We thank the reviewer for his/her interest in our work and for the thoughtful and detailed feedback provided.

Comment:

The manuscript is very well written and nicely illustrated. It deals with an interesting topic of model structure and gaining information from satellite data in a data sparse basin.

Reply:

We highly appreciate this positive assessment of our manuscript. We will in the following address all specific comments in detail.

Comment:

It makes the paper less interesting that there are essentially no major changes to the simulated spatial patterns of ET across Model A-F (e.g. Figure 11) and that the general simulated spatial pattern does not resemble the observed in any way. It seems that you are not addressing the most important issues in your set of alternative models (B-F). The general interest of the manuscript would increase greatly if some of your hypothesis would at least produce a different pattern from Model A.

The simulated spatial pattern will be a reflection of both model structure and parametrizations scheme, in my experience mainly the latter. Therefore, I strongly suggest that you add a set of model setups that reflect different spatial parameterization schemes. In your discussion you address this limitation nicely, but I also feel that the manuscript would benefit greatly from an additional analysis illustrating the importance of model parameterization and parameter distribution on the simulated spatial patterns. Basically, even the most sophisticated model structure cannot be expected to reproduce a correct spatial pattern without a sound, flexible and spatially explicit parametrizations scheme.

I think you can logically add such an analysis to your manuscript in line with the idea of learning from satellite observations, by letting the observed spatial patterns guide your spatial parametrization approach. Such a parametrization scheme could also include transfer functions or simple spatial relations to known variables such as elevation, slope, soiltype, LAI etc.

Reply:

We completely agree with the observation of the Reviewer that there are indeed only limited improvements with respect to the spatial pattern of the evaporation when looking at Figure 11 in the manuscript. However, this figure includes only a selection of models with the best results applying the second calibration strategy using multiple variables. Additional figures were included in the Supplementary Material showing the spatial pattern for all Models A – F for both calibration strategies, hence using discharge (Figure S6 in the Supplementary Material) or multiple variables (Figure S10 in the Supplementary Material). With these graphs the effect of different model structures and calibration strategies on the spatial variability of the evaporation was illustrated.

On the one hand, when calibrating with respect to discharge only, the spatial pattern of the evaporation changed depending on the model structure, but remained poorly reproduced for all Models B – F compared to the benchmark Model A (Figure S6 in the Supplementary Material). On the other hand, when calibrating with respect to multiple variables, the effect of the model structure was less significant (Figure S10 in the Supplementary Material). Only Model D showed clear differences compared to Model A, for example the dry season average evaporation was significantly overestimated in the wetland areas along the river in contrast to the observation (highlighted in Figure 1 here). In other words, the results in this study illustrate the spatial pattern of the evaporation did change with when changing the model structure or applying different calibration strategies.

However, the improvements remained very limited compared to the benchmark and the modelled spatial pattern remained poorly reproduced. We of course also completely agree with the reviewer, that at least some of the remaining problems are likely to be related to the actual distribution of parameters. As recommended by the Referee and mentioned in the discussion of our manuscript, this could, among others, be further improved by applying spatially distributed parameter sets. However, this will increase the number of calibration parameters considerably and hence also the degree of freedom such that many different combinations of parameters result in similar model performances, but do not necessarily reproduce all hydrological processes well. Therefore, it is important to have sufficient data to support spatially distributed parameters to avoid this problem of equifinality and improve the model realism. To limit the problems related to equifinalty, indeed a transfer function approach with global parameters, such as the MPR scheme developed for the mhm model (Samaniego et al., 2010;Kumar et al., 2013), could prove highly valuable. However, the design and choice of suitable and meaningful transfer functions in itself would require substantial additional analysis to assess the information content of different variables to support spatial parameter distribution (for example NDVI, LAI, topography, soil type, vegetation type or climate) and to test different distribution methods, which would warrant probably several standalone research papers. Therefore, as a first test, we analysed the effect of spatially distributing one calibration parameter related to the evaporation, the maximum interception storage ($I_{\max}$), using a linear transfer function with LAI data similar to previous studies (Samaniego et al., 2010) and using Model F as basis. While this influenced the spatial pattern of the evaporation, it did not improve significantly.

We will add an in-depth discussion on the risks and potentials of parameter distribution strategies in the revised manuscript. However, this paper focused on the added value of satellite-based evaporation and total water storage observations for model structure development and parameter selection. Therefore, additional analysis on parameter distribution strategies was considered outside the scope of this study.

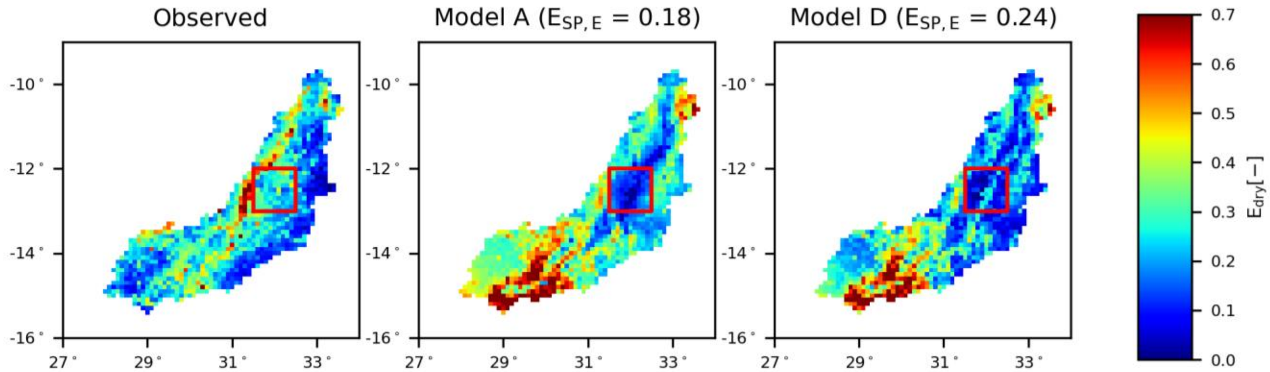

Figure 1: Observed and modelled spatial variability of the normalised total evaporation averaged over all days within the dry season for Models A and D using the "optimal" parameter set with respect to multiple variables ($D_{E,ESQcal}$).

Comment:

In sections 3.1.2 and 3.1.3 it is unclear why the different structural changes were applied. The title suggests that you are learning from satellite data, but it is not clear to me how you learn and how you used the satellite data to make new hypothesis about model structure. It is mentioned several times that you diagnose model deficiencies, however it is unclear to me how this is done. I believe this should be elaborated in a revised manuscript.

Reply:

In this study, the model structure was adapted iteratively based on the results of the benchmark Model A or subsequently developed models. Therefore, the paper was structured such that first the benchmark Model A was explained (Section 3.1.1), followed by a brief description of the model adaptations (Section 3.1.2 and 3.1.3). Based on the deficiencies of the benchmark Model A diagnosed and highlighted in Section 4.1.3, the first set of model adaptations were developed (Models B – D) as explained in Section 4.2. Similarly, based on the deficiencies diagnosed in Models B – D as explained in Section 4.2.3, the second set of model adaptations were developed (Models E – F) as explained in Section 4.3. Therefore, as the model adaptations depended on the model results, they were explained only briefly in Section 3 and more detailed in Section 4.

However, we will update the manuscript to more specifically and explicitly emphasize the role of satellite observations when diagnosing deficiencies and changing the model structure. The satellite-based evaporation and total water storage observations were used for model evaluation with respect to their spatial and temporal variability to detect model deficiencies in these system-internal variables. In addition, satellite-based evaporation data was used to evaluate whether temporal variations in the evaporation from a specific hydrological response unit, in this case wetland dominated areas, were reproduced well. In all cases, the variables were normalised to focus on dynamic fluctuations and spatial pattern rather than absolute values to avoid incorporating bias uncertainties in the satellite data.

Comment:

An issue with the use of the SPAEF metric for the water storage anomaly might be, that the histogram component of the metric, might not be so meaningful when applied to the coarse spatial resolution of 1 deg., with very few grids. You could look into this by examining the three components of the metric separately. I do not suggest to put this analysis in the paper, but it might be mentioned in a discussion.

Reply:

Thank you for pointing this out. This could indeed provide some interesting insights. Upon closer inspection of $E_{SP,E}$ and $E_{SP,S}$ for the "optimal" parameter sets for Models A — F according to the first calibration strategy using discharge only, we discovered different ranges for the individual components as indicated in Table 1 here. According to these numbers, differences in E_{SP} were mainly a result of differences in β (coefficient of variation), whereas the component with the smallest difference was α (Pearson correlation coefficient). The range in γ (fraction of histogram intersection) is indeed smaller for the total water storage where the grid size is larger compared to the evaporation. For future studies, it would be interesting to examine the different components more detailed to assess the overall information content of this model performance metric E_{SP} to identify feasible parameter sets across different spatial scales. We will elaborate on this in detail in the discussion.

Table 1: Overview of model performance ranges with respect to the spatial pattern (E_{SP}) of evaporation and total water storage including the corresponding individual components (α: Pearson correlation coefficient, β: coefficient of variation and γ: fraction of histogram intersection) using the "optimal" parameter sets according to the first calibration strategy using discharge only for Models A – F.

	Evaporation	**Total water storage**
α	0.12 – 0.23	0.43 – 0.54
β	0.55 – 1.23	0.62 – 1.16
γ	0.43 – 0.81	0.07 – 0.23
E_{SP}	-0.04 – 0.17	-0.17 – 0.08

Comment:

Did you perform any sensitivity analysis to explore which model parameters, structures or compartments were most important for simulating spatial patterns and temporal dynamics?

Reply:

While we agree that an analysis of the respective sensitivities of individual aspects in the modelling processes could indeed provide additional insights into factors influencing the spatial and temporal variability we did not explicitly perform such a detailed analysis as this would have further inflated an already long, detailed and complex manuscript. However, the model results did shed light into the importance of several different aspects. For instance, when

considering the model parameter sensitivity, of the generated parameter sets, the best 5% were selected with respect to discharge or multiple variables. Other combinations of variables to identify feasible parameter sets, for example discharge and evaporation or only evaporation, were also tested but excluded from the manuscript as they did not add further value and to keep the story concise. Regardless of the calibration strategy, the modelled spatial pattern of the evaporation and total water storage remained significantly different from the observation when using the benchmark model. Also, the evaporation from wetland areas was reproduced poorly regardless of which variables were combined with discharge in the calibration procedure. This indicated these deficiencies were more likely a result of uncertainties in the model structure, parameterization or data rather than of the selected parameter sets. That is why the model structure was adjusted stepwise. While the spatial pattern mainly improved when incorporating multiple variables in the calibration procedure (compare Figures S6 and S10 in the Supplementary Material), the evaporation from wetland areas benefited the most from the changed model structure (Figure 10 in the manuscript).
A more systematic sensitivity analysis could provide valuable information on how to further improve the spatial and temporal variability of the system-internal variables, but this was considered outside the scope of this study. We will include this as recommendation in the revised manuscript.

Comment:
3.1.2 First model adaptation (Models B – D): Please describe what made you chose to make exactly these structural changes?

Reply:
The first set of model adaptations (Models B—D) depended on the results of the benchmark Model A. Therefore, the choice of adaptions was explained after having highlighted the deficiencies of Model A (Section 4.1.3) in Section 4.2. See also our reply on a previous comment on Sections 3.1.2 and 3.1.3. We acknowledge that this is not a conventional paper set-up, but we believe an iterative analysis warrants a partly iterative description of the steps.

Comment:
Line 522: How can you argue that you significantly improve the spatial pattern of ET? Your $E_{SP,ET}$ might increase slightly from 0.18 to 0.23, but looking at the maps in Figure 11, Model F has the same pattern as Model A and none of them resemble the observed pattern.

Reply:
In this paper, the effect of using 1) multiple variables for model calibration and 2) alternative model structures on the spatial-temporal variability of among others evaporation was assessed. With respect to the spatial pattern, the results were illustrated with respect to the model performance values ($E_{SP,E}$) and figures showing the spatial pattern. In the manuscript, only a

selection of these figures was shown (Figure 11), whereas in the Supplementary Material all remaining figures were included (Figures S6 and AS10).
In Line 522, we compared both calibration strategies for Model F. When calibrating Model F using discharge only, the spatial pattern of the evaporation was poorly reproduced (Figures 2b here and S6 in the Supplementary Material). This improved considerably when calibrating using multiple variables (Figures 2c here and 11 in the manuscript) as the evaporation was lower in the south-west and east of the basin similar to the observation. We will clarify this in the revised manuscript to avoid any confusion and also tone down the language to avoid confusion and misleading interpretation of our descriptions by the reader.
We absolutely agree with the Referee that the spatial pattern in the evaporation remain poorly reproduced. However, the goal of the statement mentioned by the Reviewer was to illustrate the added value of satellite observations to improve the representation of spatial *and* temporal variability of *multiple variables*. This paper showed that only limited improvements were observed in the spatial pattern with the chosen model structures and parameterization.

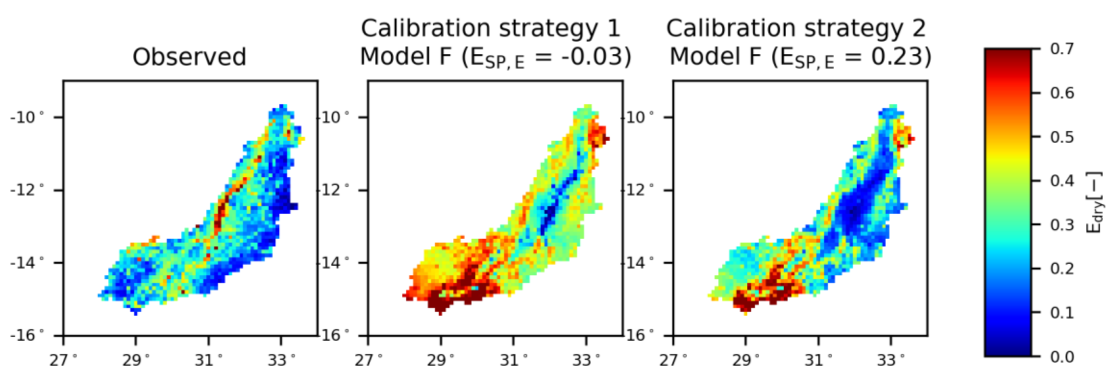

Figure 2: Spatial variability of the normalised total evaporation averaged over all days within the dry season according to WaPOR (observation) and Model F for both calibration strategies using discharge (Calibration strategy 1, $D_{E,Qcal}$) or multiple variables (Calibration strategy 2, $D_{E,ESQcal}$).

Comment:

Figure 11 and similar figures: I suggest that you condense the figures to make less white space and thereby allow the reader to make a better visual examination of the observed and simulated patterns. You can skip the lat long degree for instance, they can be added to figure 1 instead.

Reply:

Thank you for this feedback. We will condense these figures as much as possible to allow for a better visual comparison.

Comment:
Line 59: " to spatial pattern of" change to " to the spatial pattern of" or to " to spatial patterns of" Also in line 66 + 79 "spatial pattern and temporal dynamics" I suggest writing "spatial patterns"

Reply:
Thank you for pointing this out. We will correct this in the revised manuscript.

Comment:
Line 78: "for a large river systems" change to "for a large river system"

Reply:
Thank you for pointing this out. We will correct this in the revised manuscript.

Literature

Kumar, R., Samaniego, L., and Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resources Research, 49, 360-379, 10.1029/2012WR012195, 2013.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 10.1029/2008WR007327, 2010.