Manuscript Reference: NHESS-2023-152

Manuscript Title: Transferability of machine learning-based modeling frameworks across flood events for hindcasting maximum river flood depths in coastal watersheds

Summary: This paper explores the application of ML to hindcasting maximum river flood depths across events when informed by the spatial distribution of pertinent features and underlying physical processes in coastal watersheds. Trained within the same watershed, the ML model was then transferred in time to predict out-of-sample events for three major storms events. Acceptable performance was noted at 100+ gauge stations in the domain.

General Comments:

1. The content of the literature review (and motivation overall) was helpful and sound (applicable references to previous studies), highlighting gaps in the current field and therefore valid contributions that would be efficient for those wanting to hindcast flood depths for tropical storms without the need to use a hydrologic model.
2. Scientific Significance: An ML model that moves past simple categorical prediction of expecting a flood or not in a watershed to estimate actual flood depths at stream gauges; and an approach that allows for hindcasting of storm events in the watershed not previously seen or trained by the model are two main contributions of this work. While it is necessary for communities to understand expected depths along flood plains and other areas in a watershed, the study takes a key first step towards establishing a methodology for maximum flood depth estimation at point locations in a stream with an ML model that requires input features that could be harnessed from available datasets. Though retained to

point estimations of flood depths for the study in a single HUC-06 watershed, the methodology permits application/implementation to watersheds of various size and locations. The study addresses scientific questions within the scope of NHESS.

3. Scientific Quality: The PCA and SHAP analyses for assessing the importance of features on flood depth estimation will be very useful for the hydrological community. Of course these features will vary per watershed and future work may further inform potential important features currently not included that will improve performance – but, feature analyses such as these coupled with ML models are what help to make the black-box of ML models interpretable and hydrologists piece together valuable information on the physics behind flood events in a watershed without the need for setting up and running hydrologic models. Overfitting is always a concern for ML models but some steps to reduce "redundant" variables by computing correlation among variables was taken to help.
4. Presentation Quality: The overall presentation of the manuscript was well-organized and easy to follow – from the perspectives of both hydrology and ML. Figures and tables are easy to understand. There are minor grammar edits needed, but language otherwise was precise and fluent.

Response: We thank the reviewer for the constructive comments. Your review has helped us improve the manuscript quality. Detailed responses are provided to your questions. Blue text shows our response and black text shows your comments.

In the revised manuscript, we have added new features, conducted new analyses to evaluate the feature importance, performed new modeling to improve the performance and expanded our discussion to provide more insights about our results.

Minor Suggestions/Technical corrections:

1. Considering the mentioned overestimation of shallow flood depths and underestimation of high flood depths, perhaps a median metric would be a more robust test of performance (than MAE)?

Response: We have now included the median absolute error (MDAE) as an additional metric to evaluate our model performance. This enhancement aims to present a balanced view of the model accuracy across different flood depths. These new results are presented in Lines 577-582:

"The model demonstrated an excellent performance on the training dataset (R2 = 0.94, MAE = 0.64 m, MDAE = 0.44 m, and NRMSE = 24%). On the test dataset, the model achieved an R2 of 0.91, the MAE of 0.77 m, MDAE was 0.42 m, and the NRMSE was 28%, further suggesting the satisfactory performance by the model. The training history plot showed that the model performance improved with each epoch during training, indicating that the model was learning from the data. The model training process stopped at epoch 87 due to early stopping."

In Lines 621 to 635:

"The transferability of the trained and tested model (against Hurricane Ida) was examined by applying it to three other events within the same watershed. Table 4 summarizes the evaluation metrics for the three hurricanes.

Table 4. Model performance across in historical flood events. MAE: mean absolute error; MDAE: Median Absolute Error; RMSE: root mean square error; FQ: ratio of estimated over observed maximum flood depth.

Flood event	**R^2**	**MAE (meters)**	**MDAE (meters)**	**NRMSE (%)**	**F_Q (%)**
Original model					
Hurricane Ida	0.94	0.64	0.45	24.1	138.1
Transferability					
Hurricane Isaias	0.73	1.54	0.85	86.3	325.6
Hurricane Sandy	0.70	1.71	1.78	109.2	370.2
Hurricane Irene	0.85	1.12	0.85	36.7	112.6

These results demonstrated the model ability to transfer across different hurricanes within the same watershed (R2>0.70). With an MAE less than 1.71 m in all hurricanes, our model performance is consistent with the CNN model of Guo et al. (2021), demonstrating its capability for satisfactory flood depth estimates. However, when compared to the original model performance on Hurricane Ida, the R2 values and other metrics show weaker model performance for the transferability to other hurricanes, suggesting reduced estimative accuracy, but not to the extent that the model performance becomes unsatisfactory."

2. Also, given the vast difference noted between the NRMSE and the simpler FQ ratios, it would be interesting to see the scatter plot of simulated vs observed max flood depths for the storm events. This may even shed some light (on further understanding) where performance is good and not so good to help improve max flood depth estimation.

Response: We generated scatter plots of the estimated versus observed maximum flood depths for the events (Figures 6 and 8). These plots illustrate the model performance across the spectrum of flood depths, identifying areas of both strength and potential improvement in predicting maximum flood depths Lines 585 to 595:

“Figure 6 shows the performance of the ML model in hindcasting maximum water depths at stream gauges, comparing estimated values against observed values for both training and testing datasets. In the training phase (Figure 6a), points are clustered along the identity line, but tend to underestimate large water depths. This pattern suggested that the model learned the training data well, especially for smaller water depths, but did not fully capture the behavior that leads to the larger water depths. The underestimation of high values is expected due to the lower number of observations. The test data (Figure 6b) revealed a similar pattern of underestimation towards higher values; this can be since the number of observed high water depths is small.”

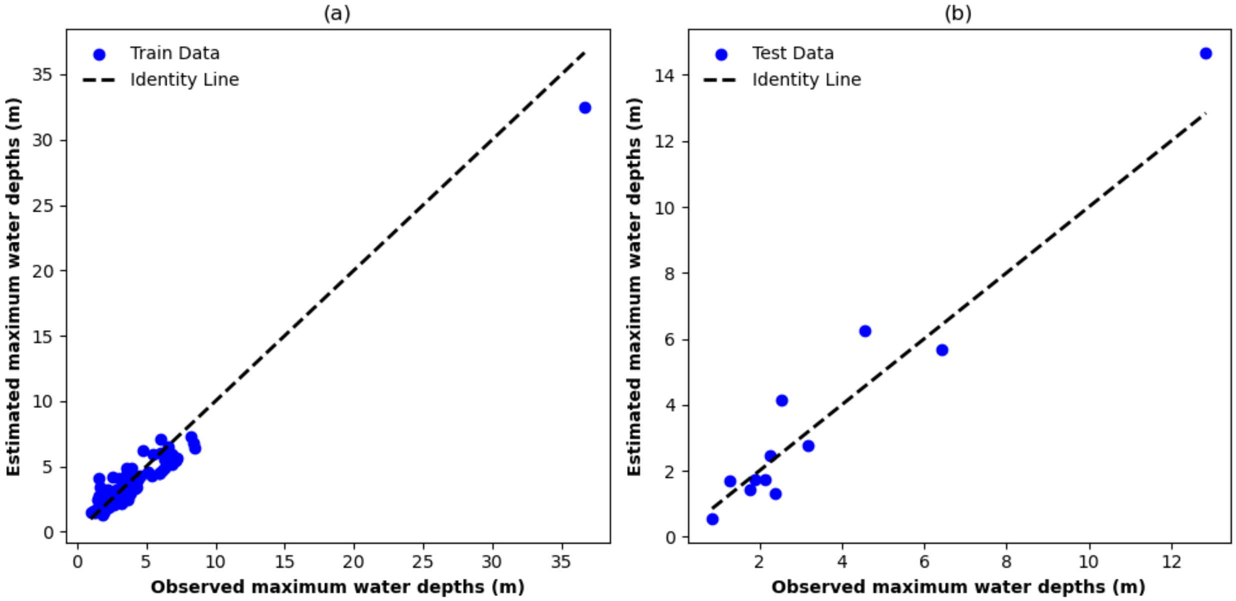

Figure 6. Scatter plots of estimated vs observed maximum water depths for: (a) train and (b) test data. The identity line represents a perfect match between the estimated and observed values.

In Lines 636 to 644:

“Figure 8 shows the relationship between observed and estimated maximum water depths for the four storm events. Most observed water depths for the hurricanes were low. For all four events, the data points suggested that the model tends to underestimate the high water depths and overestimate the low water depths (Figure 8). The plots for Hurricanes Sandy and Irene show a more dispersed set of points, suggesting a wider variance in the model estimates compared to the observations. This implied that the model is less accurate in capturing the flood dynamics of these events or that these events have unique characteristics that are not fully learned by the ML model.

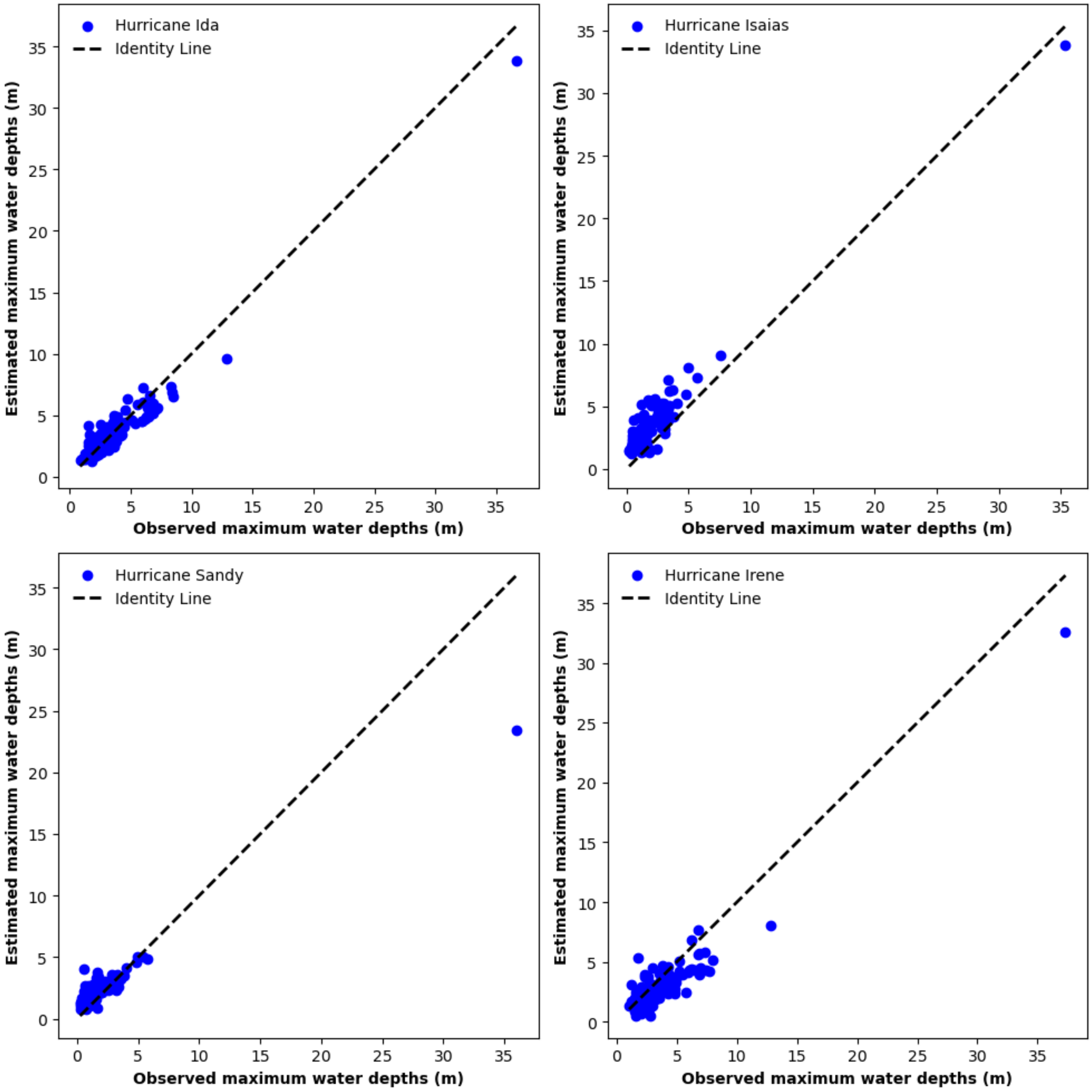

Figure 8. Scatter plots of estimated vs observed flood depth for the four hurricanes.

3. The ML model caters to multiple types of flooding (both fluvial and coastal) – as the intention was the hindcast depths at stream gauges using these events and using the same model to do locations where inland flooding is more likely versus coastal flooding (and vice versa); is there a trend or pattern noted for performance in areas susceptible to fluvial versus coastal flooding or storm surge? Finding that Hurricanes Isaias and Sandy were overestimating flood depths further from the storm tracks (tracks that were further from coastal locations for both cases) – is there a chance that it is skewed towards one flood type?

Response: Our analyses suggested a pattern where the model overestimates flood depths in areas farther from the storm track, notably for Hurricanes Isaias and Sandy. This observation indicated that the model performs better in certain flood types (fluvial vs storm surge). It also suggested that having separate models for different flood types or training a single model on diverse flood event data, can enhance the performance by accommodating the specific characteristics of each flood type. The discrepancies in our predictions, especially the overestimation for certain storm events (Hurricanes Sandy and Isaias), suggested the need for more nuanced model adjustments based on specific flood scenarios and their characteristics, such as storm tracks and primary driving factors like precipitation and storm surge. The related discussion about when our model performs better has been expanded in Lines 670 to 678:

"The other reason why the model transferability was most successful for Hurricane Irene was that the event mainly driven by significant rainfall, similar to Hurricane Ida (the event that the model was trained for). In contrast, the model performed worse for Hurricanes Sandy and Isaias because these events were mainly driven by storm surge. The original model, considered lower importance for storm surge, was not effective in predicting the water depths in Sandy and Isaias. In fact, here we see another significant advantage of strategically using physically meaningful features rather than the more commonly used black box approach. By considering the physical phenomena in our model development, we can better understand its strengths and weaknesses and more effectively evaluate its performance."

In Lines 687 to 694:

"The study underscored the complexity of efficiently predicting water depths for major hurricanes and emphasizes the necessity of refining models for better performance during such extreme events. It highlighted the importance of deeper analyses into features causing prediction discrepancies and suggested that addressing different flood types (fluvial vs. storm surge) separately can enhance the model performance. This approach, alongside adjustments for specific flood characteristics like storm tracks and similar influential factors that are distinct for each event, can improve the performance of hindcast models, aiding in the development of more transferable ML-based models."

4. Line 643 – review tense.
   The sentence does not seem to be incorrect in tense: "This pattern aligns with the southward total slope aspect, where the upper point of the river tends to have shallower depths and the mainstream exhibits deeper water depths."

   Response: We have removed the sentence as our model and the analyses have been changed based on the reviewers' comments.

5. Review Legend in Figure 6 (lowest interval overlaps with the next)

Response: This figure has been removed and instead we added MDAE based on the earlier comment by the reviewer. We limited the discussion for the F_Q metric in the revised manuscript.

6. It was useful to see the datasets and sources/references summarized in a table (e.g. Table 3).

Response: Thank you.

7. Perhaps a single note of abbreviations only at their initial mention is needed – e.g. for machine learning.

Response: We have reviewed the manuscript to ensure that all abbreviations, including those for machine learning (ML), are only spelled out at their initial appearance. The exceptions for this are figures, tables and headings that should stand alone.