The paper is interesting and presents a useful methodology for coastal flood modeling. The validation against a hydrodynamic model is OK but a bit questionable; this should be done against historical flood maps. A few assumptions should be also clarified. Limitations, uncertainties and implications need to be further discussed. I have recommended several edits and some comments in the PDF. Here are some additional comments:

Response: Thank you for the detailed review and the constructive comments.

Please provide more information on the study catchment, particularly those that affect your model results. This includes computational area, soil type, channels' size, ground slope, land use etc.

Response: We have included additional information in the revised manuscript regarding the study catchment (Section 2; Lines 148-159) and model domain (Section 3.1.1; Line 237). In short, the average slope, length, and annual discharge of the Savannah River are 0.00011 m/m, 505 km, and 320 m^3/s, respectively. Also, the river bathymetry was deepened up to 12 m for increasing the capacity of cargo transportation according to the U.S. Army Corps of Engineers. The model domain comprises an area of 1178 km^2 approximately.

Section 2: Present the source of drainage network data. Also, how detailed does that represent the drainage network?

Response: Drainage network data including river streams, tidal channels, and creeks within wetland areas can be obtained from the U.S. National Wetlands Inventory (https://www.fws.gov/wetlands/data/Mapper.html). These publicly available data are continoulsy updated by the U.S. Fish and Wildlife Service (FWS) and are derived from multiple data sources including satellite imagery and aerial photos of 1 m (or less) digital color infrared imagery. We included this information in Section 2 (Lines 157-159).

The verification of the approach against flood maps generated by a hydrodynamic model is questionable. How well is the model calibrated? For what historical events (how large/intense), it has been calibrated? Also, why not using satellite imagery like Dartmouth Flood Observatory?

Response: The primary goal of this manuscript is to propose a low complexity flood mapping (LCFM) method whose accuracy is comparable with a computationally expensive hydrodynamic model. Therefore, we compare our results with a hydrodynamic model to assess if our method can be a proper replacement of these models. This is the idea of proposing surrogate models that mimic the performance of complex physically-based models. A similar approach has been presented recently where surrogate machine learning methods are trained and validated against a well-calibrated hydrodynamic model. The hydrodynamic model has been calibrated for both non-extreme events and two major Hurricanes in the region, namely Hurricanes Matthew and Irma. Please see details of the calibration in Figure 4 and lines 371-389 in the revised manuscript.

Using satellite imagery has major limitations. 1) These maps are rarely available for the peak date of a flood event while we are looking for the maximum flood hazard maps. 2) These maps only provide the extent of flooding (HDC=0) while we need floodwater depths to

generate flood hazard maps for different levels of HDCs. For example, here we validate for both HDC=0 and HDC=0.6 m. 3. Daily satellite data, such as Dartmouth Flood Observatory uses coarse-scale satellite imagery, such as MODIS with 250-500 m spatial resolution that is not appropriate for validation. We need a much finer scale (<30 m) for validating our maps.

Please discuss the properties of the high-performance computing system that was used for simulations (Section 3.1.2).

Response: We used available computational resources of the University of Alabama (UA) for running model simulations in parallel. The UAHPC is a 87 node (1660 core) cluster featuring Dell PowerEdge M610s, M620s, and M630s. The nodes contain two Intel 8-Core E5-2650, E5-2640v2, or 10-core E5-2640v3 processors and at least 64GB of RAM per node. More information of UAHPC can be accessed in the following link: https://oit.ua.edu/services/research/. Nevertheles, we consider this information only relevant for the reviewer.

More details on the LHS application are needed. How was it informed by Hurricane Matthew peak WLs? What parameters were considered as uncertain? What probability distributions were used and how were they characterized?

Response: The Latin Hypercube Sampling (LHS) technique was used to sample 200 sets of roughness (n) values for model calibration. We considered a 7-day window around peak water levels (e.g., peak surge of Hurricane Matthew) to evaluate the model's performance. In that way, we identify the optimal combination of n-values (among the 200 model simulations) that accurately represent both non-extreme (low water) and extreme WLs. For simplicity, we only considered n-values as uncertain parameters and assumed that any errors follow a Gaussian distribution as discussed in Helton and Davis (2003). The advantage of LHS over traditional Monte Carlo approaches is that the former results in a denser stratification over the range of each sampled parameter as compared to random sampling. Hence, LHS leads to more stable results that are closer to the true probability density function (PDF) of the parameter. We included this information in Section 3.1.2 (Lines 251-254).

Please discuss how the performance of model was graded based on the fit metrics (RMSE, AUC and R^2). You may refer to Moriasi et al. (2015) and Ahmadisharaf et al. (2019) for streamflow predictions via R^2 or others for flood simulations. Neither RMSE nor R^2 measure bias. Metrics like PBIAS need to be used along to measure the model performance.

Response: We have included additional metrics to evaluate model's performance more rigorously: Kling-Gupta Efficiency (KGE), and Nash-Sutcliffe Efficiency (NSE). In addition, we have replaced R^2 by mean absolute bias (MAB). NSE measures the relative magnitude of the error variance of model simulations compared to the variance of observational data (Nash and Sutcliffe, 1970). NSE ranges between $-\infty$ to 1, where an efficiency of 1 indicates a perfect match between simulated and observed WLs. Kling-Gupta efficiency (KGE) is a robust evaluation metric that accounts for correlation, bias, and ratio of variances (Gupta et al., 2009). KGE can take values between $-\infty$ and 1, where an efficiency of 1 indicates a perfect match. Mean absolute bias (MAB) quantifies the bias of model simulations with respect to observational data. MAB of 0 suggests an absence of bias in the simulations. This information

and further discussion of model results are included in Section 3.1.2 (Lines 260-268) and Section 4 (Lines 358-360), respectively.

Please define what 'error' exactly is in the model evaluations under the Results section.

Response: Error is the summation of rate of false positives and rate of false negatives in binary classification problems. We have defined this metric in Equation 5 and lines 340-344 in the revised manuscript as follows:

"In binary classification, positive and negative refer to a value of one and zero, respectively. True positive instances are those positive cells that are correctly predicted by the classifier and false positive instances represent those negative cells that are wrongly classified as positive. The *error,* reflecting all cells that are wrongly predicted by the classifier, is a commonly-used measure for validating the performance of binary classifiers for flood hazard mapping. "

Please discuss what probability distributions exist in the MATLAB allfitdist tool.

Response: 'allfitdist' tool includes the following parametric probability distributions:

Continuous: Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale, and Weibull.

Discrete: Binomial, Negative binomial, and Poisson. We believe this information is not relevant for the main goal of this study.

The underlying assumption of a univariate flood frequency analysis is that a peak WL with a given return period leads to a flood event with the same return period. However, studies (e.g., Brunner et al. 2016) have shown that a combination of peak flow and other attributes like volume may lead to a different return period. This limitation should be at least acknowledged in the paper.

Response: Thank you for the great suggestion. We discussed this limitation and added an appropriate reference. Please refer to lines 550-553 in the revised manuscript.

Further details are needed on how TH and HDC are derived. As of now, it appears that they are subjectively derived.

Response: TH is the threshold of the hydrogeomorphic classifier that should be calibrated by optimizing the error measure calculated from the comparison of reference and simulated maps. Therefore, this variable is derived from optimization results and is not derived from subjective decisions. HDC, however, is the hazard depth cutoff that converts the continuous flood depth map to a binary flood hazard map. This is a control variable that the decision maker (emergency responder) should pick from. We use 21 HDC resulting from 0.1 increments in the range of 0-2 m and show the results (TH) for all these HDCs. This provides 21 points for generating a smooth curve (Figure 7) so that the decision-maker can simply use this curve and pick the required TH according to different values of HDCs. Please refer to lines 347-365 in the revised manuscript.

MAB has been reported in the Results section but not in the Methods section. Please either remove it from the Results section or discuss it in the Methods section.

Response: We have included a description of MAB in the revised manuscript (Section 3.1.2, Lines 249-257).

There should be a plot on calibrating w1 and w2 coefficients (for the H and D variables).

Response: Yes, we already included this plot in the manuscript. Please see Figure 6b where we show how calibrated w1 and w2 values change for different HDCs. The higher weight of w1 compared to w2 shows that feature *H* is more important than feature *D*.

L429-432: Reasons for this poor performance need to be discussed in the Discussion Section.

Response: We added more text that explains the potential reasons for the discrepancies between the hydrogeomorphic method and hydrodynamic model results. Please refer to lines 431-436 in the revised manuscript as follows:

"The main discrepancies are some noisy scattered low-hazard areas located in the east and southeast of the study area. These areas can reflect the flooded surface depressions (sinks) resulting from the pluvial impacts of extreme precipitation. Hydrodynamic models simulate the fluvial and coastal processes that occur adjacent to rivers and oceans while disregarding the pluvial impacts. The red circle in the left part of the figures shows a region that the hydrogeomorphic method cannot properly simulate, especially for higher HDCs. This can be due to the inability of the hydrogeomorphic method to properly simulate physical processes."

The comparison of computational time against the hydrodynamic model is unclear to me. Did you compare your static model against an unsteady Delft3D-FM or the steady-state? The runtime of an unsteady hydrodynamic model should not be very long; therefore, this advantage of your presented model is not as strong as it is presented.

Response: The Delft3d-FM simulates the flood in an unsteady condition. Due to the high nonlinearity and complexity of extreme floods, flood modeling in a steady state is highly erroneous. The runtime of a hydrodynamic model depends on the scale of the study area, and the number of grid cells. For a fine scale simulation (<10 m) performed for medium-large scale problems (> 1000 km^2), the computational time of hydrodynamic models can take a few days. The main goal of using LCFM methods is to reduce the computational time while providing acceptable accuracy (improve the efficiency of modeling). For emergency responders, timing is the most important factor, thus having access to more efficient models that estimate the hazardous areas in order of minutes is significantly beneficial.

Broader impacts need to be discussed. The authors should discuss what implications these results have for coastal planners and floodplain managers etc. and what existing programs in the US (e.g., FEMA FIRMs) may benefit from this research.

Response: The Discussion section already touches on this topic a bit. We have expanded this discussion on the implications for coastal planners, floodplain managers, and existing U.S. programs (e.g., the NWS) in the Discussion section. (Lines 570-599)

"Operationally, the Sea, Lake, and Overland Surges from Hurricanes (SLOSH) model (Jelesnianski et al., 1984) is the storm surge model currently used by NWS to perform storm surge forecasting and create probabilistic flood inundation maps for real-time tropical storms (Sea, Lake, and Overland Surges from Hurricanes (SLOSH), 2022). The feature of SLOSH that makes it the preferred model of the NWS for storm surge forecasting and mapping is the model's computational efficiency that allows the model to be run as an ensemble (Forbes et al., 2014). However, SLOSH is just one of several modeling options for storm surge modeling and mapping, each possessing strengths and weaknesses associated with their simulations. The inclusion of additional models that can create flood maps of storm surge for a given event should provide an enhanced understanding of the uncertainty of inundation at a given location (Teng et al., 2015). However, the higher computational burden of alternative models, such as Delft3D-FM, tend to preclude their use in real-time operations and certainly, their use in generating an ensemble necessary for probabilistic flood maps. The methodology we propose in this manuscript may offer the NWS and other agencies a means to utilize alternatives to SLOSH for flood inundation mapping and probabilistic flood inundation mapping on U.S. coastlines. Models such as Delft3D-FM can generate reference maps to train the binary classifer and build the probabilistic operating curves. The probabilistic operative curves would account for the major source of uncertainties and provide a computationally efficient and reliable decision-making tool for coastal planners and floodplain managers. The operative hydrogeomorphic threshold classifiers proposed for real-time coastal flood hazard mapping can be used as an alternative tool for the rapid estimation of hazardous areas during real-time flood events. In an operational mode, water level or meteorological forecasts can be used to estimate the return period of an upcoming coastal flood event and the methodology here can utilize this as an input to perform LCFM flood inundation mapping both deterministically and probabilistically."

Study limitations and potential areas for future research need to be expanded.

Response: We have already included three areas of research for future studies. To expand this, we added more text explaining the study limitations and potential areas for future research. Please refer to lines 511-519 in the revised manuscript.

"The proposed hydrogeomorphic index (I_{HD}) is the primary data for flood hazard mapping in this study. Thus, the quality of two main inputs of this index, namely the DEM and stream network used to calculate features H and D play a vital role in the overall accuracy of the proposed approach. To obtain maximum accuracy, here we used the best available DEM with the finest spatial resolution of 3 m that includes the bathymetry data. However, considering the limited access to such high-quality DEMs in many areas of the world, it is recommended to evaluate the sensitivity of the proposed approach to lower quality DEMs (e.g. 30 m and 90

m DEMs without bathymetry information) in future studies. Another piece of research can investigate the sensitivity of the proposed approach to the density of the drainage network used for calculating the I_{HD} index."

In general here are the areas of research we recommended for future studies:

1. Sensitivity of the hydrogeomorphic index to DEM quality and stream network density (Lines 511-519)
2. Applying the proposed hydrogeomorphic operative curves to inland floods and to other deltas across the US. (Lines 536-540)
3. Improve the flood frequency analysis, considering its uncertainties, incorporating other sources of uncertainties in the modeling to generate probabilistic operative curves (Lines 550-558)
4. A benchmark study that compares the performance of three LCFM methods (Lines 607-610)

Sources of uncertainty and how they may affect your findings need to be discussed.

Response: This has been thoroughly addressed in the discussion section. Please refer to lines 549-569 in the revised manuscript.

"The reference maps used for training the binary classifier are key components for generating reliable results. Since these reference maps are the outcomes of hydrodynamic modeling, they are prone to uncertainties stemming from unrealistic parametrization, imperfect model structure, and erroneous forcing. The design floods used as boundary conditions of the hydrodynamic model are estimated from flood frequency analysis that is prone to uncertainty as well. Here we used a bivariate approach that estimates the design flood based on the water level data. A more comprehensive flood frequency analysis that accounts for other flood attributes, such as volume can improve the reliability of flood frequency analysis in future studies (Brunner et al., 2016). With access to less than 100 years of data for flood frequency analysis, the extreme return levels (i.e. 500- and 1000-year floods) pose high uncertainties due to the extrapolation of annual maxima data. This should warn decision-makers to be more cautious about using operative curves for extreme flood events. For future studies, the uncertainty bounds of flood frequency analysis (especially extrapolations for extreme cases) can be considered in the modeling. In a real-time scenario, the forecasted WL used for flood frequency analysis is also prone to uncertainties originating from imperfect forecasting methods and nonstationary climate data. In addition, the uncertainty of model parametrization can be accounted for by running the hydrodynamic model for different combinations of optimum parameters. Model structure uncertainty can be also considered by using different hydrodynamic models and combining the results. Finally, probabilistic reference maps together with uncertainties involved in WL forecasting and flood frequency analysis can be integrated to develop probabilistic hydrogeomorphic threshold operative curves in future studies. This is in line with the report provided for the NOAA National Weather Service (NWS), showing the NWS stakeholder's preference for utilizing probabilistic storm surge inundation maps (Eastern Research Group, Inc, 2013)."

Please discuss how your presented modeling framework can be used in other study areas. What considerations should be taken to do so? Guidelines should be provided in the Discussion section.

Response: We added the following texts to address this concern of the reviewer. Please refer to lines 540-545 in the revised manuscript.

"To implement this approach, first, a hydrodynamic model should be set up for the new study area and generate reference inundation maps for different return periods. Access to observed water level data (gauges or HWMs) and flood extent maps from past floods is required to properly calibrate the hydrodynamic model. Then the I_{HD} index calculated from a DEM is utilized together with the reference maps to provide the hydrogeomorphic threshold operative curves for future floods."

Please spell out all the abbreviations in the headings, figures and tables. These need to stand alone.

Response: Done.

Please italicize all variables/parameters in the text.

Response: Done.