Dear referees, dear editor,

We would like to thank you for your comments and constructive suggestions to our manuscript. We very much appreciate the time and effort that you have invested in your reports. This letter contains the responses to the referee comments which also formed the basis for the revision of the manuscript. We sincerely hope that the revised version is now acceptable for publication in Natural Hazards and Earth System Sciences, and we are looking forward to a decision.

Kind regards,

Omar Seleem

(on behalf of the author team)

Response to referee #1

The authors have addressed my comments. I suggest accepting the paper with some suggestions for minor changes:

AR: We thank the referee for the positive feedback, and the recommendations.

Fig. 7 - it is not clear to me why feature importance is assessed on the training data. These results may not provide us any insight into which variables are relevant for generating predictions in other areas. Would it be relevant to include the NSE for SA0 in the figure?

AR: We assessed the importance of the predictive features based on the testing dataset inside the training domains (precipitation depths which were not included in the training dataset) because the importance of the predictive features for predictions in other areas outside the training domain varies with the characteristics of these areas. For example, Figure 1 shows that while the topographic wetness index (TWI) was the most influencing predictive feature in the first round for SA1 and SA2, the depth of topographic depressions (SDepth) was the most influencing predictive feature for SA0. Therefore, we would like to suggest keeping Figure 7 (in the manuscript) without modifications if the referee agrees.

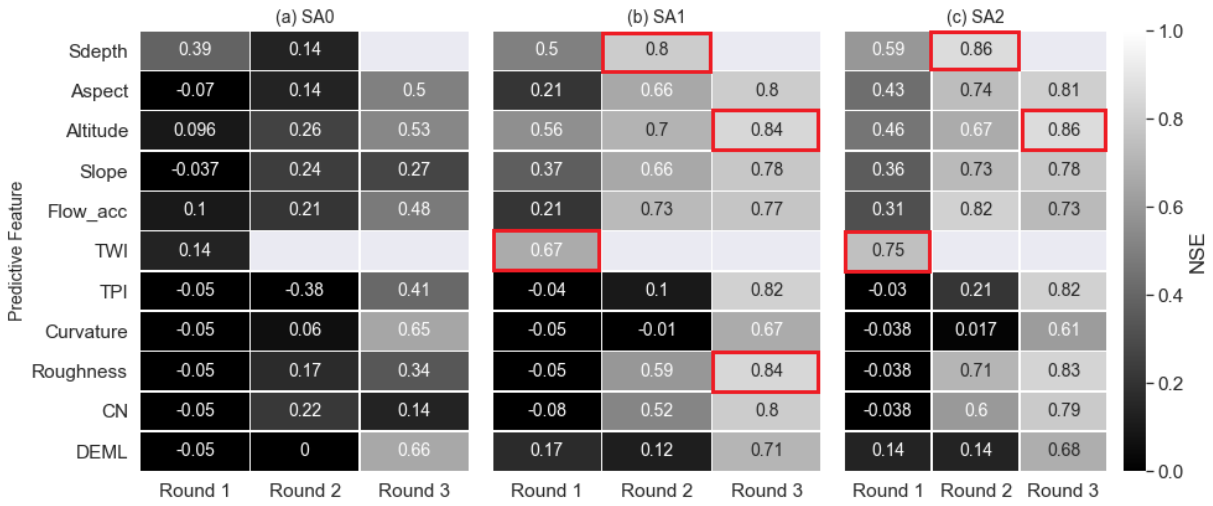

Fig1. NSE values for SA0 (a), SA1 (b) and SA2 (c) for the models trained in the forward selection process

for the best performance training data combination (U-Net - SA1&2). The best performance model in every round is marked in red.

Conclusions - I think one of the main conclusions from Fig. 4 is that an appropriate composition of the training dataset is very important for the performance of the model. This is a general principle with data-driven models where the training data need to cover the range of situations for which we want to predict (e.g. a model trained only in Berlin will not work in the Alps). I suggest adding a sentence on this.

AR: We agree with the referee. Applying a data-driven model for flood prediction is still a rising topic. The literature still lacks such a study which investigates how a data-driven model trained for a relatively topographically flat city as Berlin, performs in cities with different characteristics (cities in mountainous areas). We add the following sentence to the conclusion “Further research requires testing the data-driven model's transferability further in environments with different characteristics (particularly with cities in more mountainous environments)”.

Response to referee #2

RC: I appreciate the efforts and kind responses of the authors. I agree with most of their modification.

The only left concern of mine is the validity of research data. As alternative models to physical models, they highly rely on physical models. Therefore, the accuracy of the physical model is the key factor to determine the effectiveness of flood forecasting. As the author said, the lack of monitoring data may bring great challenges to the verification of physical models. However, the multi-source data such as social media data and satellite monitoring may provide some support for the verification of physical models. If possible, add my concerns at the end of the manuscript.

In general, I think the paper has reached the publication levels of nhess.

AR: We thank the referee for the positive feedback, and the recommendations. We added the following sentences to the manuscript “It is worth mentioning that the accuracy of the predicted flood maps by a data-driven model highly depends on the accuracy of the used hydrodynamic model simulations to train the model. While urban area lacks monitoring devices, crowd-sourced data and fine-resolution satellite images could be helpful tools to validate the hydrodynamic models.”