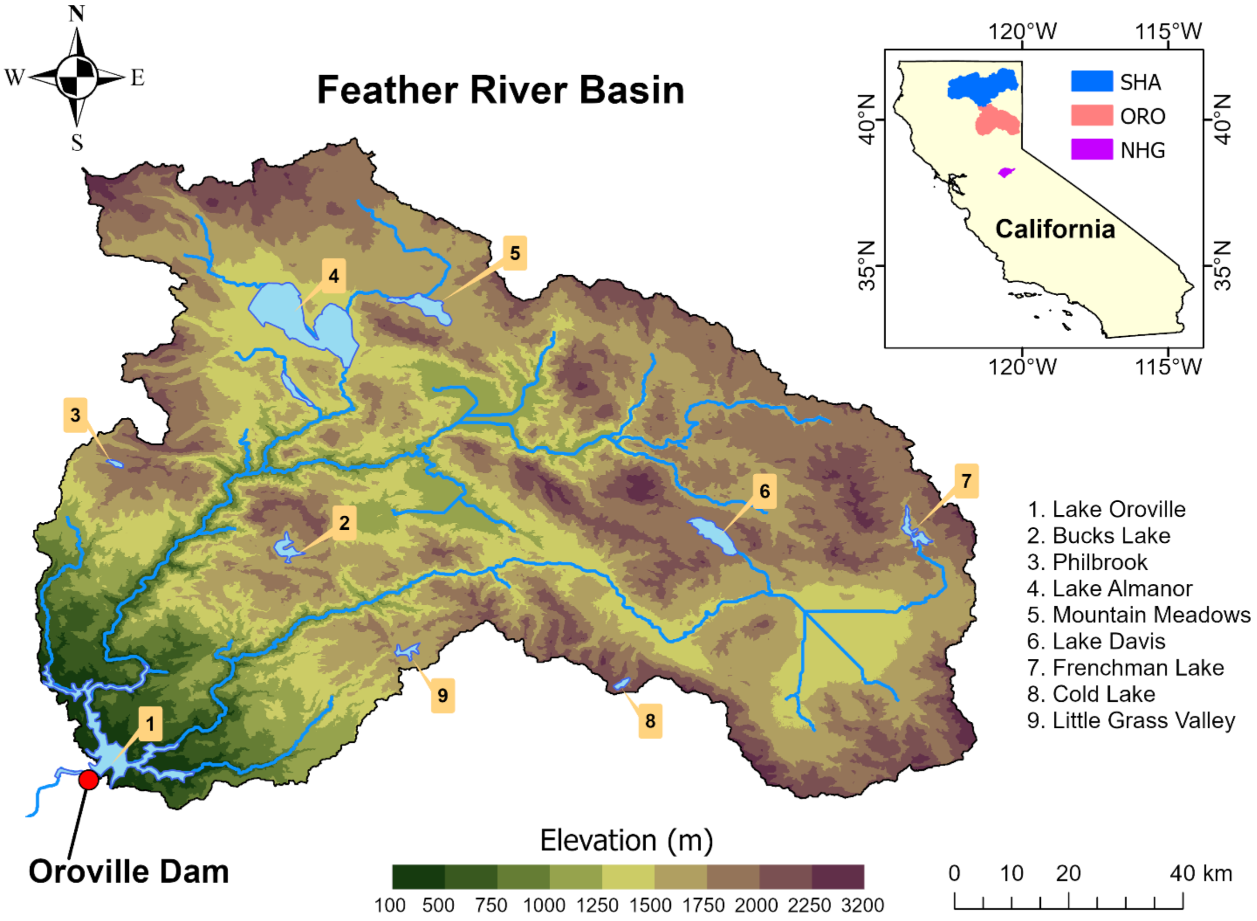


Figure 1. *Geographical area of study depicting Feather River inflow to Oroville Dam (1) as well as significant upstream reservoirs and other diversions (2-9). The inset in the upper right shows*

the locations of the primary study basin (Feather river; ORO) and two additional basins considered in part of the analysis (Sacramento River; SHA, Calaveras River; NHG).

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). To demonstrate the utility of our approach, it is necessary to establish experimental conditions where we know the truth and can therefore measure how well our methods capture it. We first select two hydrologic models, designating one as the ‘truth model’ and the other as 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. In this study, we use SAC-SMA (Burnash, 1995) as the truth model. The process model represents an (imperfect) model of the true hydrologic system that can only approximate new hydrologic responses under alternative meteorological forcing. HYMOD (Boyle, 2001) is used as the process model in this work. In the stylized experiments below, we first ensure that SAC-SMA provides a reasonable representation of the observed hydrology of the Feather River basin, but then assume the fitted SAC-SMA model *is* the real-world watershed of interest. The goal of the HYMOD model is then to predict the behavior simulated by the SAC-SMA model, and not to predict the actual observations. We use this ‘model-as-truth’ approach – akin to ‘perfect model’ experiments used in the climate modelling community (Abramowitz & Bishop, 2015; Knutti et al., 2017; Herger et al., 2018) - to overcome the challenge of having no future observations with which to compare our process model. That is, because there are no streamflow observations in the distant future under substantial amounts of climate change, there is no way to determine whether hydrologic model errors are non-stationary under highly non-stationary meteorological forcing. By using one model (SAC-SMA) as the true hydrologic system and another (HYMOD) as a model of that system, we can explore this issue in a controlled (albeit highly stylized) experiment. We describe the hydrologic models used for the truth and process models (SAC-SMA and HYMOD, respectively) in more detail in Section 3.2 below.

In the experiment, we evenly split the available record into a training period for calibration and validation (WY 1989-2004) and a test period for out-of-sample model evaluation (WY 2005-2018), following common practice. During the training period, we calibrate the truth model to observed streamflow data and then calibrate the process model to the truth model in that same period (gray dashed box in Figure 2). We then examine the distribution of hydrologic prediction errors between the truth and process models in the test period, calculated based on historical precipitation and temperature data from this period (Test), as well as historical precipitation and temperature warmed by 4^{o}C (Test+4C). The Test+4C case is produced by simply adding 4^{o}C to all temperatures in the testing period, and is reflective of the warming in California projected by a multi-climate model ensemble average by the end of the 21st century under the RCP 8.5 emission scenario (Pierce et al., 2018). We then document changes to the error distribution between the truth and process models across these two scenarios, and can attribute these changes to differences in how the truth and process models respond to warming with all other aspects of meteorology held constant.

Errors between the truth and process model in the training period, along with state variables from the process model, are then used to fit a hybrid SWM (described in Section 3.3). The hybrid SWM combines samples of error between the truth (SAC-SMA) and process (HYMOD) models with deterministic simulations from the process model to develop an ensemble of streamflow traces that is statistically consistent with flows from the truth model. We evaluate whether our proposed hybrid SWM can capture potential changes in the error distribution between the Test and Test+4C scenarios, and also compare the results of our hybrid SWM against a simpler benchmark SWM that does not depend on process model state variables (described in Section 3.3.3). We use interpretability methods to understand how the hybrid SWM uses model state variables to estimate changes to the error distribution (see Section 3.4).

Finally, because there are limitations using a model to represent the true hydrologic system in the stylized experiment above, we repeat the experiment in a real-world (non-stylized) setting. Here, we train and test the performance of a hybrid SWM against actual streamflow observations for the Feather River (ORO), Sacramento River (SHA), and Calaveras River (NHG) over the historical record, using SAC-SMA as the process model (Section 3.5). That is, the SWM combines samples of error between observations and SAC-SMA predictions with deterministic simulations from SAC-SMA to develop an ensemble of streamflow traces that is statistically consistent with observed flows.

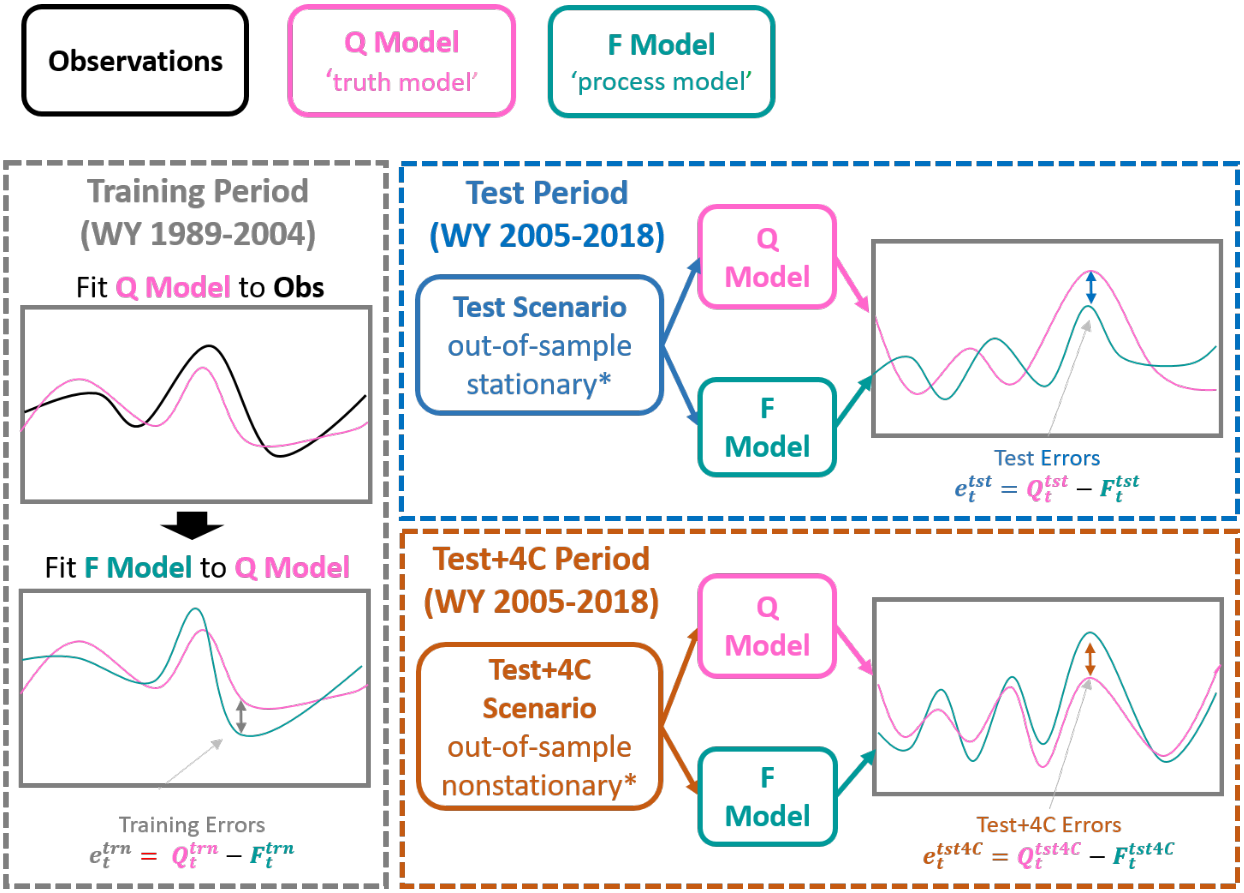


Figure 2. *Conceptual diagram showing the stylized experimental design, where a 'truth' (Q) and 'process' (F) model are designated to test the effect of alternate hydrologic model forcing on predictive errors between the two models. In the Test scenario, both models are forced with out-of-sample but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

3.2. Hydrologic Model Setup

We calibrate two daily 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. We selected these models for two reasons. First, both have previously been developed for the Feather River basin, performed very well against observations, and outperformed other process-based models like the Variable Infiltration Capacity model (see Wi and Steinschneider, 2022). Second, SAC-SMA and HYMOD have similar structures, and so any non-stationarity in the predictive errors between these models under climate change would indicate there is a high likelihood that non-stationarity would also emerge between models (or between models and the actual observations) with more significant structural differences.

Both SAC-SMA and HYMOD are built using 828 hydrologic response units (HRUs) defined for the Feather River basin by segregating each 6 km climate grid cell into different soil classes from the 1 km resolution State Soil Geographic dataset (Miller & White, 1998). The 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 the objective function. We first calibrate SAC-SMA (truth model) to the full natural flows for the period of WY1989-2004. The flows simulated by SAC-SMA were then used to calibrate HYMOD (process model) for the same period. For the training (test) periods of WY1989-2004 (WY2005-2018), SAC-SMA simulations achieved training (test) NSE of 0.92 (0.88) when fit to the observations, whereas HYMOD NSE was 0.95 (0.92) when fit to the SAC-SMA flows. We also note that in previous work (Wi & Steinschneider, 2022), HYMOD achieved training (test) NSE of 0.80 (0.78) when fit to the observations, although that calibration of HYMOD is not used in this stylized experiment.

All internal state variables simulated by HYMOD (the process model) are used to inform our error model (see Table 1, also Supporting Information S1). 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).

Table 1. HYMOD state variables

Short Name	**Long Name**	**Description**
sim	Simulation	HYMOD predicted streamflow in mm
runoff	Runoff	Upper reservoir flow of HYMOD in mm
baseflow	Baseflow	Lower reservoir flow of HYMOD in mm
precip	Precipitation	Basin averaged precipitation in mm
tavg	Average temperature	Basin averaged temperature in °C
et	Evapotranspiration	Modeled evapotranspiration (Hamon approach) in mm
upr_sm	Upper soil moisture	Basin averaged soil moisture content (mm) in upper reservoir
lwr_sm	Lower soil moisture	Basin averaged soil moisture (mm) in lower reservoir
swe	Snow water equivalent	Basin averaged snow water equivalent simulated by degree day snow module (mm)

3.3. Hybrid SWM

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

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

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. In our stylized experiment, Q_t represents flow from SAC-SMA and $F(X_t,\pi)$ represents flow from HYMOD, but in real-world applications, Q_t would be observed streamflow and $F(X_t,\pi)$ would be estimates from any hydrologic model of interest. To estimate the SWM, we follow the approach of Montanari & Brath (2004) and first train the model F to the 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.

The primary methodological contribution of this work is an adaptive, state-variable-dependent hybrid model for e_t, illustrated in Figure 3 (a simpler benchmark model is described in Section 3.3.3). There are two main components of the hybrid model. The first is an initial model updating step referred to as 'error correction' (Shen et al. 2022*)*, which is analogous to hydrologic post-processing, except that the step is targeted to'correcting' the predictive error distributions, not the hydrologic simulations directly. The error correction model *f* creates a mapping between the process model state variables $(\theta_{t,SV})$ and the raw errors (e_t), including autocorrelation in the errors through lagged error terms $(e_{t-1:t-p})$ out to lag *p:*

$$e_t = f\left(\theta_{t,SV}, e_{t-1:t-p}\right) + \varepsilon_t \qquad \text{(Eq. 2)}$$

This model corrects for conditional bias, i.e., biases in the process model predictions that are dependent on the internal states of the model and recent prediction errors. The second component is a dynamic residual model to capture the remaining stochasticity in the error correction model residual, ε_t, also as a function of process model state variables. Each of these components is described in more detail in Sections 3.3.1 and 3.3.2 below.

Importantly, we use a split-sample calibration/validation approach to fit these two components in a similar fashion to Hah et al. (2022). That is, we first fit the error correction model f to one subset of the training data (termed the calibration set), and then we fit the dynamic residual model to a separate subset of the training data (termed the validation set), after the error correction model has been applied to that validation set (see Figure 3). This strategy helps ensure that the dynamic residual model will represent the true variability of out-of-sample residuals from the error correction model. In this work we employ an approximate 70%/30% split of the training data between calibration (WY 1989-1998) and validation (WY 1999-2004) 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 added to 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 (see Section 3.3.3).

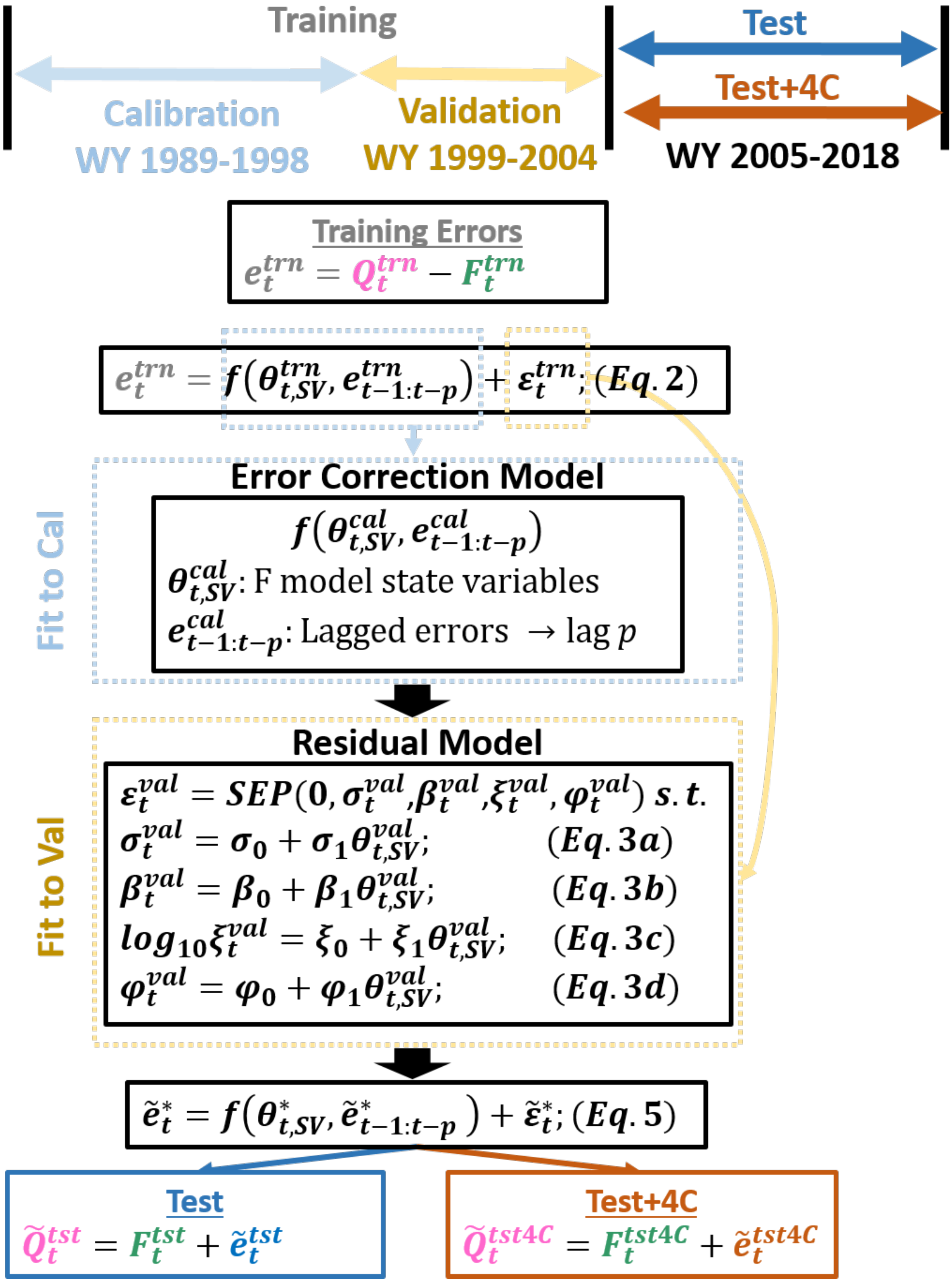

Training
Calibration
WY 1989-1998
Validation
WY 1999-2004
Test
Test+4C
WY 2005-2018
Training Errors
$e_t^{trn} = Q_t^{trn} - F_t^{trn}$
$e_t^{trn} = f(\theta_{t,SV}^{trn}, e_{t-1:t-p}^{trn}) + \varepsilon_t^{trn}; (Eq.2)$
Fit to Cal
Error Correction Model
$f(\theta_{t,SV}^{cal}, e_{t-1:t-p}^{cal})$
$\theta_{t,SV}^{cal}$: F model state variables
$e_{t-1:t-p}^{cal}$: Lagged errors → lag p
Fit to Val
Residual Model
$\varepsilon_t^{val} = SEP(0, \sigma_t^{val}, \beta_t^{val}, \xi_t^{val}, \varphi_t^{val})\ s.t.$
$\sigma_t^{val} = \sigma_0 + \sigma_1\theta_{t,SV}^{val}; \quad (Eq.3a)$
$\beta_t^{val} = \beta_0 + \beta_1\theta_{t,SV}^{val}; \quad (Eq.3b)$
$log_{10}\xi_t^{val} = \xi_0 + \xi_1\theta_{t,SV}^{val}; \quad (Eq.3c)$
$\varphi_t^{val} = \varphi_0 + \varphi_1\theta_{t,SV}^{val}; \quad (Eq.3d)$
$\tilde{e}_t^* = f(\theta_{t,SV}^*, \tilde{e}_{t-1:t-p}^*) + \tilde{\varepsilon}_t^*; (Eq.5)$
Test
$\tilde{Q}_t^{tst} = F_t^{tst} + \tilde{e}_t^{tst}$
Test+4C
$\tilde{Q}_t^{tst4C} = F_t^{tst4C} + \tilde{e}_t^{tst4C}$

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 fit 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. Consistent with Figure 2, the colors of the boxes are used to denote the time period used (Calibration, Validation, Test, Test+4C), and separate colors are also used for the 'truth' (Q) and 'process' (F) models.

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* (i.e., *f* captures conditional bias in the process model). 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 follow Shen et al. (2022) and select *f* to be a random forest (RF) model, leveraging its parsimony, demonstrated hydrologic performance, and out-of-sample robustness (Tyralis et al., 2019). More advanced models (e.g., long short-term memory networks; Kratzert et al., 2018; Frame et al., 2021) would likely work better if applied to many sites simultaneously, but our focus on a single location with limited data supports a simpler ML model less prone to overfitting (discussed further in Section 5). The primary hyper-parameters of a RF model are the number of trees in the forest ('ntree') and the number of features to randomly select at each split ('mtry'). Optimization of RF occurs on individual trees in the forest, not the overall RF output that is determined by a majority voting scheme, making hyper-parameter selection challenging.

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 (see Supporting Information S4 for more detail). 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:

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

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

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

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

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

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. We employ noise regularization during maximization of the likelihood function to smooth the parameter estimates (Rothfuss et al, 2019; Klotz et al., 2022). We also 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 the 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 $\theta_{SV,t}$.

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 a single SWM trace, we first generate new time series 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. To generate an *M*-member ensemble of SWM traces, we simply repeat the steps above *M* times, where each trace has a different random time series of deviates $\tilde{a}_t$ sampled from the SEP distribution.

Importantly, the simulation procedure above 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 (WY1989-1998) 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 linear 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. This experiment is conducted using FNF from the Feather (ORO), Sacramento (SHA), and Calaveras (NHG) River basins. SAC-SMA models are fit to the FNF at the SHA and NHG sites in the same way as described in Section 3.2 for the

ORO site, with training (test) NSE for SHA and NHG equal to 0.91 (0.90) and 0.89 (0.83), respectively. We note that in any real-world application, we would not anticipate significant amounts of non-stationarity in model residuals between training and testing periods if warming trends or other types of climate change between the two periods are modest.

4. Results

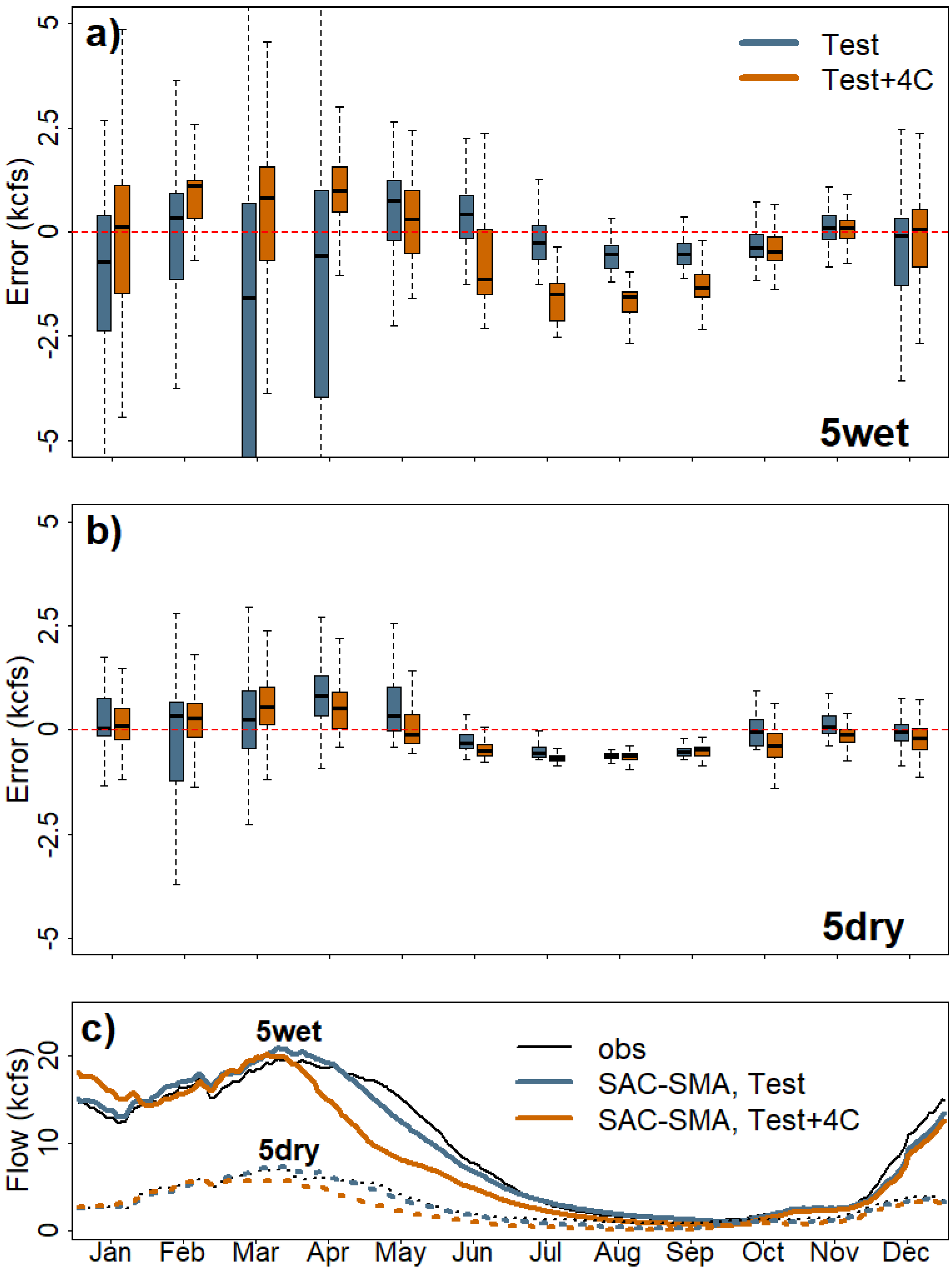
4.1. Non-stationarity in raw errors

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 in the out-of-sample test period with and without 4^{o}C warming applied to the meteorological data (Figure 4a-b). Predictive errors are defined as truth minus process model output. Note that because both models use the same meteorological data, any errors between the two models (and non-stationarity in those errors) can be attributed to differences in model structure. In Figure 4a, we show errors for the 5 wettest years in the test period (‘5wet’; 2005, 2006, 2011, 2016, 2017), while in Figure 4b we show errors for the 5 driest years (‘5dry’; 2007, 2008, 2012, 2014, 2015). We define ‘5wet’ and ‘5dry’ based on total annual streamflow.

For the wettest years (Figure 4a), there is a substantial divergence in error distributions between the historical and warmed scenarios across seasons, with the most notable differences occurring in the late winter and early spring (February-April) and summer (June-September). In March and April, when the mean daily flows are at their annual peak (Figure 4c, ‘5wet’), the errors in the two cases are biased in opposite directions, with process model outflows systematically overpredicting truth model flows in the Test case but underpredicting them in the Test+4C case.

Additionally, a reduction in the bias and dispersion of the errors in Test+4C, particularly in March, suggest less structural uncertainty between the truth and process models as snowmelt declines under warming. In April to June, when the two cases exhibit the greatest disparities in mean daily flow (Figure 4c, ‘5wet’), error biases change sign for both Test and Test+4C cases. Then in the summer, both the Test and Test+4C errors suggest process model overpredictions, but these are larger in the Test+4C case. In addition, there are several months when the error dispersion (i.e., interquartile range) differs substantially between the two cases, with February-April being the most prominent examples.

For the driest years (Figure 4b), there is little difference in the error distributions between the Test and Test+4C cases. In these years, the average streamflow with and without warming is very similar, with only small declines in the spring and summer in the Test+4C case. Overall, Figure 4 shows that the error distribution between the truth and process models can change under warming during wet years alongside shifts in key hydrologic processes, such as more precipitation falling as rain and running off in the cold season, less snowpack carrying over into the spring, an earlier start to the snowmelt season, and increased evapotranspiration in the summer (Figure 4c). This phenomenon is not observed during very dry years when there is little water to begin with and therefore little absolute change in the underlying hydrology with warming. We also compare the ‘5wet’ and ‘5dry’ subsets of the training period errors to Test and Test+4C errors, finding that the training period largely reflects the Test errors (see supporting information S5).


a)
Test
Test+4C
5wet
Error (kcfs)
5
2.5
0
-2.5
-5
b)
5dry
Error (kcfs)
5
2.5
0
-2.5
-5
c)
5wet
5dry
obs
SAC-SMA, Test
SAC-SMA, Test+4C
Flow (kcfs)
20
10
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

593

Figure 4. *a) Monthly comparison of out-of-sample (WY2004-2018) raw error distributions (e_t) between truth (SAC-SMA) and process (HYMOD) models without (Test) and with (Test + 4C) warming for '5wet' years (2005, 2006, 2011, 2016, 2017). b) As in (a) for '5dry' years (2007, 2008, 2012, 2014, 2015). c) Mean daily flow of the historical observations (WY2004-2018) and SAC-SMA truth model across WY2004-2018 for the two scenarios, smoothed with a 30-day moving average. Solid (dotted) lines are for the '5wet' ('5dry') years.*

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 (Figure 5a) and Test + 4C (Figure 5b) cases after fitting this error correction procedure to the training set and applying it to the test set, focusing on the 5 wettest years where we see the largest degree of nonstationarity. Figure 5 also shows the raw error distributions (e_t, Eq. 1) for comparison as well as scatterplots comparing raw errors to the RF predicted errors (Figure 5a-b insets) . There is a clear reduction in conditional bias across months in the Test case, where all residuals are nearly centered around zero. This reduction in conditional bias holds in the Test+4C case for October-May, but deteriorates somewhat in the summer months (June-September). Notably, the raw errors were successfully debiased in both Test and Test+4C cases in March and April, when biases in the two cases were of different sign. The scatterplots show that the RF error correction model struggles in the lower tails (large negative errors) in Test and in the upper tails (large positive errors) in Test+4C. 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.

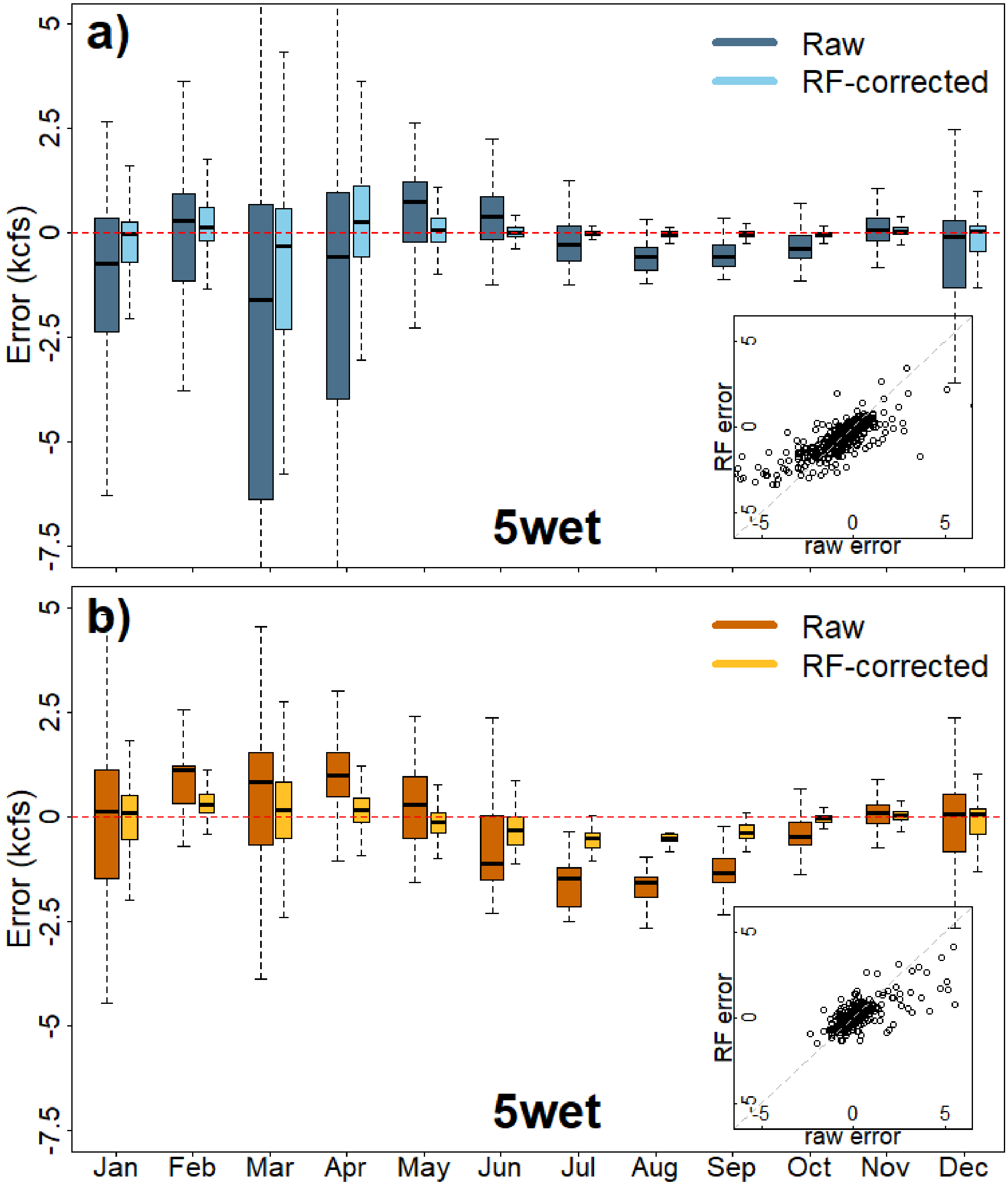


Figure 5. *a) Monthly comparison of raw error (e_t) distributions versus residual distributions (ε_t) after correction by the Random Forest (RF) model in the '5wet' years from the Test period (WY2005-2018). The inset is a scatterplot comparison of raw error vs RF predicted error. b) As in a) but for the Test+4C case.*

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 autocorrelation in the raw errors is the most influential predictor variable. 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 S6). This supports the integration of autoregressive and state variables in the RF error correction model.

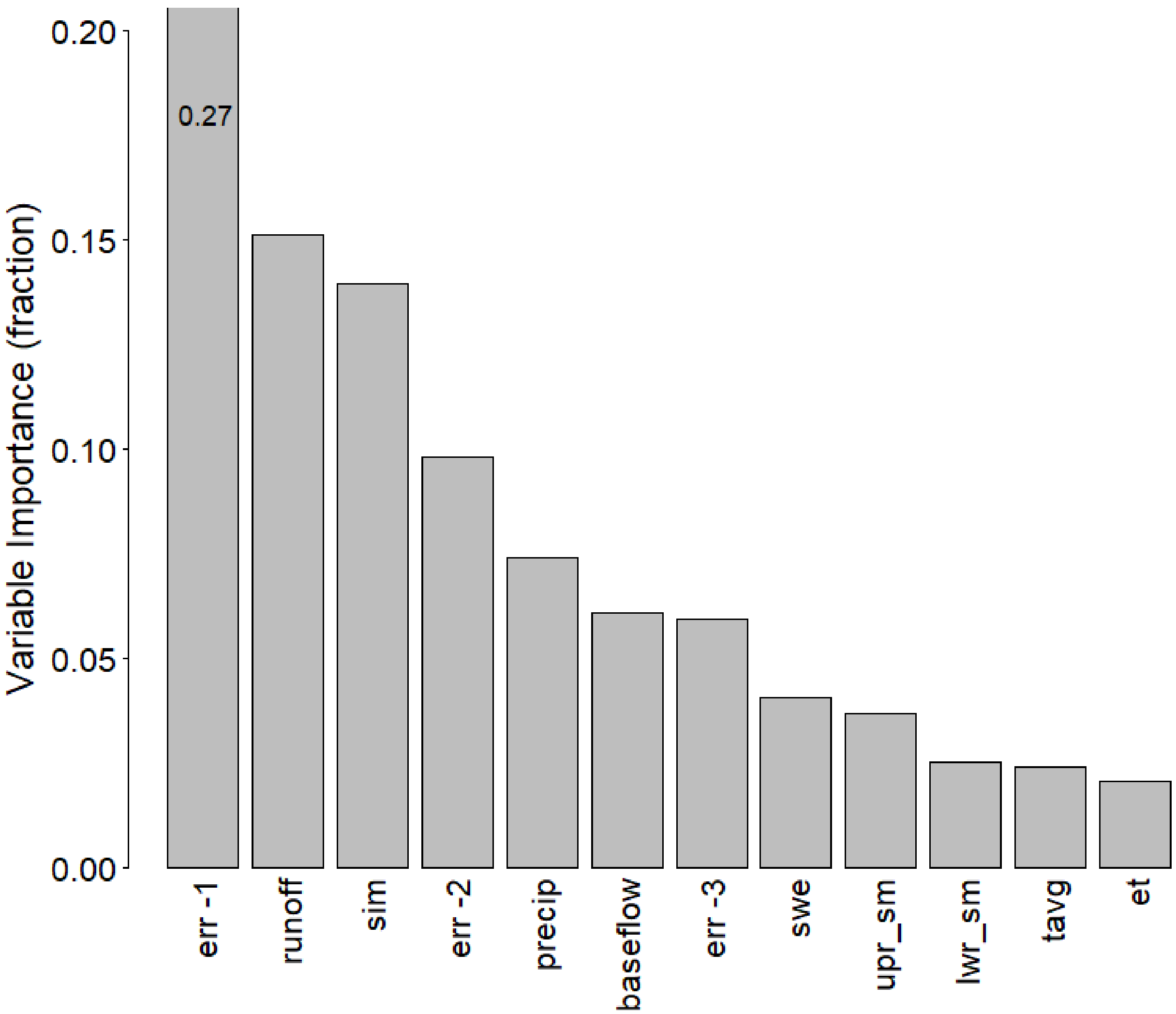


Figure 6. *Variable importance from RF error correction model fit to calibration period (WY1989-1998). Note: 'err -1' variable importance equals 0.27, which extends outside of plot bounds. Definitions for all state variable acronyms can be found in Table 1 while 'err -1' to 'err -3' reflect lag 1 to 3 errors.*

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. In a practical sense, feature weights are the normalized coefficient values of the local linear regression in the LIME procedure (section 3.4). Thus, positive feature weights imply a positive correlation between the feature (i.e. the state variable) and the local model response (i.e. the estimated error) and vice versa.

Figure 7 shows that precipitation and the process model simulated flow (sim) and baseflow exhibit the largest absolute feature weights in both the Test and Test+4C cases. These feature weights are amplified in importance from the Test to the Test+4C case, while the feature weight for snow water equivalent (swe) reduces to near zero for Test+4C. This suggests that precipitation and simulated flow partitioning dynamics better explain model error in the Test+4C case against a backdrop of decreasing SWE influence. This likely reflects changes in snowmelt dynamics under warming that lead to HYMOD overestimation (underestimation) of SAC-SMA flows in March under the Test (Test+4C) cases. While the LIME procedure cannot provide causal evidence, it’s possible that the more active role of snow accumulation and melt in determining streamflow response in the Test case, but not in the Test+4C case (where snowpack is much less prevalent in March), leads to a reversal of model predictive biases in this month.

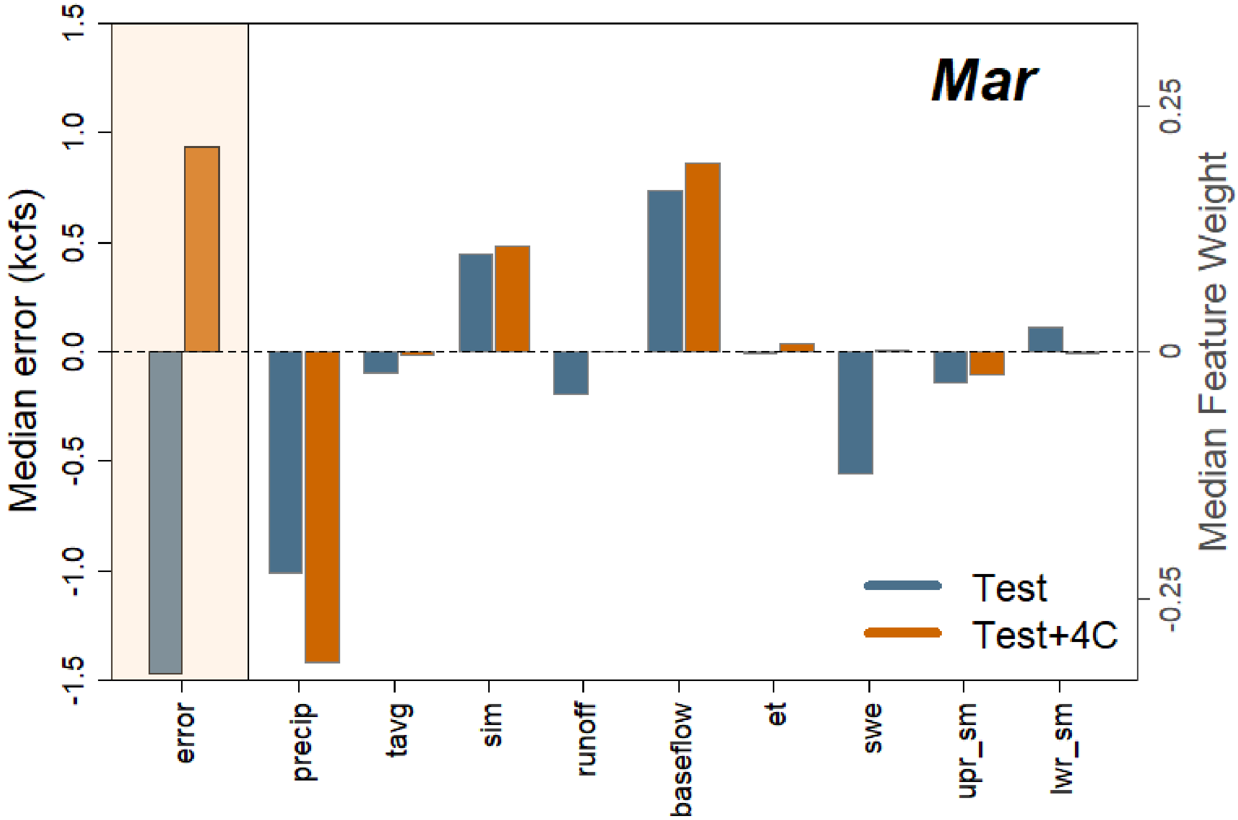


Figure 7. *Locally Interpretable Model-agnostic Explanation (LIME) median feature weight (white background, gray outline) comparison against median error (light orange background, black outline) for a selected month, where errors and associated feature weights are aggregated for errors less than (greater than) zero in the Test (Test+4C) cases. Definitions for all variable acronyms can be found in Table 1.*

4.3. Dynamic residual model performance

Overall, the error correction process yields residuals (ε_t) in both Test and Test+4C cases that are relatively 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 (February-April) 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 Supporting Information 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 and variance. In the Test+4C case (bottom row), the empirical residual distributions become more peaked in March and April compared to the Test case, and the dynamic residual model is able to infer these changes. The 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.

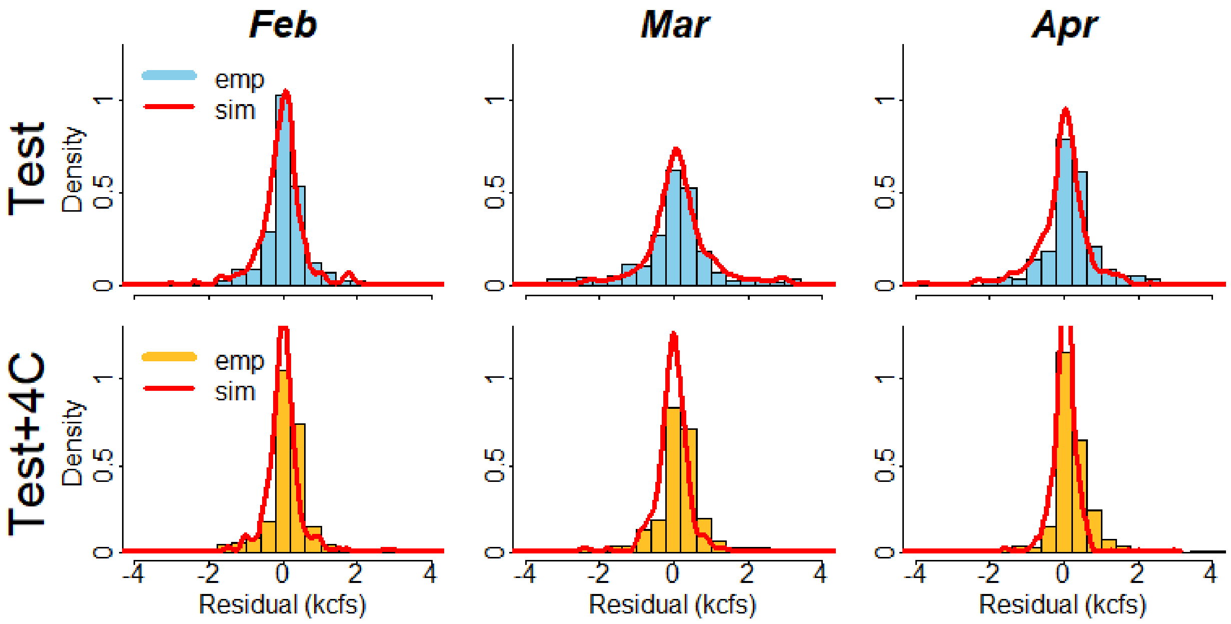


Figure 8. *Top row: Empirical distribution of RF-corrected residuals ε_t (histogram) versus the kernel density estimate of a simulated sample of residuals $\tilde{\varepsilon}_t$ from the dynamic residual model (red line) for selected months from the Test case. Bottom row: As in top row, but for the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

Table 2 shows the state variable effects for the different parameters in the SEP model (see Eqs. 3a-3d), while Figure 9 shows the seasonality in SEP parameters in Test and Test+4C cases. For the parameter σ_t (heteroscedasticity), the runoff state variable is the most influential with a small influence from SWE (Table 2). This result reflects the strong relationship between error variance and flow magnitude, as noted in previous literature (Schoups & Vrugt, 2010). This is also seen in the strong seasonal signal in both the mean and variability in σ_t (Figure 9, top left), though we note that this seasonality is truncated in Jul-Oct. This truncation results from a limit to the minimum value of σ_t that is required for model stability. Further, the greatest divergence in σ_t between the Test and Test+4C cases occurs in the late winter and spring months where mean flow magnitudes diverge most substantially (see Figure 4c) and there are important contributions from snow accumulation and melt.

Skewness (ξ_t) shows a relatively weak seasonal signal that is centered around 0 ($log_{10}\xi_t = 0$; no skew) in the warm season and slightly negative ($log_{10}\xi_t < 0$; negative skew towards process model overpredictions) in the winter and spring. Negative skew is more prominent in the Test case. As for σ_t, skewness is most strongly influenced by runoff and SWE, but also has an important contribution from precipitation (precip). In contrast, the kurtosis parameter β_t exhibits a strong seasonal signal that is primarily tied to upper and lower soil moisture (upr_sm, lwr_sm), with smaller contributions from SWE, average temperature (tavg), runoff, and evapotranspiration (et). The residual distributions exhibit values of β_t close to 1 (i.e., a Laplace distribution) in the cold season that become progressively more peaked and fat-tailed ($\beta_t > 1$) in the summer months. This reflects a concentration of probability mass around small residuals ε_t in low flow months with high probability of large (scaled) residuals (see Figure S4). Both the Test and Test+4C cases show similar seasonal characteristics, though the Test+4C β_t are uniformly larger than the Test β_t across most months except September-December.

Finally, lag-1 autocorrelation (φ_t) shows no seasonal signal but a notable increase in variability in the winter season that is most influenced by baseflow and upper soil moisture (upr_sm). 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). 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 2. *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.*

$\boldsymbol{State\,Var.}$	$\boldsymbol{\sigma_t}$	$\boldsymbol{\beta_t}$	$\boldsymbol{log_{10}(\xi_t)}$	$\boldsymbol{\varphi_t}$
intcpt	0.001	0.986	-0.047	0.528
sim	0.000	-0.087	-0.023	-0.013
runoff	0.187	-0.140	-0.113	-0.001
baseflow	0.000	0.038	0.014	-0.136
upr_sm	0.000	-0.291	-0.026	0.138
lwr_sm	0.000	0.509	0.023	0.002
swe	0.034	-0.201	0.051	0.001
et	0.000	-0.215	-0.004	-0.044
tavg	0.000	0.147	-0.017	0.022
precip	0.000	-0.003	0.040	0.007

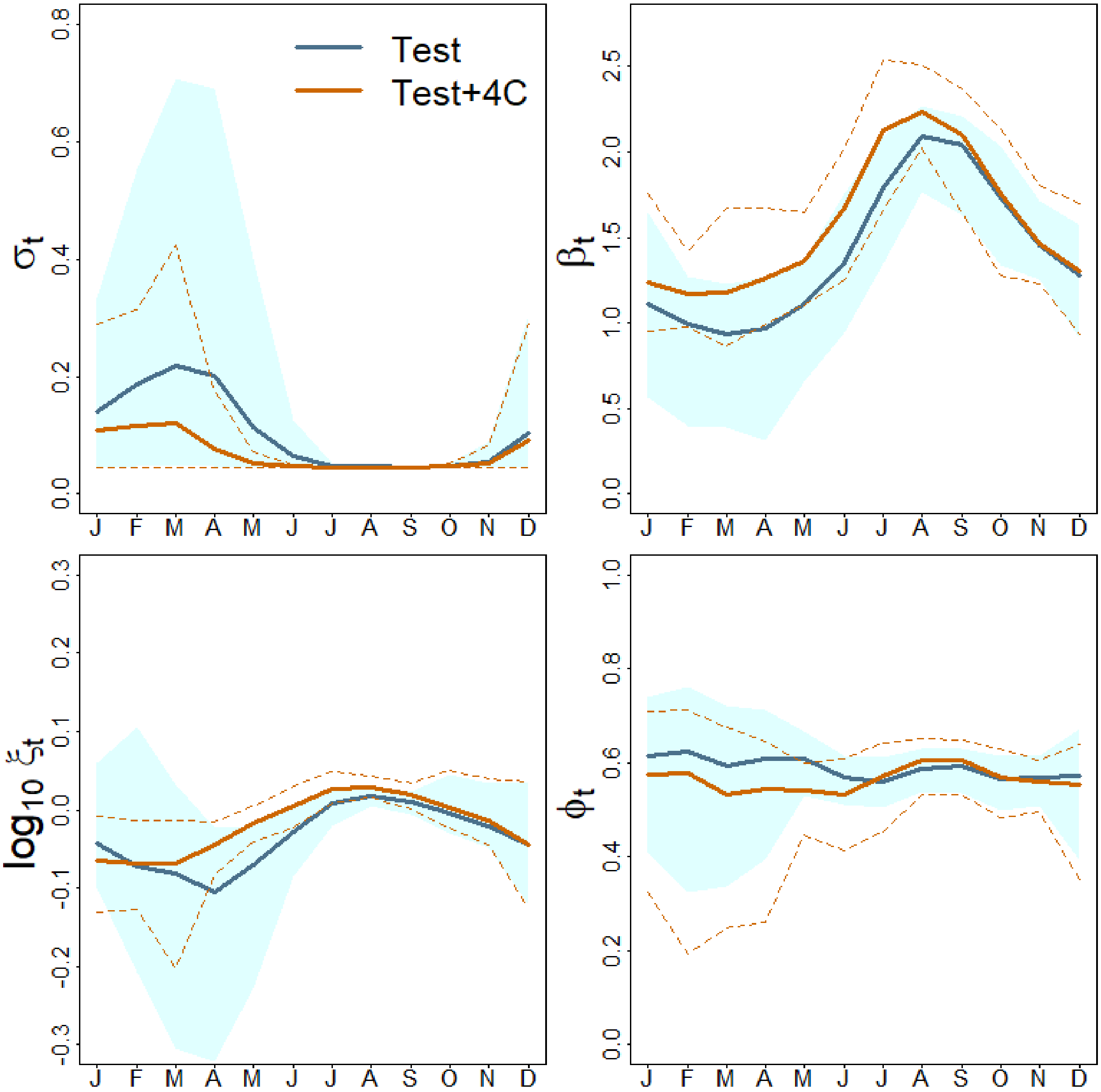


Figure 9. *Values for the four free parameters in the dynamic residual model aggregated by month for the Test and Test+4C cases. The bold lines are the mean parameter values while the blue shading is the 90% confidence interval for the Test case and the orange dashed line is the 90% confidence interval 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, again focusing on the ‘5wet’ years when error non-stationarity is more evident. 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 both March and April, the Test empirical errors are negatively biased and have substantially more variance as compared to the Test+4C errors, which are positively biased and have less variance. The model is able to simulate both of these shifts. In the summer, the hybrid model is also able to capture the negative bias in the Test case. However, this bias grows in the Test+4C case, and the hybrid model does not capture this deeper negative summertime bias, consistent with the results in Figure 5. This deficiency will lead to simulated summertime flows that are biased compared to the truth model.

In comparison to the hybrid SWM, a static SWM based purely on seasonality in the error distribution (Figure 10c-d) struggles to capture the out-of-sample changes to the error distribution under both the Test case and the Test+4C case. During the Test case (Figure 10c), the static SWM substantial overestimates error variance and is insensitive to shifts in error bias during the winter and spring. This challenge is compounded in the Test+4C case, where neither shifts in error biases or dispersion are captured faithfully. The state variable dependencies built into the hybrid model allow a more faithful emulation of these shifts, even if imperfect. We also note that both models produce relatively accurate coverage probabilities in the Test case but struggle to produce accurate coverage probabilities in the Test+4C case (see Supporting Information S8).

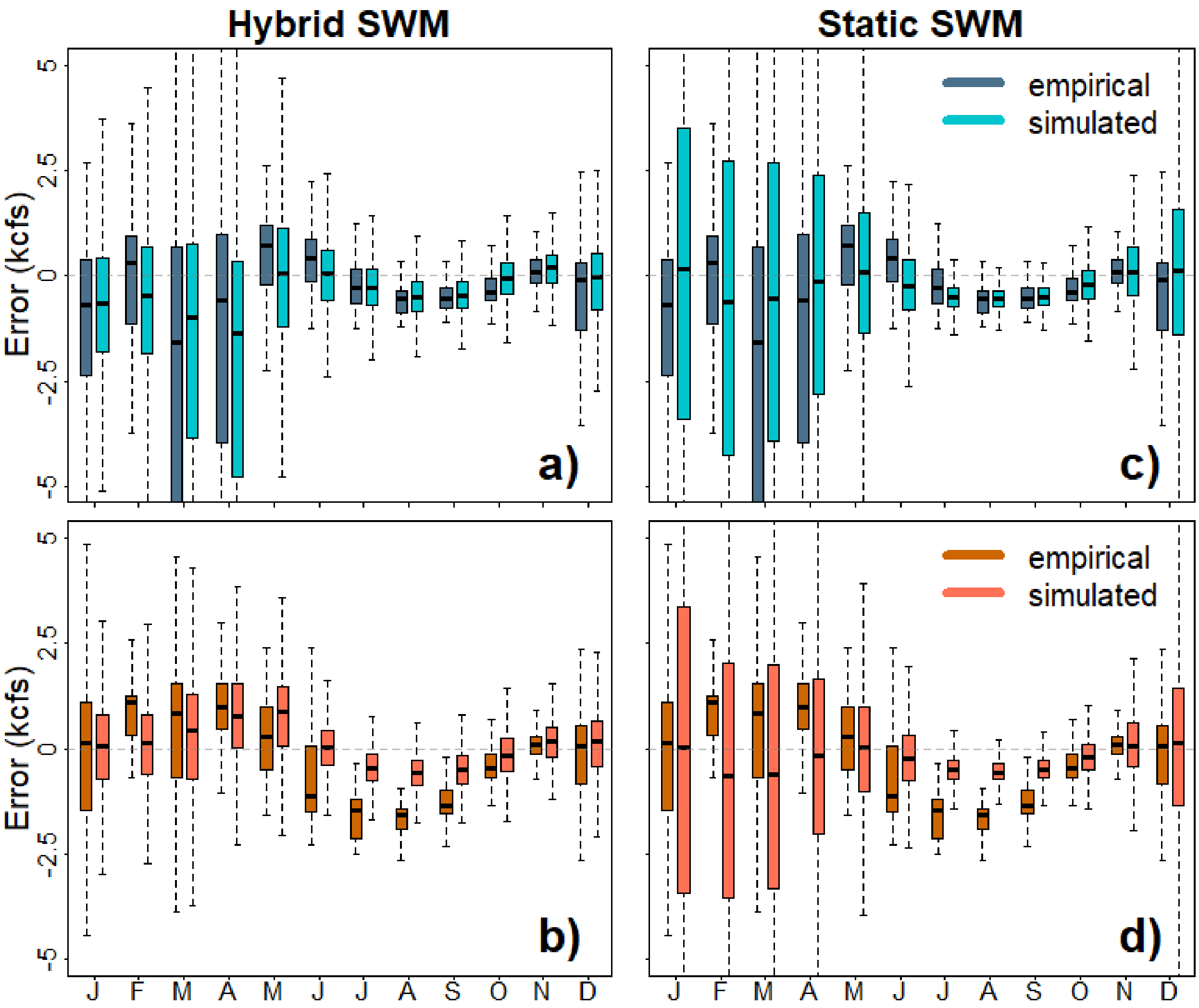


Figure 10. *a) Monthly empirical distribution of errors in the Test subset (dark blue) versus 1000 aggregated samples of hybrid model simulated errors (light blue). b) Same as (a) but for Test+4C errors, where empirical (simulated) errors are dark orange (coral). c-d) As in (a) and (b), but simulated 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 (11a) and Test+4C (11b) 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°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 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 in April (11a inset), whereas in the Test+4C case, the model helps correct for process model underprediction bias during this same time of year (11b inset). The spread in the SWM ensemble is smaller under the Test+4C case between May-July because flows are simulated to be lower (and more baseflow driven) in this subperiod, and variance is correlated most strongly with runoff (see Table 2). During March and April, the spread of the SWM ensemble is also able to capture the peak flows well for both Test and Test+4C, which is also apparent when focusing on statistics of streamflow extremes (see Figure 12). 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.

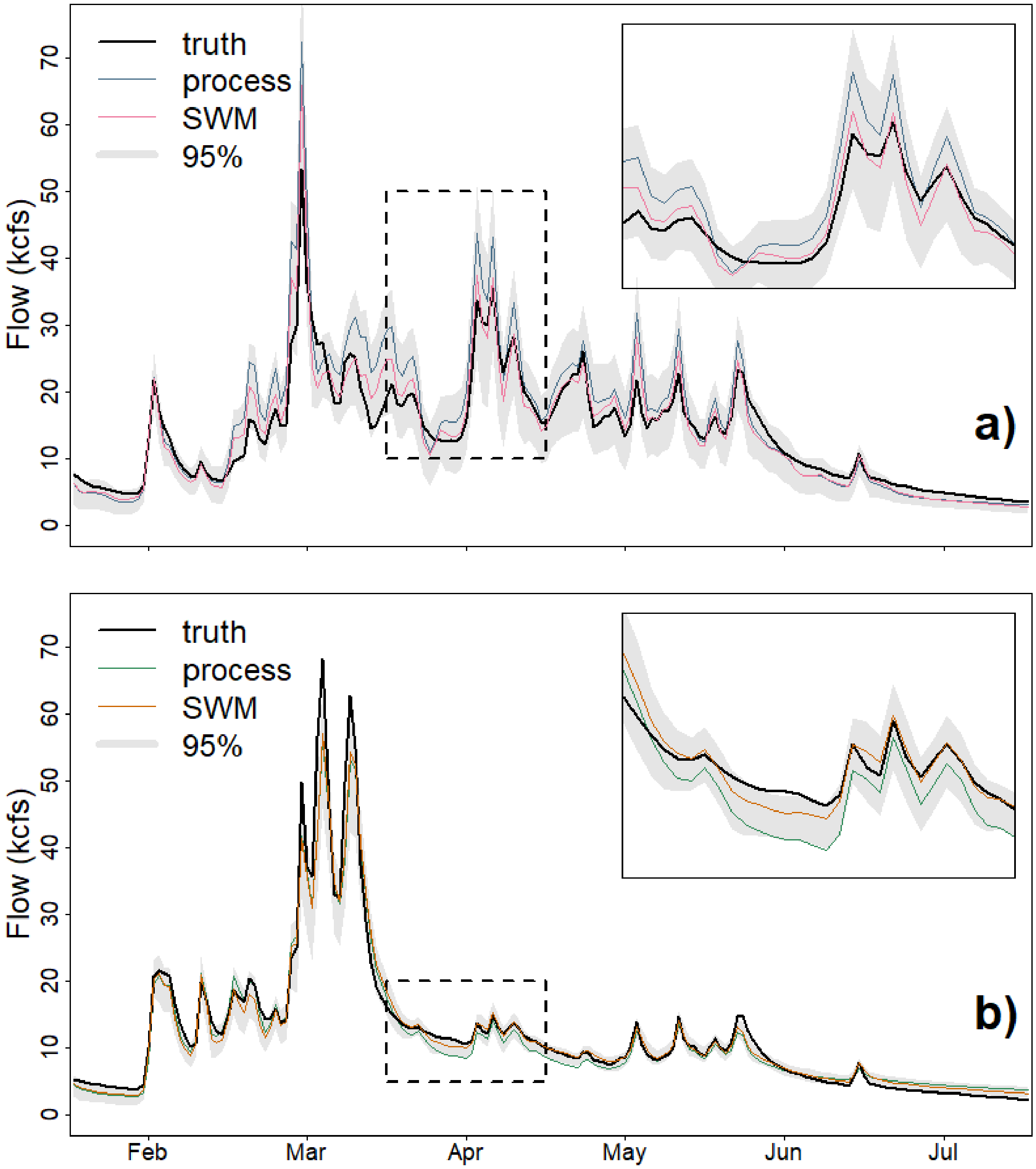


Figure 11. *a) Truth model flow (black) compared against process model flow (dark blue) and the median flow of 1000 samples (pink) from the hybrid SWM for the February-July period in 2011 from the Test scenario. 95% coverage interval for the 1000 samples are shown in light gray. Inset contains zoomed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

Finally, we verify the hybrid SWM simulations by comparing the high and low flow performance against the static SWM in Figure 12. We choose 99th percentile flows as a high flow comparison (Figure 12a-b) metric due to the short length of the Test and Test+4C period (WY2005-2018), which would lead to high uncertainty for extreme value distributional estimates of design events. In the Test scenario (Figure 12a), the truth model 99th percentile flow value is substantially below that of the process model, indicating that the process model overestimates the truth model for high flows. The hybrid SWM ensemble is mostly able to correct this overestimation with relatively low uncertainty, whereas the static SWM ensemble estimation overestimates the 99th percentile truth model flow with high uncertainty and is even slightly biased above the process model. In the Test+4C case, the behaviors of the static SWM are similar to the Test case, showing an estimation of the high flows that is biased above the process model with high uncertainty. Here, it captures the truth model value well, however. The hybrid SWM simulates the shift in bias from the Test to the Test+4C case (i.e. from a negative to a positive bias), albeit with insufficient magnitude to capture the truth model flow value well, reflecting the findings for Figure 10b.

For low flow verification, we evaluate the SWMs on their ability to emulate the lowest cumulative annual flow year in the Test and Test+4C period. We choose this aggregated metric because of the extremely low summer flow values in this basin, which result in values at or near zero for the commonly used 7Q10 low flow metric. We find that in both the Test and Test+4C case, the static SWM substantially underestimates the truth model value, again with high uncertainty. The hybrid SWM estimate in the Test case is accurate, but when the truth model model value is biased substantially high in the Test+4C case, the hybrid SWM is unable to simulate this bias. This reflects the inability of the hybrid SWM to capture the summertime shifts in Test+4C, which was also noted in the Figure 10b discussion.

Overall, these results bolsters findings in previous sections that the non-stationarity in the predictive uncertainty distributions has substantial ramifications for estimating the truth model high and low flow quantiles. The hybrid SWM is able to capture these shifts more reliably than the static SWM, but is challenged more by the low flow emulation, which has been noted in previous SWM studies (Shabestanipour et al., 2023).

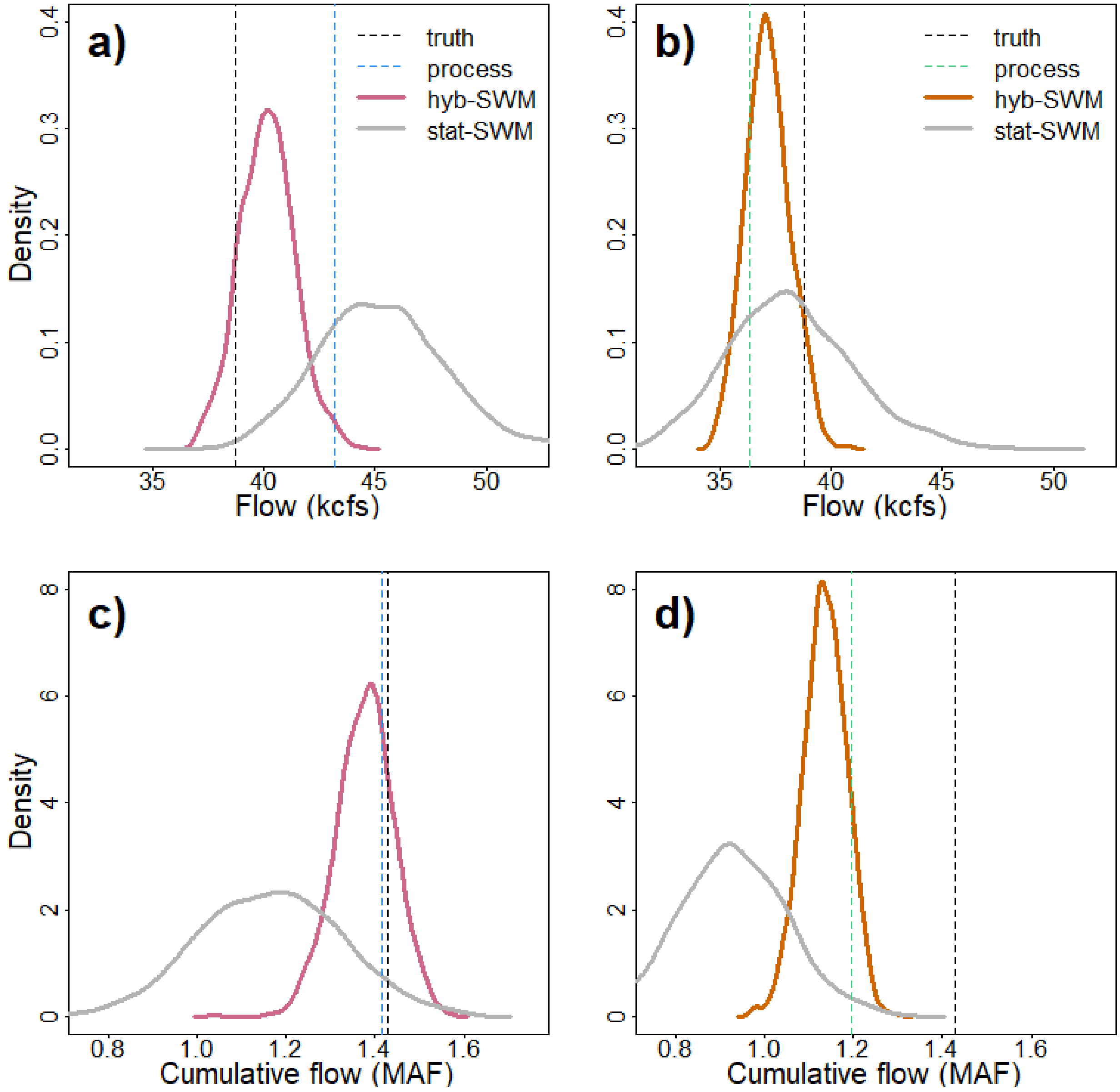
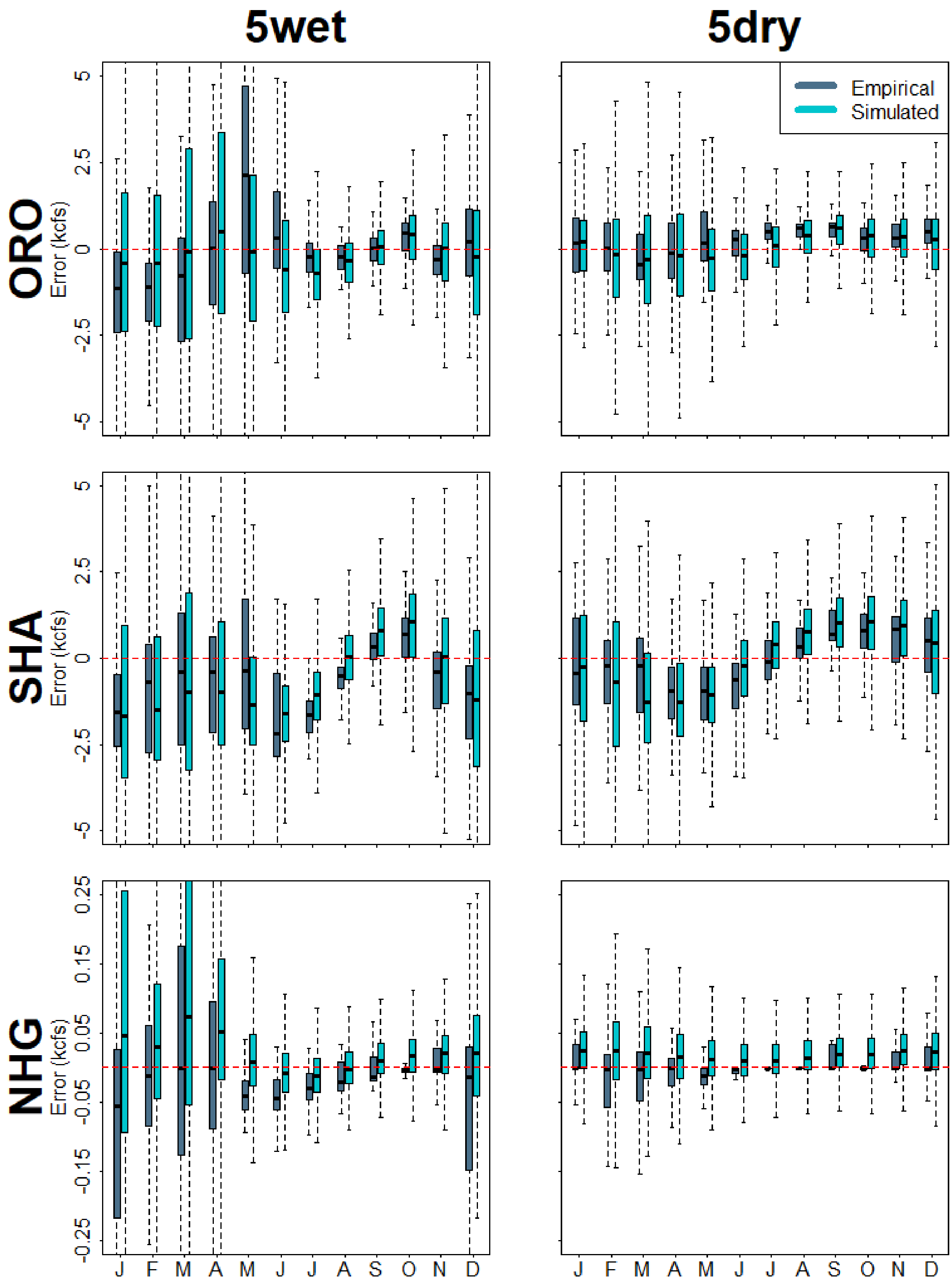


Figure 12. *a) Comparison of 99th percentile flow distributions for hybrid-SWM (pink) versus static-SWM (gray) for Test scenario. Single-value for truth (process) model simulations are shown with black (blue) vertical lines, b) As in (a) for the Test+4C scenario, c-d) as in (a-b) for the lowest cumulative annual flow year.*

4.5. Real-world application

We conclude by employing the hybrid model in a real-world setting to assess whether the model is effectively inferring conditional error distributions when the truth model is the actual streamflow observations and the process model is SAC-SMA. We fit the model to three different basins (ORO, SHA, NHG) to show the generalizability of the hybrid SWM approach across varying basin sizes and hydrologic regimes, using the same procedure outlined for the stylized case. There are statistically significant upward trends in temperature (between 1° and 1.5°C) and no significant trends in precipitation over the 1989-2018 period across the three basins (see Supporting Information S9), but we found that these modest warming trends did not drive substantial non-stationarity in the error distributions between the training (WY1989-2004) and test periods (WY2005-2018) (see Supporting Information S10). However, as shown below, error distributions due vary notably with hydrologic regime.

Figure 12 shows the results of applying the hybrid model to errors between the process model (SAC-SMA) and observed streamflow for the three basins, focusing on the 5 wettest and driest years in the Test period (an analysis across the entire Test period is shown in Supporting Information S10). Across the larger sites (ORO, SHA), there is good agreement between the hybrid SWM simulated errors and the empirical errors in terms of seasonality of the error biases and dispersion, and how this varies across wet and dry years. For example, at SHA, empirical biases during the late fall and early winter (November-February) tend to be negative during wet years and positive or near zero during dry years, and this is captured by the hybrid model. Similarly, the hybrid model captures small shifts in empirical bias across wet and dry years in certain months (November-February, August-September) at ORO. Still, there are areas where the hybrid model struggles. This is seen most prominently for NHG, which is a much smaller, flashier basin that is rainfall dominated and has many days of zero flow. During dry years, the model overpredicts the magnitude and spread of errors at NHG in most months, while during wet years bias is overestimated in some months (January, March-June). The model also has a tendency to overpredict error dispersion at ORO in the winter and spring months in wet years, and misses some shifts in bias at ORO in certain months (e.g., May). These challenges notwithstanding, the results suggest that the hybrid SWM can infer important changes to error properties from model states in an out-of-sample period and under different hydrologic conditions in a real-world setting, particularly for larger and less flashy basins.


5wet
5dry
Empirical
Simulated
ORO
SHA
NHG
Error (kcfs)
5
2.5
0
-2.5
-5
0.25
0.15
0.05
-0.05
-0.15
-0.25
J F M A M J J A S O N D

900

Figure 12. *Empirical errors between SAC-SMA and observations compared to 1000 aggregated samples from the hybrid SWM in the 5wet and 5dry subsets of the Test period (WY2005-2018). Comparisons are conducted across three separate basins (ORO, SHA, NHG).*

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 manifested 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.

This result has large implications for the use of SWMs to estimate hydrologic design events under climate change. Process-based models are one of our best tools to predict streamflow under projections of future climate, but these models exhibit systematic errors in the prediction of extreme low flows and high flows that impedes their direct use in estimating climate change impacted design events (7Q10, 100-year flood; Shabestanipour et al., 2023). One of the most important contributions of SWMs is the reduction of simulation bias in process-based hydrological model predictions at the upper and lower flow quantiles (Farmer & Vogel, 2016; Vogel, 2017), and so SWMs are generally seen as a tool that can help preserve the causal nature of process-based models while still providing the information needed for hydrologic design in long-term water resources planning efforts under climate change. However, the differences in error distributions between the Test and Test+4C cases shown in this work imply that SWMs trained to a historical period may not improve the estimation of these design criteria under future climate conditions, complicating the use of SWMs as a tool for water resources planning 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 ML error correction to remove biases conditional on hydrologic state, and then used dynamic residual modeling to capture the dependencies between hydrologic state and higher order moments of the error distribution. To better emulate out-of-sample predictive uncertainty, we introduced a training approach whereby we fit the error correction model to a calibration set and then subsequently fit the dynamic residual model to a separate validation set, before evaluating the approach on an independent test set.

We found that the hybrid model was able to capture prominent shifts in predictive uncertainty in the test set, both for historical climate (Test) and under warming (Test+4C). This included significant changes in bias during the winter and spring months, when snow accumulation and melt dynamics differed significantly between the truth and process models in the Test and Test+4C cases. Notably, a static benchmark SWM was unable to emulate these shifting biases. The hybrid modeling framework was also able to predict changes in error variance and kurtosis in the spring months under warming. Overall, predictive uncertainty estimated using the hybrid error model matched that observed between the truth and process model reasonably well, even though some attributes of predictive uncertainty (e.g., magnitude of bias; coverage probabilities) were not captured, especially in low flow months. This finding was further supported by verification of the hybrid SWM ensembles against high and low flow quantiles of the truth model. While improvements in some of these attributes should be the focus of future work, our methodology provides an important step towards addressing a gap in the hydrologic ML literature of how to adequately assess uncertainty under plausible but unprecedented future conditions (Klotz et al., 2022; Wi & Steinschneider, 2022).

Using different approaches for model interpretability (e.g., feature importance, LIME), we showed that lagged error terms, components of simulated streamflow, precipitation, and snow water equivalent 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 or non-linearly in intensity depending on the background climate and month of interest, which was of particular relevance during historical snowmelt seasons. This suggests 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), demonstrating that ourapproach to leverage relationships between model state and model error to infer these complex changes has important future work potential. We also highlight that these complex changes occurred where non-stationary was applied via a simple temperature shift to the hydrologic model forcings, suggesting future studies on how more complex forms of non-stationarity (e.g. changes to precipitation distributions) might impact the hydrologic predictive uncertainty.

We also tested the hybrid model in a more challenging real-world setting in three separate basins, where the hybrid error model had to predict changes in predictive uncertainty between a process model and actual streamflow observations. We found that the hybrid error model worked well across most sites and months, exhibiting similar performance to the stylized experiment in capturing out-of-sample, state dependent shifts in hydrologic uncertainty that varied across wet and dry years. These included prominent shifts in the sign, magnitude, and dispersion of the error distributions. However, the hybrid model struggled for the smaller, flashier, rain-fed basin (NHG), particularly in dry years, and for ORO in certain months. The results for NHG in particular suggest the model may have difficulty generalizing to flashier basins with a significant number of zero flow days, which is a challenge previously seen with other SWM approaches (Schoups and Vrugt, 2010). This challenge notwithstanding, the results showed good qualitative agreement with error seasonality and shifts across wet and dry regimes across the other two sites, suggesting the approach has potential as a generalizable SWM strategy under climate change.

In constructing the hybrid error model used in this work, we emphasized interpretability and parsimony over complexity. Future work could explore more advanced error correction procedures, potentially drawing on the forecast post-processing literature (e.g., long short-term memory networks;LSTMs; ensemble model output statistics; EMOS; Seo et al., 2006; Sharma et al., 2021; Siqueira et al., 2021), more complex optimization schemes for the dynamic residual model, or non-linear relationships to state variables in the dynamic residual model. Furthermore, this study accomplished only a limited exploration of the spatial generalizability of the approach and future work should examine in more detail how performance varies by hydroclimate regime. These two efforts should be considered in tandem, as more sophisticated error correction procedures like LSTMs perform best out-of-sample when trained simultaneously to a large set of watersheds with diverse landscape and climate characteristics (Nearing et al., 2021). In other words, to improve the spatial generalizability of our approach, future work should investigate how to train a regional hybrid SWM model, instead of the site specific training accomplished in this work.

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.

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

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

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

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

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

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

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

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

Open Research

All code and data associated with this manuscript are available in Brodeur (2023).

Acknowledgements

This project was supported through funds provided under U.S. National Science Foundation award no. 2205239.

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). Nonstationary-SWM: Nov 1, 2023 release (v1.0.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10064653

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.

California Department of Water Resources, CA DWR (2024). California Data Exchange Center (CDEC): Webservice JSON and CSV – General Data Download. https://cdec.water.ca.gov/dynamicapp/wsSensorData

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

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, *27*(12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023

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

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. *Water Resources Research*, *56*(11), 1–33. https://doi.org/10.1029/2019WR026979

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

Pierce, D. W., J. F. Kalansky, and D. R. Cayan, (Scripps Institution of Oceanography) (2018). Climate, Drought, and Sea Level Rise Scenarios for the Fourth California Climate Assessment. California's Fourth Climate Change Assessment, California Energy Commission. Publication Number: CNRA-CEC-2018-006.

Pierce, D.W., Su, L., Cayan, D. R., Risser, M. D., Livneh, B., & Lettenmaier, D. P. (2021). An extreme-preserving long-term gridded daily precipitation dataset for the conterminous United States. *Journal of Hydrometeorology*, 22(7), 1883-1895.

PRISM Climate Group (2014). Oregon State University, https://prism.oregonstate.edu, data created 4 Feb 2014.

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

Reichert, P., Ma, K., Höge, M., Fenicia, F., Baity-Jesi, M., Feng, D., and Shen, C. (2023). Metamorphic Testing of Machine Learning and Conceptual Hydrologic Models. *Hydrol. Earth Syst. Sci. Discuss.* [preprint], https://doi.org/10.5194/hess-2023-168, in review.

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

Rothfuss, J., Ferreira, F., Walther, S., & Ulrich, M. (2019). Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks. Retrieved from http://arxiv.org/abs/1903.00954

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

Siqueira, V. A., Weerts, A., Klein, B., Fan, F. M., Paiva, R. C. D. D., & Collischonn, W. (2021). Postprocessing continental-scale, medium-range ensemble streamflow forecasts in South America using Ensemble Model Output Statistics and Ensemble Copula Coupling. *Journal of Hydrology*, *600*, 126520. https://doi.org/10.1016/j.jhydrol.2021.126520

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

Wi, S., & Steinschneider, S. (2022). Assessing the physical realism of deep learning hydrologic model projections under climate change. *Water Resources Research*. https://doi.org/10.1029/2022wr032123

Wi, S., and Steinschneider, S. (2023). On the need for physical constraints in deep learning rainfall-runoff projections under climate change, *Hydrol. Earth Syst. Sci. Discuss*. [preprint], https://doi.org/10.5194/egusphere-2023-1744, in review.

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