Dear Anonymous Referee #2,

Thank you very much for your feedback on our paper "Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in data scarce regions". To respond to your comments:

Comment 1: My main comment is that the general strategy followed for the model calibration lacks legibility. The 7 (?) calibration strategies tested are explained at various places in the manuscript (3.1, 3.3.1-4, again in 4.1.1-4) with a lot of redundancy and at the same time partial information here and there. We don't really understand how the strategies interact (are they all independent from each other). For example it is not clear in section 3 whether the altimetry and water level strategies were applied after the GRACE strategy or independently. Is there a reference strategy to which all other strategies are compared? We lack also information about the objectives behind the technical setup of each strategy (what are the assumptions tested, why)? I think that a synthetic table presenting the strategies and how they are linked to each other would be very informative.

Response: This is an excellent suggestion - we agree that including a table presenting the strategies and their links will be very helpful; therefore Table 1, as shown below, will be added. Each strategy was explained detailed in the methods section (section 3), but we will adjust these descriptions and directly link them to the new table. When explaining the results for each strategy, the individual strategies were briefly summarized in the results section to help the reader. However, this might not be necessary anymore when including the new table. We hope that with this table it becomes clearer how the different calibration strategies build on each other and interact with each other: the overall objective of this paper is to explore how well we can select parameter sets for hydrological models in catchments when *no* flow observations are available. The sequence of strategies is therefore meant to follow the potential thought process of a modeler in such an ungauged situation: first remove parameter sets that cannot reproduce the seasonal signal as indicated by GRACE. As this set of solutions still (at least in our case) contains many solutions that cannot reproduce river flow in a reasonable way, the set is subsequently further constrained by water level data from altimetry observations. This, in itself, has similarly little additional constraining power. Thus, water levels were converted to flow using different methods, including calibrated rating curves and the Strickler-Manning formula.

Comment 2: I have doubts on the interest of the "water level" strategies presented in the paper. They don't correspond to the title of the paper that mentions only GRACE and altimetry data. If I understood correctly, these strategies correspond to using the water level time series of the gauging station instead of the discharge data. Since the discharge data are available, what is the interest of these strategies? Is it just about reconstructing a rating curve using Google Earth cross – sections? Why not, but there is really no need to involve a hydrological model in that case. I think that the authors should question the interest of presenting these strategies in the paper, and if yes explain how they relate to the other strategies and what they bring for the use of satellite altimetry data.

Response: Thank you for this comment. The "water level" strategies indeed used water level time series at the gauge station. The objective of including this strategy was to illustrate the importance of incorporating more accurate cross-section information. At the locations where altimetry observations were available, cross-section information was extracted from high-resolution terrain maps available on Google Earth. This, unfortunately, has a low accuracy, leaving us with inaccurate

cross-section information at these locations. Unfortunately, accurate cross-section information from in-situ surveys was only available at the gauging station where, in turn, no altimetry observations are available. That is why water level time series were used to illustrate the importance of using more accurate cross-section information. We will clarify that in the manuscript.

Comment 3: About water level based calibration : as shown by the results (Altimetry strategies 1 and 2) and discussed by the authors (p 25, l. 620-625; p26 l. 649-653), calibration of models directly on water level data generates additional uncertainties associated to the level – discharge transformation. Have the authors considered separating the problems by 1/ tackling the altimetry water level – discharge transformation issue (without hydrological model) 2/ considering the model multi-station calibration on discharge. It would bring a clearer theoretical framework, by separating the uncertainty sources (see for example Renard et al., 2010). Moreover, there is already a rich literature corpus on each subject, to which the authors could relate. I think this could be worth a discussion. Renard, B.; Kavetski, D.; Kuczera, G.; Thyer, M. & Franks, S. W. (2010), 'Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors', Water Resources Research 46(5), W055521.

Response: It would indeed be interesting to separate the uncertainties related to the discharge – water level conversion from the hydrological model. For this conversion, there are several calibration parameters. So when disentangling this from the model, alternative information sources would be needed to estimate these parameters. Unfortunately, there was no further really useful information available at the virtual stations to estimate these parameters. This would be very interesting for a follow-up study doing exactly that in a more data rich region to look into these uncertainties more detailed. This aspect will be added to the discussion section

Comment 4: More information should be provided in the paper about GRACE, for the readers not familiar with satellite products. In particular, readers need to understand how the GRACE water storage anomalies (what is it exactly)? can be compared to total water storage in the model (not even speaking about calibration).

Response: Thank you for pointing this out. A section explaining GRACE more detailed for those not familiar with this product is indeed missing and will be added to Section 2.1.2.

Comment 5: Many performance indicators are used in the paper and not always explained / justified. The use of NSE on variables like water storage of flow duration curve seems a bit strange, as these variables behave very differently from discharge time series for which NSE is defined. Similarly, the general performance indicator for signatures combines NSE values and relative error values. Again, it is not clear to me how this indicator can be interpreted. What is the added value of using such complex indicators instead of more direct relative errors?

Response: In this paper we indeed used the Nash-Sutcliffe efficiency for the discharge time series, but also for other signatures such as the flow duration curve, and other variables such as total water storage or water level. Even though the Nash-Sutcliffe efficiency was originally defined for discharge time series as pointed out by Referee #2, it was assumed this performance metric can also provide valuable information for other signatures or variables. We agree additional study is required to confirm this assumption and to assess which performance metric would be most suitable, but this was beyond the scope of this study. We will include this issue in the discussion. Furthermore, we

wanted to incorporate multiple signatures of the discharge in the performance metric, instead of focusing on only part of the information available in the discharge time series. That is why these performance metrics for each signature were combined, in a similar way as in many earlier studies doing multi-objective calibration. However, we also agree that in the choice of objective functions there is always a strong subjective component and one error metric may be able to capture some aspects of the response better than another one. The reason we did not only show the individual performance indicators, but also decided to provide the combined ones is that we think that a good model should be able to reproduce all indicators simultaneously as well as possible. As there is quite some difference between the performance levels of different indicators, it is difficult to see the overall effect, when only analyzing their individual values.

Comment 6: In the model presentation it is not clear how the flow routing in the hydrographic network is computed – or is there any channel routing at all? This is quite important to know in the context of calibration with water level data (see also Comment 3).

Response: The flow routing scheme was indeed explained only briefly in the manuscript. For the flow routing, the mean flow length of each model gird cell to the outlet was derived based on the topography using a digital elevation map. In addition, it was assumed the flow velocity was constant in space and time; this velocity was calibrated. With this information on the flow path length and velocity, the accumulated flow in each grid cell was calculated at the end of each time step. This will be explained in more detail in the manuscript.

Minor comments: - A table of presenting the parameters (+ how many parameters and which ones were calibrated for each strategy) would be useful in the main text, instead of the detail of all model equations - Provide a table with a clear list of signatures + associated performance criteria – the reader is left to guess what goes with what when it comes to presentation and interpretation of results. - p 11 l 253: what are type II errors? - p 13 l345-350: the authors present a Distance as performance criterion like Eq 3, but there are only water levels in this strategy? Were signatures calculated here as well? - Table 4 is confusing. Why are the criteria different for each strategy in the "model efficiency" column?

Response: We will include a table to create an overview of the different calibration strategies and performance metrics; we hope that this will also make it clearer why there are different performance metrics for each calibration strategy in Table 4 in the manuscript. A table of the parameters was presented in the supplements. A type I error is the rejection of a true hypothesis (e.g. a good parameter set that was supposed to be accepted got rejected), while a type II error is the non-rejection of a false hypothesis (e.g. a bad parameter set that was supposed to be rejected got accepted). When calculating the model performance with respect to altimetry, the Euclidian distance was used to combine the model performance of each individual virtual station into one error metric; we agree though the reference to Eq. 3 is confusing.

Table 1: Overview of the calibration strategies applied in this study

Calibration strategy name	Calibration data	Objective function	Nr. of calibration parameters	Comments	Discharge – water level conversion method	Benefits (+) & limitations (-)
Discharge (reference)	Discharge (at basin outlet)	D_E	17	Traditional model calibration on observed flow data Combination of 8 different flow signatures	-	-
Seasonal water storage	GRACE	$E_{NS,Stot}$	17	No discharge data used	-	-
Altimetry Strategy 1	Altimetry (at 18 virtual stations) & GRACE	Altimetry: $D_{E,R,WL}$ GRACE: $E_{NS,Stot}$	17	No discharge data used Combination of 18 virtual stations Combined with GRACE	-	+ No extra parameters or data needed + Assumption: monotonic relation between discharge and river water level - Focus on dynamics only, not volume
Altimetry Strategy 2	Altimetry (at 18 virtual stations) & GRACE	Altimetry: $D_{E,NS,RC}$ GRACE: $E_{NS,Stot}$	25	No discharge data used Combination of 18 virtual stations Combined with GRACE	Calibrated Rating curve	+ No extra data needed - Two extra parameters per cross-section
Altimetry Strategy 3	Altimetry (at 18 virtual stations) & GRACE	Altimetry: $D_{E,NS,SM}$ GRACE: $E_{NS,Stot}$	18	No discharge data used Combination of 18 virtual stations Combined with GRACE	Strickler-Manning	+ Only 1 extra parameter - Cross-section data needed - Assumption: constant roughness in space and time
Water level Strategy 1	Water level (at basin outlet) & GRACE	Altimetry: $E_{NS,SM,GE}$ GRACE: $E_{NS,Stot}$	18	No discharge data used Combined with GRACE	Strickler-Manning	+ Only 1 extra parameter - Cross-section data needed - Assumption: constant roughness in space and time
Water level Strategy 2	Water level (at basin outlet) & GRACE	Altimetry: $E_{NS,SM,GE}$ GRACE: $E_{NS,Stot}$	18	No discharge data used Combined with GRACE	Strickler-Manning	+ Only 1 extra parameter - Cross-section data needed - Assumption: constant roughness in space and time