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

Here are our replies to the general comments:

1. 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. However, while measurement errors might heteroschedastically be something like 10%, the variance in measurement errors can be large, so it's easy to imagine specific cases where the numerical error would exceed or fall short of this assumption. What's more, it could be that measurement error is related to precipitation intensity or discharge, so potentially one could expect different measurement errors for the different synthetic forcing data in this paper. On the other hand, finding more ways to contextualize errors makes the paper more accessible. Therefore, we intend to add 1-3 paragraphs and possibly a new figure which compare NRMSE to a 10% error in discharge in the results section. This will also imply a slight change to the methods section and perhaps an expanded discussion. We will also look further into heteroschedastic error in hydrology and incorporate appropriate references.
2. 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. Related to our response to comment 8, we believe that the precipitation intensities used in this manuscript are extreme but not exaggerated, due to the fact that they are based on observed events. So, while this would be another interesting angle from which to sell the manuscript, we think that it is a somewhat different topic. Also see our response to comment 5.
3. 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. Nonetheless, in Section 2 we will discuss the methods mentioned by the reviewer that we do not use. Midpoint methods are quite similar to Heun methods, and the difference between the two is easy to explain, which we will do in Section 2. The lack of Runge-Kutta methods is mentioned in the discussion, but we will also briefly mention these in Section 2. As for backward Euler, we will highlight the difference between backward Euler and implicit Euler, as implicit Euler is clearly used.
4. This is an important point! We will cite the inventors of these methods.
5. 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. While we don't think that these issues are a central focus of this paper, we of course do have an opportunity to mention these. Therefore, we plan to add a paragraph in the discussion which describes the difference between (convergent) hydrological systems and less numerically stable ones (for example, meteorological systems of equations). Here, we can clearly show how perhaps convergence has 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. We will also consider discussing opportunities in hydrologic education with respect to numerics.
6. We will add a discussion of Schoups et al (2010) and that their recommendation of numerical technique was quite similar (from a perspective of calibration rather than on synthetic data, which is encouraging). We also plan to mention other models that use sophisticated numerics (such as in or based on the Schoups et al paper), but we do not intend to change the models in the literature review. The models present in the literature review are based on the work by Addor and Melsen (2019). These are selected to avoid cherry-picking for models that support or refute the idea that proper numerical methods are

now widespread, and because doing a literature review of all hydrologic models would be a much larger project. Still, we like the idea that some progress has been made, and it is important to note this.

7. We plan to add this to the discussion or introduction; it indeed further underscores the importance of numerics and is easy to cite (Schoups et al, 2010; Kavetski and Clark, 2010).
8. We do not think the precipitation intensities used in this paper are exaggerated. It 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. 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 (probably should) receive its own manuscript, given that the current manuscript is quite long. We will therefore recognize this as an unanswered aspect of our analysis in the discussion.

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

-Line 2: we will remove 'approximate'.
-Line 3: we will change 'like' to 'such as'.
-Line 5: we will remove 'events'.
-Line 6: we will change 'is' to 'are'.
-Line 7: we will change 'expense' to 'cost'.
-Line 8: we will add a sentence wherein we introduce root mean square error and also remove 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 will change 'low' to 'small'.
-Line 12: we will delete 'basic'.
-Line 13: we will change 'might' to 'may'.
-Line 14: we agree that the last sentence could be made more concrete. We will change 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 will be deleted.
-Line 18: we will delete 'the' and change '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. So the highlighted text will be 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 will be 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: We will indicate that the change in precipitation intensity from 20.1 mm/d to 60.2 mm/d is extreme and the diagram is just used to illustrate the relationship between precipitation intensity and numerical error. Do note however than 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 ares might be subject to the depicted magnitude of numerical error.
-Line 55: we will change ‘when precipitation becomes intense’ to ‘under relatively intense precipitation regimes’
-Line 57: we will change ‘might’ to ‘may’.
-Line 59: we will change 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 will make the notation easier to read, starting with the given suggestions in LaTeX.
-Equation 9: very good point! We will include the units of all variables in the text after Equation 9.
-Line 169: the sentence will be reworded such that the word ‘defaulting’ or ‘default’ occurs once.
-Line 170: ‘pass this check’ will be changed to ‘satisfy this threshold’.
-Line 172: the suggested text will be implemented; it is more technical wording, which we feel is appropriate here.
-Line 196: we will change ‘in generating’ to ‘to generate’.
-Line 227: that interpretation is correct; we do refer to intensities applied to the spin-up period.
-Line 228: we will change ‘from’ to ‘resulting from’.
-Line 230: we feel as if our language here is sufficiently descriptive without mentioning control volumes.
-Equation 10: we will check the notation connections for HESS.
-Line 252: we will change ‘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).
-Line 276: we will delete ‘the’.
-Figure 5: see our response to comment 3.
-Figure 8: We will experiment with other color combinations and line widths in order to improve figure readability. Also, we now note that the ‘x’ representing the median values in computational cost and RMSE for the 120% of world record intensity events and the ‘x’ used to describe the guidelines might be ambiguous, so we will change the ‘x’ on the plot to be a different symbol. Further, we will remove the text “in the style of a Pareto front” from the caption. Finally, we will make it clear that ‘x’ and ‘y’ in the second to last line of the caption are dummy variables simply used to create the gray lines for visual aid.
-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.
-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.

Finally, we appreciate the grammatical corrections provided - these certainly help the paper to read more smoothly. Conveniently, the lead author is a native English speaker and will re read the manuscript with a special attention to grammar.