Reviewer #1

The authors have developed a global geomorphic model of fluvial floodplains, notably at 90 m resolution. The authors made use of global elevation, flow direction, and drainage area models along with the global HydroBASINS boundaries for their analysis.

The methods presented here closely follow the methods of Nardi et al., 2019 (i.e., GFPlain250) which uses height above nearest drainage (HAND) with a floodplain hydraulic geometry (FHG) thresholding scheme. FHG suggests that potential inundation depth can be represented as a function of a river's upstream draining area. Here, the authors argue that FHG parameters optimized for each basin, as opposed to global values, will better represent the spatial heterogeneity of global basins and ultimately be of benefit to floodplain delineation.

The authors propose an iterative process with starting values based on previous knowledge that converges on suitable parameter values for each global basin. The authors found that Parameter *b* in the FHG model loosely corelates with a basin's aridity index. Finally, the authors use two global hydrodynamic inundation maps (JRC and GAR) and another geomorphic floodplain model (GFPlain250) as reference data for comparison.

The authors have a logical claim; basins across the world are heterogenous and locally optimized FHG parameters could produce better models of floodplains when compared to global parameters. Of note, Nardi et al., justified the use of global coefficients by finding reasonable measure-of-fit values with varying *b* parameters. However, they also supported the notion that regional values for the scaling law parameterization could be further refined to capture local climatic variations.

SHIFT data and code were easily accessible. In North America, SHIFT aligns well with GFPlain with some noticeable differences. Specifically, in North America, SHIFT tends to estimate a narrower floodplain in comparison to GFPlain. Both products have notable examples of areas identified as floodplains that are omitted by the other.

I have several comments I would encourage the authors to consider.

Reply: Thanks for your clear and comprehensive comment. You'll find our point-to-point Reply below.

- 170 - I'm unclear on why 34 'major' river basins were selected for further analysis. The authors rely on the results in these 34 basins as evidence throughout their paper. Please explain the selection of these basins, what is significant about them, and what is the justification for analyzing them independently.

Reply: Thank you for pointing out this. The selection of the major river basins was based on the largest river basins aggregated by MERIT-Basin that are hydrologically connected. Specifically, we traced from all the outlet basins inland (and aggregated the inland rivers accordingly) to identify the largest basins. We performed a separate investigation into the floodplain hydraulic scaling relationship for major river basins, based on the assumption that larger river basins may exhibit more consistent floodplain hydrological processes and formation mechanisms, thus having stronger scaling relationships. This consistency may also provide a clearer context to investigate the relationships between relevant factors and the scaling exponent. However, some of the Arctic basins fall beyond the boundary of our reference datasets, so we manually excluded those basins. In this revision, we have refined our selection process by detecting whether the centroid of a basin is within the Arctic to make the definition clearer. As a result, we have identified 7 Arctic basins to exclude, leaving 33 major river basins for our analysis. We have added the necessary explanations to our revised manuscript to clarify the selection criteria and the justification for analyzing these basins independently.

- 233 – This could use more explanation. Why did the authors choose to define river as a function of UPA versus using the delineated river network in MERIT Hydro? Why select 1000 km2? The authors touch on this at the end of the paper.

Reply: The MERIT-Hydro dataset does not originally provide vector-based river network datasets[1], but MERIT-Basins does[2]. MERIT-Basins is delineated by setting a threshold on the UPA of the river, but it uses a 25 km² threshold, which would include too much small streams for global floodplain delineation. Conversely, a larger threshold, such as the 5000 km² used by the JRC dataset, imposes a stricter criterion on river streams, leading to fewer river networks and reduced floodplain boundaries in areas like deltas or other medium- to large-size rivers. We chose to define rivers as a function of UPA with a threshold of 1000 km² because it balances how many rivers are incorporated in the global-scale floodplain delineation process. This threshold ensures significant river networks are included without overwhelming the analysis with smaller, less relevant streams, which was also adopted by Nardi et al. (2019). Additionally, while MERIT-Hydro provides Height Above Nearest Drainage (HAND) data, it is based on a 0.5 km² threshold, which contains too many small streams not relevant for our purposes. Thus, we have chosen to use the 1000 km^2 and we have expanded our explanation in the methods section (2.2) and provided relevant discussion in section 4.4 to clarify this choice.

- 350 – What method was used to resample to 1-km?

  Reply: For continuous data like UPA and HAND, we used median as the resampling method. As for categorical data like the reference datasets and watershed division, we used mode as it accounts for the majority of information in the selected area. The details are added to the corresponding section.

- 352 – I see many permanent water bodies in the final SHIFT product. (e.g., the North American Great Lakes).

  Reply: In the previous version of our 1-km product, we've marked permanent water bodies with identifier 2, which can be easily filtered by putting a mask. We've also put the identifier for water bodies in the version 2 of our updated data. We will put our revised data in the zenedo repository.

- 450 - Use of overall accuracy overly rewards correctly classifying the 94.5% (author's estimates) of the world's land area that is not a floodplain. I would be more persuaded by overall accuracy if the authors were to limit their accuracy analysis to some reasonable distance from your river network (e.g., 1km, 10km).

  Reply: Thank you for bringing up this good point and we agree with your suggestion. To assess the consistency of our data with two other maps, in this revision, we've changed it to MAI because OA will overly reward non-floodplain areas as the reviewer noticed (Fig. 7). For the pairwise comparison, we've added the OA within a buffer of 20-km in the pairwise comparison section to provide a better understanding. The buffer here is calculated by the hydrological distance, that is by d-8 flow direction, and we've tried different buffer threshold from 5-km to 50-km. Overall, the OA statistics shown by all buffered threshold showed quite similar patterns, and since 5-km to 10-km may be too small since it'll be continuously 10-km of floodplain in some large basins, so a 20-km buffer is finally decided. The buffered patterns generally align with the un-buffered OA because OA considers non-floodplain areas, and the buffer merely adjusts the extent of these areas considered. In contrast, MAI focuses exclusively on overlapping floodplain areas and does not consider non-floodplain areas, resulting in different patterns. Results and descriptions are revised correspondingly.

- 454 – I would think to prove “the effectiveness of our parameter estimation scheme in capturing information from the reference maps”, I would need to see this same accuracy measurements but with global values used (e.g., the Ndari et. al., values: *b*= 0.3, *a* = 0.01) and the deltas.

Reply: Thank you for your suggestion. It is indeed a great idea to include this comparison. We have now added a comparison with MERIT-Hydro using universal parameters (UP) in this revision. Statistically, the total area of UP is 50.85% larger than SHIFT. We further conducted a pairwise comparison including UP, and the spatial distribution is shown in Fig. 8b (shown below). Among all pairs, SHIFT-JRC performed best in 62 basins, SHIFT-GAR in 74, UP-JRC in 8, and UP-GAR in 37. This result offers evidence that the derived parameters are effective in deriving better floodplain maps.

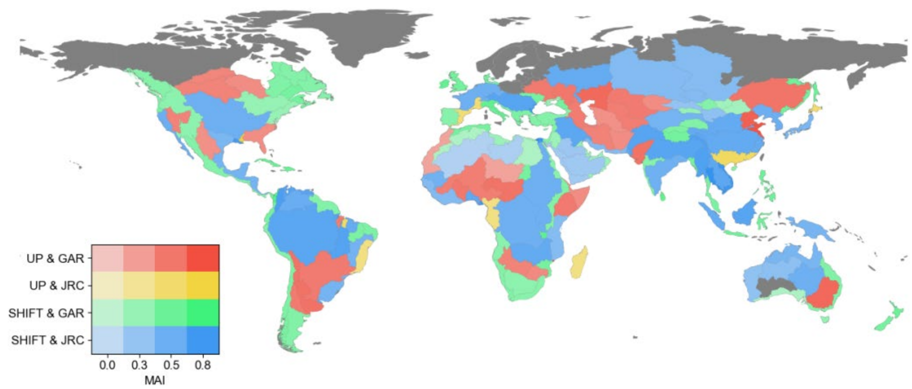

Fig 1 (Fig. 8b in the revised manuscript). Bivariate choropleth map of the highest-performance MAI pair among four pairs (SHIFT & JRC, SHIFT & GAR, UP & JRC, UP & GAR) and the corresponding MAI value for each basin Different pairs are represented by different hues, with higher MAI values shown in higher saturation. Basins where a SHIFT-pair performs best are marked in cool colors, while those where a UP-pair performs best are represented in warm colors.

In the boxplot showcasing consistency between the hydrodynamic maps and the geomorphic maps, however, we observed vast statistical difference between GFPlain and UP larger than that of SHIFT and UP. The explanation for these observations is twofold. First, the terrain inputs of GFPlain and UP are different. MERIT-Hydro includes hydrological corrections where all water body values are manually lowered, resulting in significantly lower HAND values and subsequently larger inundation extents under the same parameter applied. Second, the difference in SHIFT and UP is underrepresented in the boxplot, as SHIFT-JRC pair usually has high consistency where UP-GAR agrees better, and vice versa.

We have documented these interpretations objectively in the revised texts.

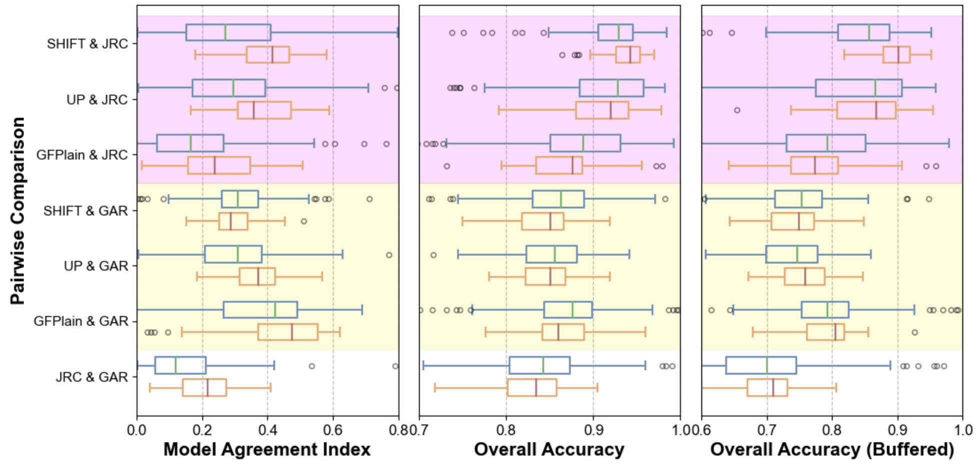

Fig 2 (Fig. 8a in the revised manuscript). Boxplots of pairwise analysis among SHIFT, GFPlain, UP (MERIT-Hydro but with Universal Parameters), JRC and GAR across three metrics: MAI (left), OA (middle) and OA within a 20-km buffer (right). Two group comparisons are marked in different colors (magenta for JRC and yellow for GAR). Statistics for all basins with valid data inputs (see Methods) are shown in blue boxes, and those for the 33 major river basins are shown in orange.

- 472 – “Superiority” is an overstatement. Agreement does not equate to superiority.

  Reply: We agree and this was removed in this revision. We stepped back and more objectively mentioned the improvement of agreement of SHIFT.

- 560 – I’m not sure I would call FHG correlation with hydroclimatic conditions ‘reasonable’. There is a loose correlation. Earlier the authors described it as ‘statistically significant but weak’. That is a more apt description.

  Reply: Thank you for raising this reasonable concern. We agree that "significant but weak" is a more accurate description. By "reasonable", we meant that the correlation is within our expectations, as it is important to notice that we did not anticipate a perfect correlation between exponent 'b' and hydroclimatic conditions for several reasons. First, the scaling relationship is a simplified theory that summarizes floodplain-forming processes, and the strength of this relationship itself warrants further investigation. Second, there are inherent noises in the two maps (as inherited from their own model chain errors), which can also limit the strength of the observed relationship. Since 'b' is an empirical parameter, and its physical interpretation remains an area of study [4], our goal was to identify any observable patterns between our optimized b with other factors, rather than expect perfect correlations.

We'd like to argue though, that while the correlation with the Aridity Index is not strong, its statistical significance supports the effectiveness of our parameter estimation methods, which helps to derive spatially-varying parameters that are physically meaningful. The largest basins, being hydrologically connected and are internally consistent in hydrological characteristics, also result in stronger correlations with the factors as expected.

To address our focus on interpreting the exponent b and its linkage with physical factors, we conducted two additional experiments in the hope of more comprehensively identifying factors that can be related to the geomorphic floodplain-forming processes:

i. **Additional Variables**: We included more variables in our analysis, such as LAI, terrain (mean and deviation), and soil factors (soil components in a river buffer). Our hypothesis was that AI would be the most significant factor, with LAI inherently related to AI, while terrain and soil might also be related but with less clear mechanisms. The results showed that AI was indeed the most significant, with LAI only significant in large basins. Other factors exhibit inconsistent correlations with b, also as expected.

ii. **Different Scales**: We also tested the estimation of the parameters at different scales (i.e., Level-4 and Level-5 basins) to increase the sample size. The results showed that AI and LAI have statistically significant relationships with the exponent b, while terrain factors showed significant but much weaker relationships, followed by soil factors that do not show statistically significant relationships with b.

While the correlations shown in the above analyses may not be very strong, they meet our expectations: AI is significant as the primary factor for explaining the spatial variability of b, LAI plays a role, and terrain might be related but not showing readily detectable correlations with the exponent b. We've revised our manuscript accordingly. Please refer to the newly performed analyses in Supplementary Table 1, and more objective statements of our parameters and hypothesis in Section 4.1.

- 561 – I'm not convinced this loose correlation proves effectiveness of the methods.

Reply: Thank you for your comment. We agree that it is challenging to prove the effectiveness of our methods without ground truth. However, we'd like to argue that the correlation analysis is only an indirect way of demonstrating that it is meaningful to derive spatially-varying exponent b for floodplain delineation, because these spatially-varying parameters are related to possible floodplain-forming factors, but not random values leading to floodplain maps matching with the hydrodynamic model outputs. To be more specific, we believe the effectiveness is indirectly supported by: 1) the parameters have weak but statistically significant relationships with variables like the Aridity Index, which aligns with our hypothesis; 2) the filtered data conforms to a relatively stable power law, indicating a relatively robust scaling relationship, and 3) changes in parameters result in improved consistency, and they are observed in spatial patterns. For a detailed explanation of our hypothesis and the correlation, please refer to Section 4.1 and our previous reply. For the spatial pattern of improvement, please see our subsequent reply.

- 566 – I'm not convinced of "superior consistency". Sometimes SHIFT is part of the highest agreement pair in a basin and sometimes it is not (Fig 7). The authors mention "superior consistencies" in the abstract as well. I'm not sure how to interpret that phrase.

Reply: We agree and this was removed in this revision. We stepped back and more objectively mentioned where improvements occur and what leads to the change of consistency. Descriptions are now better delivered in section 3.3.

- Fig 7 – It looks like GFPlain has higher agreement with GAR and SHIFT has higher agreement with JRC. Any explanations as to why this is?

Reply: Thank you for pointing that out. Indeed, GFPlain has higher agreement with GAR, while SHIFT shows higher agreement with JRC. In our original manuscript, we suggested that this might be because GAR tends to overpredict in certain regions such as in dry regions, similar to GFPlain. Now that we've performed the analysis with UP (i.e., universal parameters), we also observed that UP tends to provide larger estimates in some areas, whereas the estimates of MERIT-Hydro are inherently smaller. As seen in Figure 8b, although taking up a small percentage, UP still aligns better with GAR as both of them tend to over-predict floodplains, such as areas around the Caspian Sea. SHIFT aligns well with JRC as it highlights the inundation of large rivers, especially in the major river basins.

- Why include JRC & GAR and SHIFT & GFPlain combinations in the choropleth map? I'm less interested in where the two hydrodynamic models (JRC & GAR) or the two geomorphic models (GFPlain & SHIFT) agree and I'm more interested in where SHIFT outperforms or underperforms against GFPlain. That is, where does GFPlain better align with hydrodynamic models and where does SHIFT better align with hydrodynamic models?

> Reply: Thank you for your suggestion. We have redesigned our experiments in the pairwise comparison accordingly. Now, group comparisons were conducted with JRC and GAR, where each hydrodynamic map was tested against SHIFT, GFPlain, and UP (Universal Parameters on MERIT-Hydro, see above). The JRC-GAR pair serves as the baseline. For the choropleth map, we only compared SHIFT and UP to the two hydrodynamic maps individually to see which pair performs best and to identify any patterns in their performance. GFPlain is excluded as to address your comments on showing results in localized parameters. These results show that SHIFT outperforms UP for majority areas in Fig. 8a (see blue and green areas where they have the highest MAI), which is a proof that the estimated parameters of SHIFT are useful.

- Fig 7 - The color combinations for SHIFT + GAR and JRC + GFPLAIN are indistinguishable.

> Reply: Thank you for your advice and we've changed the color combinations accordingly. Please see our revised Fig.8.

General: The authors argue that locally optimized FHG parameters better represent the climatic heterogeneity of the world's basins than using global parameters. I would be more persuaded by a direct comparison of the two methods. That can be accomplished either by comparing SHIFT to the results of the author's methods but with global FHG parameters (e.g., Fig 6 using b = 0.3, a = 0.01 globally) or a direct comparison of SHIFT and GFPlain to reference data (e.g., Fig 7 without the JRC & GAR and SHIFT & GFPlain250 combinations)

Reply: We have performed the analysis as suggested, and please see our reply above and the revision in Section 3.3 that addresses this comment.

Essentially, the question is: Do locally optimized FHG parameters meaningfully improve the delineation of floodplains over global parameters and is there a spatial pattern of where those improvements occur? Any answer to those questions would be useful information for the community.

Reply: Thanks for highlighting again the key contributions of this study (scientifically in addition to contributing to data), that we should better discuss how locally optimized FHG parameters can help better delineate floodplains. We have revised our main texts and figures with additional analyses, attempting to more objectively document the pros and cons of SHIFT.

References:

1. The website of MERIT-Hydro: https://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_Hydro/

2. MERIT-Basins: https://www.reachhydro.org/home/params/merit-basins

3. Nardi, F., Annis, A., Di Baldassarre, G., Vivoni, E. R., and Grimaldi, S.: GFPLAIN250m, a global high-resolution dataset of Earth’s floodplains, Sci. Data, 6, 180309, https://doi.org/10.1038/sdata.2018.309, 2019.

4. Nardi, F., Vivoni, E. R., and Grimaldi, S.: Investigating a floodplain scaling relation using a hydrogeomorphic delineation method: HYDROGEOMORPHIC FLOODPLAIN DELINEATION METHOD, Water Resour. Res., 42, https://doi.org/10.1029/2005WR004155, 2006.