This manuscript presents a modeling framework for evaluating the joint influence of non-tidal residuals (NTR), rainfall (RF), and mean sea level variability on coastal flooding in the Gloucester City area, using the SFINCS hydrodynamic model and a copula-based statistical approach. The topic is timely and relevant, and the study is generally well-structured with a strong emphasis on scenario-based risk quantification. However, several methodological choices—particularly regarding data selection, parameter thresholds, and model assumptions—require further clarification or justification. Issues such as the generalization of AORC performance, the treatment of tropical versus non-tropical events, and simplifications in the SFINCS physics raise concerns about the robustness and generalizability of the findings. Despite these limitations, the study offers valuable insights into compound flood risk assessment. Detailed comments are provided below, which I hope will be useful in clarifying and strengthening the manuscript:

We thank the reviewer for their constructive and insightful comments, which have helped enhance the overall quality of our manuscript. Below, we provide detailed responses to each point and outline how we plan to address them in the revised manuscript. The changes to the existing text are highlighted using track changes with line numbers in the original manuscript.

Page 5, Line 146: The sentence claiming that AORC has "higher accuracy" than other gridded rainfall datasets seems too general. For example, radar-based products like MRMS have been shown to perform as well as or better than AORC in some events, including Hurricane Harvey (e.g., Gao et al., 2021; Gomez et al., 2024). I suggest the authors either include MRMS in their comparison or rephrase the sentence to clarify that AORC's performance advantage may depend on the region or event type.

We thank the reviewer for bringing up this point and agree that MRMS products have demonstrated higher accuracy than many other datasets, while offering higher resolution (1 km and hourly for historical). However, our primary reason for selecting AORC over MRMS rainfall data is its longer temporal coverage (1979 to present), which allows the extraction of rainfall fields for a sufficiently large number of observed events required in our event-generation process. In contrast, MRMS precipitation data are only available from 2012 onwards (NOAA/NSSL, 2023), which limits their applicability for our framework.

We will revise the relevant section in the original manuscript as follows:

L135 : "We use both gridded RF data from the Analysis of Period of Record for Calibration (AORC) from 1979 to 2021 and hourly RF gauge data at the Philadelphia International Airport from 1900 to 2021 (Kitzmiller et al., 2018). Although radar-based quantitative precipitation estimates, such as the Multi-Radar Multi-Sensor (MRMS) products, often provide higher accuracy compared to other gridded rainfall products, their temporal coverage is relatively short (Gao et al., 2021; Gomez et al., 2024). We use AORC rainfall data because of its availability from 1979 onward and its demonstrated higher accuracy among products with similar temporal coverage, while offering hourly data with ~4 km spatial resolution (e.g., Hong et al., 2024; Kim and Villarini, 2022) AORC RF data has demonstrated higher accuracy compared to other gridded data sets while offering an hourly temporal resolution and ~4 km spatial resolution (e.g., Hong et al., 2024; Kim and Villarini, 2022). To leverage the long-term in-situ observations and obtain more robust results from the statistical analysis, we apply a bias correction to the hourly RF gauge data, to match with the hourly basin-average RF values calculated from AORC. The bias correction is performed using the

Additional Ref.

NOAA/National Severe Storms Laboratory. (2023). Multi-Radar Multi-Sensor (MRMS) system. National Oceanic and Atmospheric Administration. https://www.nssl.noaa.gov/projects/mrms/

Page 6, Line 160-166: In Section 4.1, several choices such as the 3-day pairing window for NTR and RF, the 5-day declustering period, and the 350 km radius for identifying TC events are not clearly explained. It would be helpful to clarify whether these are based on physical reasoning, prior studies, or simply assumptions made for this analysis. Providing brief justifications or references would improve transparency and reproducibility.

Both the choices of peak pairing window and declustering windows depend on the location-specific storm climatology and hydrologic response of the catchment. The choice of a 3-day pairing window for NTR and RF was based on the typical timescales over which storm surge (here represented by NTR) and rainfall peaks can occur in a compound flood event. We manually checked the time series of RF and NTR of POT (peaks over threshold) events and found that a 3-day pairing window was sufficient to capture both NTR and RF peaks of most events. Using a longer window could capture peaks from two different storm events and treat them as a single event in the bivariate frequency analysis. Furthermore, a 3-day pairing window has been found to be appropriate in similar previous studies (Couasnon et al., 2020; Kim et al., 2023; Maduwantha et al., 2024).

Different declustering windows have been used in previous studies to assure sampled extreme events are independent (e.g., 3 days (Haigh et al., 2016), 7 days (Santos et al., 2021), 10 days (Kim et al., 2023), and 14 days (Terlinden-Ruhl et al., 2025). Longer declustering windows are often adopted when the influence of river discharge is present, as its effects can last several days (Terlinden-Ruhl et al., 2025). Considering the duration of the storm events at our tide gauge location of interest, we use a 5-day window, since elevated NTR rarely lasted more than 5 days.

Previous studies have applied various search radii to identify TC events using a similar approach to ours (e.g., ~400 km (Kim et al., 2023) and 500 km (Towey et al., 2022)). In this study, we tested the sensitivity of the correlation between peak NTR and peak accumulated rainfall to the TC search radius, following Kim et al. (2023). Increasing the search radius captures more nearby TC tracks but also introduces events that are too distant to influence flood drivers at the study site, thereby reducing the overall correlation between rainfall and NTR of the TC sample. This is also influenced by TC intensity; for example, a Category 5 hurricane is more likely to generate significant storm surge even when passing far from the tide gauge, compared to a Category 1 hurricane. In our analysis, we considered all cyclones listed in the HURDAT2 database as TC events, regardless of their sustained wind speed or whether they were hybrid systems during part of their lifetime. We therefore selected a 350 km search radius, as it provided a higher correlation between drivers while still retaining a reasonable number of TC events in the sample.

To support these choices in the main manuscript, we add the following paragraph:

L 415: "A detailed description of the procedure for estimating the joint probability distribution applied in this study is provided in Maduwantha et al. (2024). When applying the two-way sampling to extract POT events, we use a 3-day pairing window to capture peak NTR and RF, following similar studies (Couasnon et al., 2020; Kim et al., 2023). We also manually checked the RF and NTR time series of POT events and found that a 3-day window was generally sufficient to capture both peaks. To ensure event independence within the POT samples, previous studies have applied various declustering windows (e.g., 3 days (Haigh et al., 2016), 7 days (Santos et al., 2021), 10 days (Kim et al., 2023), and 14 days (Terlinden-Ruhl et al., 2025)). Longer declustering windows are often adopted when the influence of river discharge is present, as its effects can persist for several days (Terlinden-Ruhl et al., 2025). In this study, we use a 5-day declustering window (2.5 days before and after the event peaks), as elevated NTR rarely lasts more than 5 days at the Philadelphia tide gauge location. Previous studies have applied various search radii to identify TC events, such as ~400 km (Kim et al., 2023) and 500 km (Towey et al., 2022). In this study, we tested the sensitivity of the correlation between peak NTR and peak accumulated rainfall to the TC search radius, following Kim et al. (2023). Increasing the search radius captures more nearby TC tracks but also introduces events that are too distant to strongly influence flood drivers at the study site, thereby reducing the overall correlation between rainfall and NTR of the TC sample. We selected a 350 km search radius, as it leads to a higher correlation between flood drivers while still retaining a reasonable number of TC events in the sample."

Additional Refs:

Towey, K. L., Booth, J. F., Rodriguez Enriquez, A., and Wahl, T.: Tropical cyclone storm surge probabilities for the east coast of the United States: a cyclone-based perspective, Nat. Hazards Earth Syst. Sci., 22, 1287–1300, https://doi.org/10.5194/nhess-22-1287-2022, 2022.

Couasnon, A., Eilander, D., Muis, S., Veldkamp, T. I. E., Haigh, I. D., Wahl, T., Winsemius, H. C., and Ward, P. J.: Measuring compound flood potential from river discharge and storm surge extremes at the global scale, Nat. Hazards Earth Syst. Sci., 20, 489–504, https://doi.org/10.5194/nhess-20-489-2020, 2020.

Terlinden-Ruhl, L., Couasnon, A., Eilander, D., Hendrickx, G. G., Mares-Nasarre, P., and Antolínez, J. A. Á.: Accelerating compound flood risk assessments through active learning: A case study of Charleston County (USA), Nat. Hazards Earth Syst. Sci., 25, 1353–1375, https://doi.org/10.5194/nhess-25-1353-2025, 2025.

Page 8, Line 217-228: Could the authors clarify the physical justification for uniformly scaling entire NTR and RF time series based solely on peak values? For example, does this approach preserve key timing or intensity ratios in cases with asymmetrical hydrographs or localized RF bursts?

We thank the reviewer for raising this important point. In flood hazard modeling, the approach of scaling the observed flood driver time series to match design peaks and then considering them as design hydrographs/hyetographs with known return periods is already widely used. Examples include applications for NTR (Kim et al., 2023; Amorim et al., 2025), RF (Li et al., 2020; Kim et al., 2023), and streamflow (Yue et al., 2002). The assumption is that uniform scaling (also referred to as "same frequency amplification" in some studies) of the flood-driver time series produces another realistic time series.

Even with asymmetrical NTR time series/hyetograph shapes, our linear scaling approach preserves the relative timing and intensity ratios of sub-peaks to the main peak, which is a key factor in determining the resultant flooding. Linearly scaling a storm-tide hydrograph where tidal influence is more visible would create an unrealistic total water level time series. However, here we utilize NTR time series for scaling and then add tides and MSL to the scaled NTR time series in a consistent manner with their seasonal variability accounted for. Although we cannot validate synthetic events on a one-to-one basis against real events, we compared statistical properties of key time series characteristics, such as durations, peaks, intensities, and lag times, between observations and synthetic events (Fig. 9). These comparisons show that the synthetic events preserve the dependencies found in the observed events.

We acknowledge that scaling smaller events to much higher peaks can still produce unrealistic storm events. To mitigate this, we select observed events whose peak values are close to the target peak, so that the scaling factor remains close to 1 in most cases (with medians, for NTR: 1, for RF: 0.99), and we impose an upper limit on the scaling factor when generating events.

To clarify this further, and based on another reviewer's comments, we will revise the discussion section as follows:

L 493: "One key assumption of the framework is that uniform scaling (also referred to as "same frequency amplification") of flood-driver time series creates a realistic compound event. This approach has been widely adopted in previous studies to construct design hydrographs and hyetographs (e.g., Serafin et al., 2019; Moftakhari et al., 2019; Zellou and Rahali, 2019; Kim et al., 2023; Liu et al., 2024; Xu et al., 2024). However, assessing whether each generated synthetic storm event is physically plausible is not possible. Instead, we validated the framework by comparing statistical properties of key time series characteristics between observations and synthetic events (Fig. 9). Another ~~One key~~ limitation of the proposed framework is that certain characteristics of synthetic events, such as RF duration and lag times, are limited to the observed values. To generate more diverse lag times, the observed lag times could be fitted to a parametric distribution (or alternatively to a copula that accounts for the dependence between peak values and lag times) and sample lag times from the fitted distribution during the event generation process. This would introduce unobserved lag times into the synthetic events, enhancing their diversity. Additionally, the stratification of POT events utilizes a simple yet commonly applied approach (e.g., Kim et al., 2023; Maduwantha et al., 2024) as discussed in Section 4.1. However, this method may fail to capture all TCs, particularly those that produce significant RF and storm surges from distances greater than 350 km. Such events are classified as non-TC events, meaning the analysis in Section 4.2 may not fully reflect the true characteristics of TC and non-TC events. "

Page 10, Line 273-275: The use of SFINCS is well-suited for handling large scenario sets; however, two model limitations warrant further discussion. First, SFINCS does not explicitly model nonlinear tide–surge interactions, which can influence the timing and amplitude of water levels in estuarine environments (e.g., Arns et al., 2020; Dullaart et al., 2023). Second, the omission of advection in the local inertia formulation may affect surge dynamics in narrow tidal channels like those surrounding Gloucester City. I recommend the authors provide a brief sensitivity analysis or comparison illustrating the impact of including vs. excluding the advection term, as SFINCS offers both options (Leijnse et al., 2021).

We agree with the reviewer that the tide-surge interaction can produce important effects on estuarine water levels. However, the limitations of SFINCS simulating these effects do not affect our analyses since the boundary condition of our model is placed along the Delaware, near the coastline of our study site. Therefore, we are not simulating the generation and propagation of the surge into the estuary, in which interaction between tides and surges occurs. Rather, we account for the tide-surge interaction within the presented framework to generate the boundary conditions for the flood model. The tide-surge interaction is accounted for by considering the dependency between NTR peaks with tidal high water peaks (i.e., the time difference between the NTR peak and the peak of the tide). As mentioned in L: 235 in the original manuscript, we use the observed time difference between peak NTR and the subsequent high tide of the sampled NTR time series to combine it with the sampled tidal signal. Therefore, the synthetic events reflect the effect of tide surge interaction, which was observed in the historical events. Fig. 1 below shows a histogram of the time difference between the peak NTR and the subsequent high tide of observed events (orange) and synthetic (blue) events.
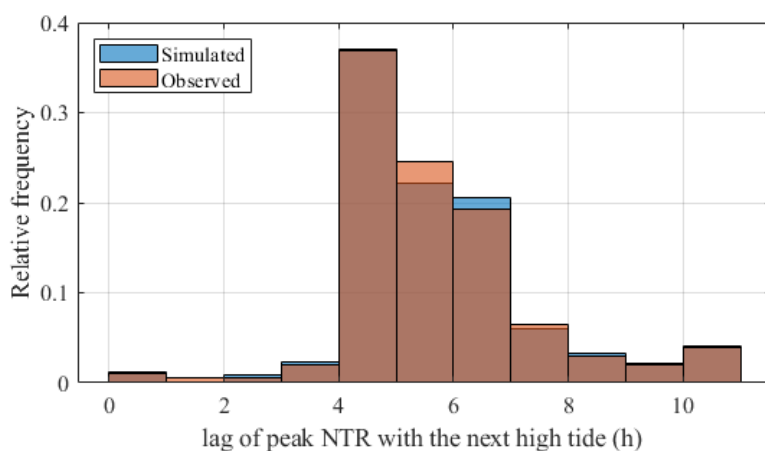

Fig.1 The distribution of time difference between the peak NTR and the next high tide of observations (orange) and simulations (blue). Positive values indicate the peak NTR occurred before the next high tide.

As requested by the reviewer, to assess the potential impact of omitting the advection term in the flood model runs, we conducted an additional set of simulations with the advection term enabled. The results (see Fig.2 below) show negligible differences in resultant flooding compared to our original simulations without advection, even in the narrow tidal channels surrounding Gloucester City. Given these results and the fact that the main conclusions of our study are based on the statistical framework of boundary condition generation rather than fine-scale hydrodynamic sensitivities, we are confident that the omission of the advection term does not largely affect the main results and conclusions of the paper.

To further clarify this and address a related concern raised by another reviewer, we will add the following sentences.

L 465: "At the Philadelphia tide gauge, peak NTR often occurs 4–7 hours before the next high tide (see Fig.S2 in the supplementary material). To account for this, we combine scaled NTR time series with tide time series using the observed lag between peak NTR and subsequent high tide of the sampled NTR event (see Section 4.3.3)."

We will revise the following sentence.

L 274: The model is run with the advection term neglected, solving the local inertia equations (we tested the sensitivity of the results when the advection term was enabled, but changes were negligible).
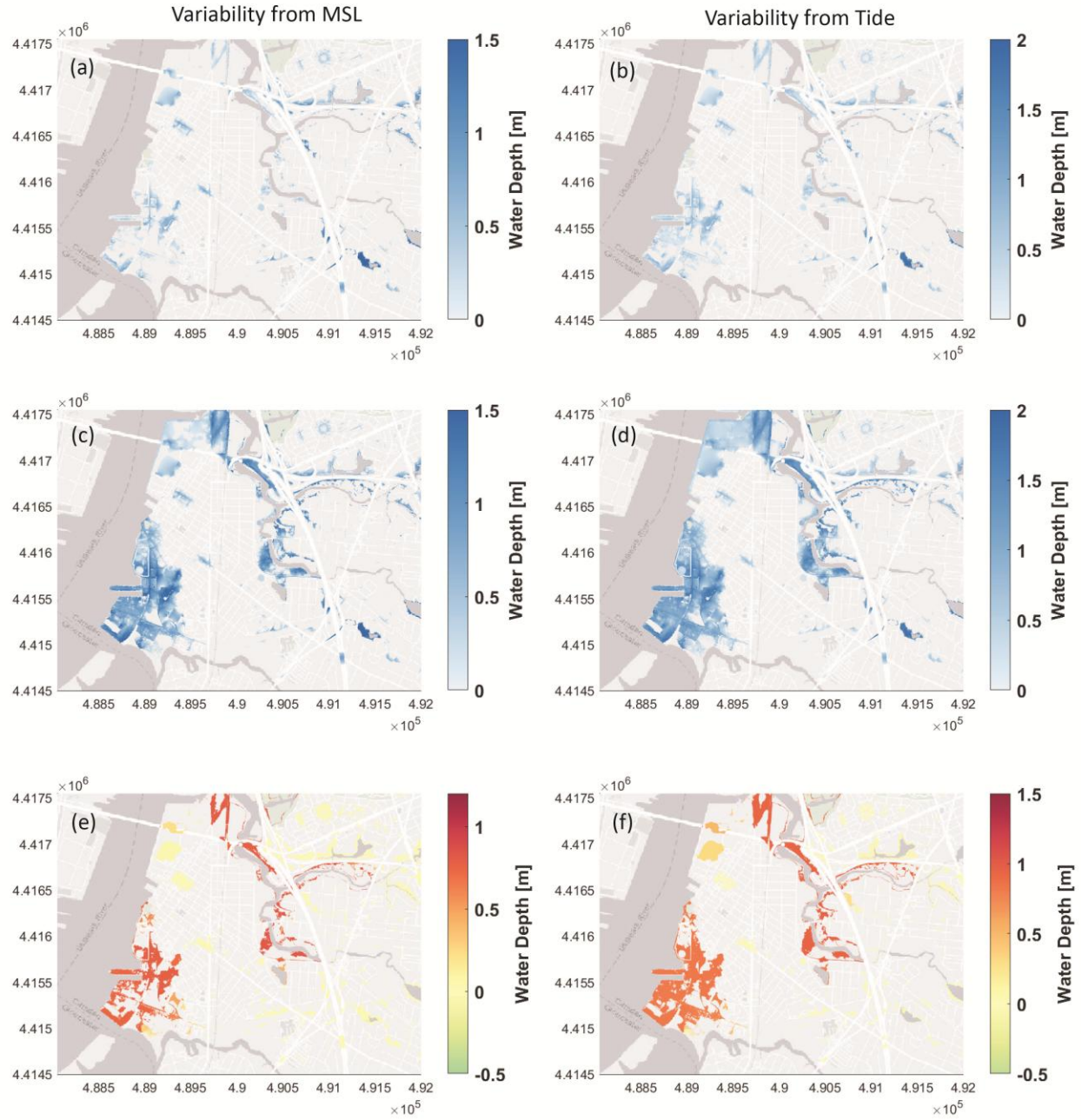
Fig 2. Changes in flood depths associated with variability when the advection term is enabled, in MSL (left) and Tide (right) of a selected 0.01 AEP most-likely event. (a) When considering the lowest 30-day MSL, (b) when considering the tidal segment with the lowest high tide, (c) when considering the highest 30-day MSL, (d) when considering the tidal segment with the highest high tide, (e) the difference between (a) and (c), (f) difference between (b) and (d). (X, Y coordinates system: UTM- zone 18N).

Page 11, Line 300-301: The authors use a fixed NTR threshold of 0.63 m to yield ~5 exceedances per year, which is reasonable and aligns with past compound flood studies. However, the threshold selection could be strengthened by applying one of several recent automated, data-driven approaches developed for POT analysis, such as the Sequential Goodness-of-Fit method (Bader et al., 2018), the Extrapolated-Height Stability method (Liang et al., 2019), the L-moment Ratio Stability method (Silva Lomba & Fraga Alves, 2020), or the comparative multi-method approach applied in a coastal flood design context by Radfar et al. (2022).

We thank the reviewer for the comment and agree that more advanced approaches, such as those mentioned, offer more robust methods for objectively determining thresholds. However, these methods are primarily designed to optimize the fit of the distribution tail. In the context of compound flooding, extreme flood events are not always driven by extreme peak flood driver combinations. With a suitable combination of timing, duration, intensities, and tidal conditions, extreme flooding can also occur under moderate peak NTR and RF combinations (Santamaria et al., 2025). To generate synthetic storm events with moderate NTR and RF peaks, the copulas must therefore be fitted using POT events with flood potential and still large enough to be well described by a GPD (Generalized Pareto Distribution). Accordingly, we selected a threshold that balances both, small enough to capture moderate conditions, yet large enough to represent the upper tail of the distribution. Rather than applying more robust threshold selection methods, we followed the approach used in similar studies and considered the average number of historical flood events to guide our choice.

We will add the following text to the manuscript.

L 161: Recent data-driven threshold-selection methods such as the Sequential Goodness-of-Fit method (Bader et al., 2018), the Extrapolated-Height Stability method (Liang et al., 2019), L-moment ratio stability (Silva Lomba & Fraga Alves, 2020), and a comparative multi-method approach (Radfar et al., 2022) provide robust POT thresholds but primarily optimize tail fit. Extreme compound flood events are not necessarily generated by extreme flood driver peaks. With favorable timing, duration, and tidal conditions, extreme flooding can occur even under moderate flood-driver conditions (Santamaria et al., 2025). Therefore, we set thresholds on NTR and RF, targeting an average of five exceedances per year to extract POT events, following similar studies (Jane et al., 2020; Kim et al., 2023).

Page 12, Line 328-336: While the authors maintain stratification for joint probability estimation, they combine TC and non-TC time series for event generation based on overlapping confidence intervals and similar time series shapes. Given the well-established physical differences between tropical and extratropical systems (precipitation structures, spatial scales, storm tracks), could the authors clarify how confident they are that this approach adequately preserves the distinct characteristics of these storm types?

We agree that it is well established that tropical cyclones, extratropical cyclones, and other non-cyclonic locally generated systems differ in terms of physical properties. However, for our study area, factors such as the limited sample size of TC events, the small size of the catchment, and the hydrologic response of

the upstream Delaware River reduce the extent to which the inherent large-scale storm characteristics are reflected in the time series of NTR and basin-average rainfall.

The overlapping confidence intervals of Kendall's tau, along with comparisons of distribution parameters fitted to the time series characteristics of RF and NTR, indicate no statistically significant differences between the TC and non-TC samples (see Fig. 6). Similarly, the shapes of the NTR and basin-average RF time series produced by TCs are similar to those generated by non-TC events (see Fig. 7). Based on these results, we conclude that generating event time series (i.e., water-level hydrographs and RF hyetographs) separately for the two storm types would yield results similar to those obtained without differentiating between them. However, we acknowledge that in other locations, significant differences in time series characteristics between TC and non-TC samples may exist, especially when model domains are large. In those cases, we recommend conducting event generation separately for each storm type. This limitation is already discussed in the original manuscript (Section 6, Lines 440–455).

Additionally, given the limited number of TC events, how do the authors assess whether their analysis has sufficient statistical power to detect meaningful differences? Would alternative approaches like physics-based conditioning (e.g., storm track or seasonal constraints) potentially better preserve known meteorological distinctions while addressing sample size limitations?

We thank the reviewer for this thoughtful comment. We acknowledge that the limited number of TC events in our dataset reduces the statistical power to detect subtle differences in time series characteristics between TC and non-TC samples. As noted in the manuscript (Section 6, Lines 440–455), this is recognized as a limitation of our application.

Methods involving physics-based models to generate synthetic flood drivers from TC events (often using synthetic TC tracks) can better preserve the inherent spatio-temporal characteristics of storms (e.g., Emanuel et al., 2006; Gori et al, 2020). However, these methods are computationally demanding, as they require both hydrologic and hydrodynamic modeling, which limits their ability to generate large numbers of events. We agree that such approaches could help address the limited TC sample size in our study. Implementing this type of framework is beyond the scope of the current work, but we plan to incorporate it in future studies.

We plan to add the following sentences to the main manuscript.

L 452: "One option to overcome the limited TC sample size in our analysis is employing physics-based models to generate time series of flood drivers from synthetic TC tracks (e.g., Emanuel et al., 2006; Gori et al., 2020). We plan to explore this in future work."

Page 20, Figure 10: The authors demonstrate substantial flood depth due to MSL and tidal variability using a single most-likely 0.01 AEP event. While this effectively illustrates the potential importance of these factors, could the authors comment on whether this sensitivity pattern is representative across different event types and return periods?

The most-likely 0.01 AEP (100-year return period) event was selected to show an illustrative example to highlight the potential influence of MSL and tidal variability on resultant flooding. While traditional approaches often neglect this variability in their modeling frameworks, we explicitly account for it. We use the 0.01 AEP, as it is commonly adopted in flood hazard mapping for planning purposes (e.g., FEMA Special Flood Hazard Areas; FEMA, 2020).

We agree that this single event may not fully represent the sensitivity patterns across the full range of event types and return periods. In a separate study (Santamaria et al., 2025, under review), we simulated flooding for the same study site using a set of 5,000 synthetic storm events that were generated from the developed statistical framework. Resultant flooding shows a wide range of variability, even among events with similar joint return periods. However, attributing this variability to a single factor like MSL or tides is challenging due to the complexity of their interactions. This is discussed in detail in Santamaria et al. (2025).

For discussing this in the manuscript, we will modify the section starting from L 488 as follows:

L488: "Analyzing only the most likely event, even if it appears to be the most plausible based on observations, does not capture the range of flood levels that could be generated by different combinations of flood drivers (i.e., NTR and RF) with different time series properties. Therefore, the flood model simulations presented here are aimed at evaluating the importance of explicitly accounting for the variability of MSL and tides, and not to produce comprehensive probabilistic flood maps. In a separate study (Santamaria et al., 2025), we simulated flooding of 5,000 synthetic storms at this site and found large variation in resultant flooding, even for events with similar joint return periods. However, attributing this variability to a single factor like MSL or tides is challenging due to the complexity of their interactions."

Additional Ref.

Federal Emergency Management Agency (FEMA). (2020). Flood Insurance Study Guidelines: Guidelines and Specifications for Flood Hazard Mapping Partners. https://www.fema.gov/flood-maps/guidance-partners/guidelines-specifications

Santamaria-Aguilar, S., Maduwantha, P., Enriquez, A. R., and Wahl, T.: Large discrepancies between event- and response-based compound flood hazard estimates, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-1938, 2025.

Additionally, given that the most pronounced effects occur along the Delaware River and Newton Creek boundaries, could the authors discuss whether the model's spatial resolution, boundary condition placement, or coastal setup might be influencing the magnitude of these sensitivities?

In the flood model, the coastal boundary is placed along the Delaware River to match the water levels at the Philadelphia tide gauge and includes the entire catchment of Newton Creek to capture all contributing runoff. We used the Coastal National Elevation Database (CoNED) digital elevation model with a horizontal resolution of 1 meter and a vertical accuracy of 10 cm. We use the subgrid approach of SFINCS with a dual resolution of 10m and 1m. This level of detail is relatively high compared to many other compound flood hazard assessments, which often use coarser grids. Given the size of the Delaware River and surrounding creeks, we believe the model adequately resolves hydrodynamics and that the observed sensitivities reflect physical processes rather than numerical limitations.

Refs.:

Arns, A., Wahl, T., Wolff, C., Vafeidis, A. T., Haigh, I. D., Woodworth, P., Niehüser, S., & Jensen, J. (2020). Non-linear interaction modulates global extreme sea levels, coastal flood exposure, and impacts. Nature Communications, 11, 1918.

Dullaart, J. C. M., Muis, S., de Moel, H., Ward, P. J., Eilander, D., & Aerts, J. C. J. H. (2023). Enabling dynamic modelling of coastal flooding by defining storm tide hydrographs. Natural Hazards and Earth System Sciences, 23, 1847–1862.

Leijnse, T., Dazzi, S., Yu, D., & Bates, P. D. (2021). Efficient coastal flood hazard mapping with a 2D non-inertia model. Coastal Engineering, 170, 103994.

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection in extreme value analysis via goodness-of-fit tests with adjustment for false discovery rate. Annals of Applied Statistics, 12(1), 310–329.

Liang, B., Shao, Z., Li, H., Shao, M., Lee, D., 2019. An automated threshold selection method based on the characteristic of extrapolated significant wave heights. Coast. Eng. 144, 22–32.

Radfar, S., Shafieefar, M., & Akbari, H. (2022). Impact of copula model selection on reliability-based design optimization of a rubble mound breakwater. Ocean Engineering, 260, 112023.

Silva Lomba, J., Fraga Alves, M.I., 2020. L-moments for automatic threshold selection in extreme value analysis. Stoch. Environ. Res. Risk Assess. 34 (3), 465–491.

Gao, S., Zhang, J., Li, D., Jiang, H., & Fang, Z. N. (2021). Evaluation of multiradar multisensor and stage IV quantitative precipitation estimates during Hurricane Harvey. Natural Hazards Review, 22(1), 04020057.

Gomez, F. J., Jafarzadegan, K., Moftakhari, H., & Moradkhani, H. (2024). Probabilistic flood inundation mapping through copula Bayesian multi-modeling of precipitation products. Natural Hazards and Earth System Sciences, 24(8), 2647-2665.