Improving Model Calibrations in a Changing World: Controlling for Nonstationarity After Mega Disturbance Reduces Hydrological Uncertainty

Elijah N. Boardman, Gabrielle F. S. Boisramé, Mark S. Wigmosta, Robert K. Shriver, Adrian A. Harpold

Response to Reviewer Comment RC1

We greatly appreciate the reviewers' helpful comments on this manuscript, which we will address in a revision. Our comments are interspersed into the review in blue.

The authors should define counterfactual. On page 2, paragraphs 2 and 3, the authors use this term, but the meaning is not clear. It may need to be rephrased to increase the manuscript's accessibility.

We added a sentence immediately after the first use of "counterfactual" to clarify its meaning:

In this context, a "counterfactual" refers to a hypothetical scenario in which a particular disturbance did not happen, so comparing the actual post-disturbance behavior to the counterfactual scenarios enables attribution of disturbance effects.

On page 3, in the first paragraph, the word "since" should be replaced by "because" due to the authors' intent.

This sentence has been restructured and no longer includes either word.

On page 3, paragraph 1, the logical connection between the lack of landscape-scale observation of many environmental properties and model properties is that we cannot infer the parameters from data and therefore need to calibrate. We suggest making this explicit to increase the accessibility of the text.

Agreed—the start of this paragraph has been reworded to make this connection more explicit.

On page 3, line 71, "since" is used instead of because.

Changed.

On page 3, line 72, the authors hypothesize that equifinal parameter sets may produce divergent predictions. Could work be cited here to state this as a fact rather than formulating this as a hypothesis?

While we agree that this is a commonly discussed idea, we have not been able to find any studies explicitly demonstrating it as a fact, specifically in the context of post-disturbance changes. Thus, we believe that the current formulation as a hypothesis is most appropriate, though we are open to suggestions for references we might have missed.

Figure 1 on page 4 could be improved by including more details. This would reduce the number of assumptions readers will have to make in order to understand the image. -"Initial water balance": for the sake of argument, we'll assume that all models perfectly fit current streamflow, even if given different inputs (P), the simulated ET_tree and ET_other components are such that Qsim = Qobs = 1. "Disturbance response": some disturbance reduces ET_tree (line 79-80).

We have expanded the caption for this figure to include additional explanation as follows:

In the "initial water balance" panel, we assume that all three models closely match pre-disturbance streamflow, with uncertain precipitation and evapotranspiration (ET) components counterbalanced to produce $Q_{Observed} = Q_{Modeled} = 1$ (normalized annual units). After a disturbance (e.g., a wildfire) reduces ETTree, the different models predict various degrees of streamflow change, which is mediated by the potential for compensating increases in ETOther (e.g., soil evaporation and understory ET). In the "disturbance response" panel, the arrows illustrate the direction and magnitude of the water balance changes predicted by each model. In the "streamflow bias" panel, the resulting model predictions are compared to measured streamflow, showing how some models could exhibit a bias after disturbance due to uncertain estimation of water balance changes.

The image could better explain what the up and down arrows in this column show. Readers can speculate that the downward ET_tree arrows mean that (to match reality) the models simulate no ET_tree anymore. In model 1, this means Q_sim goes up by 2, but why would ET_other in models 2 and 3 go up by either the negative change in ET_tree or half that? If these are meant as examples of what could happen with a given model (rather than what will happen), it would be good to state this in the text explicitly. E.g. "[..] increased soil water availability. The three examples show cases where ET_other does not change (model 1), where ET_other fully compensates the reduction in ET_tree (model 2), and where ET_other only aprtly compensates the reduction in ET_tree".

The reviewers have indeed given the correct interpretation of the figure, and we have clarified with an adapted version of the suggested text:

The three examples show cases where ET_{Other} does not respond to the disturbance (Model 1), where ET_{Other} fully compensates for reduced ET_{Tree} (Model 2), and where ET_{Other} only partly compensates for reduced ET_{Tree}. Intuitively, these hypothetical model responses are connected to the pre-disturbance balance of ET_{Tree} and ET_{Other}, which primes some models to predict a larger or smaller compensation effect.

"Streamflow bias": here seems to have an underlying assumption that Qobs increases by 1 after the disturbance. This is key to understanding why the models are assumed to underpredict/overpredict/no change and should be explicitly stated somewhere.

Agreed—we added the following sentence to the figure caption:

In this hypothetical example, we assume that $Q_{Observed}$ increases by 1 unit after disturbance, matching the prediction of Model 3.

Given that this example is intended to give the reader an easy intro into the concepts used in the paper, it's worthwhile to make sure all assumptions are clearly stated. Without this figure, it may raise more questions than it answers. A possible solution is to move the paragraph with lines 92 to 100 before the figure.

We believe that Figure 1 makes the most sense immediately after the first paragraph of text where it is mentioned. However, we have substantially expanded the caption (see above), which we believe makes the figure much easier to interpret even without the rest of the text.

Figure 2 should have labels in a better font. The current font choice looks out of place. The caption also needs to define the acronyms RCMAP and UTM.

All figure fonts used throughout are Arial, which is an extremely widely used and easily readable font recommended for figures by the APA style guide:

https://apastyle.apa.org/style-grammar-guidelines/paper-format/font

We determined that the acronyms were unnecessary for the figure caption because the datasets are specified elsewhere, so they have been removed.

On page 7, lines 139 to 141, the description of the methods would be strengthened if the resampling technique used to aggregate the 30 m resolution data to the 90 m resolution model were explicitly stated. Was it averaging, weighted aggregation or majority rule? An explanation of how this mapping works in cases where there are differences in burned area at the original 30m resolution would help reproducibility.

The canopy cover data are reprojected using nearest-neighbor because the dataset is a combination of continuous (canopy cover %) and categorical (trees present: true or false), and nearest-neighbor reprojection is the appropriate technique for categorical data. This has been added to the text.

On page 7, lines 143 to 144, the authors should provide a quick justification for why October 1 was chosen to update the vegetation maps. A quick sentence saying it corresponds to the water year, or the availability of vegetation maps at the end of the fire season, or any other justification, should be stated. An October 1st update could introduce artificial bias, so it would be good to justify why this approach was followed.

We added the following justification:

The October 1st date is used for annual model updates because this date represents the start of a new water year, and Sierra Nevada watersheds are typically near their driest condition around this time of year, which limits the impact of model changes on simulated hydrological fluxes.

On page 7, lines 147 to 148, the authors should explain why estimation is preferable to using LAI observations from something like. Are the empirical estimates close to satellite observations?

Current satellite technology cannot observe LAI directly; rather, satellite observations (of reflected light) are converted into estimates of LAI through various models and empirical relationships (e.g., see Table 1 summary in Yan et al. 2018).

> Yan *et al*., "Generating Global Products of LAI and FPAR From SNPP-VIIRS Data: Theoretical Background and Implementation," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2119-2137, April 2018, doi: 10.1109/TGRS.2017.2775247

Lidar surveys show weaknesses in satellite-based optical LAI estimates, including saturation in dense forests and the inability to resolve 3-dimensional canopy structure that is important for LAI. Pre- and post-fire lidar surveys are not publicly available in the study area, so lacking high-resolution canopy structure data, we opt to estimate LAI heuristically through calibration.

> Winsemius, S., Babcock, C., Kane, V. R., Bormann, K. J., Safford, H. D., & Jin, Y. (2024). Improved aboveground biomass estimation and regional assessment with aerial lidar in California's subalpine forests. *Carbon Balance and Management*, *19*(1), 41. https://doi.org/10.1186/s13021-024-00286-w

> Zolkos, S. G., Goetz, S. J., & Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment*, *128*, 289–298. https://doi.org/10.1016/j.rse.2012.10.017

We have added the preceding justification to the methods.

On page 8, it is important to speak about calibrating correction factors for meteorological inputs, which has the potential to compensate for any deficiencies in the model itself. For a study trying to investigate how calibrated models predict process changes after a disturbance, calibrating bias-correction parameters for the forcing could introduce a lot of complexity (in other words, substantially enhance equifinality) that will complicate later analysis. A broad statement like "gridded meteorological data can have considerable uncertainty" is insufficient to justify this. Is there any concrete evidence that the specific meteorological data chosen are biased in this particular watershed? (Even) If so, the correct approach would be to bias-correct the forcing before calibration starts, so that every model in the ensemble uses the same inputs. This seems the only way to get a clean comparison between models later.

We are not primarily interested in "a clean comparison between models," but rather, we are interested in the effect of the fire on real-world annual streamflow. This effect is uncertain due to both the model and the forcing data. Failure to propagate meteorological uncertainty (by only using one set of inputs) would systematically overestimate our confidence in the combined model/weather system. It is helpful to think of this in a Bayesian context: the forcing data themselves are uncertain, but this uncertainty can be constrained by performing simultaneous inference over the model/data system.

We recognize that a fully Bayesian perspective is rare in distributed process-based hydrological modeling, so we have substantially expanded our explanation of this approach as follows:

While most of these parameters are widely recognized as suitable for calibration (Cuo et al. 2011, Du et al. 2014), precipitation and temperature biases are less frequently included in the calibration of distributed process-based models despite considerable uncertainty in gridded meteorological data. Among gridded meteorological datasets, there is mean relative difference of 21% for annual precipitation in our study watershed (Henn et al. 2018), and misestimation of large storms can lead to yearly biases of about 20% across the Sierra Nevada (Lundquist et al. 2015). Similarly, different meteorological datasets have mean air temperature differences as large as ±8 °C in the Sierra Nevada, and basin-average uncertainty is lower but still on the order of several °C (Schreiner-McGraw and Ajami 2022). Compensation between unknown errors in the meteorology data, model structure and calibration, and reconstructed streamflow can potentially contribute to spurious goodness-of-fit metrics with hidden physical deficiencies. Moreover, we expect that interactions between meteorological uncertainty and parameter equifinality may contribute to the overall uncertainty of disturbance simulations (Fig. 1), but this uncertainty would remain hidden if meteorological biases were assumed to be negligible. Because perfectly resolving the weather data with infinite precision is not feasible across a large, rugged mountain region, the robust approach is to propagate meteorological uncertainty into our final results (the post-fire hydrological change in this case), so that our conclusions include the quantified uncertainty caused by the model-data-calibration interaction. We view the combined meteorological data and hydrological model as a single inferential system, thereby acknowledging that the meteorological data themselves are based on various uncertain observations and empirical model assumptions (Abatzoglou 2013). In a Bayesian context, the goal of our calibration can be understood as sampling the probability of the streamflow and snow observations given a particular combination of model parameters and meteorological assumptions: P(streamflow + snow | model + meteorology). Because we lack a closed-form likelihood function for spatially distributed hydrological models like DHSVM, we estimate the unknown parameters of this whole weather-model system using an informal approximation based on traditional hydrological model calibration objective functions (Beven and Binley 1992).

Lundquist, J. D., Abel, M. R., Henn, B., Gutmann, E. D., Livneh, B., Dozier, J., & Neiman, P. (2015). High-Elevation Precipitation Patterns: Using Snow Measurements to Assess Daily Gridded Datasets across the Sierra Nevada, California. *Journal of Hydrometeorology*, *16*(4), 1773–1792. https://doi.org/10.1175/JHM-D-15-0019.1

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Schreiner-McGraw, A. P., & Ajami, H. (2022). Combined impacts of uncertainty in precipitation and air temperature on simulated mountain system recharge from an integrated hydrologic model. *Hydrology and Earth System Sciences*, *26*(4), 1145–1164. https://doi.org/10.5194/hess-26-1145-2022

The correct approach would be to bias-correct the forcing before calibration starts, so that every model in the ensemble uses the same inputs.

We disagree, and this is actually one of the fundamental contributions of our study (Figure 1).

It is not possible to bias-correct the forcing in the absence of ground-truth spatially representative meteorology data, which do not exist. Moreover, this proposed "bias-correction" seems to just be a type of calibration by a different name, which should be included in the formal quantification and propagation of calibration uncertainty. Otherwise, we would end up with a single, infinitely precise estimate of meteorological biases, which is not realistic. Our conceptual model (Figure 1) illustrates why meteorological uncertainty needs to be propagated through calibration because different forcing assumptions lead to different pre- and post-disturbance water balance configurations, and it is impossible to determine the "true" spatially distributed mountain weather with infinite precision.

We have clarified this in the introduction as follows:

As illustrated in Fig. 1, meteorological uncertainty (e.g., a precipitation bias) can interact with uncertain model parameterizations, contributing to uncertainty in the streamflow response to disturbance. Basin-scale meteorology data are highly uncertain in mountain regions (e.g., Lundquist et al. 2015, Henn et al. 2018, Schreiner-McGraw and Ajami 2022), and whatever assumptions we make about the meteorology may cause the model to compensate for inaccuracies with other calibrated parameters (Elsner et al. 2014). For example, assuming a larger precipitation bias correction may cause the model to simulate a larger ET_{Tree} component and a corresponding large post-fire change (Model 3 in Fig. 1), whereas a smaller precipitation bias correction could limit the predicted post-fire streamflow change since the pre-fire P-Q residual is smaller (Models 1-2 in Fig. 1). From a Bayesian perspective, we can treat the data and model as a combined inferential system, which enables us to constrain uncertainty in the interactions between uncertain meteorology and uncertain hydrology by generating suitable ensemble samples (Kavetski et al. 2003).

Kavetski, D., Franks, S. W., & Kuczera, G. (2003). Confronting input uncertainty in environmental modelling. In Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, & R. Turcotte (Eds.), *Water Science and Application* (Vol. 6, pp. 49–68). American Geophysical Union. https://doi.org/10.1029/WS006p0049

Elsner, M. M., Gangopadhyay, S., Pruitt, T., Brekke, L. D., Mizukami, N., & Clark, M. P. (2014). How Does the Choice of Distributed Meteorological Data Affect Hydrologic Model Calibration and Streamflow Simulations? *Journal of Hydrometeorology*, *15*(4), 1384–1403. https://doi.org/10.1175/JHM-D-13-083.1

On page 8, lines 194 to 195, some extra explanation about why these three objectives are based on reconstructed daily streamflow is needed and how it works would be good to add. A reference to a more in-depth explanation would be enough. How does the reconstruction approach deal with the disturbance if this is reconstructed? Is that accounted for somehow?

We have clarified what is meant by "reconstructed" streamflow. This is not a model output, but rather a combination of observations, which are added and subtracted to "undo" the effect of upstream artificial storage and diversion. We have added additional information to the text about the streamflow dataset, including a citation to the California DWR documentation as follows:

Streamflow is estimated at the outlet of Millerton Lake (Fig. 2) by reconstructing observations to remove the effects of artificial flow regulation (California Department of Water Resources 2024). Millerton Lake unimpaired outflows are estimated assuming sub-daily surface routing times by summing the daily change in storage, canal and dam releases, surface evaporation, and storage changes at eight smaller upstream reservoirs (Huang and Kadir 2016). Note that the reconstructed streamflow timeseries used in this study (called "full natural flow" by the California Data Exchange Center) is based on an explicit mass balance equation applied directly to the respective storage and flow measurements, not model output, unlike various other meanings of "natural flow" that are sometimes applied to California water datasets (Huang and Kadir 2016). The reconstructed streamflow timeseries (hereafter "observed streamflow") constrains the actual effects of the Creek Fire (and other disturbances) because the reconstruction procedure is directly based on measurements at specific diversion, storage, and outflow points, which are themselves responsive to the basin hydrological conditions. Three objective functions are based on a cleaned version of this daily streamflow timeseries (spurious negative values during low-flow periods and other missing values are imputed).

Huang, G., & Kadir, T. (2016). *Estimates of Natural and Unimpaired Flows for the Central Valley of California: Water Years 1922-2014* (pp. 1–256). Department of Water Resources, Bay-Delta Office. https://data.ca.gov/dataset/estimates-of-natural-and-unimpaired-flows-for-the-central-valley-of-california-wy-1922-2014

On page 9, the caption for Table 2 states that NSE is identical to R^2 for statistical models. It would be better if it stated that NSE is analogous to R2 and not identical, and if a citation from the original Nash and Sutcliffe paper from 1970 were provided.

We have changed this wording to avoid unnecessary confusion. However, the calculations are indeed numerically identical:

$$R^2 \equiv \left(1 - \frac{SS_{res}}{SS_{tot}}\right) = \left(1 - \frac{\sum(meas - obs)^2}{\sum\left(meas - meas_{avg}\right)^2}\right)$$

$$NSE \equiv \left(1 - \frac{\sum(meas - obs)^2}{\sum\left(meas - meas_{avg}\right)^2}\right)$$

On page 10, line 217, 600 samples for a 14-dimensional space does not seem a lot, and a large number of these might already be Pareto-optimal just because of the low number of samples. A table that shows how many of each sample fulfills each individual criterion would be good.

This is actually one of the key advances of our underlying methodology, though we agree it needs to be better highlighted in the manuscript. Specifically, our calibration approach uses parallel expected hypervolume calculations obtained from surrogate machine learning models to search for Pareto-efficient parameter sets extremely efficiently. In such a context, a few hundred samples is actually quite a lot, as these techniques are intended to work with even just tens of parameter samples. A good overview of this methodology is provided by one of the included citations for the expected hypervolume indicator:

> Binois, M., & Picheny, V. (2019). **GPareto**: An *R* Package for Gaussian-Process-Based Multi-Objective Optimization and Analysis. *Journal of Statistical Software*, *89*(8). https://doi.org/10.18637/jss.v089.i08

We have also clarified in the text as follows:

While this number of tested parameter sets may seem small by conventional standards considering the 14-dimensional search space, we note that each new parameter sample is selected after an independent optimization procedure using 100 to 1,000 particle swarm samples from the objective function surrogate models. Thus, our overall calibration explores the objective function tradeoffs across more than 160,000 parameter sets, though only 600 of these are actually tested in DHSVM. Because testing hundreds of thousands of parameter sets directly in DHSVM would require prohibitive amounts of computational expense, the Bayesian surrogate optimization procedure is essential for efficiently selecting parameter sets that have the best likelihood of substantially improving the Pareto frontier.

On page 10, line 234, given that streamflow is reconstructed, the authors should use the term “reconstructed streamflow” instead of calling this "observed" to help the reader understand that we're comparing results from one model to another.

Please see prior comment—the “reconstructed streamflow” is indeed calculated directly from observations. It is only reconstructed in the sense of removing upstream diversion/storage effects using a simple water balance equation, and it is not a model output.

On page 11, line 242, the authors should state the importance of the bias shift metric. The co-reviewers came to differing conclusions about its importance. Is it a metric to inform people about models, or is the bias shift something that needs to be corrected, and how is the correction applied?

The bias shift metric is useful in both contexts. We clarified this immediately at the start of Section 2.4, where we introduce the metric:

The bias shift metric is useful in two contexts. First, it is useful for understanding and refining the behavior of models, potentially including reducing equifinality by preferring models with near-stationary bias. Second, it is useful for correcting model predictions to estimate what a hypothetical model with stationary error would have predicted.

The reasoning on page 11, lines 249 to 250, is a bit difficult to follow. Are there any tests that can be done to see if this assumption is valid, or can the authors speculate how the results would change if it didn't hold?

We have clarified that this reasoning is supported by the scatterplots in Fig. 5:

The linear relationship between bias shift and yearly ΔQ_{Fire} is supported by graphical analysis of bivariate scatterplots, as illustrated subsequently in Fig. 5. In other watersheds or disturbance scenarios, it might be necessary to posit a nonlinear relationship with the bias shift, which could again be detected from analogous bivariate scatterplots.

Figure 3 shows some variability in the simulations, particularly around the middle flow magnitudes. Can it be clarified why these simulations are called "satisfactory"? Presumably, there's an implicit middle step that says "these NSE scores are [something], therefore these parameter sets all satisfactorily reproduce observed streamflow". I encourage the authors to add their reasoning for thinking that these simulations are good enough for this study. Are these simulations of typical quality for this particular watershed (i.e., comparable to other modelling efforts)? Are these ranges of scores particularly high for this specific watershed?

We have reworded this sentence as follows:

All behavioral parameter sets also achieve NSE of 0.80-0.87 and log-scale NSE of 0.76-0.84 considering just the four years after the Creek Fire, which is considered satisfactory because the model skill is similar on pre- and post-fire periods

We have also added a comparison to a similar study in a different Sierra Nevada basin:

Additionally, the post-fire daily NSE of at least 0.80 achieved by all behavioral DHSVM parameter sets is substantially higher than the post-fire daily NSE of -0.13 to 0.60 achieved by a different distributed hydrological model (with dynamic vegetation and other fire-aware updates) after a megafire in other Sierra Nevada sub-watersheds (Abolafia-Rosenzweig et al. 2024).

Figure 4 needs more explanation. For example, for both x-axes, are we looking at different parts in space? In other words, does this figure show how different parts of the watershed respond? The reviewers are asking because the caption says that the values on both x-axes are derived from data and not calibrated, but if this is so, the reviewers are unsure how this figure shows parameter uncertainty. By counting we can summarize that each symbol stands for a behavioural parameter set, but this should be stated somewhere.

Thank you for bringing this to our attention—it was indeed ambiguous as written. We have clarified the caption to indicate that each point represents a behavioral parameter set, with all values spatially averaged within the watershed.

The data and code availability statement on page 22 is incomplete. See the guidelines at https://www.hydrology-and-earth-system sciences.net/submission.html. The co-reviewers would like to take a look at the streamflow series for this basin but cannot easily find the location of this data either in this section or through the California DWR reference mentioned in the text.

Our understanding from the options presented during the HESS submission process was that final datasets could be archived after acceptance, thus incorporating any changes suggested during review. At this stage, we are comfortable that our methods and results are largely final, so we have created the archive and it has been added to the paper.

The original data, as well as the cleaned dataset with imputed values that is used in our study, can be downloaded from the new Zenodo archive:

https://doi.org/10.5281/zenodo.16972670

Additionally, the raw streamflow data (sensor number 8) can be downloaded here:

https://cdec.water.ca.gov/dynamicapp/staMeta?station_id=SBF

In the references, line 586, this link does not go to the dataset but to the landing page. A DOI, an accurate link, and the access date are needed.

This has been updated to the station metadata page above. However, a DOI is not available for this dataset, which is hosted on a California state webpage.