Replies to Reviewers' comments

In this document, we provide detailed answers to the comments raised by the Reviewers on our manuscript "On the Value of Satellite Remote Sensing to Reduce Uncertainties of Regional Simulations of the Colorado River" by Mu Xiao et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-204-RC1, 2022.

Reviewers' comments that have been numbered and reported in bold. In our answers, when referring to the revised manuscript, we reported the lines of the manuscript version with track changes unless otherwise specified.

Anonymous Reviewer #1

Reviewer 1 provides the following general comment:

I enjoyed reading the manuscript. This study combines several remote sensing products to improve the hydrologic model's physics together with streamflow performance. I only have several concerns regarding the presentation of the work and the framework.

Thanks for working with us to improve this paper. We appreciate the constructive comments and we have provided detailed responses below.

1. **L159: USBR dataset needs a reference (url/doi)**

The original U.S. Bureau of Reclamation (USBR) webpage that provides the flow data records: https://www.usbr.gov/lc/region/g4000/NaturalFlow/documentation.html.

This link is now reported in the Open Research section of the paper. The naturalized flow data we obtained from USBR has been also uploaded onto the same Zenodo online archive where we stored the model parameters. This has been also mentioned in the Open Research section in the updated paper (lines 559-566 of the revised version).

2. **L239: why monthly and not daily streamflow performance was targeted in calibration? daily water balance is key for hydrologic models. Monthly fit is easier and reducing the value of baseline simulation.**

This is a very good point that was mentioned by both Reviewers. The river is heavily regulated and the highest resolution of the naturalized flow records from USBR that is currently available is monthly. This is the main reason why we could not extend our calibration and validation against discharge at a daily scale. To address this comment, we added this sentence on lines 161-162:

"Note that this is the largest available resolution for the reconstructed naturalized flow since the river is highly regulated".

Given the lack of daily streamflow data, we might also argue that for the Colorado River Basin adding daily remotely sensed products to the model testing phase is even more critical to capture daily dynamics.

3. **L248-Fig3: in this section (3.3) I read what has been done but I couldn't find answers for the question "how". Framework needs elaboration. Baseline simulation is clear but other steps are not clear.**

We acknowledge that we have not provided an in-depth description of the "how" in each step mentioned section 3.3. The reason is that we considered section 3.3 as an overview of the calibration methods, while we preferred reporting the details of the Forcing-adj, Veg-adj, and Snow-adj steps in sections 4.3.1, 4.3.2, and 4.3.3 of the Results, respectively. To address the Reviewer's comment, we have:

(1) changed the name of Section 3.3 to "*Model improvements with remote sensing products: overview of the stepwise calibration strategy*";
(2) added the following sentence in section 3.3: "here, we provide an overview of the steps and describe the details of each step in the corresponding sections in the Results" (lines 256-257).

4. **Most importantly, model calibration is an exercise of fine tuning of the model parameters. Before calibration a robust sensitivity analysis (SA) must be applied for such sophisticated models with many parameters to reduce the search space. Did authors apply SA in their study?**

We did not apply a systematic robust SA, but we adopted a hybrid approach based on the physics and sensitivity of single parameters. Specifically, we first carefully considered the physical equations implemented in the model to simulate the processes related to the observed variable (i.e., land surface temperature, LST, snow cover, SC). These equations are reported in the manuscript Appendix. We then identified the set of key parameters that are (1) involved in the equations, (2) spatially variable, and (3) not derived from any type of direct or indirect observation. We then computed the spatial correlation between these parameters and the pattern of the errors between simulated and remotely sensed LST. The outcomes of these analyses are reported in Figure 7 of the manuscript. For the parameters with the largest correlation, we performed a sensitivity analysis to verify that that parameter importantly affects the simulation of LST. We ultimately focused on the subset of parameters that exhibited the largest sensitivity.

5. **The authors followed a stepwise approach but sensitive analysis (sobol's, LHS O-A-T, Morris etc) may reveal parameter interactions which can be important to consider during calibration. The authors should discuss the implications of parameter interaction in their framework.**

We agree with Reviewer 1 that a sensitivity analysis targeting multiple parameters at the same time can ultimately lead to improved performance. However, this type of analysis would be very computationally expensive for a model like VIC and the size of the Colorado River Basin and would most likely require an entirely separate study. Here, our focus is instead to highlight the

importance of accounting for spatially variable observations from remote sensors in the calibration process. As discussed in the answer to the previous comment #4, we focused the calibration on a set of parameters identified based on model physics and the impact of these parameters on the spatial variability of the errors between simulated and remotely sensed LST and SC. The calibration was then performed by focusing on one parameter or forcing variable at a time within each step.

To address this comment, in the Summary and Conclusions section of the revised manuscript version, we have added a number of future research avenues that we believe will help improve the fidelity of hydrologic models, including the need to perform a systematic robust sensitivity analysis as suggested by Reviewer 1 (lines 547-550):

"First, once the key parameters involved in the physical equations simulating a variable observed by satellite sensors have been identified as done here, a robust multiparameter sensitivity analysis could be conducted to investigate possible interactions among the parameters; this effort will help further refine the calibration".

6. **It would be good to simultaneously use LST and snow RS data on uncertainty reduction via model calibration.**

We agree with the Reviewer that this would be an interesting idea. However, it would require the use of an automatic calibration routine with an optimization function that accounts for both daily LST and monthly SC, which would require significant computational power and would be out of the scope of this study, as for the case of the multiparameter sensitivity analyses (see answer to comment #5). To our knowledge, studies that investigated the utility of remotely sensed products in large basins like the Colorado River are still very few; therefore, it is still necessary to separately gain insights into the utility of each remotely sensed dataset, along with the associated parameters and equations. Once this knowledge based on single variables is built, calibration strategies that target multiple variables could be better designed in future work.

To account for this useful suggestion, in the Summary and Conclusions section of the revised manuscript version, we have added the sentence (lines 550-552):

"Second, automatic calibration strategies could be designed and applied to simultaneously target the simulation of multiple variables (here, LST and SCF)".

7. **My biggest concern is about the spatial structure of the selected hydrologic model (VIC) which is a semi-distributed model. In such model parameters get the same value in the same subbasins which inevitably leads to uniform parameter fields and resultant uniform flux maps. One way to avoid this, is using fully distributed models with parameter regionalization tool based on pedo-transfer functions using soil and vegetation properties.**

We appreciate the Reviewer's comment and point out three issues that, hopefully, address this concern.

First, VIC is a macroscale hydrologic model with a gridded domain where most of the parameters and all outputs do vary spatially. For example, all parameters identified in Figure 7 vary in space. See also the maps in Figs. 1d and 1e that report the vegetation fraction, f_v, and soil depth that are used in the baseline simulation.

Second, the parameters of the baseline simulations are mainly based on the gridded products derived by Bohn and Vivoni (2019), who utilized high-resolution (from 500 m to 1 km) remote sensing datasets to generate several model parameters in the same grid at 1/16˚(~6-km) resolution used in our manuscript. This is mentioned in Section 3.2 of the manuscript.

Finally, we have more than 15,000 pixels for the Colorado River Basin which allow the spatial variability to be appropriately captured.

8. **The authors used bias-sensitive error metrics (rmse, bias) and CC as bias insensitive metric. CC must be used with cautious it can be affected by outliers in the sample. High CC values are not always informing. Instead spatial metrics (SSIM, FSS etc) could be preferred.**

We would like to clarify one important issue. As mentioned in lines 255-258 of the original manuscript: "The first two steps [*of the calibration*] were guided by metrics quantifying the agreement between simulated and remotely sensed LST, including the correlation coefficient (CC), root mean squared error (RMSE), and Bias (mean LSTV - mean LSTM) between: (1) time series of daily LST_V and LST_M at each grid cell, and (2) daily spatial maps".

In particular, the maps and metrics shown in Figures 5, 6, 7, and 9 that drove the calibration effort are based on RMSE, CC, and Bias between simulated and observed time series at each pixel. Therefore, for these cases, it is not possible to compute spatial metrics like SSIM, FSS, etc. The only case where we computed metrics between maps (also called "fields" in the paper) is Figure 8, which we used as additional measures of calibration accuracy.

Regarding the role of CC, we fully agree with the Reviewer. When we carried out our analyses, we noted that the CC alone cannot be a good indicator for model evaluation because (1) it is not robust, as pointed out by the Reviewer, and (2) it is always high (>0.8) even in the baseline (Lines 319-320). Because of this, our model adjustments (Forcing-adj, Veg-Adj, and Snow-adj) are mainly based on either RMSE or Bias of the time series.

To address the concerns of both Reviewers, in the revised manuscript, we have:

1) Reported in Figs. S10 and S11 of the Supporting Information the maps of simulated and observed long-term climatology of monthly SCF in the snow season and LST, respectively, over 2003-2018.
2) Reported in Table S2 the values of Structural Similarity Index Measure (SSIM) and Spatial Efficiency (SPAEF) between these long-term maps. These metrics were introduced in lines 266-270 of the revised Section 3.3.

These changes are reported in lines 452-457.

9. Fig6: the readers can be curious why median night time bias for baseline is usually less than other 3 cases.

As shown in Figure 6, the median bias for nighttime LST in the baseline case is negative. This result was ascribed to the negative bias of the air temperature forcings from the Livneh dataset, which we removed using the PRISM long-term normal products in the "Forcing-adj" calibration step (see Section 4.3). The resulting median bias for nighttime LST becomes close to 0 or slightly positive in the "Forcing-adj" step, thus improving the simulation (see also Figures 8 and 9). The bias does not change significantly in the other steps because they do not involve changes in parameters that affect nighttime LST. These details are described in Section 4.3.1, where we concluded that "…the Forcing-adj simulations improved Bias, which was reduced in most subbasins" (see Lines 370-371 of the original manuscript version).

10. Fig7 should be better explained. How correlation between parameters is assessed?

The correlation coefficients in Figure 7 are the Pearson correlation coefficient between two spatial fields in each subbasin. The first spatial field is either T_{air} or any of the parameters shown in the rows of the heatmaps (e.g., Elevation, Porosity, …, LAI, and f_v), while the second spatial field is either RMSE (heatmaps on left) or the Bias (heatmaps on right) between time series of LST_V and LST_M at each domain pixel. The columns in each heatmap report the correlation coefficient in each subbasin. For example, the first pixel on the top left is showing the correlation coefficient between T_{air} and Daytime RMSE in the Green subbasin.

In the revised version of the paper, we have updated the caption of Figure 7 and added a clarification on the use of time series for the errors in line 330 of the revised version.

Anonymous Reviewer #2

First of all, we thank Reviewer 2 for their time and useful comments. Reviewer 2 provides a general comment:

The authors present a well-written and well-motivated study on the value of remote sensing data to improve hydrological simulations. I enjoyed reading the manuscript and only have a few comments that will require some attention by the authors.

Thanks for the nice overview. Below, we provide point-to-point responses to the specific comments.

1. **The authors chose to evaluate the model against monthly runoff. In my point of view this cannot be justified since the remote sensing data are used at daily scale to evaluate the model. In order to achieve a balanced evaluation of the model runoff should also be evaluated at daily scale.**

This is a very good point that was mentioned by both Reviewers. The river is heavily regulated and the highest resolution of the naturalized flow records from the US Bureau of Reclamation (USBR) that is currently available is monthly. This is the main reason why we could not extend our calibration and validation against discharge at a daily scale. We also note that other modeling studies of the Colorado River Basin are based on the same monthly dataset. To address this comment, we added this sentence on lines 161-162:

"Note that this is the largest available resolution for the reconstructed naturalized flow since the river is highly regulated".

Given the lack of daily streamflow data, we might also argue that, for the Colorado River Basin, adding daily remotely sensed products to the model testing phase is even more critical to capture daily dynamics.

2. **More details are required with respect to the "adjustment of the VIC parameters" as presented in section 3.2. Did the authors conduct a manual calibration or were the parameter values estimated via automatic calibration? Please specify.**

It is a manual calibration. For the baseline simulation of section 3.2, we manually changed the soil parameters to improve streamflow performance in the "traditional" way. In the revised version of the manuscript, we have:

- added "manually" on line 241 of the new manuscript version in Section 3.2;
- added the sentence "All modifications of the model parameters were performed via manual tuning" at the end of Section 3.3.

3. **One of my main concerns relates to how the fit between observed and simulated spatial patterns was assessed. The authors chose to do a grid-wise evaluation of the simulation. I think this makes sense for the forcing adjustment where the temporal dynamics of**

simulated LST are strongly linked to the forcing data (air Temp). However, when it comes to model parameters, I would suggest to evaluate the model against spatial pattern that are aggregated over time, for example a long-term average annual (or summer) LST map. Evaluating a model at daily scale will always be very much affected by the quality of the forcing data and the model parameters have a limited affect here. Nevertheless, the imprint of the model parameters emerges when aggregating the simulation results over time and quantifying the spatial pattern match (e.g. with help of the SPAEF metric (https://doi.org/10.5194/gmd-11-1873-2018)) instead of the grid-to-grid comparison. Along these lines, the spatial patterns of RMSE and Bias presented in Figure 9 do not show a clear improvement of the model developments. Maybe a spatial pattern oriented evaluated of the long term average LST patterns is more insightful.

Thanks for pointing this out. Reviewer 1 had also a similar comment related to the need to use metrics that quantify the match of spatial patterns. However, there is a detail of our approach that, perhaps, was did not properly explain. As mentioned in lines 255-258 of the original manuscript: "The first two steps [*of the calibration*] were guided by metrics quantifying the agreement between simulated and remotely sensed LST, including the correlation coefficient (CC), root mean squared error (RMSE), and Bias (mean LSTV - mean LSTM) between: (1) time series of daily LSTV and LSTM at each grid cell, and (2) daily spatial maps".

In particular, the maps and metrics shown in Figures 5, 6, 7, and 9 that drove the calibration effort are based on RMSE, CC, and Bias between simulated and observed time series at each pixel. Therefore, for these cases, it is not possible to compute spatial metrics like SPAEF and SSIM (as suggested by Reviewer 1). The only case where we computed metrics between maps is Figure 8, which we used as additional measures of calibration accuracy.

Regarding the option to use long-term averages of LST, we decided not to do so because we wanted to assess the model's ability to capture dynamics at higher temporal resolutions, especially considering that we do not have streamflow data at a daily scale, as mentioned in the answer to comment #1.

To address the concerns of both Reviewers, in the revised manuscript, we have:

1) Reported in Figs. S10 and S11 of the Supporting Information the maps of simulated and observed long-term climatology of monthly SCF in the snow season and LST, respectively, over 2003-2018.
2) Reported in Table S2 the values of Structural Similarity Index Measure (SSIM) and Spatial Efficiency (SPAEF) between these long-term maps. These metrics were introduced in lines 266-270of the revised Section 3.3.

These changes are reported in lines 452-457.

4. The authors only present maps of the three selected metrics. For interested readers, observed and simulated maps of the actual variables will provide insightful information. The authors could select single days or long-term averages of LST (night

and day) and snow cover to illustrate the catchment characteristics and how the model represents those.

Thanks for the suggestion. As mentioned in the answer to the previous comment #3, in the revised version we have included maps of long-term averages of LST and SCF in Figures S10 and S11, respectively, in the Supporting Information.

5. **It would be interesting to see how the three steps of model development affect the water balance of the model. The authors only present the simulated runoff of the various simulations, but aggregated numbers of evapotranspiration, runoff, groundwater recharge, etc. would provide relevant alternative information.**

This is another good suggestion. To address it, we computed the climatological monthly mean of the water balance components for the Upper Basin, where most runoff is generated. Results are presented in Figure R1, which shows in panel (a) fluxes (P, ET, and RO; see caption for their definition) and changes in state variables (ΔSM and ΔSWE) for the Baseline simulations, and in panels (b)-(d) the difference between a given variable simulated in each calibration step and the variable from the Baseline simulation. The Forcing-adj and Veg-adj steps lead to small changes in ET and RO with a decrease of both fluxes in the summer months and an increase in the other months. The modification of these fluxes is due to a change in the storage components with (1) lower SWE (i.e., negative ΔSWE) and higher SM from November to February, and (2) higher SWE and lower SM from March to July. The Snow-adj step modifies the seasonality of SWE compared to Baseline by increasing this storage component in February and March and reducing it in April and May. This, in turn, leads to an opposite behavior for SM, which is ultimately translated in a positive (negative) change of RO in May and June (July and August). In all cases, the changes in runoff occurred in a similar way for both the surface and underground components.

These considerations have been added at the end of Section 4.4 of the revised manuscript that is now called: "*4.4. Impacts on VIC streamflow performance and water balance".* We have also added Figure R1 that is the new Figure 13.

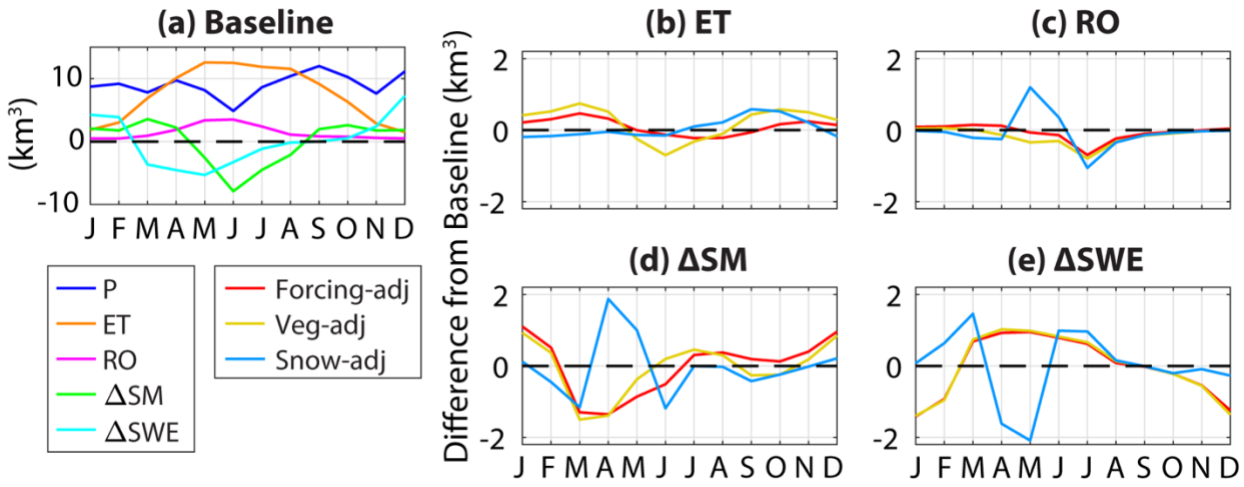

Figure R1. (a) Climatological monthly mean of the water balance components for the Baseline simulations in the Upper Basin. P is precipitation, ET is evapotranspiration and sublimation, RO is surface and underground runoff, and ΔSM (ΔSWE) is the differences between soil moisture (snow water equivalent) at the end and beginning of the month. (b)-(e) Difference between each variable for the Forcing-adj, Veg-adj, and Snow-adj simulations and the Baseline simulations.