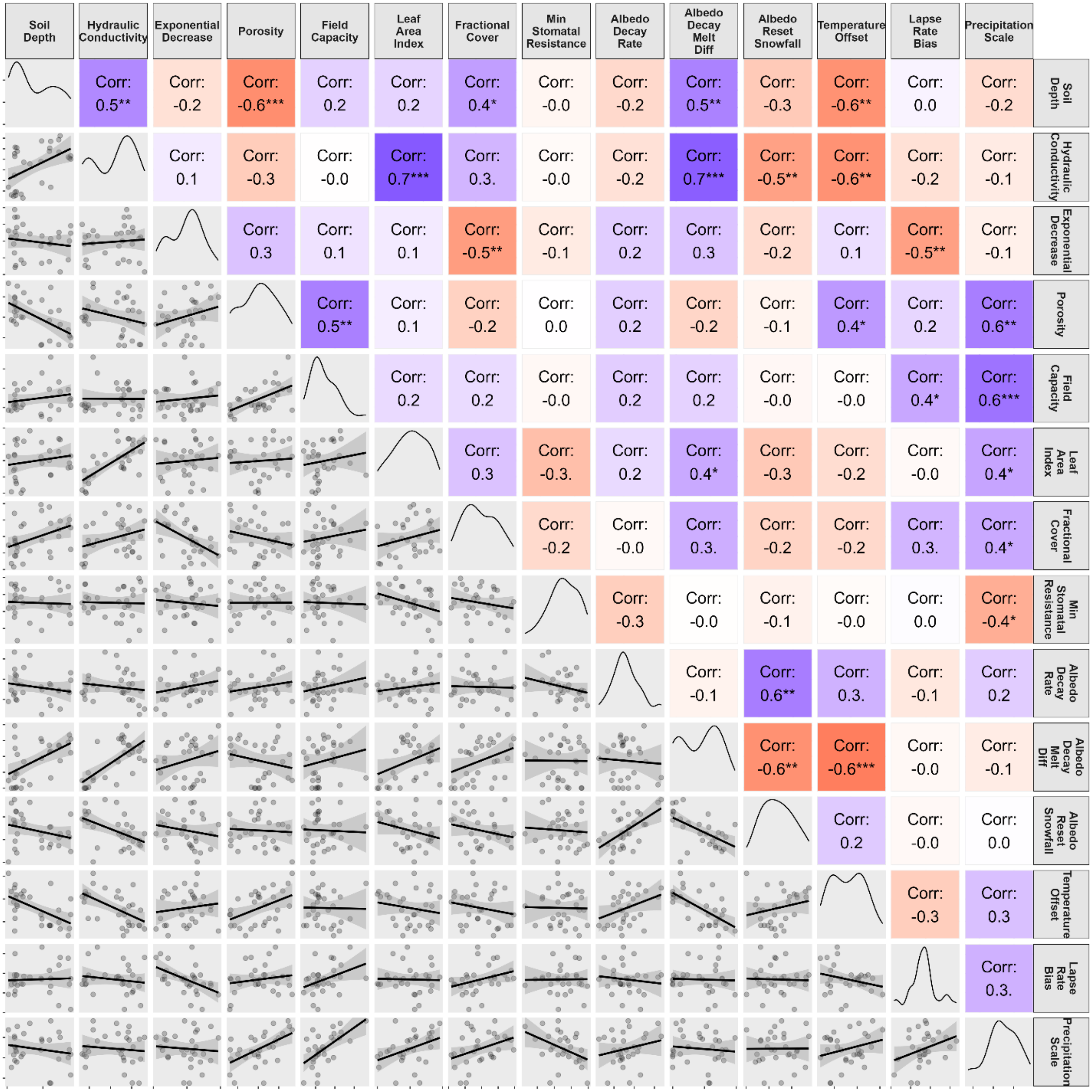

Figure S1: Correlogram of 14 calibrated model parameters for the 30-member behavioral ensemble of Pareto-efficient parameter sets. Three, two, or one stars (*, **, *) indicate statistically significant correlations at $p < 0.001$, 0.01, and 0.05, respectively.**

Across all 91 pairs of parameters, 5 correlations are significant at $p < 0.001$, 16 correlations are significant at $p < 0.01$, and 23 correlations are significant at $p < 0.05$.

Axes scales are rescaled to the residual uncertainty in the behavioral ensemble and do not account for nonlinear transformations applied to some parameters. See Table 1 for actual parameter calibration ranges and transformations. See Figure S4 for behavioral parameter ranges relative to the total calibration range.

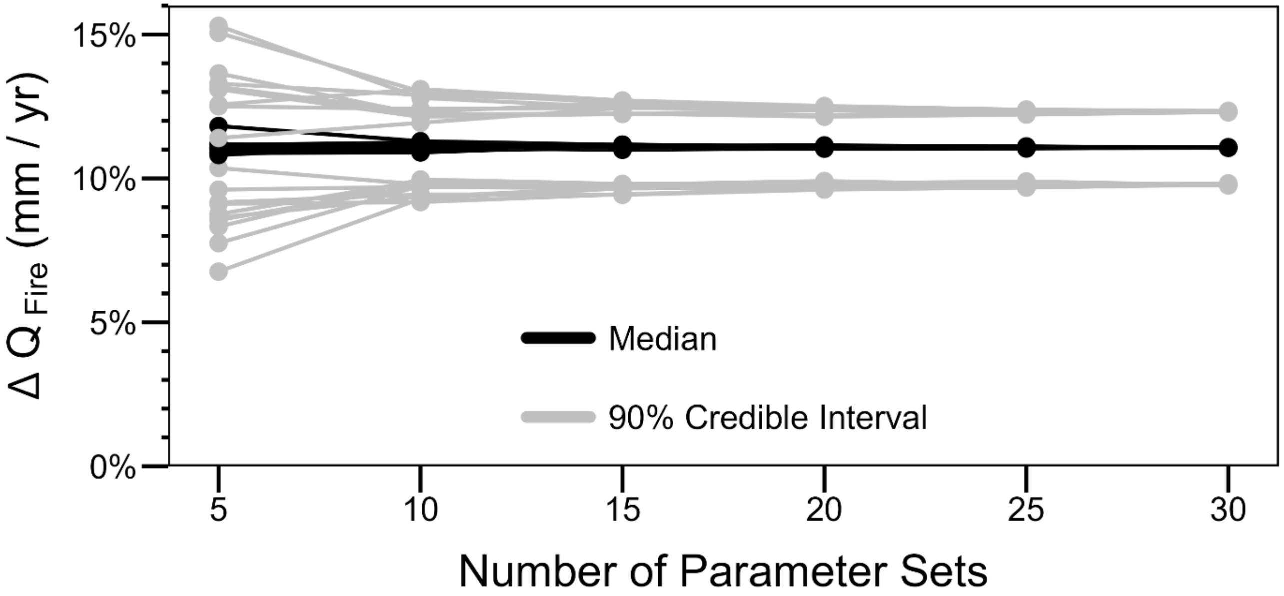

Figure S2: Sensitivity of the post-fire streamflow change uncertainty to the number of selected DHSVM parameter sets used to constrain Eq. 3. Each line is generated by re-fitting Eq. 3 with different sized subsets of N different parameter sets randomly selected from the 30 behavioral parameter sets. The median and 90% credible interval of the conditional metamodel (Figure 5) both stabilize for random sub-selection of different parameter sets as long as at least ~10 parameter sets are used. Results become numerically unstable for < 5 parameter sets (not shown).

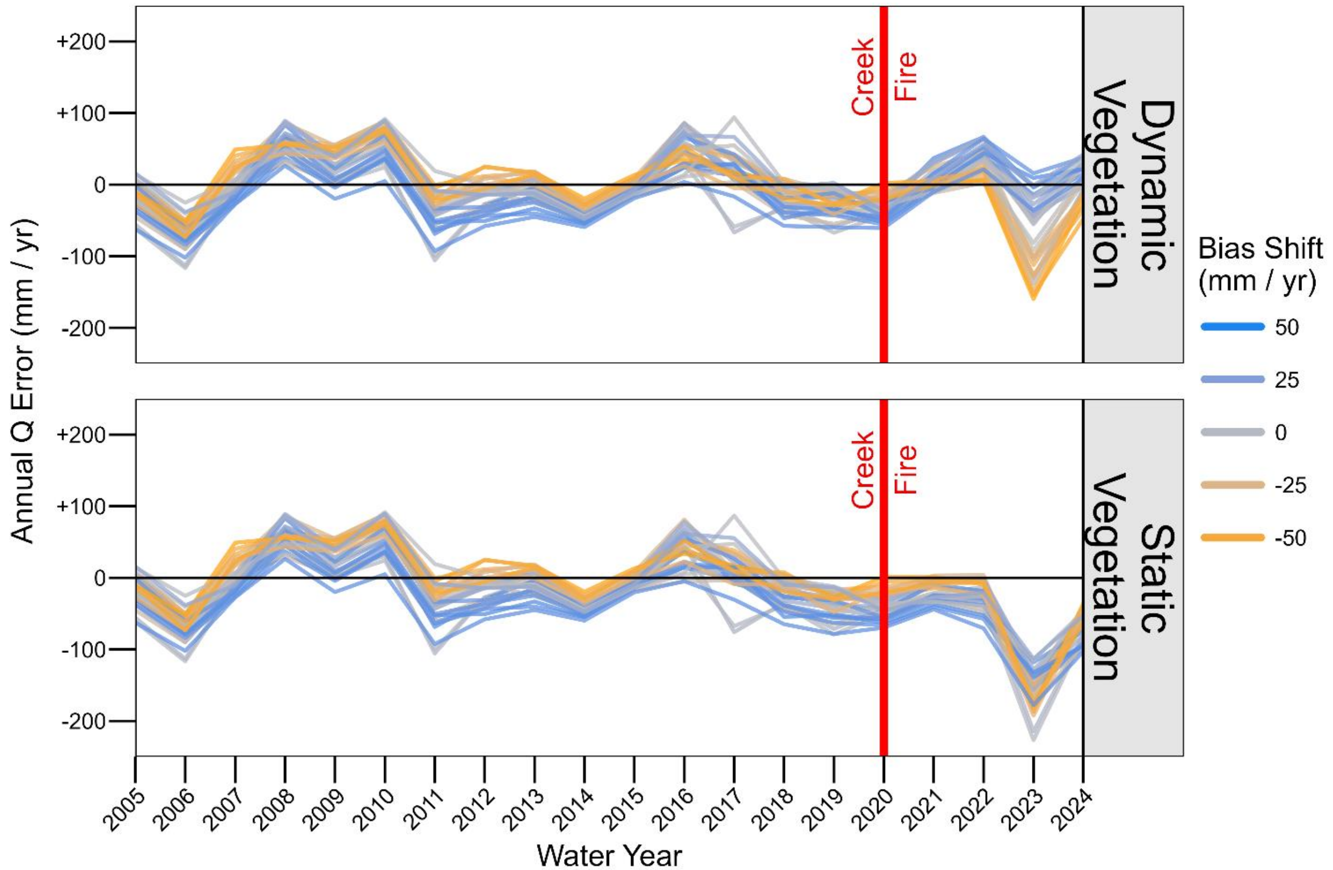

Figure S3: Timeseries of annual streamflow error ($Q_{model} - Q_{measured}$) for all 30 behavioral parameter sets using either dynamic vegetation updated in fire years (top panel) or static vegetation from 2011-era with no fire (bottom panel). With dynamic vegetation, all parameter sets have mean absolute percent error of 4-9% (Table 2) across the 10 calibration years (2015-2024). Both model configurations produce stochastic errors on the order of ±100 mm/yr on the pre-Creek Fire period. After the Creek Fire, the static vegetation models universally underpredict streamflow because they fail to account for the post-fire streamflow increase. Some dynamic vegetation models over- or under-predict post-fire streamflow in different years.

Importantly, the stratification of model error (dynamic vegetation scenario) reverses on pre- and post-fire periods. Orange-colored lines indicate models that shift from over-predicting to under-predicting the true streamflow. Blue-colored lines indicate models that shift from under-prediction to over-prediction. Gray lines indicate models with near-stationary bias.

Compare with Figure 1 and 5 in the main manuscript.

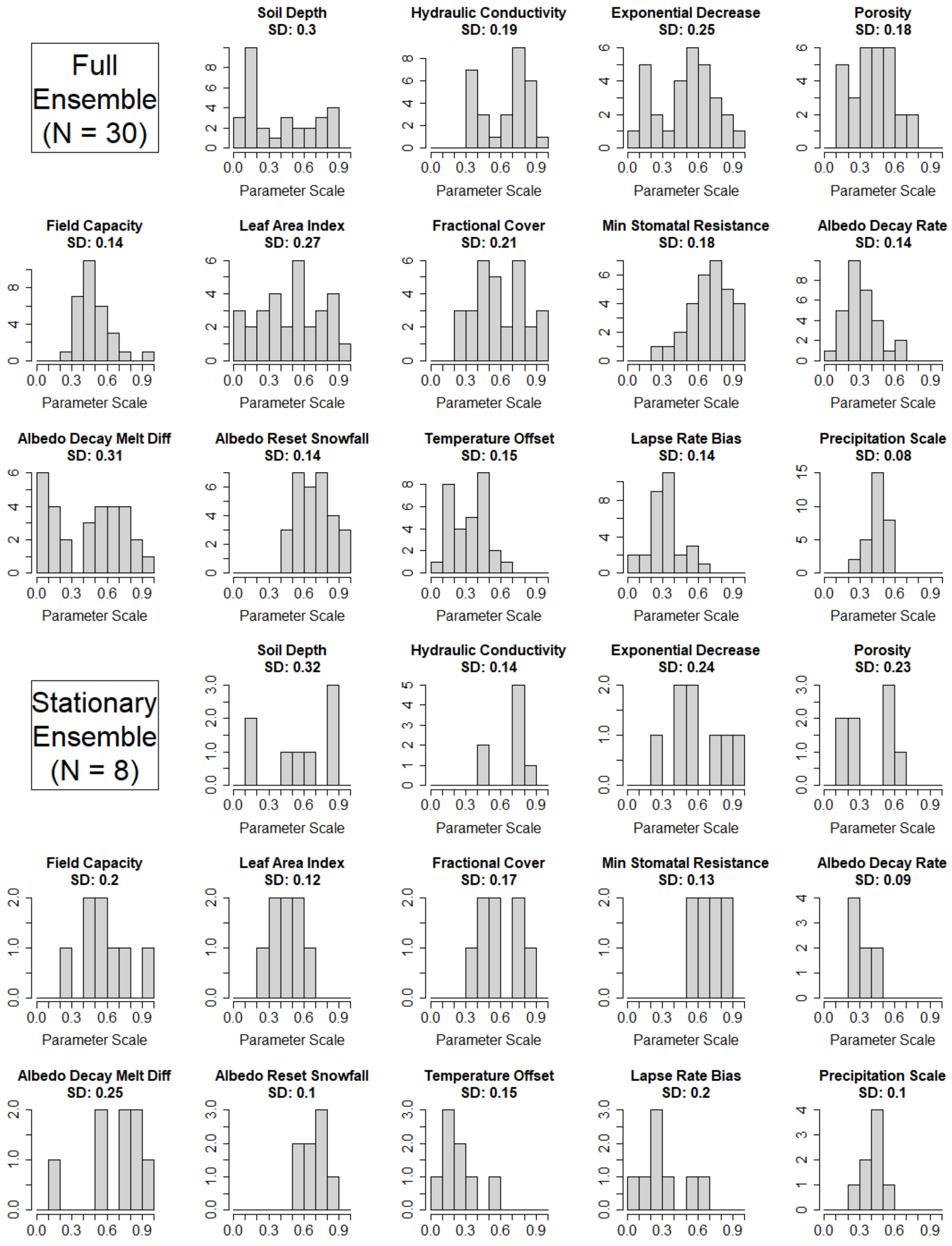

Figure S4: Parameter distributions for the 30-member behavioral ensemble and the 8-member near-stationary sub-ensemble. Horizontal axis scale is indicative of the total calibration range and does not account for nonlinear transformations applied to some parameters. See Table 1 for actual parameter calibration ranges and transformations.

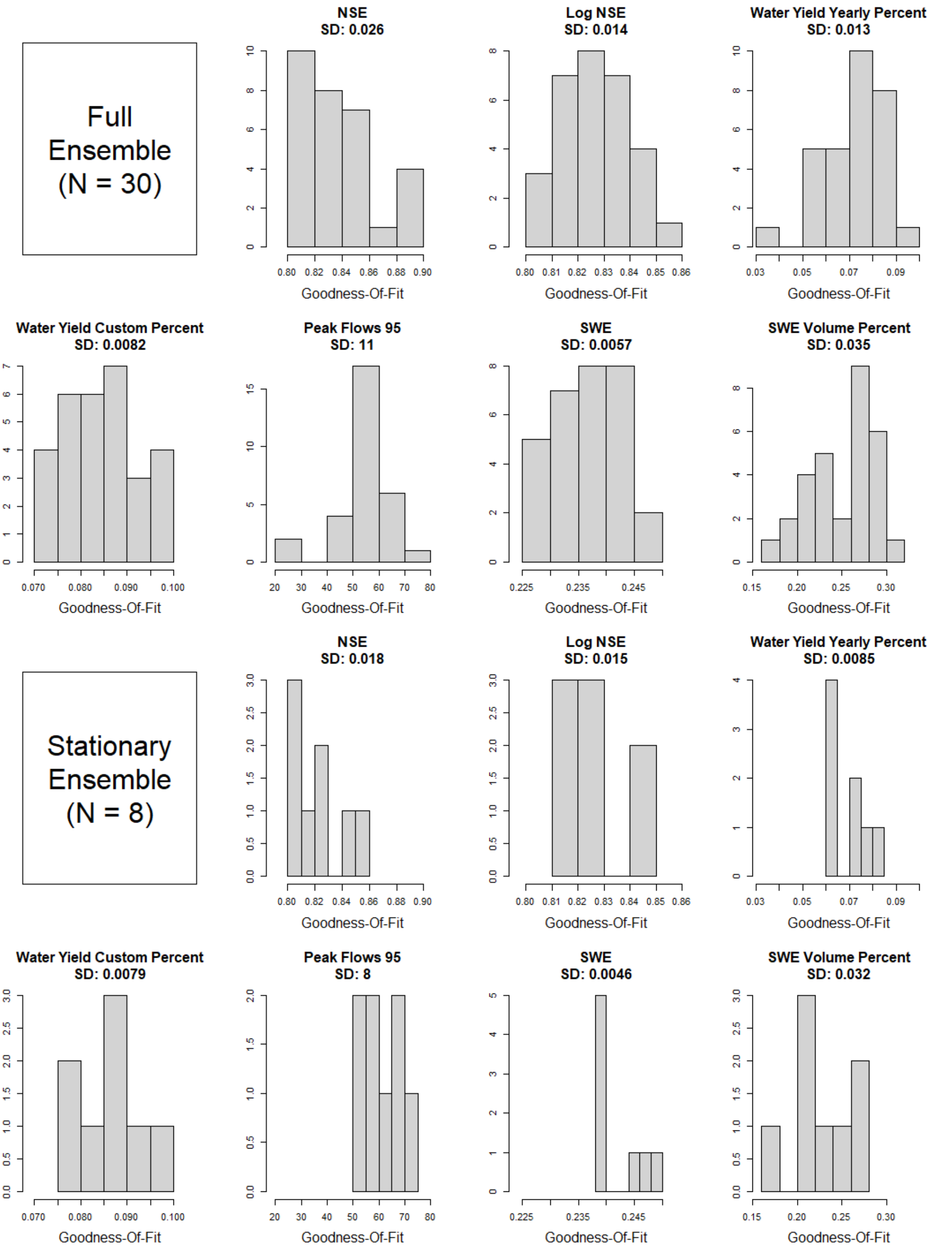

Figure S5: Objective function distributions for the 30-member behavioral ensemble and the 8-member near-stationary sub-ensemble. See Table 2 for descriptions and units.

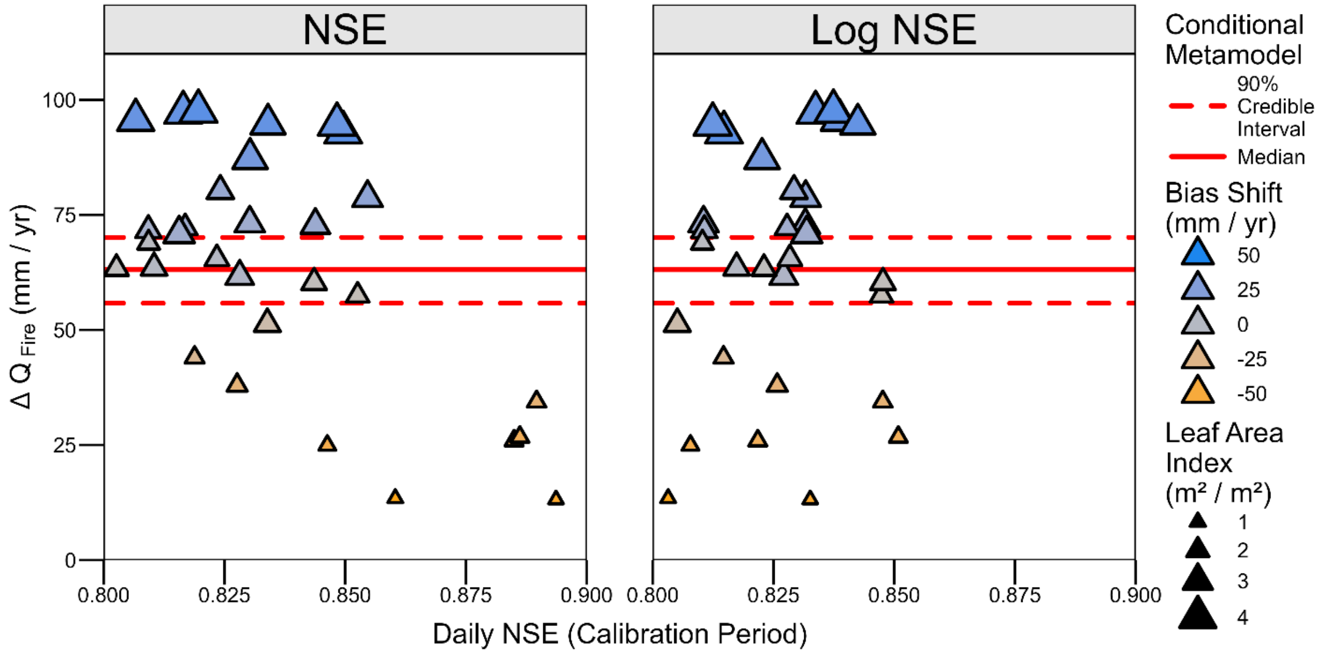

Figure S6: Relationship between calibrated daily NSE or log-scale NSE and the predicted streamflow change after the Creek Fire. Red lines indicate the median and 90% credible interval (5th-95th percentile range) predicted by the conditional metamodel (Eq. 3, Figure 5) for a hypothetical DHSVM simulation with stationary bias. Parameter sets with the highest NSE are outliers with unrealistically small predictions of the streamflow change. Log NSE appears to vary independently of the predicted streamflow change.

In this study, neither of NSE nor log NSE could be used to distinguish between realistic and unrealistic simulations of the hydrological response to disturbance. Instead, a bias shift metric is necessary; parameter sets with near-stationary bias (gray color) generally fall within the 90% credible interval.