REVIEW 2

This paper integrates synthetic storm generation (rainfall and storm surge) with a reduced-complexity hydrodynamic model to assess the impact of between-storm variability in flood hazard estimates. The authors find that using a single nominal design storm can provide a very limited understanding of flood hazard or risk, suggesting important directions for future work.

Overall, this paper makes a valuable and substantial contribution to the field of coastal flood risk assessment by highlighting the limitations of traditional design storm approaches and advocating for the use of large ensembles and probabilistic methods. However, the paper needs significant revisions to clarify its contributions and limitations, and to improve its readability.

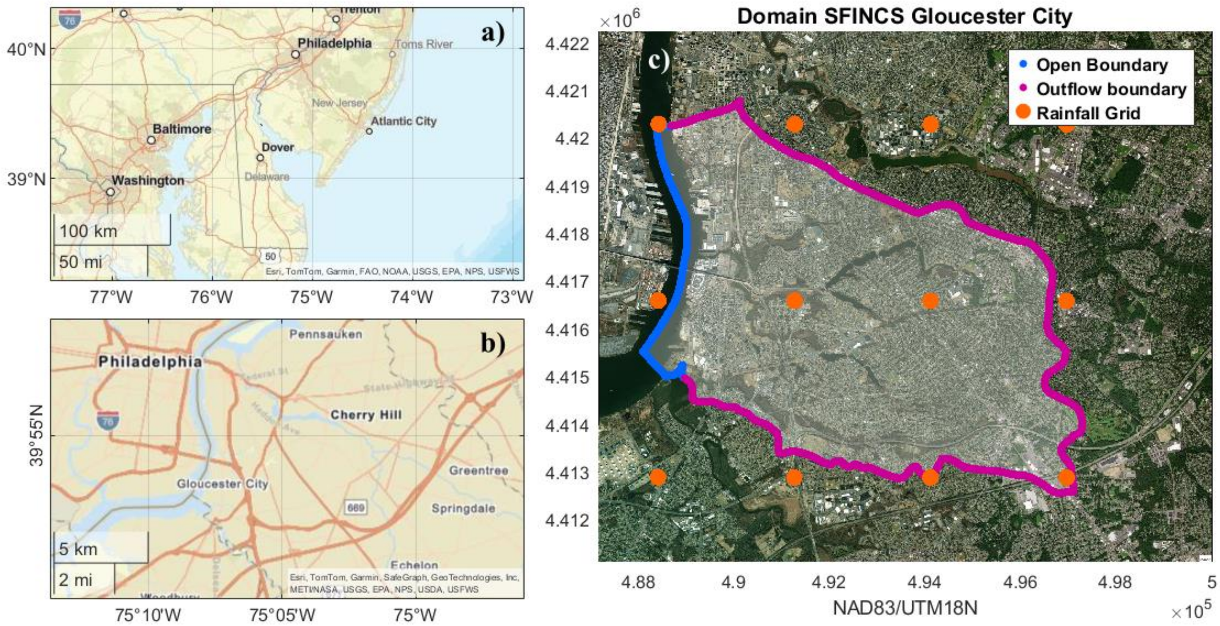
The authors would like to thank the Reviewer for the valuable and insightful comments. We address each point in detail below and describe how the corresponding suggestions will be implemented or considered in the revised manuscript.

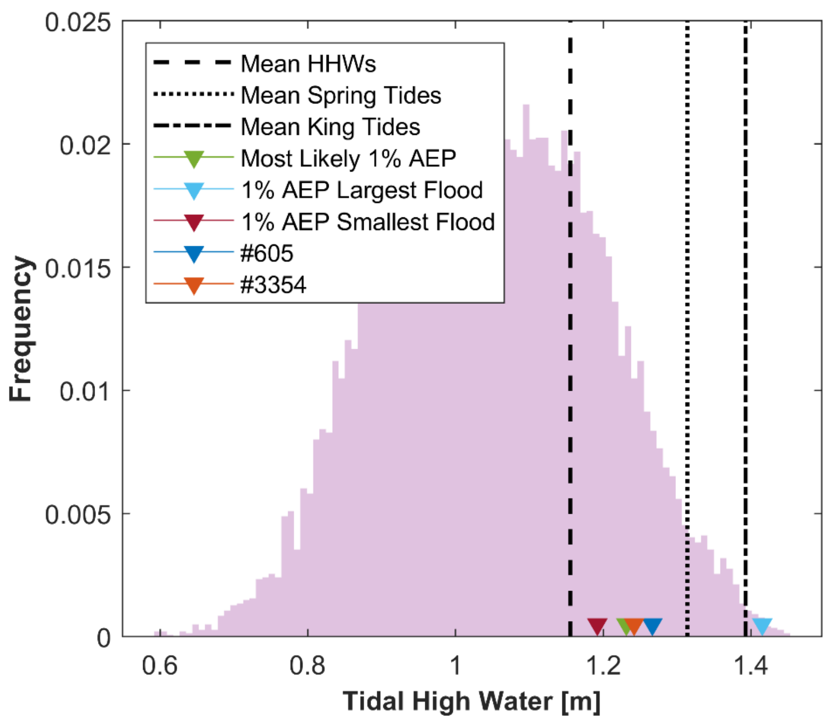
My primary concern is that the paper "buries the lede" by spending a lot of time on the specific test case and model development, which detracts from the main contribution of the work. The key findings about the limitations of design storm approaches and the benefits of large ensembles and probabilistic assessments are somewhat obscured by the detailed discussion of the modeling setup and validation. Fundamentally, I don't think the conclusions of the paper are particularly sensitive to specific details about how well-validated the SFINCS model for this watershed is (although there are of course many questions for which that would be important!) I'd suggest moving some of the testbed results and figures to an appendix or supplementary material.

We appreciate the reviewer's comment. While we initially considered including the flood model validation in the Supporting Material, the authors view model validation as a critical component of any modeling study and have a critical view on studies that provide limited information about validation or insufficient detail. We also aim to include as much information as possible about the model setup to ensure transparency and reproducibility. That said, given the manuscript's primary focus on advancing flood hazard modeling and in line with the reviewer's recommendation, we have moved the validation section to the Supporting Material to better emphasize the core analyses and findings in the main text.

A related concern is about the level of polish of the figures. Improving the clarity and presentation of the figures would greatly enhance the overall readability and impact of the paper. Just to pick the first figure as an example, fig 1 spends a large amount of space on the x and y axis labels, wasting space, is very low-resolution, lacks (a) (b) (c) markings, and the use of satellite imagery in the larger figure could more usefully be topography or land use. Similar critique applies to most of the figures in the paper. For another example, I found the number of different vertical lines with different colors and line styles in fig 3 to be confusing and distracting.

We thank the reviewer for this helpful comment and have revised the figures to enhance their readability. Specifically, we updated Figure 1 to make more efficient use of space by reducing the font size of axes and labels to the minimum permitted by the journal and replacing the original maps with higher-resolution versions. In response to feedback from another reviewer, we have also added a map of the topobathymetric and land cover data used in this study to the Supporting Material. We chose to retain the satellite basemap in Figure 1, as we believe it offers valuable context about the study area. For the remaining figures, we reduced the font sizes of axes and labels to maximize the area available for data visualization. Additionally, in Figure 3, we replaced the lines representing tidal levels with triangle markers to reduce visual clutter while preserving key information.





A third concern is about the discussion of probabilities in this paper, which I similarly find to be confusing and distracting to the reader. As an example, figure 6 talks about the "most likely 100 year event." This is particularly confusing, as a standard interpretation of the "100 year event" in a bivariate context is that there is a 1% chance of a draw from the bivariate distribution landing "outside the curve". While some definition is provided, it is confusing and I don't think that it necessarily supports the development of the paper. Section 3 says "We investigate differences in flooding between the event- and response-based approaches by simulating flooding from a large number (5,000) of compound events that allow estimating the empirical distribution of flooding and comprise several 1% AEP events." This discussion is tortured. Although I understand what you're saying, I would suggest being more specific and explicit with your language. For example, instead of calling something a 25-year compound event or 4% AEP event, I would take a little bit more space and describe it as "an event whose rainfall or storm surge would be expected to be exceeded once every 25 years," which is a bit of a longer text but will greatly enhance the readability of the final work.

We appreciate the reviewer's observation regarding the clarity of terminology. To improve this, we have added a brief definition of the 1% Annual Exceedance Probability (AEP) or 100-year event in the introduction. Throughout the manuscript, we have revised several instances of "1% AEP" to more intuitive phrasings such as "event with a 1% chance of occurring in any given year," particularly in the lines noted by the reviewer. However, we retained the terms "1% AEP" and "100-year event" in other parts of the text, as they are widely recognized in the field and help maintain conciseness in more technical or complex sentences.

Regarding the use of the term "most likely" event, we agree it can be misleading. However, as this terminology is commonly adopted in multivariate design frameworks to identify a "design event" (i.e., representative event) along an isoline based on joint probability density, we have chosen to retain it to ensure consistency with existing literature and to facilitate comparisons with standard event-based approaches in multivariate analyses (see e.g., Gori et al., 2020; Jane et al., 2022; Moftakhari et al., 2019). Nevertheless, we have extended the explanations related to this approach in the introduction, methods and discussion sections.

The following text has been modified and added in the manuscript:

"For the latter, selecting a single 1% AEP design event is particularly challenging, as multiple combinations of flood drivers can yield the same joint exceedance probability. This challenge has sometimes been addressed through the use of ambiguous constructs, such as the "most likely" event, which attempts to identify a representative scenario among equally probable combinations based on the density of observed events (Jane et al., 2022; Moftakhari et al., 2019b; Salvadori et al., 2011a)."

"To further investigate differences in flood hazard estimations between approaches, we also define a "design event" from all the 100-year events following the "most likely" approach for multivariate events (Jane et al., 2022; Moftakhari et al., 2019a). This approach selects one event in the isoline based on the density of observed events along it (Salvadori

et al., 2011b), identifying this event as the most representative scenario ("most likely") among the equally probable combinations along the isoline."

My last concern is about the methodology used to generate the synthetic storms in a probabilistic framework. The authors provide a brief overview and largely refer to a previously published paper. It would be helpful to have a bit more detail here. In particular, the ways in which the authors sample from this bivariate distribution, and how it conceptually relates to other related approaches such as JPM-OS, climate model downscaling, synthetic storm generation, etc. are not clear (to be clear; those methods solve problems different from what the authors are doing here, so I am not suggesting that they need to defend and justify their choice so much as explain conceptual links and differences). The caveats and limitations of this approach ought to be more clearly discussed in the discussion section, in particular the discussion of how the assumption of uniform rainfall affects the finding that pluvial flooding is relatively insensitive to between-storm variability, which is at odds with other previously published work.

In response to this comment, along with related feedback from the other reviewer, we have expanded the description of the multivariate statistical framework used to generate the synthetic events. This includes additional details on the selection of extreme compound events, the definition of NTR, and the validation procedures. We have also extended the discussion of methodological limitations and clarified how this approach compares with alternative methods, as suggested by the reviewer.

The following text has been included:

Introduction:

"For inland flooding, FEMA applies the event-based approach that starts by defining a design rainfall storm, typically derived from NOAA Atlas 14 which provides rainfall depths for specific probabilities and durations (e.g., 1% AEP, 24-hour storms). The design storms are used in hydrologic models to simulate runoff, with the resulting hydrographs then routed through hydraulic models to estimate flood depths and extents. Similarly for coastal regions, a design event is selected from the distribution of coastal water levels to estimate the 1% AEP regulatory floodplain. In regions affected by tropical cyclones (TCs), FEMA further implements the Joint Probability Method (JPM) to construct synthetic storm climatology. This involves statistically sampling combinations of key storm parameters (e.g. central pressure deficit, radius to maximum winds, forward speed) based on their joint probability distributions. These synthetic events are then dynamically downscaled to the coast and exceedance probabilities of coastal water levels are calculated based on the

probabilities of the storm characteristics. Although the JMP approach might reduce the uncertainties related to estimating the likelihood of low-probability coastal water level events by increasing the sample size of this events, in both cases, the probability of the event is assumed to approximate the probability of flooding."

"Similarly, for rainfall and river discharge, traditional approaches defined a single "design storm" or "design event" to represent the temporal and spatial patterns of these drivers (i.e., a representative event structure). However, some recent studies have shown that relying on a single "design storm", overlooking the variability in event structure across multiple storms, can underestimate flood hazards and associated impacts (Baer, 2025; Perez et al., 2024) ."

"To our knowledge, the differences in flood hazard estimates between these two approaches have only been evaluated for rainfall flooding (Baer, 2025; Perez et al., 2024; Winter et al., 2020)(Baer, 2025; Perez et al., 2024; Winter et al., 2020), but remain unexplored for compound coastal flooding. "

Discussion:

"Likewise, differences in rainfall-induced flooding between the event-based approach and the use of synthetic storms that capture the breadth of temporal and spatial variability of rainfall fields have been shown to significantly influence flood hazard estimates in the U.S. East and Gulf coast regions (Baer, 2025; Perez et al., 2024) and in Austria (Winter et al., 2020) , among others.

Another limitation of our study is that we use a synthetic event set developed using a data-driven statistical framework, which is limited to observed events. Although the statistical framework used to generate the synthetic events accounts for more dependencies between parameters that characterize the events (e.g. time lags) than other previous frameworks (Couasnon et al., 2018; Moftakhari et al., 2019a), it may not fully capture the full range of the potential spatio-temporal variability of flood drivers. Tropical cyclones might also be underrepresented in the historical sample since their frequency of occurrence is very low. This limitation can be overcome by using synthetic tropical cyclones that are dynamically downscaled to the study site . (e.g., Gori et al., 2020). Methods such as the JPM, which expand the storm climatology, enable the generation of a larger set of tropical cyclones, and capture greater variability in their spatio-temporal characteristics compared to historical records. However, these methods are computationally demanding, as flood drivers must be generated in advance of the flood assessment using hydrodynamic models. Further research is needed to evaluate how different synthetic event generation approaches affect flood hazard estimates. Given the *high computational demands of JPM, its application across large coastal areas may be impractical, making data-driven approaches like the one used in this study a more efficient alternative. Similarly, other data-driven techniques, such as stochastic storm transposition, are increasingly being adopted to generate synthetic rainfall fields for assessing rainfall-driven flood hazards (Baer, 2025; Perez et al., 2024; Winter et al., 2020). However, further investigation is needed to ensure that this method adequately preserves the interdependencies between coastal and rainfall processes when generating synthetic compound events for coastal flood assessments. A potential source of uncertainty in the variability captured by our synthetic event set arises from not disaggregating river- and coastal-driven components of the NTR. In our mid-estuarine study area, both processes contribute to the NTR, along with their nonlinear interactions. Separating these contributions would introduce considerable complexity due to their tightly coupled dynamics. Our approach is supported by recent work from McKeon & Piecuch (2025), who investigated the relative influence of coastal and fluvial drivers in the Delaware Estuary above flood thresholds. They found that most events observed at the Philadelphia tide gauge were primarily driven by coastal processes (e.g., tides and storm surge), but others resulted from river discharge alone or a combination of both mechanisms. Another limitation of the synthetic event set used is the reliance on mathematically defined thresholds for event selection, rather than thresholds based on actual flood impacts. This approach may exclude relatively frequent, lower-magnitude events that fall outside the statistical tails of the drivers' distributions but are still capable of causing localized flooding, potentially influencing response-based flood estimates. In our study, we evaluated the flood response of events near the selected thresholds and found that several produced no flooding, while others resulted in only minor inundation, with empirical return periods between 1 and 2.8 years. As a result, the selected thresholds did not affect our response-based flood estimates; however, this may not hold true in other regions with different hydrologic or exposure characteristics."*

Conclusions:

"Additionally, future research should aim to evaluate how different methods for generating synthetic events influence the resulting flood hazard estimates. Such comparisons can help inform best practices for generating more reliable flood hazard assessments under both current and future climate conditions."

Minor Points

I have a few specific suggestions that illustrate the above points, though they do not provide comprehensive wordsmithing for the paper.

- page 1: avoid "hinterland" word choice

Changed to "inland areas"

- page 5: Maduwantha.....characteristics: this seems like a lot of work to go through to avoid generating synthetic, realistic events

We are a bit unclear on what exactly the reviewer meant here. We believe (and show through different validation steps) that our synthetic events are "realistic". As discussed in previous comments and our responses, there are other ways of generating synthetic events. If the reveiwer refers to the approach that generates synthetic storms, that approach is actually a lot more work to go through than what we use here, because a storm surge model has to be used with the wind and pressure fields, a hydrologic model with the rainfall fields, and a hydrodynamic model to link the two. While SFINCS can do many of those things it hasn't been tested (much) for it's ability to generate and propgate storm surge from the open ocean to the coast. It would also require to cover a very large domain (in our case essentially the entire Delaware basin) making it computationally very demanding as opposed to our approach. - page 11: "We find that the floodplain of each of these 1% AEP events is different, resulting in very large differences in both flood extent and depth between some of the events." this is perhaps an even better example of how the discussion of return periods here is confusing. What is an AEP event? I eventually was able to figure this out. A key finding of the paper is that there is not a 1:1 link between the return period of the rainfall rate or the storm surge and the return period of the flooding or the damage, in line with other studies. Calling something a 1% AEP event is not helpful here. I don't think the methods need to change, just the presentation.

Following the response to the previous comment, we have defined what 1% AEP means in the introduction and modified the use of 1% AEP for "event with a 1% chance of occurring in any given year" in several sentences along the manuscript. We have also modified the objective of the analysis in the introduction: "Here, we explore the degree of linearity in the relationship between events with 1% chance of occurring any year and flooding of equal probability, from compound events of precipitation and estuarine water levels in a case study for Gloucester City, New Jersey"

And in the discussion:

"Here, we have assessed the relationship between the probability of the event and the probability of flooding for a case study in Gloucester City (NJ, U.S.) by comparing the flood hazard with a 1% chance of happening in any given year (1% AEP) based on the event- and response-based approaches. We find that the 1% AEP water depth can be produced by different events in different parts of the city and that the AEPs of these events are often

much larger than 1%. This means that the relationship between the probability of the event and the probability of flooding does not follow a one-to-one relationship."

- I appreciate that the authors have pushed their code to GitHub. However, I visited <https://github.com/CoRE-Lab-UCF/MACH-Compound-Flooding/tree/main/Scripts> and was not able to figure out how to reproduce the results in the paper. It would be helpful for them to include a README file with instructions, a `main` script that runs the analysis, clear instructions for setting up the data, etc. I also was not able to figure out how the SFINCS code was run or where the input files for it are. It might be helpful to have a colleague check for reproducibility, as I would likely not be able to reproduce the results of this work.

"The reviewer is correct in noting that the final analysis codes have not yet been uploaded to GitHub. We are currently modifying the original scripts because the analyses were conducted using SFINCS outputs in a different file format than those provided in the Zenodo repository. Due to file size constraints, we converted the original SFINCS simulation outputs to a more compact NetCDF format before uploading them to Zenodo. Given that we do not anticipate users attempting to rerun all 5,000 SFINCS simulations and subsequent downscaling to 1-meter subgrid resolution, we are updating the analysis scripts to work directly with the NetCDF files available on Zenodo. Once these modifications are complete, we will ask a colleague to test the full reproducibility of the workflow, as recommended by the reviewer, and add a readme file to guide users.