The manuscript “Generating Boundary Conditions for Compound Flood Modeling in a Probabilistic Framework” by Maduwantha et al. introduces a statistical framework designed to generate many synthetic but physically plausible compound events, including storm-tide hydrographs and rainfall fields, which can serve as boundary conditions for dynamic compound flood models. The framework is later applied to the case of Gloucester City in New Jersey.

The topic of the manuscript is highly relevant for quantifying flood hazard, particularly water depth resulting from the joint occurrence of storm surge and rainfall. However, the proposed framework requires further clarification and additional information to assess its validity better and ensure reproducibility by others.

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.

The novelty of the proposed framework should be better highlighted. If I understand correctly, the essence of the proposed framework is to select joint events of NTR and rainfall from historical observations, and then fit a bivariate copula to generate “unseen” pairs. Pairs are used to amplify historical time series of NTR and rainfall over a short period of time with a temporal resolution consistent with the one required by the hydrodynamic model. An almost identical workflow was proposed by Xu, H et al (2024) "Combining statistical and hydrodynamic models to assess compound flood hazards from rainfall and storm surge: a case study of Shanghai", Hydrol. Earth Syst. Sci., 28, 3919–3930, https://doi.org/10.5194/hess-28-3919-2024. How does this framework differ from Xu et al 2024? How does this framework differ from previous studies? The novelty is hidden in the introduction.

We thank the reviewer for raising this important point. While our framework shares some conceptual similarities with Xu et al. (2024), it introduces several methodological and practical novelties that distinguish it from previous studies.

We agree that the general workflow of (i) modeling the joint probability distribution of flood drivers, (ii) sampling unseen pairs from the fitted distribution, (iii) assigning corresponding time series, and (iv) propagating them through a compound flood model is not unique to our study. This general approach has been adopted in several previous works (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), and we do not claim novelty in this regard. These studies, including Xu et al. (2024), were primarily tailored to generate design events with specified joint return periods (e.g., 100-yr,50-yr) of flood-driver peaks. This only supports an “event-based” hazard analysis, where one or a few design events are simulated through flood models, and the joint probability of the flood drivers is assumed to be the same as the probability of the flood response.

In contrast, our framework is designed to generate a large and diverse set of synthetic storm events. These are then propagated through the hydrodynamic model to produce a broad ensemble of flood responses. Extreme value analysis is then applied directly to the resultant flood depths, thereby supporting a response-based flood hazard assessment. This shift from generating boundary conditions for event-based to response-based analysis is a key novelty of our work, as it allows us to capture the full spectrum of possible flood outcomes, rather than relying solely on the flood response of a single design event or a small subset of events. Additionally, our framework can also be used to support event-based hazard modelling by generating many diverse synthetic storm events with a unique (or known) joint return period. In addition to the objective differences, we have listed other key advancements of our modeling framework over Xu et al. (2024):

1. Accounting for distinct storm types (populations), thereby capturing the unique dependence structures of flood-driver peaks
2. Generation of total water level time series by combining scaled NTR, tides, and MSL consistent with seasonal variations
3. Offers a more flexible framework, by incorporating different rainfall accumulation times and employing a two-way conditional sampling method that enables modeling of a wide range of storm events, including non-extremes

To better highlight the novelty of our work, we have revised the following paragraphs in our introduction.

L 75: Rescaling of total water level (or non-tidal residuals (NTR) time series) of observed events is another deterministic approach that leverages observed event data (Dawson et al., 2005; Kim et al., 2023; Xu et al., 2024).

L 100: "Among the applications of uniform scaling of flooding drivers, Xu et al. (2024) applied the "same frequency amplification" method to construct a 200-yr storm surge hydrograph and rainfall hyetograph for their flood simulations. However, their approach was limited to point rainfall and assumed uniformly distributed rainfall across the catchment. Kim et al. (2023) proposed a framework for generating synthetic time series of RF fields and associated NTR by scaling time series of observed TC events. The framework was used to capture different spatial patterns of RF fields as this aspect was shown to significantly contribute to compound flood hazard (e.g., Gori et al., 2020). However, their analysis exclusively focused on TC events and the methodology only produces NTR time series and does not extend to producing complete storm-tide hydrographs; this is because it was applied to the Texas coast where the tidal range is small, and where compound flooding is primarily driven by TCs. Other types of storms can produce compound flooding in many other areas and tides often contribute significantly to the resulting still water levels.

The existing statistical approaches that generate time-varying boundary conditions for dynamic compound flood models are primarily intended to construct design events with specified joint return periods (e.g., 50-yr, 100-yr) (Serafin et al., 2019; Moftakhari et al., 2019; Zellou and Rahali, 2019; Kim et al., 2023; Liu et al., 2024; Xu et al., 2024). This method supports the "event-based" flood hazard analysis, where a single (or only a few) synthetic event with known joint return periods is simulated through a flood model, and it is assumed that the (joint) probability of the flood drivers directly translates into the probability of the flood response. However, this neglects the range of potential different flooding scenarios that may arise from variations in temporal and spatial patterns, differences in the relative timing of multiple flood drivers, and other complex interactions (for example, tide-surge interactions).~~Additionally, their approach is~~

~~designed for creating specific design events with known joint return periods of the peaks of the flood drivers, i.e., it supports "event-based" flood hazard analysis where one or few events with a given return period are routed through an inundation model and it is assumed that the joint probability of the drivers translates to the probability of the flood response~~. For a more complete characterization of flood hazard and risk, the flood response of many synthetic events needs to be modeled, allowing the derivation, for example, of return levels of flood depth at all points within the model domain (i.e., "response-based" flood hazard analysis)."

In lines 502-503, the Authors say that "Although measures are taken to prevent the generation of physically unrealistic events (see Section 4.3), it cannot be fully ruled out." If this statement is true, then the Authors cannot claim that the framework generated physically plausible compound events (Abstract - Lines 16-17). This is quite an important point, and the Authors need to be transparent about the potential of the framework to generate physically plausible events or not.

We thank the Reviewer for the comment. Our claim in the Abstract (Lines 16–17) that the framework generates physically plausible compound events comes from a statistical and process-based perspective since the framework explicitly preserves the dependence structures of peaks of flooding drivers and adheres to ranges and interdependencies of observed timeseries characteristics, as described in Section 4.3.

However, we acknowledge that we cannot fully eliminate the risk of generating extreme storm events that may be physically "implausible", at least under current climate conditions (we believe this is not unique to our approach but true for most statistical methods that generate extreme unseen environmental data). For instance, when deriving peak NTR–RF combinations from the multivariate statistical model, the use of unbounded marginal distributions can yield large peaks that might be physically impossible to occur. However, the absence of such events in the observational record does not in itself indicate that the generated synthetic events are unrealistic, as the record length may be insufficient to capture their occurrence. Extremes are often wrongly perceived as impossible until they happen. Moreover, even with the assumption that uniform scaling produces realistic NTR and RF time series individually, it still requires the additional assumption that their combined representation is also physically realistic. Ideally, a direct one-to-one validation of synthetic events against observed events would provide a rigorous test of these assumptions, but such validation is impossible. To evaluate the performance of our framework, we compared the statistical properties of key time-series characteristics, including durations, peaks, intensities, and lag times between observations and synthetic events (Fig. 9).

To further acknowledge these limitations, along with comments from another reviewer, we will add the following to the discussion:

L 494: "One key assumption of the framework is that uniform scaling (also referred to as "same frequency amplification") of flooding-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 validate the framework by comparing statistical properties of key time series characteristics between observations and synthetic events (Fig. 9).

The data selection procedure and its effects on the results need further clarification.

We appreciate the Reviewer’s comment and agree that the data selection procedure requires a clearer explanation. The framework is designed to generate high-resolution rainfall fields over the catchment and storm-tide hydrographs at the coastal boundary. For water levels, we use the nearest tide gauges to the study site (Philadelphia (St. ID: 8545240) and Philadelphia Pier 11-north (St. ID: 8545530)). Although some rainfall gauges exist near Gloucester City, the Philadelphia Airport rain gauge provides the longest hourly record (from 1900). For rainfall fields, radar-based products such as MRMS offer higher resolution and accuracy compared to other gridded RF data (1 km, hourly), but their temporal coverage is limited (MRMS: from 2012; Stage IV: from 2002). We therefore use AORC precipitation data (1979 to present), which balances high resolution (4 km, hourly) with sufficient temporal coverage (from 1979) to extract observed RF fields from many events. For identifying TC-induced POT (peak over threshold) events, we use HURDAT2, the most widely used and reliable TC catalog for the Atlantic basin.

While using different datasets may introduce some changes to the joint probability distributions and resulting synthetic storm events, we emphasize that our choices represent the best available balance of resolution, accuracy, and temporal coverage. Therefore, we believe our results are robust. A full sensitivity analysis of synthetic storms to different datasets would be valuable, but it's beyond the scope of this study.

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

L 135: “For the statistical analysis, we consider RF and NTR as flood drivers. We use hourly water level data from the nearest tide gauges to the study site from the ~~the~~ National Oceanic and Atmospheric Administration (NOAA):~~ tide gauges at~~ Philadelphia (St. ID: 8545240) and Philadelphia Pier 11-north (St. ID: 8545530).

L 144: “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 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)~~.”

First, synchronous NTR are selected, and later on astronomical tide and mean sea level are added. Did the Authors consider using the concept of skew surge? If not, why? In addition, tide-surge interaction is mentioned but not really discussed. How relevant is it? Would the concept of skew surge solve it?

We thank the Reviewer for the comment. We agree that skew surge would implicitly account for tide–surge interactions and offers a straightforward way to combine it with Tides. However, as skew surge provides only two values per tidal cycle (semidiurnal in the Philadelphia region), it cannot capture the full temporal evolution of NTR at hourly scales or the timing between flood drivers, which are critical for compound flooding (Gori et al., 2021). Previous studies (e.g., Terlinden-Ruhl et al., 2025) assumed a constant skew surge over a 12-hr period, which we consider insufficient to reproduce realistic hydrographs for our analysis. Additionally, estimating the duration of elevated NTR is difficult from skew surge alone, but it is a key time series characteristic that we want to examine and make sure it is consistent in the synthetic events.

At the Philadelphia tide gauge, peak NTR often occurs 4–7 hours before the next high tide (see Fig.1 below). To account for this, we combined 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), ensuring that synthetic events preserve the same dynamics.

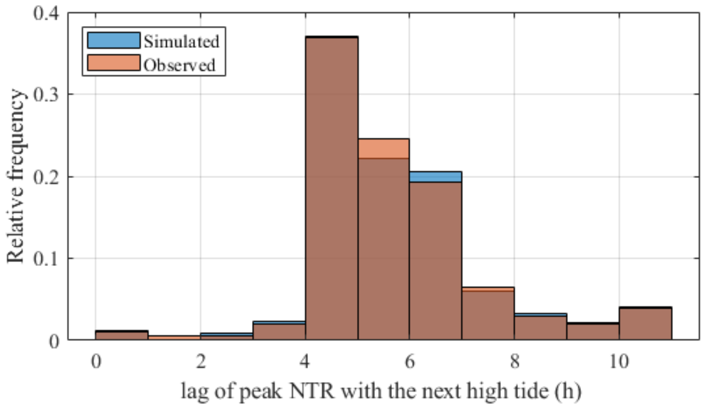

Fig.1 The distribution of time difference between the peak NTR and the next high tide of observations (orange) and simulations (blue). Positive values indicate the peak NTR occurred before the next high tide.

In addition to the explanation in the main manuscript in Section 4.3.3, we plan to further clarify this in the discussion by adding the following:

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)."

Second, multiple rainfall measurements are considered. However, it is unclear how such measurements are aggregated and how this aggregation affects the dependence between NTR and rainfall, and so the fitted copula. Moreover, which rainfall measurement is used as a reference for the lag time between peak surge and peak rainfall?

A detailed description of the rainfall (RF) data treatment is provided in our previously published work (Maduwantha et al., 2024), but we summarize the main steps in the Methods section. Rain gauges measure very local weather conditions. However, the assumption that such point RF quantities are uniformly distributed over the entire catchment could lead to mischaracterization of the flood hazard potential. Therefore, we apply a bias correction to the hourly RF gauge data to match the hourly basin-averaged RF quantities calculated from AORC. The quantile mapping method is used for the bias correction, fitting both hourly measured gauge data and hourly AORC basin-averaged data to gamma distributions. This procedure allows us to increase the POT sample size, thereby reducing uncertainty in the marginal distributions. As discussed in Maduwantha et al. (2024), the correlation between peak NTR and rainfall shows nonstationarity with an increase in recent decades. Hence, we estimated the copula parameters using POT events from the last 30 years, ensuring that the dependence structure reflects current climate conditions, and we avoid potential underestimation of the flood hazard.

Lag times and other time-series characteristics (e.g., durations, peaks, and intensities) are derived directly from the historical events, using AORC rainfall fields combined with tide gauge records (those are then scaled to match the target peaks sampled from the NTR-RF joint distribution). This part is explained in the original manuscript from L 185 to L 190.

To further clarify how the aggregation of RF gauge data and AORC data affects dependence, we will revise the following paragraph:

L 147: "Rain gauges measure highly localized rainfall. Assuming that these point measurements occurred uniformly distributed across the entire catchment can misrepresent the compound flood hazard. To address this, we apply a bias correction to the hourly gauge data so it matches the basin-averaged hourly rainfall estimates derived from AORC. This correction is performed using the quantile mapping method, in which both the gauge-based and AORC-based rainfall distributions are fitted to gamma functions. ~~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 quantile mapping method, fitting both the hourly measured gauge data and the hourly AORC basin-average data to gamma distributions~~ (for more details see Maduwantha et al. (2024))."

We plan to add the following:

L 171: "Maduwantha et al. (2024) found significant non-stationarity in Kendall's τ between peak NTR and RF over the analysis period. To capture most recent climate conditions and avoid underestimating compounding effects, we model dependence using only the last 30 years of data."

Finally, the distinction between TC and non-TC leads to copulas with different asymmetries. How do the Authors justify such differences? What happens when the TC and non-TC are combined together? I would say this is mostly relevant in paragraph 5.3 "Event Generation Process". How do the Authors know that the 106-year event (which is also quite an interesting number!) corresponds to a TC? Given the length of the data, the 106-year event is inferred from the copula and not observed. Regardless of how the Authors track whether it is a TC or a non-TC, how sensitive are the results to the type of event? For example, what is the difference between a TC and non-TC event with the same return period? What about in terms of the drivers' magnitude and water depth? I suggest adding some sensitivity analysis to the assumptions made, including the lag time between peak rainfall and peak surge.

We thank the Reviewer for the comment. A key objective of our framework is to account for the distinct dependence structures of different storm types (TCs vs. non-TCs), which are often overlooked when all POT events are treated as a single population. In our analysis, peak NTR and peak RF were found to be more strongly correlated in TC-induced events than in non-TC events, in line with previous studies (e.g., Kim et al., 2023). Therefore, as the reviewer mentioned, TC and non-TC lead to copulas with different asymmetries. We derive 5,000 combinations of peak NTR and RF by sampling from the fitted copulas to each sample such that the relative proportion of extremes is consistent with their historical occurrence. 351 combinations from the copulas fitted to the TC sample and 4,649 from the fitted copulas to the non-TC sample.

The illustrative "106-year event" presented in the paper was drawn from the TC sample (one of the 351 combinations). Since the events are generated from the fitted copulas through the Monte Carlo approach, the resulting combinations do not correspond to standard return periods (e.g., exactly 50- or 100-year), instead, reflect the joint probability distribution of peak NTR and RF.

Based on the analysis we conducted in section 4.2, we found no significant differences in statistical features nor various time series characteristics in NTR and RF from TC events and non-TC events. Given these results, we conclude that generating the event time series (i.e., water level hydrographs and RF hyetographs) separately for the two storm types would produce similar results compared to the ones we derive without stratifying. Yet, we acknowledge that with sufficient data, the most effective approach would be to use observed time series of flood drivers from TC events exclusively for generating synthetic TC events, while using those from non-TC events separately to generate synthetic non-TC events.

The NTR and RF peak magnitudes of an event with the same joint probability can differ substantially between storm types and depending on where it lies along the probability isoline. For example, the selected event (NTR = 1.75 m; 18-h RF = 80 mm) has a ~106-yr joint return period when joint probability distributions of TC and non-TC storms are combined. The same event has a ~111-yr joint return period in the TC sample and ~2431-yr joint return period in the non-TC sample. Conversely, an event with relatively lower peaks (NTR = 1.0 m; 18-h RF = 50 mm) has a ~5-yr joint return period in the non-TC sample but ~15-yr joint return period in the TC sample. In conclusion, small to moderate events are more frequent in the non-TC storms, whereas the most extreme events are more likely generated by TCs. This aspect is discussed in detail in Maduwantha et al. (2024).

As discussed in L 464, lag times between peak NTR and peak RF can vary considerably, and more extreme events tend to exhibit shorter lag times (see Figs. 4 (e) and 4 (i)). However, this behavior is evident in both TC and non-TC events. To reflect this behavior in synthetic events, we select nearby historical events for scaling and adopt the associated lag time from one of the selected time series. This ensures that synthetic compound events retain the same timing dynamics as similar observed events (See Fig. 9). Extending this analysis to evaluate how differences between storm types influence flood depths would require flood model simulations of a broader set of events, which is beyond the scope of this study.

For further clarification, the following sentences will be revised:

L 239: "Target events are derived from copulas fitted to the TC sample and the non-TC sample. If the target event is derived from a copula fitted to TC (non-TC) events, we sample the month from the distribution of TC (non-TC) events (Fig. 2 (a)). Once the month is selected, we randomly sample a MSL value and a tidal signal segment from the selected month (Fig. 2 (g))."

The following will be added:

L 420: "The joint probability distributions of peak NTR and peak RF of TC and non-TC storms are substantially different. Small to moderate compound events are more frequent in the non-TC storms, whereas the most extreme compound events are more likely generated by TCs. This aspect is further discussed in Maduwantha et al. (2024).

Minor comments.

The "target event" is never explicitly defined, and from my personal perspective, this creates some confusion. How is it how is it selected?

We thank the reviewer for raising this point. We refer to a "target event" as a synthetic peak NTR–RF combination selected from the 5,000 combinations generated via Monte Carlo sampling of the fitted copulas. While some studies use the term "design event", we used "target event" since our framework does not rely on standard design return periods (e.g., 50- or 100-year). We will define it in L 179.

L 179: "We generate an event set of 5,000 combinations of NTR and RF ("target events") by sampling from the fitted copulas such that the relative proportion of extremes is consistent with the empirical distribution."

ETC and non-TCI seem to be used interchangeably. I suggest checking the notation for consistency.

Checked and will be corrected.

Line 409: the Authors say that flood depth varies in some regions. However, the case study seems to concern only one region. I would suggest checking this sentence.

The term region refers to the different areas within the catchment. We will revise the sentence as follows:

“The difference in flood depths between using the highest and lowest 30-day MSL reaches up to 1 m in some ~~regions~~areas of the city”.

The comparison between rainfall and surge is done considering duration. How did the Authors handle discrete variables when assessing correlation?

We thank the reviewer for raising this point. We have used the MATLAB function “corr” for calculating Kendall’s tau. When handling ties, this function uses an adjustment, calculating tau-b, which still varies in the range of -1 to +1 and leads to a stable estimation. Tau-b is calculated as follows:

$$Tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

Where,

C = number of concordant pairs
D = number of concordant pairs
T_x = number of pairs tied only in X
T_y = number of pairs tied only in Y