September 23, 2024

Prof. Yue-Ping Xu
Editor
Hydrology and Earth System Sciences

Ref. No.: HESS-2024-63

Dear Prof. Yue-Ping Xu,

We would like to thank you once again for handling our manuscript. We are pleased that both Referee #1 and Referee #2 have recommended accepting the revised version of the manuscript after we exhaustively addressed their comments in the first round of revisions. Additionally, we appreciate the opportunity to receive further feedback from another reviewer.

The newly revised manuscript entitled '*Enhancing the generalizability of data-driven urban pluvial flood models*' is enclosed, along with a newly revised supplementary material and our response to all new remarks.

We have addressed all the concerns raised by the reviewer point by point, with explanations where needed. The manuscript has been revised following the recommendations while ensuring that the modifications are in line with previous comments from Referee #1 and #2. We believe the revised manuscript is now ready for publication in *Hydrology and Earth System Sciences*. You may find below our responses (in blue) to the comments and suggestions of Referee #3, where line numbers refer to the 'track changes' version of the newly revised manuscript.

We are looking forward to your feedback.

Sincerely,

Tabea Cache, on behalf of all co-authors

Letter of response

Referee #3

Overall:

The paper presents a timely contribution to the rapid pluvial flood mapping using machine learning. The main novelty lies in the integration of larger "contextual terrain" information with high-resolution local "patches" (context-aware data-driven model), based on concepts from geospatial image segmentation. The methodology is solid and well-researched with multiple case studies (Zurich, Luzern, Singapore), different topographies, DEM resolutions, and rain events with varying return periods (2-100 years). "Transfer learning" from one terrain to another and a parsimonious retrain for new catchments indicates strong generalization potential. The results are well illustrated, using both visual and statistical criteria to evaluate the performance for unseen test data. However, some sections require clarification or additional details to strengthen the paper.
We sincerely thank the reviewer for dedicating their time to evaluate our manuscript and for offering constructive feedback. We greatly appreciate the thoughtful suggestions, which have been addressed in the responses below and were implemented in the newly revised manuscript.

General/structural comments:
1. L46-49: You describe the need for rapid flood mapping mainly due to the long computational times of hydrodynamic models. To enhance this perspective, consider mentioning other issues that machine learning models can address, such as calibration flexibility and handling uncertainties in large or complex catchments. Also, acknowledge that while advances in computational power can reduce time for hydrodynamic models, they still don't match the speed and parsimony of machine learning for rapid flood mapping.
We kindly thank the reviewer for this comment. In response, we have added the following remark in lines 46-48 (line number refers to the track-changes version of the manuscript): 'While recent improvements in computational power and more efficient algorithms have reduced the burden of hydrodynamic models, their run times are still insufficient for applications requiring a high number of simulations.' This is linked to the need to address uncertainty, which is further developed in lines 49-51: 'This is problematic, as multiple runs of these models are required per city to account for the large degree of uncertainty in future climate projections and urban development scenarios (Hirsch, 2011; Miller and Hutchins, 2017), necessitating the development of alternative models.'

2. L63: Please spell out "CNN" as "Convolutional Neural Network" the first time it is mentioned for readers unfamiliar with the acronym.
We thank the reviewer for pointing out this issue. We introduced the acronym in line 65, where CNN is mentioned for the first time in the manuscript.

3. Section 2 and Fig. 2: You reference Fig. 2 only once in Section 2. Some references to details in Fig. 2 in relevant parts of Section 2 would enhance their presentation. Ensure that the main figure and its caption in the text are as complete as those in the supplementary materials.
We agree with the reviewer's suggestion and have added multiple references to Fig. 2 in the text (lines 113, 116, 122, and 139). Additionally, some of these references refer to specific sections of the figure, as suggested by the reviewer. These references include: '(see the 'Inputs' and 'Spatial features' panel in Fig. 2)' in line 113 and '(see 'Scaled dot-product attention mechanism' in Fig. 2)' in line 122. We believe that these references will help the reader link the text with the specific sections of the figure.

4. L125: Fix the formatting error for the subscript “norm”.
We kindly thank the reviewer for noticing this formatting mistake, which has been corrected (line 134).

5. L203: Clarify which model you are referring to in “the model to generate flood maps” (is it the proposed model or WCA2D?). Be consistent in using "generation" for target data preparation and "simulation" for the outputs of the data-driven model.
The terms ‘generation’ for the target data and ‘simulation’ for the output of the machine learning model have been used more consistently in the revised manuscript. To ensure this consistency, changes have been made in the following lines: 17, 212, 219, 301, 346, 350, 478, and 480.

6. L225: Clarify that “18 1-hour uniform rainfall” refers to spatially uniform rainfall.
We agree with the added value of the suggested clarification, which has been added in line 235: ‘18 1-hour spatially uniform rainfall’.

7. L227: Specify that “shapes” refers to hyetograph shapes.
We thank the reviewer for highlighting that the term ‘hyetograph’ was missing. The term has been added (line 237).

8. Table 1: This table seems incomplete or needs redesign. Additionally, ensure it is referenced in the text (it currently isn't).
We have redesigned Table 1 and believe it enhances the table’s readability, thereby addressing the reviewer’s concern. Furthermore, we would like to kindly thank the reviewer for pointing out that the table had not been referenced in the text. The references to Table 1 have been added in lines 294 and 323.

Figures:
1. Fig. 5: It’s difficult to understand the transparency codes without reading the caption. Add a description, such as “(lighter color for shallower depth)” at the end of the text above the pie charts. Also, mention the case study in the caption, or in the title similar to Fig. 7.
We agree with the reviewer’s suggestion and have revised the description of the pie charts accordingly: ‘Proportion of cells in each water depth range (lighter colors indicate shallower depths)’ in Fig. 5 and Fig. 7. Additionally, we specified the case study in the caption of Fig. 5: ‘Violinplot of the simulation error in Zurich [...]’.

2. Fig. 7: Apply the same suggestions as for Fig. 5 regarding the transparency. Additionally, use similar y-axis limits for Figs. 5 and 7a-b for easier comparison.
As mentioned in the response hereabove, the description in Fig. 7 was revised to help readers understand the transparency codes. Furthermore, the y-axis limits in Fig. 5 were adjusted to allow for an easier comparison of the results between the case studies, as suggested by the reviewer.

3. Fig. 8 and Section 6: Introduce the heatmap (Fig. 8) first and use it to strengthen your justification for choosing the P31-2 rain event for transfer learning.
We would like to sincerely thank the reviewer for their suggestion. We considered introducing the heatmap earlier in Section 6 (as suggested) but prefer to maintain the current structure, as we believe it provides greater clarity for readers. Additionally, Fig. 8 presents results for models retrained using various rainfall events, not just $P_{31\text{-}2}$, which could make the explanation misleading. Moreover, the performances of the models retrained for different rainfall events are overall comparable, and the models retrained for $P_{31\text{-}2}$ do not show the best performance. This is

another reason why we believe that introducing the results presented in Fig. 8 earlier on in Section 6 could be misleading.

Specific clarification:

1. L248-252 (CSI Calculation): Clarify how "correctly identified cells" are defined, particularly whether this applies only for depths < or > 0.1 m. Given the high CSI values (0.98), it seems no threshold was applied for significant differences between target and simulated depths, but this should be briefly clarified.

We agree with the reviewer that a clarification was needed. The text now reads: 'The CSI for water depths below 0.1 m, i.e. where positive values in Eq. 1 correspond to water depths below 0.1 m, are 0.98 for P_{19-1} and 0.97 for P_{46-1}' (lines 264-265). To avoid any confusion with the CSI values for wet cells, we also clarified the text in lines 309-310: 'Considering a wet cell depth threshold of 0.1 m and a flood depth threshold of 0.3 m (i.e. positive values in Eq. 1 correspond to flood depths above 0.1 m and 0.3 m respectively) [...]'.

2. RMSE and CSI Values: Summarize RMSE and CSI values across all case studies and events in a table for transparency and easier comparison.

We have summarized the $RMSE_{0.1}$, $CSI_{0.1}$ and $CSI_{0.3}$ values across all case studies (i.e. in Zurich, Luzern, and Singapore) for the model trained in Zurich as well as for the models retrained in Luzern and Singapore for P_{31-2}. The RMSE and CSI values were evaluated for the following rainfall events: P_{19-1}, P_{31-2} and P_{46-1}. The table has been added to the supplementary material and referenced in the manuscript in lines 314, 353, and 390. Additionally, we have included the $RMSE_{0.1}$ in the figure's titles in Fig. 4, similarly to Fig. 6.

3. RMSE0.1: Define this index when first introduced (L333?).

The meaning of the notation $RMSE_{0.1}$ was indeed missing and is now defined in line 268, where it is first introduced.

4. Transfer Learning (Section 6): The section title "terrain adaptation" is a bit misleading. Consider renaming it to "Improving generalizability via transfer learning and parsimonious retraining" to better reflect its focus.

We thank the reviewer for the suggested section title revision. We have modified the title accordingly to: 'Improving generalizability to terrain through transfer learning and parsimonious retraining' (lines 317-318).

5. Additionally, explain your choice of using only one rainfall event (P31-2) earlier in the section, building on Fig. 8 to justify this selection.

As mentioned in comment n° 3 of the 'Figures' section above, we sincerely thank the reviewer for their suggestion. However, we believe that the current structure makes the text and methodology easier for readers to follow and understand, and therefore prefer to maintain it as it is.

Discussion enhancements:

1. Discuss the potential incomparability across case studies since rainfall ranges differ by location, but seemingly you used same events from Zurich for Singapore. Singapore may receive more intense rainfall over longer period, resulting in higher water depths for larger areas (due to flatter terrain) and potentially higher modeling errors as you suggested for higher water depth ranges. This opens opportunities for further studies on transfer learning for various rainfall lengths and more diverse intensities. If my assumptions are wrong, you could possibly discuss based on right assumptions.

We thank the reviewer for pointing out the need to elaborate on this point in the discussion. While the need to further train and test the model for rainfall events with different characteristics (e.g.

multiple peaks and longer durations) was mentioned in the paper, we have developed the discussion and made it more specific to the case studies of the paper. The new discussion elements were incorporated in the previously existing discussion in lines 452-456: 'Additionally, the model was neither trained nor tested for rainfall events with multiple peaks, intermittency, or events with rainfall on more than 1 h. The model should be further tested to account for different types and durations of design storms, reflecting the hyetograph patterns and variability specific to each city. Namely, the model should be evaluated using design storms characteristic of Singapore, such as intense events of up to 2 hours in duration, which align with the island's characteristic heavy rainfall.'

2. Generalization Discussion: Consider adding a reflection on combining hydrodynamic models and AI, or training models based on actual flood events and measured water depths. You could also mention the potential of integrating in-situ and remote sensing data to improve model performance when trained against actual observations.
We agree that this reflection is relevant to our study and have added it in lines 471-473: 'These databases could include not only simulated flood data but also observed data from real flood events, made more accessible by recent advancements, such as versatile flood level detection from images (Moy de Vitry et al., 2019), which offer new opportunities for in-situ flood data collection.' We intentionally kept this reflection brief to ensure alignment with the comments from Reviewer #2 from the previous round of revisions.

3. Hybrid AI and Hydrodynamic Models: As your model uses multiple data-driven methods, a brief discussion on how future work could benefit from hybrid approaches—where machine learning augments traditional hydrodynamic models—would strengthen the paper's relevance for broader applications. If there is a need for such approach, etc.
We kindly thank the reviewer for their valuable suggestion. This is indeed an interesting area of research, and we have therefore revised the text accordingly in lines 72-80, where we believe the discussion fits well: 'Lastly, another promising application of machine learning for rapid flood mapping is the use of hybrid approaches, which combine the advantages of different model types. The advantages of hybrid approaches have been demonstrated in recent studies, including Fraehr et al. (2023), where a fast model was developed by integrating a simplified, physics-based hydrodynamic model, optimized for speed through a coarse computational grid and long computational time steps, with a mathematical model that transforms the flood patterns from the low-fidelity model into those of high-fidelity, non-simplified models. The model's generalizability was tested in two study areas with distinct topographies, using a temporal resolution of 1 h and a spatial resolution of 20 m.'