[Water Resources Research]

Supporting Information for

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

[Zachary P. Brodeur[1], Sungwook Wi[1], Ghazal Shabestanipour[2], Jon Lamontagne[2], Scott Steinschneider[1]]

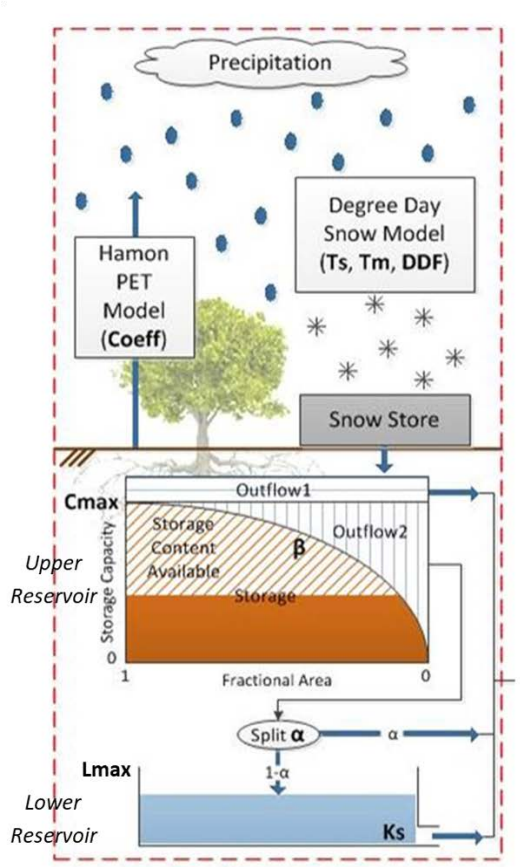
[[1] Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY
[2]Department of Civil and Environmental Engineering, Tuft University, Medford, MA]

Contents of this file

Introduction

The supporting information contained in this file provide background information and justification for various choices in our analysis. Section S1 visualizes the HYMOD hydrologic model. Sections S2 and S3 provide justification for the choice of RF model hyper-parameters and the specific split-sample use of calibration/validation procedures for the hybrid error model. Section S4 is background information on the SEP distribution. Section S5 shows the impact of RF error correction without using lagged error terms. Section S6 provides residual histogram plots for all months. Section S7 shows additional performance diagnostics for the hybrid SWM, including coverage probabilities for the hybrid vs static SWM and the ability of the hybrid SWM to capture streamflow extremes. Section S8 details information about trends in temperature and precipitation over the period of record. Section S9 examines hybrid SWM performance in more detail for the non-idealized 'real-world' case across 3 sites.

S1 – Visualization of HYMOD



Parameter	Description
Coeff	Hamon PET coefficient
Cmax	Maximum upper reservoir capacity
α	Runoff and baseflow split factor
β	Shape of distribution function for upper storage capacity
Ks	Lower reservoir release coefficient
Lmax	Maximum lower reservoir capacity
DDF	Degree day snowmelt factor
Ts	Snow/rain temperature threshold
Tm	Snowmelt temperature threshold

Figure S1. Graphical depiction of HYMOD hydrologic model, with description of parameters.

S2 – RF hyper-parameter selection

Figures S2.1-S2.3 compare the error correction results between the baseline settings of the RF model and a tuned model. We tuned the model through a grid search of the primary hyper-parameters ('ntree' and 'mtry') with integer values between 10 and 500 (by increments of 10) for 'ntree' and integer values between 1 and the total number of variables for 'mtry'. We used 'out-of-bag' prediction error as our selection metric and repeated the grid search 100 times due to randomization in the RF model fitting, yielding an optimal seeded configuration of 20 for 'ntree' and 9 for 'mtry. Figure S2.1 illustrates that only marginal performance gains can be achieved in both the Test and Test+4C cases, as evidenced by a slight reduction in bias and variance for most months. Figure S2.2 shows that substantially more variable importance is apportioned to the lag-1 variable in the tuned model versus the baseline model. In simulation (Figure S2.3), this results in a poorer emulation of the conditional bias and an overestimation of variance.

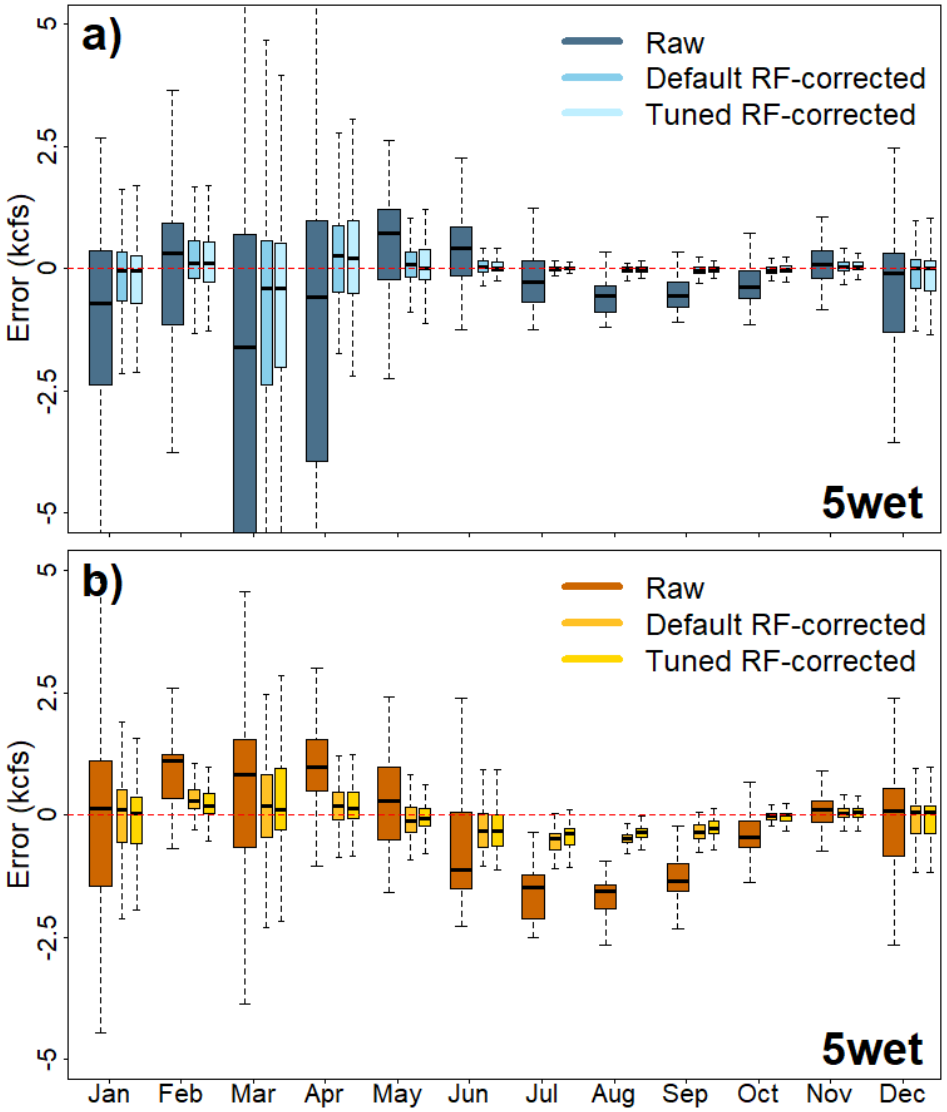


Figure S2.1. Comparison of RF error correction results in the Test (a) and Test+4C (bottom b) between the baseline setting of the RF model (Default RF-corrected) and a model with tuned hyper-parameters (Tuned RF-corrected) according to the procedure detailed in section 3.3.1.

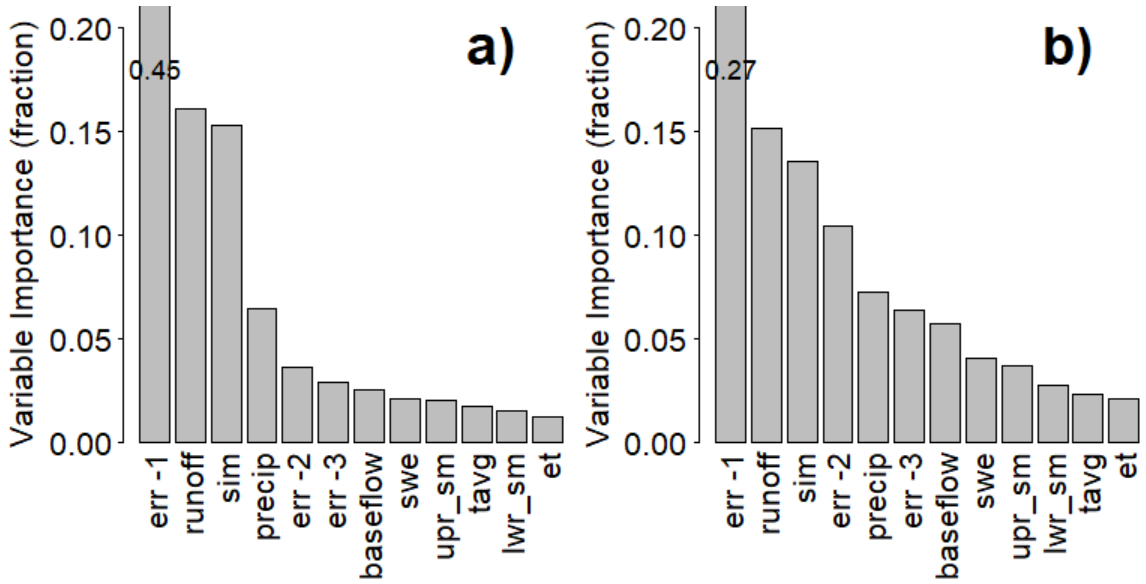


Figure S2.2. Comparison of variable importance between 'tuned' RF error correction model (a) versus 'baseline' RF error correction model (b). In both plots, variable importance for lag-1 error extends beyond the axis limits, but this value is larger for the tuned RF error correction model (0.45) compared to the baseline RF error correction model (0.27).

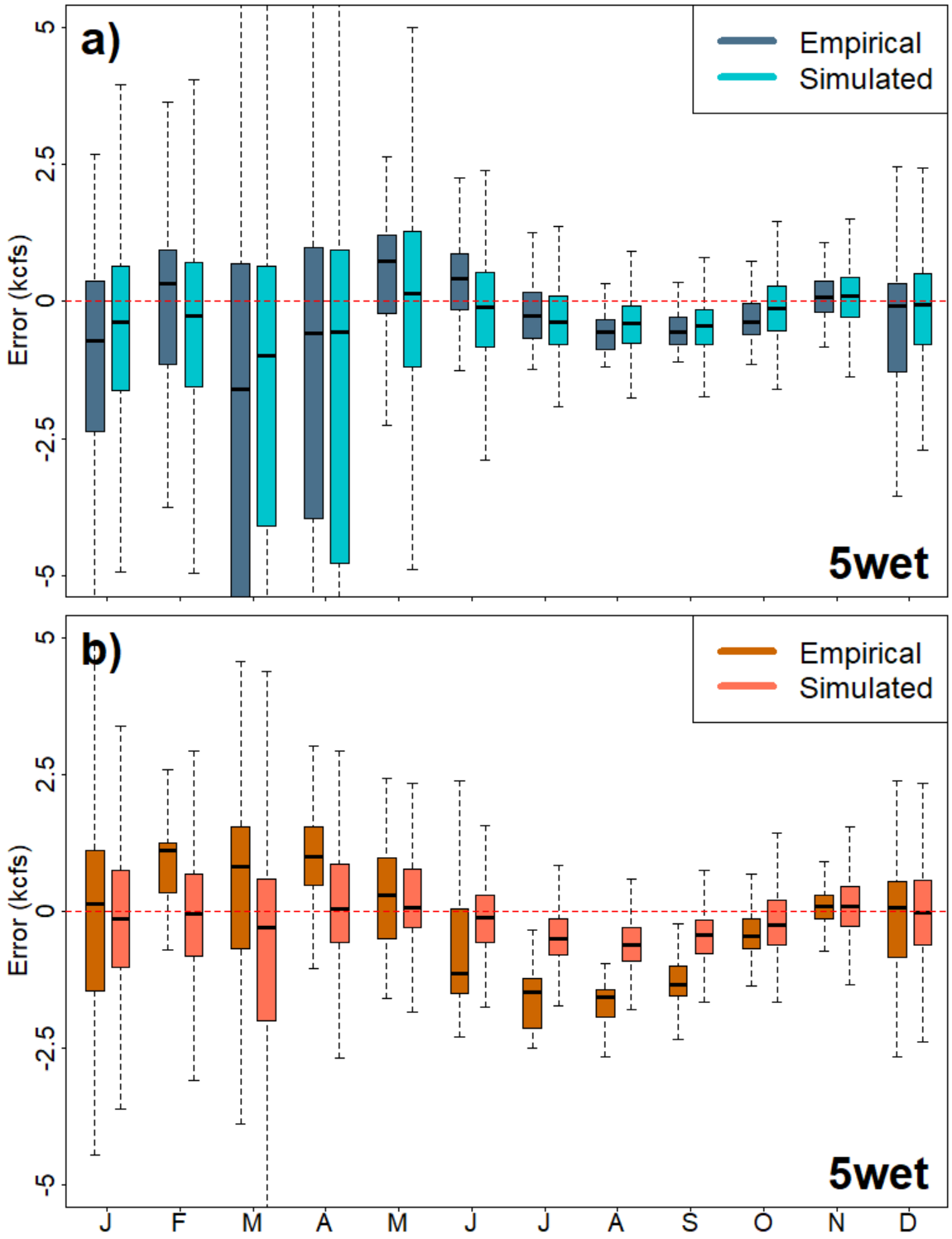


Figure S2.3. a) Monthly empirical distribution of errors in the Test set (dark blue) versus 1000 aggregated samples of hybrid model simulated errors using the tuned RF error correction model (light blue). b) Same as (a) but for Test+4C case.

S3 – Fitting residual model to calibration or validation period

Figures S3.1-S3.2 visualizes simulated residuals from the dynamic residual model when that model is fit to validation (Figure S3.1) versus calibration (Figure S3.2) residuals, and compares the distribution of simulated residuals against the Test and Test+4C empirical residuals. Fitting the model to calibration residuals (Figure S3.2) often underestimates variance, which justifies our choice of fitting the model to validation period residuals (Figure S3.1) for optimal out-of-sample performance.

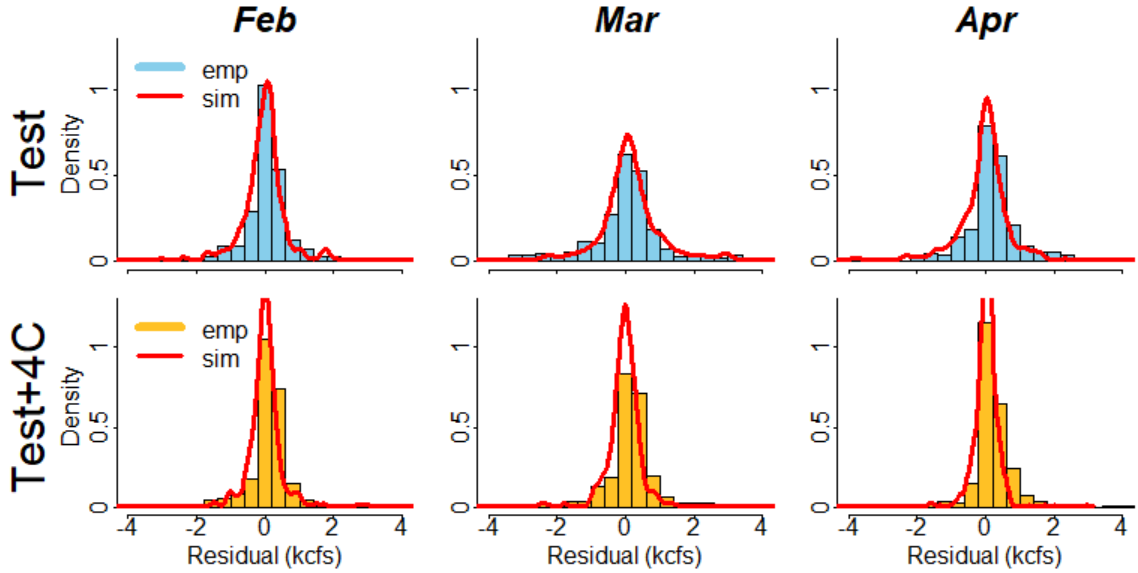


Figure S3.1. Replicate of Figure 8 from the primary manuscript showing the dynamic residual model density estimation (red) versus the empirical distribution for Test (Test+4C) in blue (yellow). The dynamic residual model in this version was fit to the validation residuals as described in the manuscript.

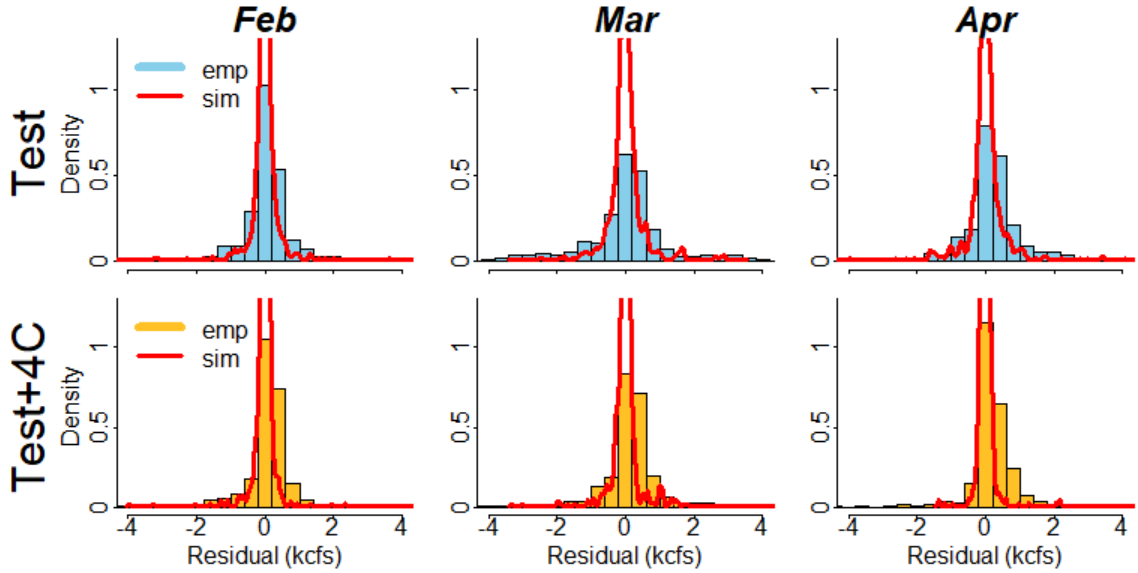


Figure S3.2. As in Figure S3.1 above, but the dynamic residual model in this version was fit to the calibration residuals. Note how the simulated residuals (red density) underestimate the variance of the empirical residuals (histogram) compared to Figure S3.1.

S4 – SEP distribution

Figure S4 shows parameterizations of a standardized SEP distribution, illustrating the distributional form with different values of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. The dynamic residual model samples from this distribution via conditional estimation of these two parameters and we show values approximately within the range of the case study estimates (see Figure 9 in the main article for value ranges). For reference, $\boldsymbol{\beta = 0}$ corresponds to a Gaussian distribution and $\boldsymbol{\beta = 1}$ corresponds to a Laplace distribution. Values of $\boldsymbol{\beta > 1}$ indicate progressively more peaked and fat-tailed distributions.

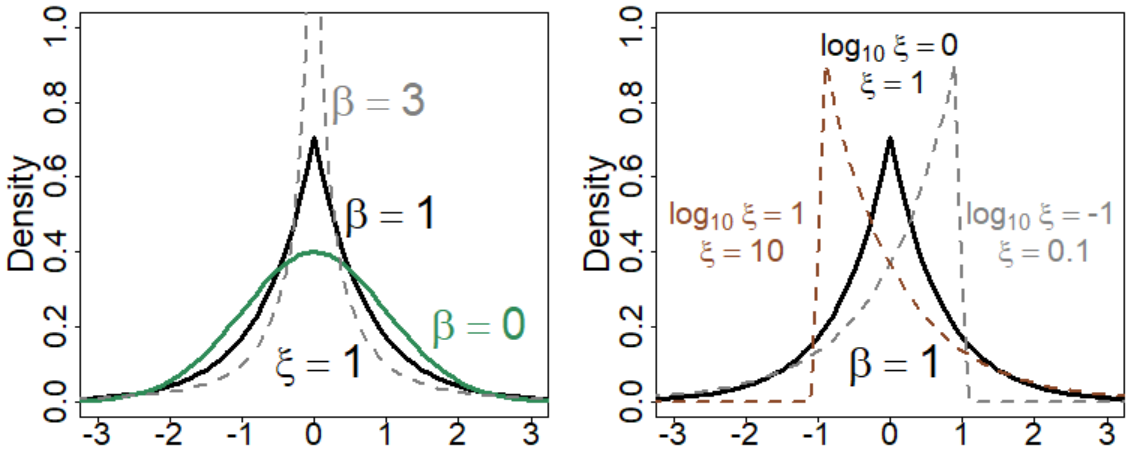


Figure S4. Left panel: Comparison of $\boldsymbol{\beta}$ parameterizations of the SEP distribution with $\boldsymbol{\mu = 0}$, $\boldsymbol{\sigma = 1}$, and $\boldsymbol{\xi = 1}$. Right panel: Comparison of $\boldsymbol{\xi}$ parameterizations of SEP distribution with $\boldsymbol{\mu = 0}$, $\boldsymbol{\sigma = 1}$, and $\boldsymbol{\beta = 1}$. $\boldsymbol{log_{10}\,\xi}$ values are also shown for reference.

S5 – Comparison of Training to Test and Test+4C errors

Figure S5 compares the '5wet' and '5dry' subsets of the Training period (WY1989-2004) to the Test and Test+4C errors that are discussed in section 4.1 and Figure 4. We note that the training period errors largely reflect the Test period errors, although in a number of cases of interest to the analysis (Jan-Feb), they are actually somewhat more extreme in their difference from the Test+4C errors in terms of bias.

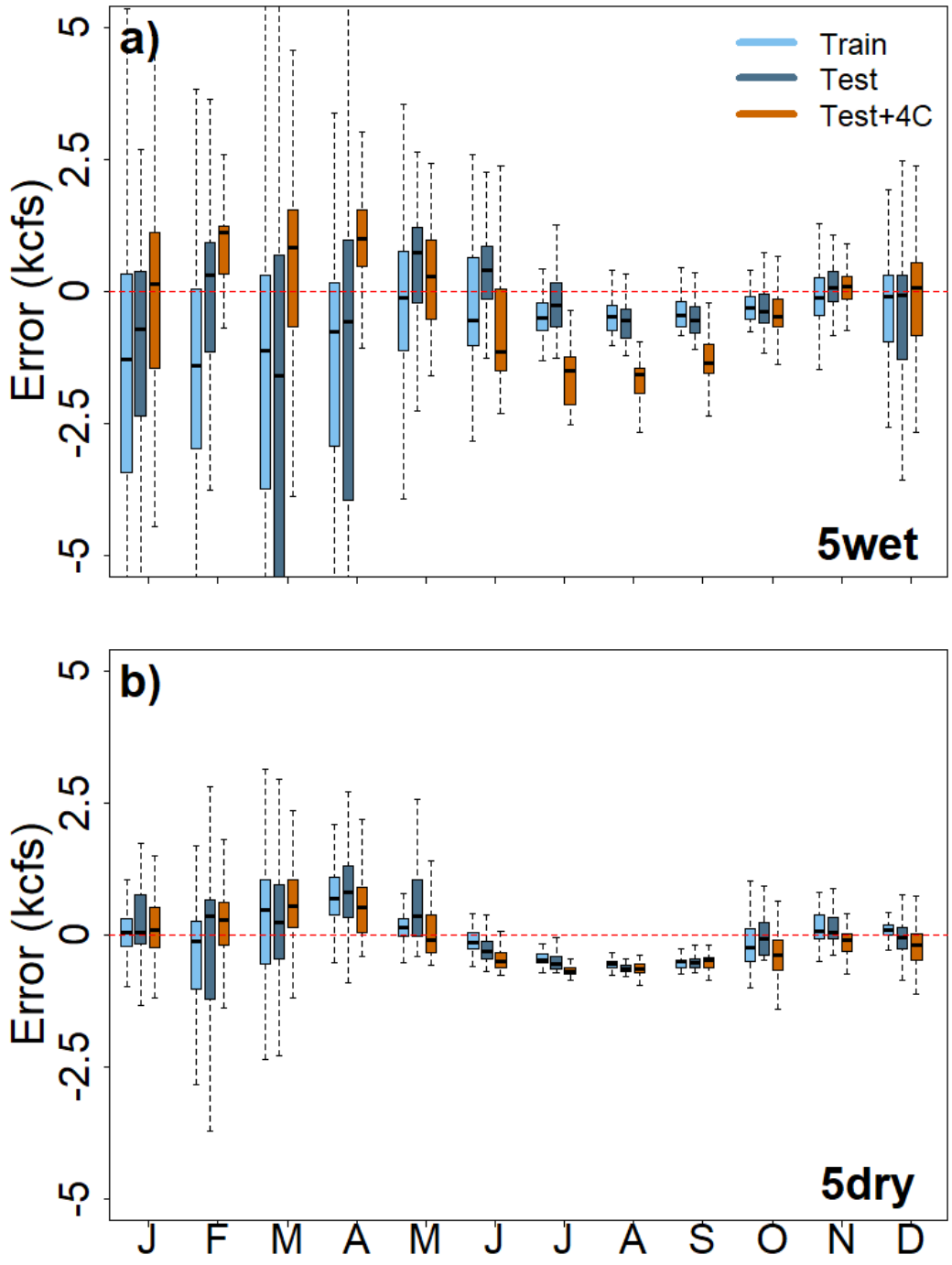


Figure S5. As in Figure 4, but comparing also '5wet' and '5dry' subsets of the training period (WY 1989-2004).

S6 – Results of RF correction without lag terms

Figure S6 shows the result of the RF error correction procedure without including lagged error terms as predictors. This figure shows that an RF error correction model based solely on state variables can generally infer the direction of bias in both the Test and Test+4C cases, but underpredicts the magnitude of bias. Inclusion of the lagged error terms allows a more complete debiasing of the residuals, as shown in Figure 5 of the main article.

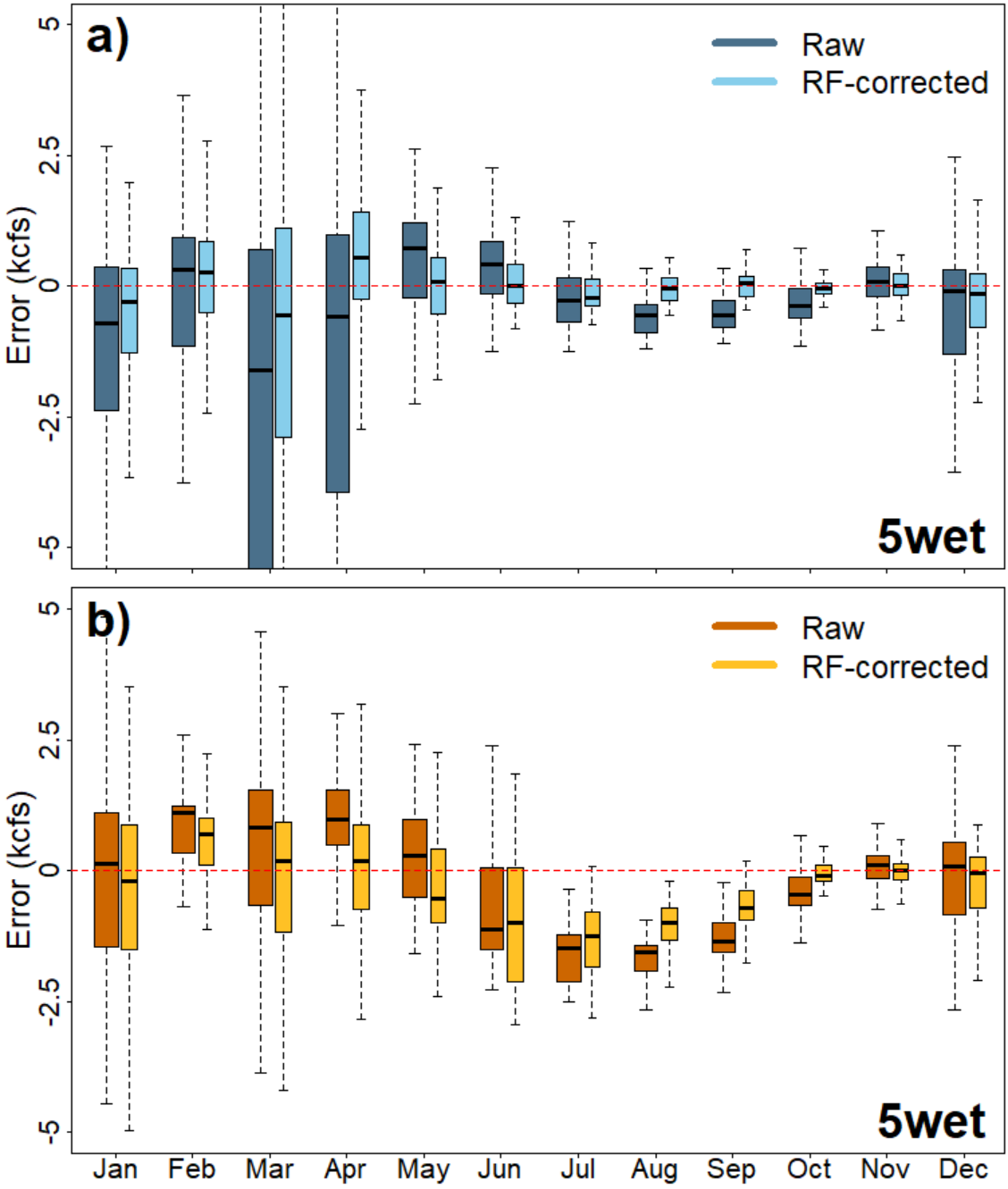


Figure S6. As in Figure 5, but showing RF error correction without incorporating lagged error terms as predictors.

S7 – All month performance of dynamic residual model

In Figure 8 of the primary article, we selected three months to illustrate the performance of the dynamic residual model in monthly subsets that changed substantially between the Test and Test+4C case. In Figure S7.1 and S7.2 we show all monthly subsets for reference.

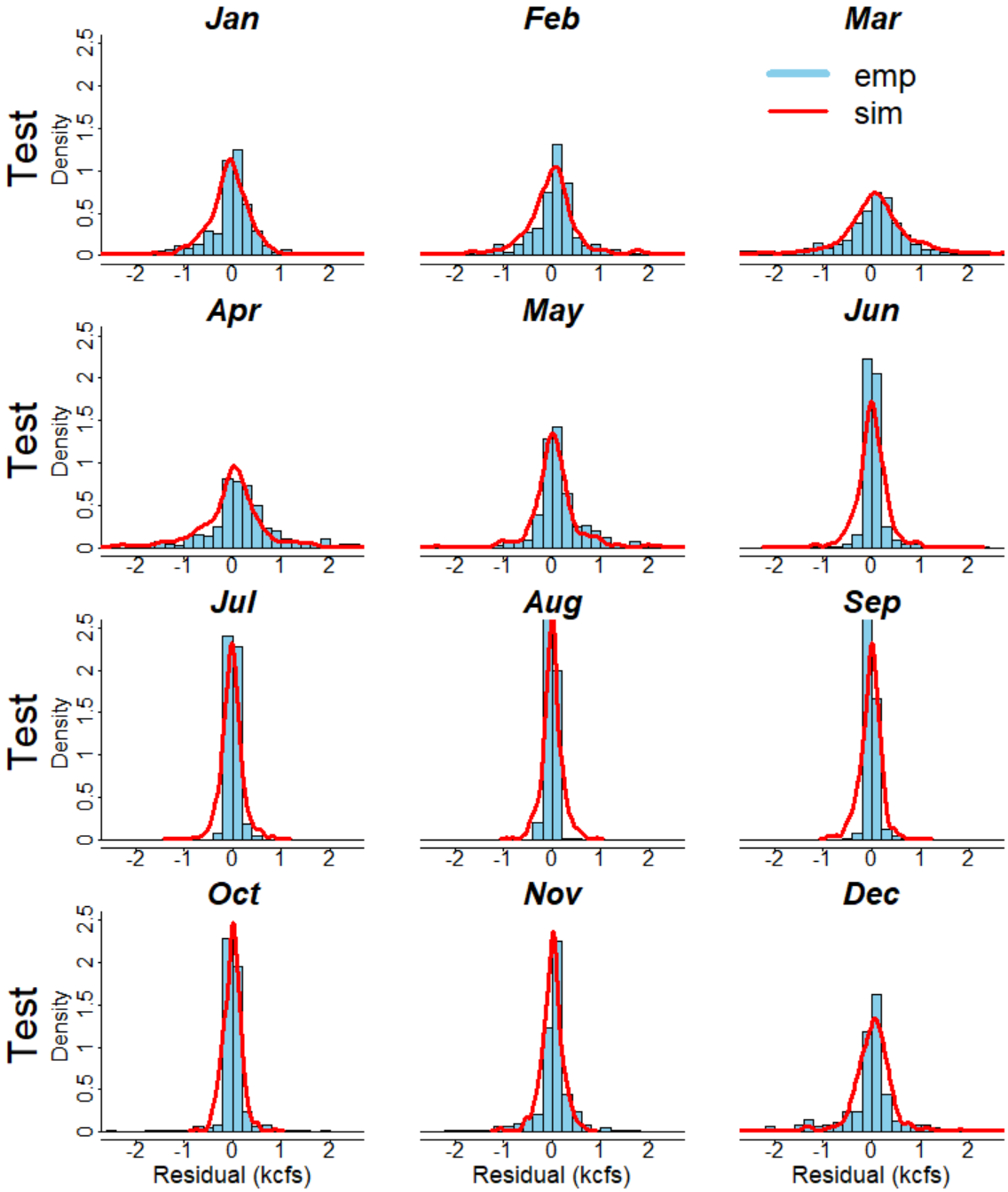


Figure S7.1. As in Figure 8 of the main article (top row) but for all monthly subsets in the Test period.

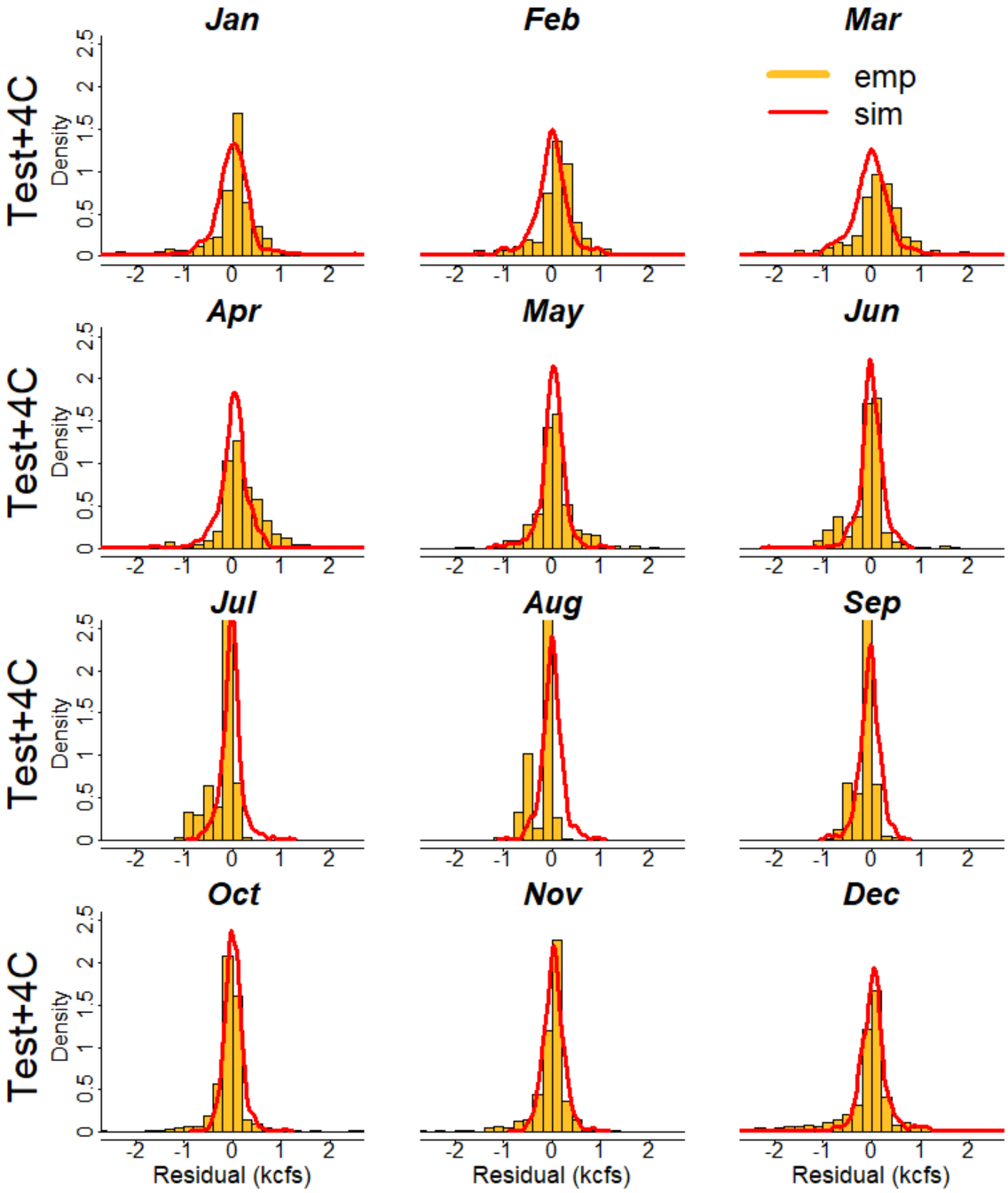


Figure S7.2. As in Figure 8 of the main article (bottom row) but for all monthly subsets in the Test+4C period.

S8 – Additional performance diagnostics of the hybrid SWM

We show median errors and coverage probabilities in March for the hybrid and static SWMs in Figure S8.1. March was selected because it is the month with the largest shift in error bias and magnitude (see Figure 4 of main article) between the Test and Test+4C cases. Figure S8.1 shows that the hybrid model does better at capturing the median error bias change between the Test and Test+4C case than the static SWM, but both struggle to produce accurate coverage probabilities in the Test+4C case.

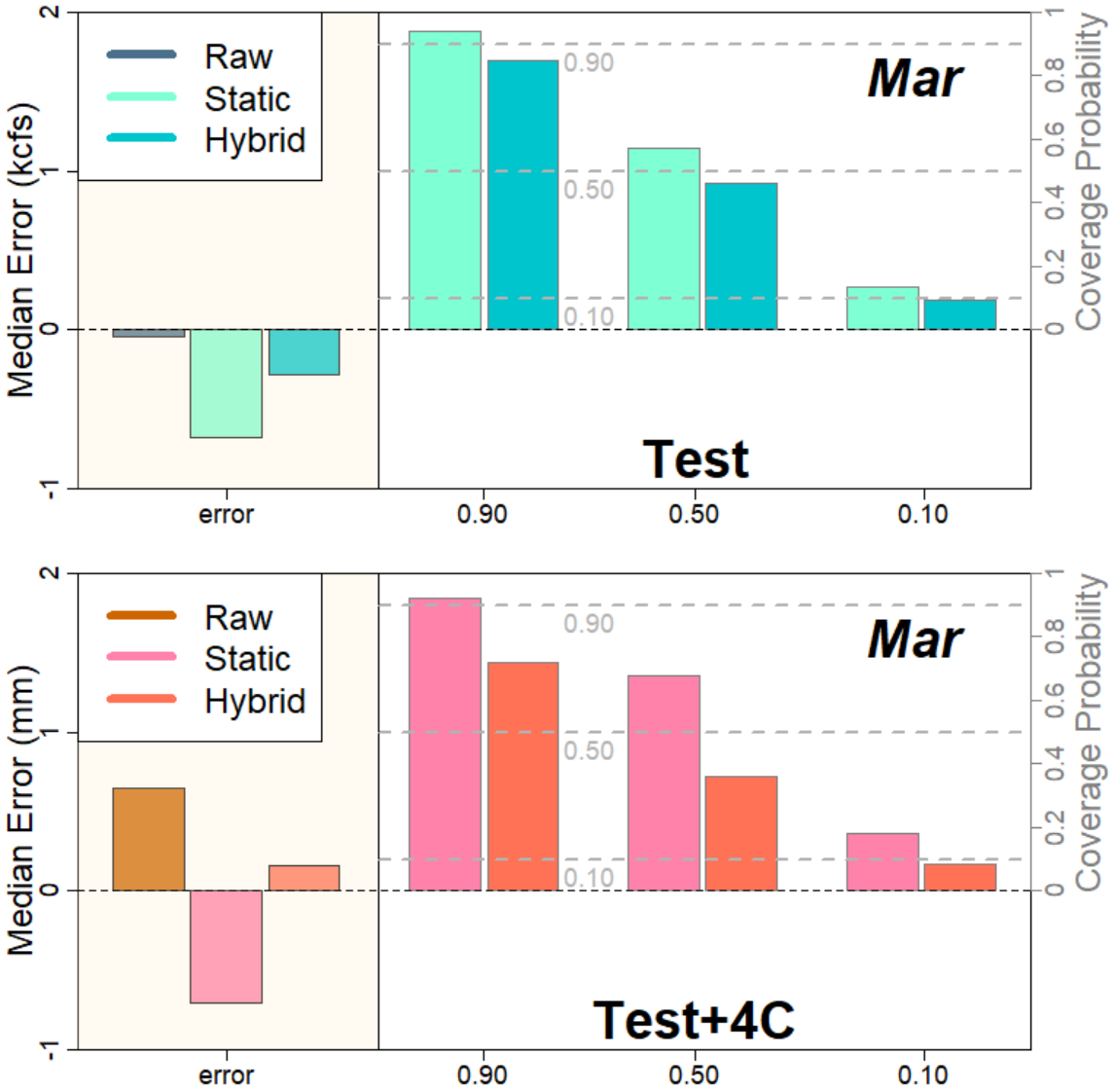


Figure S8.1: Top: Comparison of median error and coverage probabilities for March in the Test case for both hybrid and static SWMs. The median error for the raw errors is also shown. Coverage probabilities are shown for three selected confidence intervals (90%, 50%, 10%), with gray dashed lines showing the target values. Bottom: As in top, but for the Test+4C case.

S9 – Temperature and precipitation trends in meteorological input data

Figure S9.1 shows a significant warming trend across the historical period in the Feather River (ORO), Sacramento River (SHA), and Calaveras River (NHG) basins. Figure S9.2 shows no significant change to annual precipitation over the historical record.

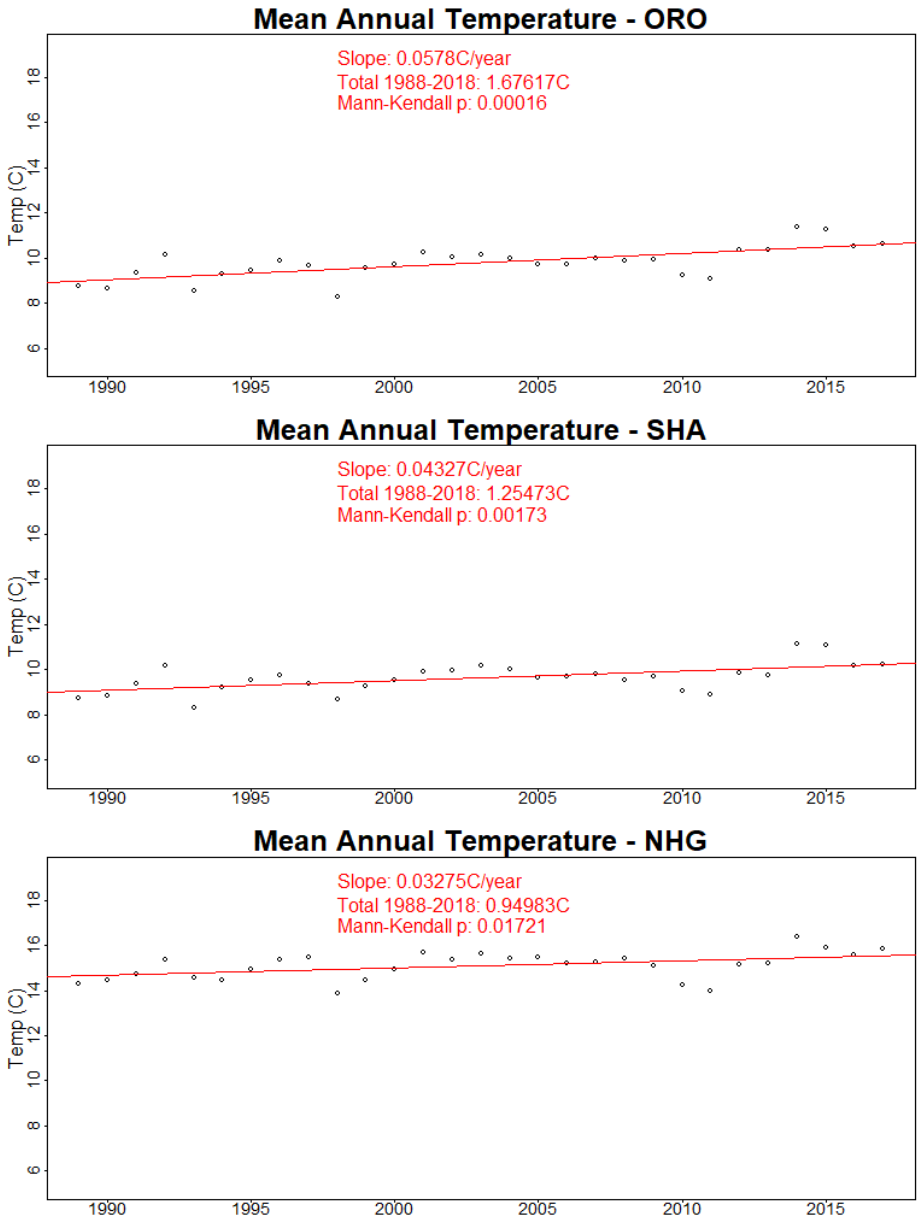


Figure S9.1: Scatterplot of mean annual temperature for three study basins over the historical period of 1988-2018 with ordinary least squares (OLS) fitted linear trend (red line). Red text shows the slope of the linear regression, the estimated total amount of warming over the historical record, and a two-sided Mann-Kendall trend test p-value.

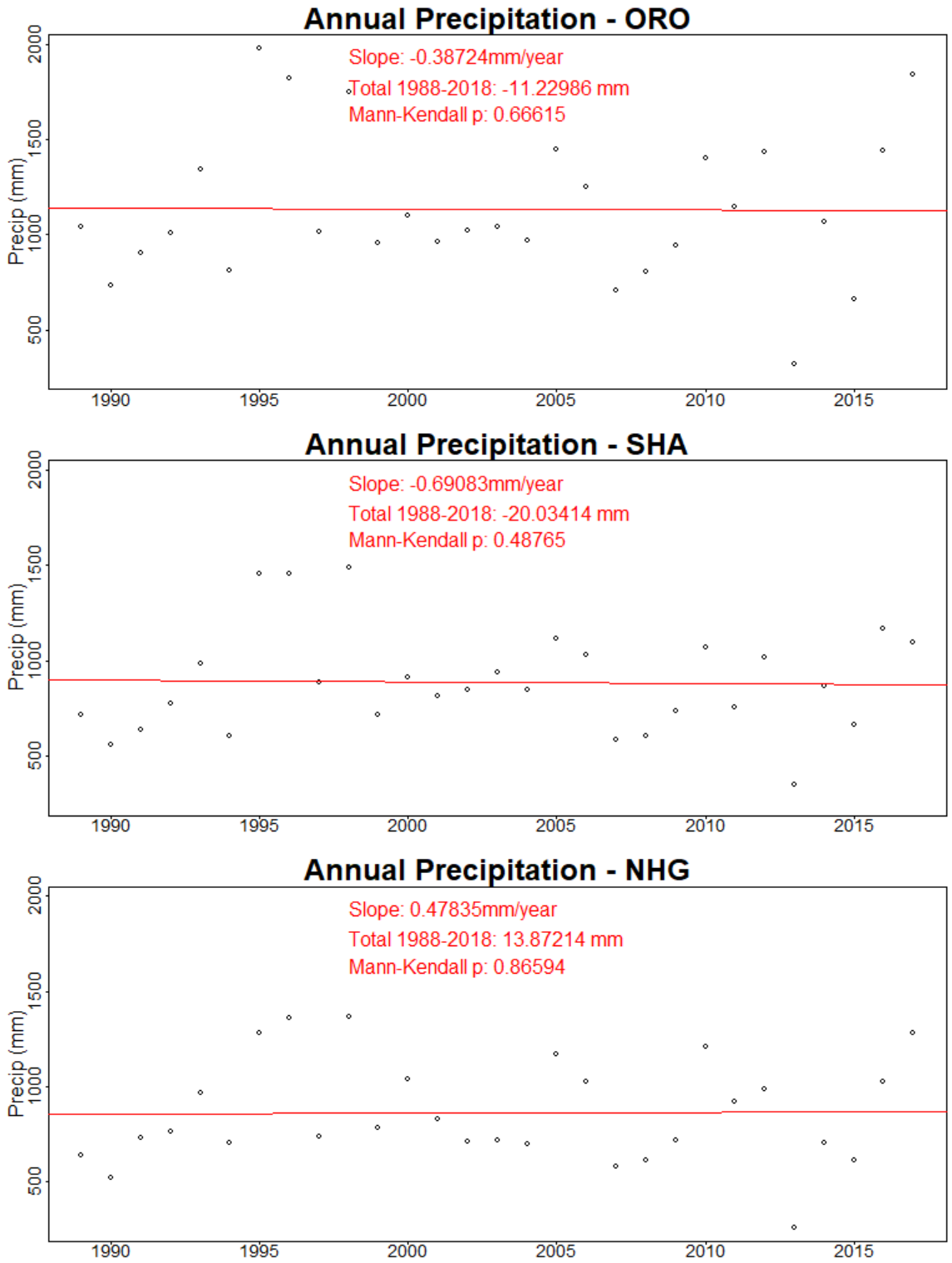


Figure S9.2: Scatterplot of annual precipitation for three study basins over the historical period of 1988-2018 with ordinary least squares (OLS) fitted linear trend (red line). Red text shows the slope of the linear regression, the estimated total amount of precipitation change over the historical record, and a two-sided Mann-Kendall trend test p-value.

S10 – Additional results for real-world application

The following sections provide a more detailed examination of the real world application of the hybrid SWM model. First, we show properties of the real world errors (i.e. SAC-SMA vs observations) for the three sites split between the calibration, validation, and test subsets. Figure S9.1 highlights that there is little evidence of shifts between the training (i.e. cal/val periods) and the test period errors across all three sites, with only a few exceptions (e.g., May for ORO). We also note that SHA exhibits a shift from May-Nov in the validation subset of the training period. Overall though, we find no significant shifts in the error distributions on the order observed in the stylized experiment, possibly because the positive trends in temperature noted in Figure S8.1 are not yet large enough. This finding supports our use of a stylized experiment to explore potential non-stationarity under drastic temperature shifts.

Figure S10.2 shows the results of the hybrid SWM as in Figure 13 but across the entire Test period. This analysis shows aggregate hybrid SWM performance that generally aligns with our findings in the main article.

Figures S10.3.1-10.3.3 highlight properties of both the error correction model and the dynamic residual model in the real world application, similar to Figures 5 and 8 in the primary article. These results show that the performance of the error correction and dynamic residual model in the real world case broadly aligns with the results shown for the stylized experiment. Nevertheless, the results also highlight that the real world application is more challenging than the stylized experiment, which is evidenced by incomplete debiasing in certain months and somewhat poorer fits of the conditional density estimates in some months across the sites.

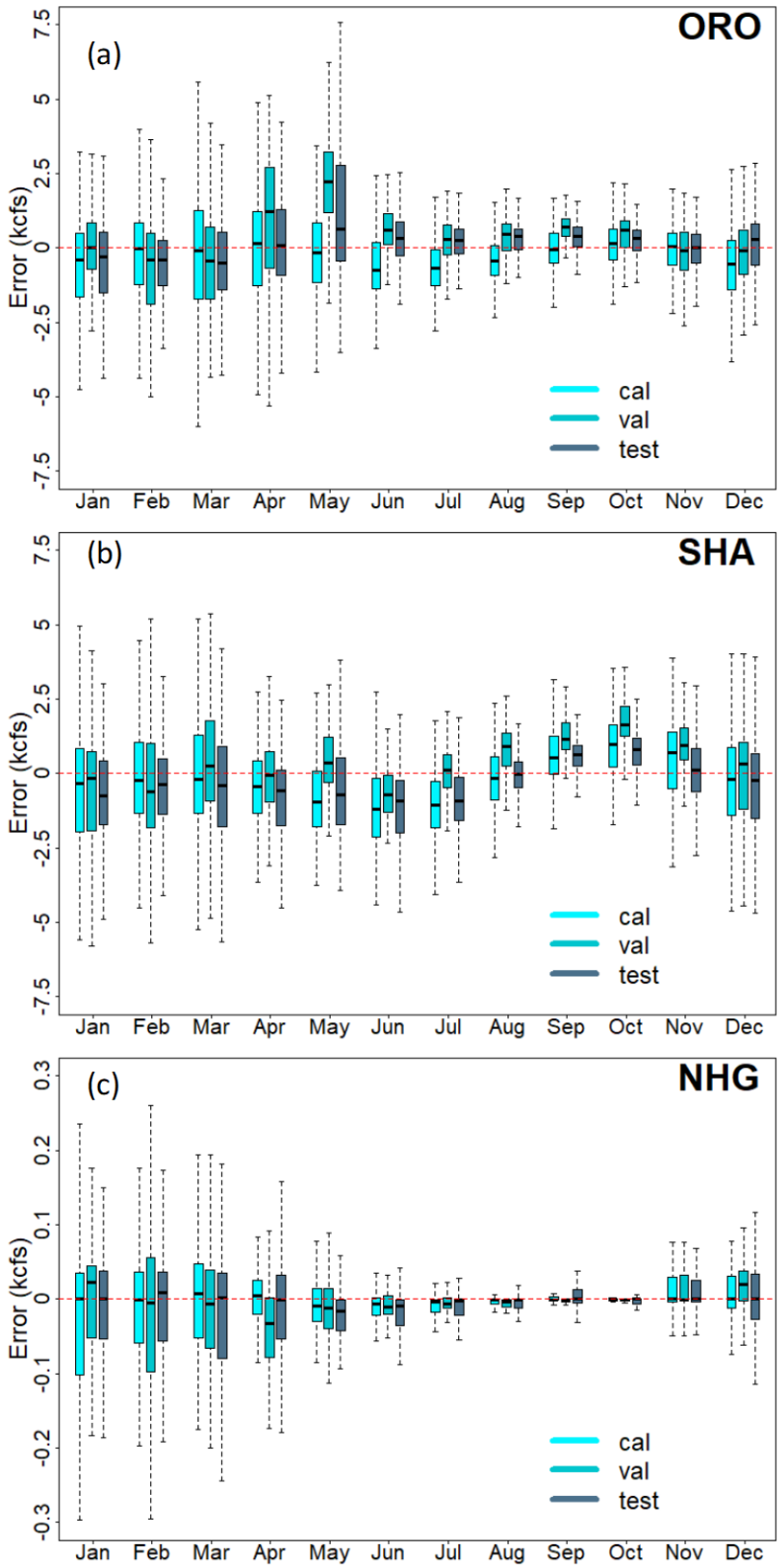


Figure S10.1. Comparison of calibration (WY 1989-1998), validation (WY 1998-2004), and test period (WY 2005-2018) errors for the (a) Feather River, (b) Sacramento River, and (c) Calaveras River basins across all months.

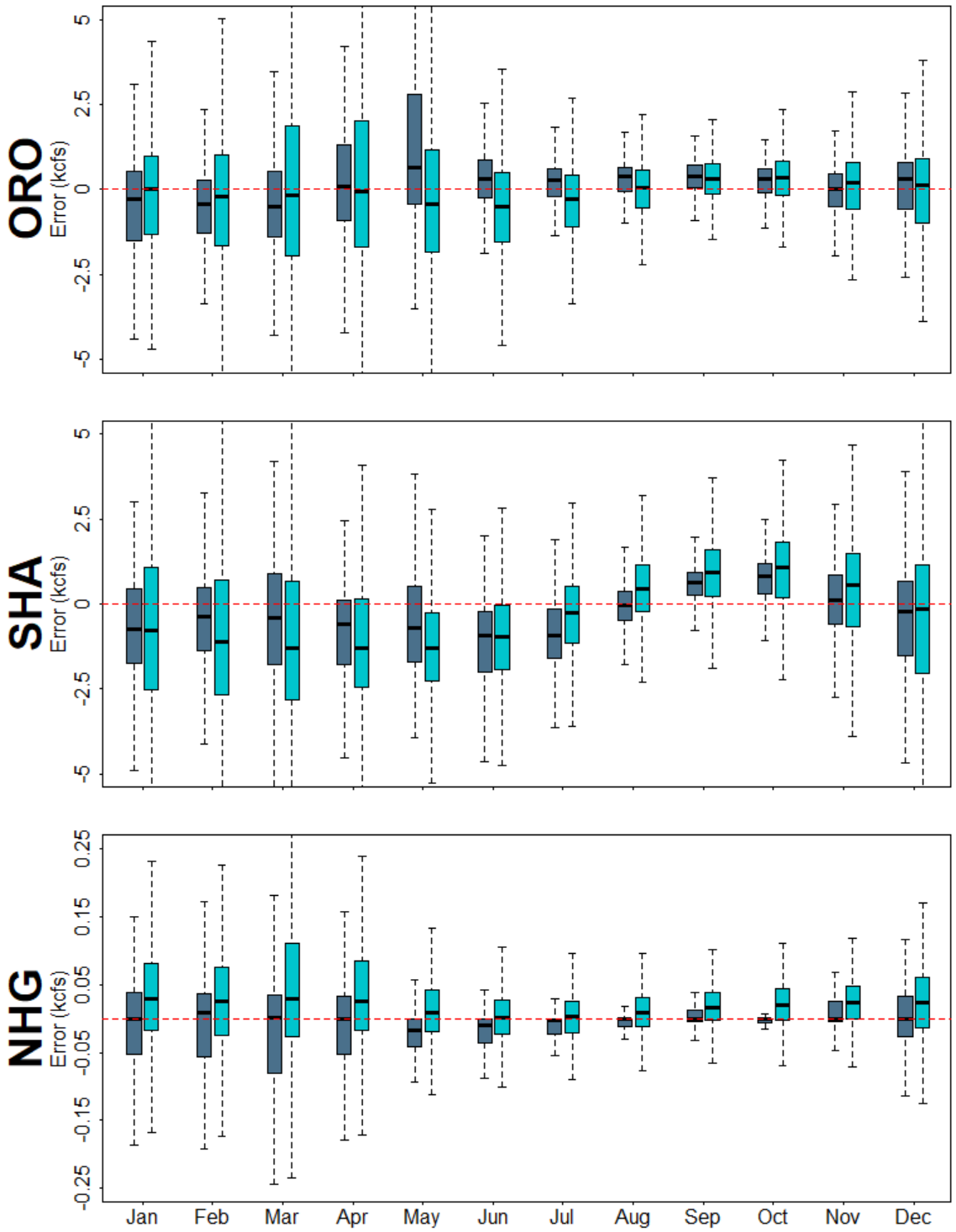


Figure S10.2. 1000 aggregated samples of hybrid SWM error simulations compared to the empirical error distributions in the test period for the Feather River (ORO), Sacramento River (SHA), and Calaveras River (NHG) basins across the entire Test period.

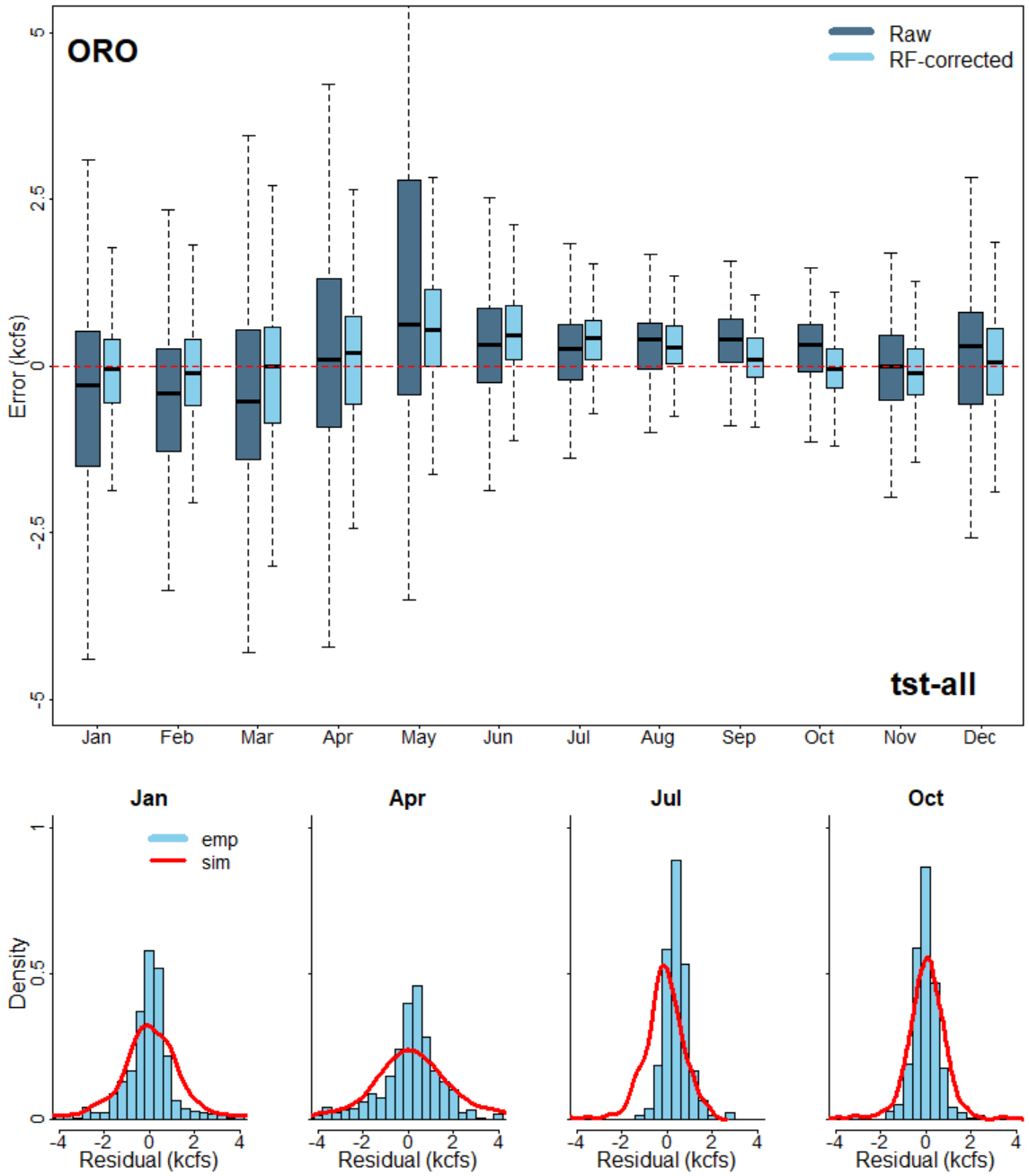


Figure S10.3.1. Top panel: As in the top panel of Figure 5 in the main article, but for the process model (SAC-SMA) fit to the observations at ORO in the Test period. Bottom panel: As in Figure 8 in the main article, but for the four selected months shown.

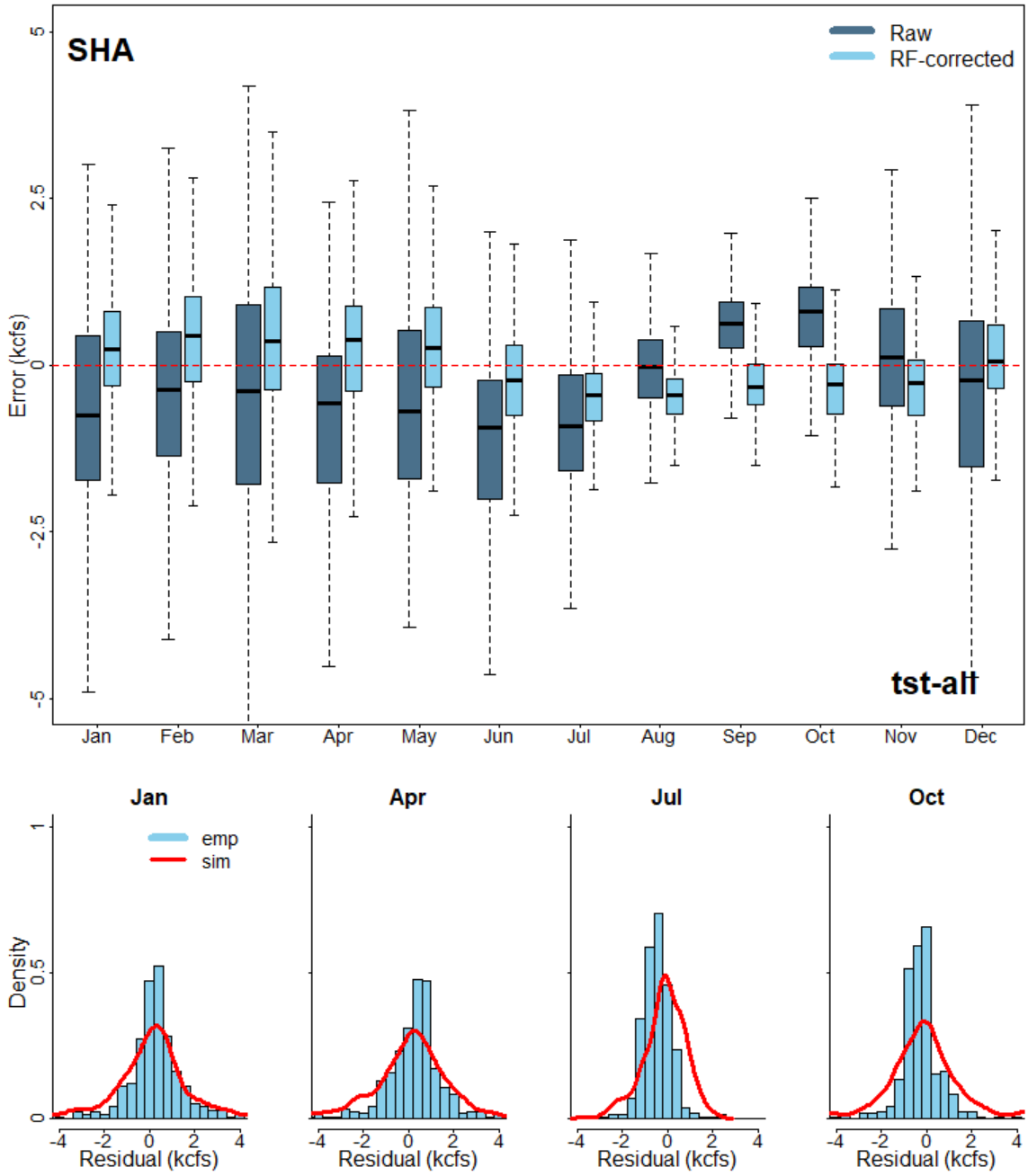


Figure S10.3.2. As in Figure S10.3.1, but for SHA.

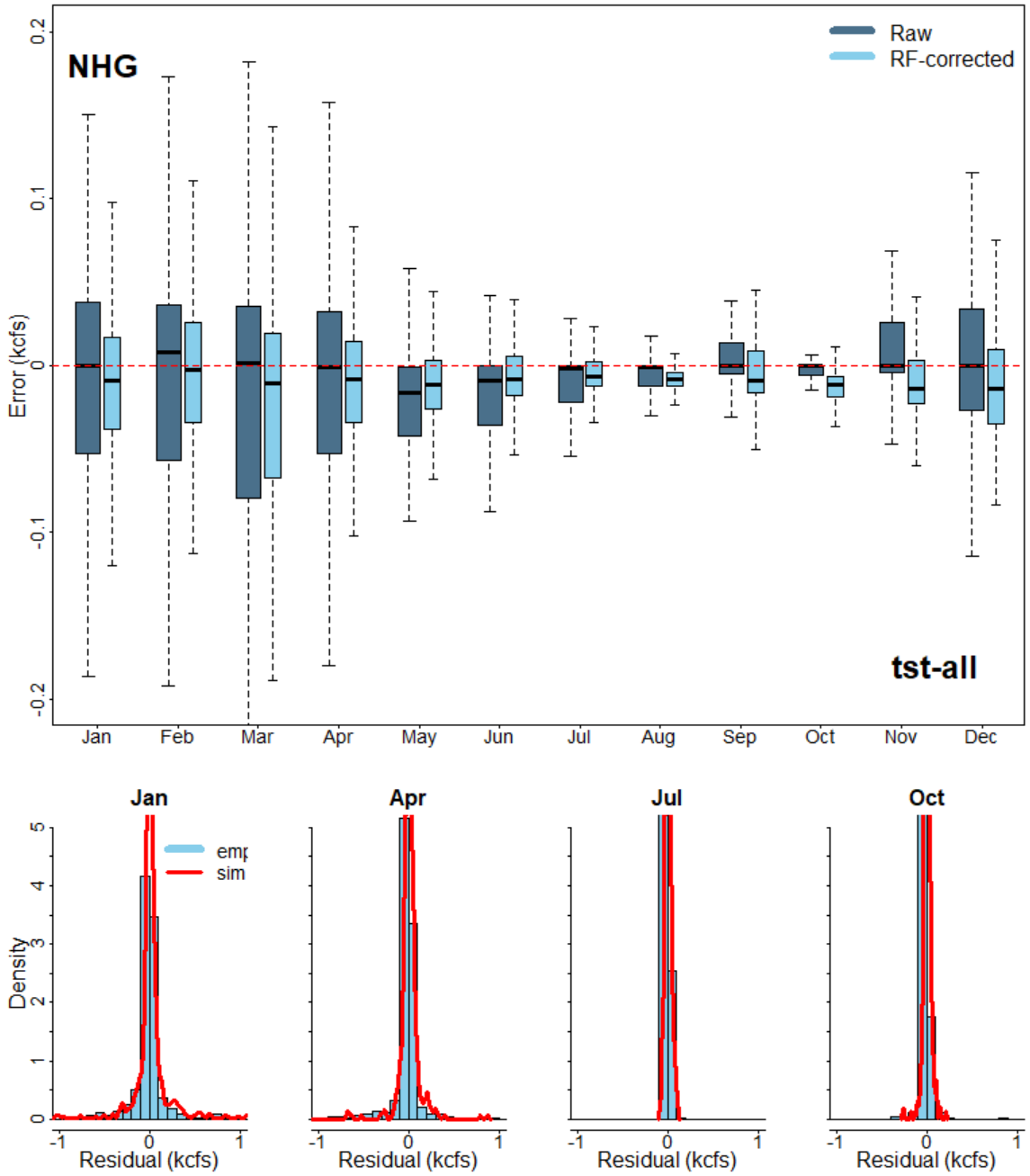


Figure S10.3.3. As in Figure S10.3.1, but for NHG.