Reply to reviewer 1:

We appreciate the reviewer's great attention to detail; we feel as if the large number of suggestions (general and textual) has directly led to a significantly improved manuscript. We also appreciate that the reviewer agrees with our perspective that proper numerics in hydrology are important.

Here are the general comments and our replies:

1. Is it possible to express the numerical error as percentage of the rainfall error? If you make some simple assumptions about the rainfall error. Then you can use this as metric and show how it increases with rainfall intensity/duration. Same question about discharge measurement error. If one makes an assumption about 10% error (heteroscedastic) then one can express the numerical error as percentage of the discharge measurement error. One can show how this error evolves with time.

   Reply: We think this is an insightful point; putting the numerical error in the context of other sources of error would certainly help readers get a feeling of the magnitudes of numerical errors. Therefore, we have expanded section 4.1 with a paragraph where we place the numerical errors that we found in context with errors observed for discharge and rainfall measurements. Diving into the literature, we found that it is hard to provide a single number (such as the suggested 10%) to observation uncertainty, given that it depends on many aspects such as measurement device, spatial coverage, intensity, climate, etc. The added text gives some more insight in how numerical errors compare to observation errors, and demonstrates that they have the same order.

2. Then on a related note, can you investigate the relationship between numerical error and flow level (=discharge) ? May be interesting to see - as this, I believe, is not explicitly addressed in earlier studies. This should show that numerical errors are relatively large during rainfall events, and these errors dissipate during a subsequent drying period. This, dissapation is one reason numerics has not got the attention it deserves from the community. I will revisit this point in a later comment.

   Reply: We feel that this would be an interesting question to answer, but it is just a bit outside the scope of this already fairly long manuscript. Also see our response to comment 5.

3. Did you consider midpoint methods? Why/why not? Similar question for Runge-Kutta methods by the Carl Runge and Wilhelm Kutta? They developed a whole toolbox of explicit/implicit/fixed/variable time step integration methods. What about backward Euler? I know this implies more work, nevertheless, maybe there was a reason not to include these methods in your analysis - then, it would be good to know the reasons.

Reply: This experiment uses all 8 numerical methods available in FUSE; we assume (and the literature review tentatively suggests) that these represent a broad variety of popular choices among hydrological models. This is also supported by the original numerical daemons papers. Nonetheless, in Section 2 we have added a brief description of all of the additional numerical methods that the reviewer indicates. Further, this paragraph contains the justification for why we examine the numerical methods we do (we find them to broadly represent popular choices).

4. When it comes to Heun and Euler some papers are cited but the original inventors of these solution methods are not mentioned! I would include a reference to Leonhard Euler (Institutionum calculi integralis) and Karl Heun, among others.

   Reply: This is an important point! Euler's publication is now cited, and the inventors of the other methods which have a surname are textually indicated.

5. One reason proper numerics receives little attention among surface hydrologists is the nature of hydrologic systems. Negative feedback loops ameliorate differences in initial states and promote convergence to a stable state. Indeed, for such systems one can simply use a spin-up period (as you use in the paper) to prepare the initial states of the model for subsequent simulation. In systems with positive feedback, numerical errors will have a devastating effect on long term behavior as model runs will diverge rapidly and suggest a very different system behavior later on. Thus, one reason that numerical errors have historically received relatively little attention is the nature of hydrologic systems, that is, negative feedback loops induce stable attractor states - hence, why we can solve poor knowledge of the initial states with a spin-up period. Note, in some fields, differential equations are so incredible sensitive to numerics that these small errors can induce chaos (example: two-predator-two-prey systems).

   Reply: We agree that numerics in hydrology receive relatively little attention due to the negative feedback loops inherently present in hydrological systems. Further, we very much agree with the previous comment that the need for this paper likely underscores a dearth of numerics courses in hydrologic education. Therefore, we added a new Subsection (5.4) that offers potential reasons for the lack of numerics attention in hydrologic modeling and describes other systems whose behaviors are inherently chaotic, necessitating more sophisticated numerical methods. In contrast, convergence in hydrologic systems has possibly led to less numerical attention, while this paper demonstrates that peak (as in Section 4.4) and intense precipitations can lead to large numerical errors. Finally, we comment that there is a clear opportunity in hydrologic education for increased numerics education.

6. The paper of Schoups et al: doi:10.1029/2009WR008648 draws similar conclusions as herein, that is, the use of a second-order integration methods is preferred. I believe the text should address this earlier paper and those possibly related to it more properly. The paper is cited, but the text does not address that similar findings have been reported elsewhere. On a related note, there are more surface hydrologic model codes that use

proper numerics. For example, the hmodel of Schoups et al. (same paper) has been used in various publications. This model uses 2nd order adptive Heun for its numerical solution.

Reply: Indeed the paper of Schoups et al. is an important contribution to numerical analysis in hydrologic modeling. Therefore we have added a paragraph in recommendation of numerical method section (5.3) describing that Schoups et al. demonstrate the preference for second order adaptive explicit methods. Three papers which introduce models or MMFs that use sophisticated numerical methods and cite the Schoups et al. paper have been included in Section 5.4, in order to acknowledge the impact this paper and the original numerical dameons papers have had. We however do not plan to expand the literature review; we use models studied by Addor and Melsen because these models were selected in that paper for their popularity, rather than their numerical choices. This method therefore avoids cherry-picking for models which support or refute the idea that sophisticated numerical methods are now widespread. We have updated Section 3.4 with a paragraph containing this justification.

7. You may want to emphasize in the paper that poor numerics not only affects the simulated discharge, but compromises tasks such as parameter estimation, prediction, simulation of related state variables (groundwater table, soil moisture), etc. Those familiar with proper numerical procedures are aware, but not all those others reading this manuscript.

   Reply: This is very true and indeed another highly relevant implication of numerical errors. A few sentences have been added to convey this message, in Section 4.1:

   The effect of measurement uncertainty in forcing and discharge observations on parameter and model structure inference has already been explored in literature (Kavetski et al., 2006a,b, Vrugt et al., 2008). This study shows that numerical errors, having the same order of magnitude, can also hamper the process of parameter identification, model structure identification, uncertainty estimation, or in short, in testing hydrologic theory.

   We also feature more on this topic in Section 5.3, in the same paragraph where we discuss the relevance of Schoups et al. to our current work.

8. I do not see anything else that is wrong with this paper (see my written comments on pdf), except that the questions listed above may help find a second hook to 'sell' this work. Climate change is an interesting hook, yet, one may argue that the precipitation intensities as used herein are a bit exaggerated. Maybe a more detailed investigation into how numerical errors behave in a simulation may be interesting, their dependence on simulated flow level and simulated state variables. Perhaps even better, can you pinpoint which process in the model contribute most to the numerical error during precipitation extremes. This must be the most nonlinear process - or that process (its flux) that changes most rapidly in a time step.

Reply: We do not think the precipitation intensities used in this paper are exaggerated. It is our aim to test the robustness of numerical methods (and thus, hydrological models) under extreme conditions. Naturally, the most extreme are the world records, but for a more general picture we also look at downscaled intensities. Consider, for example, that in 2019, the lead author's home region experienced more than 500 mm of precipitation in less than 2 days, which is around 25% of the world record intensity for 2 day events. At least for the simulated 5 day events, NRMSE was maximized for some methods at around 25% of the world record intensity. In response to a comment from the second review, we updated Figure 3 in order to place the precipitations in this manuscript in the context of real events. Finally, we do agree that pinpointing which processes are more responsible for numerical error would be an interesting topic, but we also feel as if this could (and probably should) receive its own manuscript, given that the current manuscript is quite long. We have therefore recognized this as an unanswered aspect of our analysis in Section 5.1.

Here are the responses to the textual comments. We agree with and have implemented the vast majority of the suggestions.

-Line 2: we removed 'approximate'.
-Line 3: we changed 'like' to 'such as'.
-Line 5: we removed 'events'.
-Line 6: we changed 'is' to 'are'.
-Line 7: we changed 'expense' to 'cost'.
-Line 8: we added a sentence wherein we introduce root mean square error and also removed the text '(root mean square error)'.
-Line 9: we believe that this question is interesting, but somewhat outside the scope of the paper. See our replies to comments 2 and 5. Note that we do examine the effect of a variable rainfall signal (not necessarily strictly increasing) in Section 4.4.
-Line 11: we changed 'low' to 'small'.
-Line 12: we replaced 'basic' with 'small'.
-Line 13: we changed 'might' to 'may'.
-Line 14: we agree that the last sentence could be made more concrete. We changed it to: "We conclude that relatively large numerical errors may be common in current models, highlighting the need for robust numerical techniques, in particular in the face of increasing precipitation extremes."
-Line 17: we agree that this is layman's language. We also have found that the highlighted text is unnecessary and it was deleted.
-Line 18: we deleted 'the' and changed 'determining the security of a city's water supply' to 'assessing water supply security'.
-Line 22: this is a good point - some lumped models (for example recession analysis) don't have particular state variables corresponding to specific physical regions. The highlighted text was deleted.

-Line 28: We agree that this sentence is technically imprecise, due to the fact that we currently omit that hydrological models contain relationships between state variables. The sentence was changed to: “Therefore, hydrological models contain mathematical relationships between state variables and fluxes (as well as relationships between state variables themselves) that need to be solved numerically (approximately) rather than analytically (exactly).”
-Figure 1: The change in precipitation intensity from 20.1 mm/d to 60.2 mm/d is indeed extreme and the diagram is just used to illustrate the relationship between precipitation intensity and numerical error. Do note however that while the change might be unrealistic, many parts of the world routinely experience 5 day events at 60 mm/d and modeling efforts with poor numerical choices in these areas might be subject to the depicted magnitude of numerical error. The caption of Figure 1 has a new sentence which describes how the change depicted in Figure 1 is unlikely to happen in any given climate.
-Line 55: we changed ‘when precipitation becomes intense’ to ‘under relatively intense precipitation regimes’
-Line 57: we changed ‘might’ to ‘may’.
-Line 59: we changed the commented sentence to “Given a sufficiently large time step or large enough flux, fixed-step numerical solvers are not equipped to handle large precipitation events.”
-Equation 1: we made the notation easier to read, using the given suggestions in LaTeX. /bigl and /bigr were applied in the appropriate places for all equations. Also, in Equations 5 and 6, we added the missing ‘n’ subscript below ‘f’.
-Equation 9: very good point! Units now explained shortly after listing of Equation 9.
-Line 169: the sentence was reworded such that the word ‘defaulting’ or ‘default’ occurs once.
-Line 170: ‘pass this check’ was changed to ‘satisfy this threshold’.
-Line 172: the suggested text has been implemented; it is more technical wording, which we feel is appropriate here. Note however that we replaced the suggested “the step size” with “it”
-Line 196: we changed ‘in generating’ to ‘to generate’.
-Line 227: that interpretation is correct; we do refer to intensities applied to the spin-up period.
-Line 228: changed ‘from’ to ‘which are characteristic of’.
-Line 230: we feel as if our language here is sufficiently descriptive without mentioning control volumes.
-Equation 10: We used a \textrm on RMSE
-Line 252: changed ‘on’ to ‘for’.
-Line 266: we do not intend to change the models in our literature review, but we do intend to mention that higher order adaptive explicit methods have been used in the context you suggest. (See our response to comment 6.) We do not want to formally expand the literature review because we do not want to cherry-pick for studies that confirm or deny the idea that proper numerics have been widely adapted; this could go both ways. The models in the literature review are the same as in the work of Addor and Melsen (2019). We have added a paragraph at the end of section 3.4 justifying our choice of models in the literature review. We also indicate in the discussion (5.3) that there are models which use sophisticated numerical methods that are not included in our literature review.
-Line 276: deleted ‘the’.
-Figure 5: see our response to comment 3.

-Figure 8: Figure has been adapted: colors have been changed for ease of identifying specific methods, line widths and point sizes have been increased, and the ‘x’ representing the most intense 5 day event has been changed to a triangle. The grayscale of points has also been slightly adjusted. Pareto text removed. Caption updated to clarify purpose of ‘x’ and ‘y’.
-Figure 9: we acknowledge the debate on the use of p-values in science; indeed, the establishment of non-significance at a value greater than 0.05 seems arbitrary. However, in this case, the p-values for groupings based on numerics are many orders of magnitude lower than those for groupings based on structure, where values based on numerics groupings are always lower than 1e-20. While we believe the p-value metric is somewhat imprecise and its use might be limited in future publications, in this case we find it a convenient way to demonstrate the importance of numerical technique choice on numerical error, in contrast to structural choice. Therefore, we have removed the figure, which added little to the analysis, where the text indicates the large difference between p-values for groupings based on numerics vs structure. This many orders of magnitude difference is now clearer without focusing on the figure, and removal of this figure takes attention somewhat away from p-values.
-Line 583: see our responses to comments 2,5, and 8. Further: even if the climate was not changing, the ‘hook’ that numerical method choice matters even with our current climate (as demonstrated by this paper, the Clark and Kavetski papers, and the Schoups et al. paper) is still quite prevalent.

Reply to reviewer 2:

We appreciate this review; we believe that it has enhanced the apparent relevance of our manuscript via incorporating additional contexts and perspectives.

General comments and our replies:

1) The manuscript as a whole is well structured, the methods are generally thoroughly described or supported by relevant sources and/or equations, the discussion is rather exhaustive, and actually covers some of the critical points that popped to my mind while reading the results. However, dealing the paper with a source of error and its relevance in hydrological modelling, specifically focusing on extreme precipitation, I miss putting this source of error into context, as compared with all the other sources of error a modeler would expect, and specifically the ones the modeler would expect when dealing with extreme precipitation and possibly with floods. Per se we are dealing with an ill-posed problem most of the time in hydrology, without even the need to go into other interrelated dynamics (see Di Baldassarre et al. 2016).. but anyway, I think it would be important to mention the uncertainty and errors the authors would expect to be related for instance to discharge measurements in a modelling application (see e.g. Westberger et al. 2020) – being this the most common variable used for calibrating and validating hydrological models-, or errors in the input data themselves, and in particular in precipitation, being this the main input the authors are focusing on. Errors in precipitation estimates can be considerable – if not deviating- and can occur in the original precipitation measurement itself, and further in the extrapolation over a larger area.

Reply: We agree that the manuscript misses an opportunity to compare numerical errors to other sources of error (e.g. observational, structural); this was also noted by Prof. Vrugt, who suggested comparison of simulation in discharge to heteroscedastic errors in discharge observations. Therefore, we have examined relevant literature to determine what kinds of errors in precipitation and discharge observations can be expected. An extra paragraph has been added to Section 4.1 which compares our results in numerical error to observational errors in literature, further explaining the relevance of numerical error.

2) Another aspect which might be worth to spend a few (more) words on is the time scale: why do you only look at the daily time scale, and what would you expect to be difference by going down to the hourly time scale – which is the time scale at which e.g. flood forecasting is performed ? While the daily time scale is more relevant and common in climate change application, this might not be the case for present applications, which however also deal with the difficulties of making intense precipitation and hydrological models getting along, and for which a correct representation of the process is of primary importance.

Reply: We examine the daily time step because it is commonly used in hydrological models. However, as the reviewer indicates, simulations are often performed with a smaller time step. If this experiment was repeated with a smaller time step, we would expect to see the same trends in evolution of numerical error with respect to precipitation extremeness, simply with lower magnitudes of error. This is supported by the theory described in Section 2, where numerical error is described as proportional to the time step to some integer power. Therefore, we can expect lower magnitudes of error for all methods, and especially for those methods which are proportional to the time step to the second power, when a smaller time step is used, in the context of this experiment (which ignores potential variation of precipitation forcing data at the hourly time scale). We added a new paragraph added in Section 5.1 (under "Choices in temporal and spatial discretization") which describes this.

3) I am not a native speaker, and as such not the best judge, but I dare to say the paper is well written, and it mostly reads fine, just sometimes it becomes a bit' cumbersome to read. I think this is the case mainly because of the close repetition of words in some paragraphs (e.g. P15 L 328 The numerical techniques were sorted into one of the three ranked groups based on their rank,...).

Reply: In L 328, we removed a "ranked" to make it more readable and have re read the manuscript with a special attention for cumbersome language.

Specific comments and our replies:

-Abstract: While the authors and probably many hydrologists are familiar with FUSE, it might be advisable and useful to specify you are looking at conceptual models only in the abstract of your paper (as Clark & Kavetski 2010 and Kavetski & Clark 2010 do in the abstract, but also actually in the title of their two papers).

Reply: We agree that we examine conceptual models only and that this information should be included in the abstract. Therefore, we added a sentence in the abstract clarifying that all models are conceptual and lumped.

-Introduction: P3-L62: You choose to use the same numerical methods Clark & Kavetski (2010) and Kavetski & Clark (2010) applied: why didn't you consider expanding the numerical methods to include more of these?

Reply: As the results of our literature review tentatively indicate, the eight numerical methods used by FUSE are the most common methods used in conceptual hydrological models in general. In other words, of our surveyed modeling codes, none used a numerical method that is completely different from those that FUSE can offer. We have emphasized that the methods used by FUSE represent popular, current choices; this comment has also been addressed by responses to Prof. Vrugt's comments. Justifications for our choices of numerical method can

now be found at the end of Section 2 (and acknowledgement of future work via incorporation of more numerical methods can be found in Section 5.1, “numerical details”.)

-Methods, P9-L218-219: Being a swiss hydrologist and mainly familiar with European rivers and hydrology I must say I do not agree with the assumption 5,10 and 20 days are the time frames usually used for modeling floods. This is really highly depending on the catchments’ scale and other catchments’ features resp. catchments’ type. For mesoscale catchments 3-4 days are already relevant event durations (see also Froidevaux et al.2015), for some smaller alpine or quickly reacting catchments 1-2 days can also be enough. So here I would specify you consider event durations that are relevant for larger catchments, such as the Rhine for instance?

Reply: We agree with the reviewer’s comment that the relevant flood modeling period depends on catchment size. Therefore we incorporated the suggested edit, indicating that 5, 10, or 20 days are relevant time scales for modeling floods in larger catchments.

-Methods, Fig. 3: what is the meaning from the climatological perspective and where would Katrina be in this graph?

Reply: This manuscript does not consider world records resulting from different climatologies, rather we scale precipitation data from individual world records in order to have a baseline of rainfall intensities that are possible at a given duration. Therefore, no climate-specific conclusions about numerical error can be drawn, rather we simply use precipitation datasets that span a large number of different climates. However, we have indicated the location of a few notable events, including hurricane Harvey, in order to make the figure more concrete and give context to the synthetic precipitation events used in this experiment. The caption of Figure 3 has also been updated, describing these real events.

-Results, Fig. 9: It is just a suggestion, but wouldn’t it be more interesting to show the figure using stretched colours resp. a colours’ palette, using hatches of something else to indicate if 0.05 is exceeded? p=0.05 is commonly used in literature but it is still a subjective threshold.

Reply: We appreciate that this figure could be somewhat better explained, but we don’t believe that using a stretched color palette would be appropriate, considering that all p-values for numerics groupings are below 1e-20, and p-values for groupings based on structure are typically greater than 0.05. Since the purpose of this subsection is to demonstrate that numerical method is vastly more impactful than controlling numerical error than structure is, and we feel as if the figure adds little, we have removed the figure.

-Results, Fig. 10: it is OK to show the grouping of the reviewed codes as a mere visual, but it be more interesting for the readers to actually see a table with the exact numerical methods found resp. applied for the different codes? You could also attach it as supplement, if you consider it too much or not that informative, but as a matter of transparency and as you already did the job of extracting that piece of information, I think it would be a pity to not show it somewhere.

Reply: This information is now available as a supplement. Also, we added a sentence to the end of the Appendix explaining that there's more info on numerical methods used by the various surveyed codes in the supplement, and another one stating that codes have possibly been updated since the time of the literature review.

-Discussion, P 23-L488-492: what would you expect to change if you used a (semi-)distributed model? What do you think is the role of numerical error by allowing water transport in space?

Reply: In a distributed model, numerical error is proportional to the chosen spatial discretization raised to some power, as well as to the time step raised to some power. This theoretically implies that numerical error could potentially increase in terms of maximum value in the context of distributed models. However, a fully distributed model (with distributed forcing data) could have its local extremes in space smoothed out given the choice of spatial discretization, while this option is not available for lumped conceptual models. More generally, it is possible for numerical errors due to spatial and temporal discretizations to interact. Therefore, we believe that this is an interesting question which is not straightforward to answer. We indicated (using this reasoning) in the discussion (Section 5.1, "Choices in temporal and spatial discretization") that our results are not directly applicable to distributed models. How large the net numerical error is in case of a spatially explicit model, and what controls the magnitude of the errors involved, are interesting subjects for future research (also stated).

-Discussion, P24-L496-501: You might want to reinforce your findings with some literature? See e.g. Müller-Thomy & Sikorska-Senoner 2019

Reply: In the reference suggested by the reviewer, it is found that the impact of differing rainfall signals is to some extent diminished by the lumped rainfall-runoff model structure, and that flood peaks are sensitive to temporal aggregation method of forcing data. This clearly supports our claim that rainfall signal geometry can affect model results; we therefore included the suggested reference. Also we deleted the sentence beginning on line 488 and consolidated the first two paragraphs of section 5, because now there is an entire paragraph dedicated to discussing spatially distributed models.

-Appendix: P27-L609: maybe instead of TOPMODEL => (dynamic) TOPMODEL would be more correct? Metcalfe et al.2015 applied the dynamic TOPMODEL – even though I am aware that here the main difference between the two models is how water is distributed, what you are not applying here.

Reply: We incorporated the suggested edit: TOPMODEL -> dynamic TOPMODEL.

Technical corrections

-P 8: "Appendex" has been corrected to "Appendix".

-P 12: The reviewer is correct; we have changed the right hand label in Figure 4 to “Precipitation”.

Finally, we would like to thank the reviewer for the thoroughness of the review, the comment that the manuscript is mostly well written, and the compliment on the title.

Reply to reviewer 3

We appreciate this encouraging review.

Specific comments and our replies:

1) The type of hydrological models assessed (i.e., conceptual and lumped) appears very late in the manuscript. I expected to find this information much earlier, and it should even be mentioned in the abstract.

Reply: Consistent with the review from Dr. Kauzlaric, the abstract has been adapted to reflect that all models are conceptual and lumped.

2) Numerical methods, l.104: “which in some fields is desirable”. An example could be a nice-to-have small addition.

Reply: An example regarding operator splitting in convection-diffusion problems has been added.

3) Methods: I would expect event durations inferior to 5 days to be relevant for flood modeling in smaller catchments in many places across the World.

Reply: We agree with this comment, which is essentially the same as the first Methods specific comment from Dr. Kauzlaric. We have indicated that 5, 10, or 20 days are relevant time scales for modeling floods in larger catchments.

4) Results, l.306-307: Median (relative) errors of several methods are not always increasing with precipitation intensity as stated (for ex: adaptive explicit Heun, adaptive implicit Heun, adaptive semi implicit Euler). Or have I misunderstood something?

Reply: We would like to indicate that the reviewer has indeed missed a small detail (as potentially suggested by the reviewer). The reviewer does correctly identify that normalized median errors do clearly decrease with some numerical methods past a given precipitation intensity for 5 day events. We also identify (and explain) this trend in the final paragraph of subsection 4.1; the rest of subsection 4.1 which describes monotonic increases in numerical errors for all methods refers to RMSE rather than NRMSE. This is indicated in the second sentence of subsection 4.1.

5) Results, Section 4.2: This section is clearly the part I like least. First, it requires quite some effort to follow the ranking strategy and the details of the results. Then, the establishment of the groups seems to be a bit arbitrary, with some groups containing two methods and others three, which is also changing between low and high precipitation intensity. Why are there differences in the size of the groups between low and high precipitation? Also, the same colors were used to

represent different groups of methods between both panels of Fig. 6, which makes it confusing. Finally, I am not convinced that the conclusion of this analysis is worth the effort it takes to follow all the details of the procedure. I will let the authors judge what they want to do with it, but I would suggest simplifying this section.

Reply: We agree that this section is somewhat technical, but we believe that its thorough establishment of the robustness of our results justifies its presence. Still, we attemped to further clarify why the ranked groups occur as they do. To this end, we added a reference to Figure 5 in the first paragraph in Section 4.2; in this figure, the described ranked groups apparently emerge, which can be seen by looking at median errors. Also, in this paragraph, “fixed step implicit or semi” has been corrected to “fixed step implicit and semi”.

6) Results, Section 4.3: The impact of the reduced daily intensity with increasing duration is, in my opinion, a likely explanation for the decrease of the errors. I would have thus expected to find this hypothesis earlier in this section.

Reply: We believe that this subsection is appropriately placed, as it is the first time in the results that varying duration of precipitation event is discussed; subsections 4.1 and 4.2 deal solely with 5 day events.

7) Results, Section 4.4: This section would benefit from an introductive or transitional sentence.

Reply: We agree that the addition of a transitional sentence at the start of subsection 4.4. would improve readability and have added one.

8) Results, Section 4.7: I agree with Martina Kauzlaric that the link between the models assessed and the numerical methods implemented would benefit the reader.

Reply: We have added this information as a supplement (as indicated by our response to Dr. Kauzlaric’s review, in the specific comment on Figure 10). The appendix also indicates that this information is in the supplement.

Technical corrections

1) Intro, l.60-61: This sentence might be rephrased for more clarity. This is only a suggestion.

While this sentence is large, we feel as if it is sufficiently descriptive of the knowledge gap that our manuscript addresses.

2) Intro, l.83-85: It might also be rephrased for more clarity (again, a suggestion).

This sentence is indeed imprecise and has been changed to: “When the numerical techniques used by the reviewed codes are placed in the context of this study, an estimate of the potential magnitude of numerical errors arising from these codes is obtained.”

3) Numerical methods, l.137: Citation: the names of the authors should be in the parenthesis.

We have implemented this change.

4) Methods, l.197: The model here named 563 should be 536, according to Clark and Kavetski (2010)

We have implemented this change. We would also like to compliment the reviewer on catching this subtle detail.

5) Results, l.474-475: This sentence is not so clear and might be rephrased.

We have changed “extremes easily exceed 100 % NRMSE” to “extremes in NRMSE when using this numerical technique easily exceed 100 %” for increased precision.

Finally, we would like to indicate a few miscellaneous changes that we deemed appropriate:

The sentence beginning on line 66 has been removed because we deem it unnecessary.

Line 189: “This lets us test for generality” has been altered to “This allows to test for generality”

Line 243: a missing opening parenthesis was added and a space was added between a number and its unit.

Line 245: “coarser” was changed to larger.

Line 246: “coarser” was changed to “larger tolerance”

Line 438: “differences between the exact and approximate solutions” was changed to “as the difference between the exact and the approximate solution”

Line 439: “that result from the choice of numerical method used to find an approximate solution” has been deemed unnecessary and has been deleted.

Line 440: “in different structural contexts” has been changed to “in a different model structural context”.

Line 563: we changed “processes” to “process”.

Line 565: “paper” was changed to “study”.

Under Data availability, the DOI of our data has been added.

The citation for the data for this project has had its year of publication corrected from 2020 to 2021.

We would also like to indicate that we have changed the corresponding author from Prof. Teuling to Prof. Melsen.