Response to the referee 1

The authors have addressed all major concerns. The manuscript is significantly improved. I recommend acceptance pending only one minor revision:

1. The parenthetical addition about advection testing could reference the new Figure 2 in supplementary material: "(we tested the sensitivity of the results when the advection term was enabled and found negligible differences; see Fig. S5)" Currently, it is missed.

We sincerely thank the reviewer for noticing this. Although we have included that figure in our previous reply to the reviewer, it is missing in the submitted supplementary material. This has now been corrected, and the updated figure is included in the revised supplementary file.

Response to the referee 2

First of all, I would like to thank the reviewers for carefully addressing my comments. However, two points remain unclear.

1. First, about the process in itself. The Authors say that "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[…] separately for the two storm types would produce similar results compared to the ones we derive without stratifying". Yet, you mentioned 3 different copulas in the generative process, with opposite tail dependence in the case of TC and non-TC (from line 340). How do the Authors justify that the results with and without stratification are the same? What is then the added value of the methodology proposed? As a more general comment, it would help to use different colours for the simulated pairs in Figure 3 panel b, as done in panel a.

We sincerely thank the reviewer for these insightful questions and thoughtful feedback. We acknowledge that some confusion may remain regarding the characterization of storm events, and we appreciate the opportunity to clarify this aspect of the framework.

In this study, we consider two main categories of storm-event characteristics derived from historical observations:

(a) Peak values of non-tidal residuals (NTR) and rainfall (RF)
(b) Other time-series characteristics (e.g., temporal evolution, duration, lag time, intensity)

Accordingly, the proposed framework consists of two main steps. First, we derive the joint probability distribution of peak NTR and peak RF, and simulate peak combinations from it, which addresses category (a). Second, we assign a time series to the simulated peak combination, which addresses category (b).

We check how both categories of characteristics vary between the two storm types, tropical cyclones (TCs) and non-TC events. Notably, the dependence (here we quantified using Kendall's tau) between peak NTR and peak RF is substantially stronger in the TC sample than in the non-TC sample (see Figure R1 below). As a result, treating all events as a single sample when estimating joint probabilities would mischaracterize the dependence structure. For this reason, the joint probability analysis is performed separately for TC and non-TC events. Within each storm type, we further consider two conditional samples (conditioning on NTR and conditioning on RF), and we identify and fit the copula family that best represents each conditional dependence structure. Figure R2 illustrates these conditional samples and the corresponding joint probability distribution for each conditional sample and combined for storm type. This part of the analysis has already been published in Maduwantha et al. (2024), so we do not plan to include these figures in the main manuscript.

When examining the other time-series characteristics of the two storm types, we find no statistically significant differences in lag times, durations, or intensities (see Figures 4–7 in the main manuscript). This finding supports our conclusion that, when selecting an observed historical time series for scaling, it is not critical whether the event originates from a TC or non-TC storm, as the temporal characteristics are comparable. Thus, we claim, "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". However, for estimating the joint probability distribution of peak values, stratification by storm type remains necessary due to the different dependence structures.

In other catchments, where significant differences exist in the time series characteristics between TC and non-TC samples (as discussed in Section 4.2 of the main manuscript), we suggest that the event generation process could also be conducted separately for each storm type using the developed framework.

We have revised the manuscript to improve clarity on this point, as follows. Additionally, we have updated Figure 3 (panel b) to show simulated pairs in separate colors for each storm type for better comparison, as suggested (see Figure R3).

L496: Given these results, we conclude that ~~generating~~ sampling and scaling of the observed 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. Note that stratification is still applied when deriving the joint probability distribution since the dependence structure of the peaks of NTR and RF is substantially different. Importantly, this applies to the specific study location. In other places, significant differences may exist in the time series characteristics between TC and non-TC samples (as discussed in Section 4.2), warranting that the event generation process is also conducted separately for each storm type.

2. Second is about the 106-year event. The Authors say: "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." Based on the sampling, most of the events come from the non-TC case (roughly 93%). I would have expected a return period from the combined sample closer to the non-TC case. This is also what the Authors say about the fact that a TC event is less frequent than a non-TC. However, the return period of the event from the combined set (106 years) is closer to the return period of the TC sample (110-yr). How do the Authors justify this result?

While it is true that approximately 93% of our events are related to non-TC sources, the return period of the selected event (~106 years in the combined distribution) is much closer to the return period estimated using only the TC sample (~111 years), rather than the non-TC sample (~2431 years). This result arises from the shape of the joint probability distributions of individual storm populations and how the combined joint probability distribution is constructed.

Although non-TC events are more frequent in the historical record, TC events are often more extreme and exhibit stronger dependence between peak NTR and RF. This behavior is reflected by the structure of the joint probability density, particularly in the upper tail. The selected event (NTR = 1.75 m; RF = 80 mm) lies in a region of the joint probability space where TC events have a higher contribution to the total exceedance probability (see Figures R2 and R3 where the selected event is shown as a star). As a result, the combined annual exceedance probability (AEP) at this point is more influenced by the TC distribution, leading to a return period that is numerically closer to the TC-only estimate.

This can also be explained using the equations that we used to derive the total joint probability distribution of two storm type populations. Assuming these two populations are independent from each other, the total annual non-exceedance probability (*ANEP*) of a given pair of (NTR, RF) is calculated as follows (Maduwantha et al., 2024):

$$ANEP_{(NTR,RF)} = ANEP^{TC}_{(NTR,RF)} \times ANEP^{non-TC}_{(NTR,RF)} \quad (1)$$

$$ANEP_{(NTR,RF)} = \left(1 - AEP^{TC}_{(NTR,RF)}\right) \times \left(1 - AEP^{non-TC}_{(NTR,RF)}\right) \quad (2)$$

The associated return period (*RP*) is calculated as

$$RP_{(NTR,RF)} = \frac{1}{1 - ANEP_{(NTR,RF)}} \quad (3)$$

Therefore, the combined return period (RP) is always closer to the return period of the storm type that contributes the higher exceedance probability (lower return period) at the given (NTR, RF) combination. In this case, that contribution comes from the TC sample.

As this part of the study and the governing equations are discussed in detail in our prior publication (Maduwantha et al., 2024), we don't plan to elaborate on it in the current manuscript to avoid redundancy and additional length. We have added the following text to the main manuscript to clarify this point.

L455: "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. For example, the selected event for demonstrating the event-generation framework (NTR = 1.75 m, 18-h RF = 80 mm) corresponds to a ~106-yr joint return period in the combined joint probability distribution. The same event has a joint return period of ~111 yrs in the TC sample and ~2431 yrs in the non-TC sample, highlighting that rare events are primarily associated with TCs."

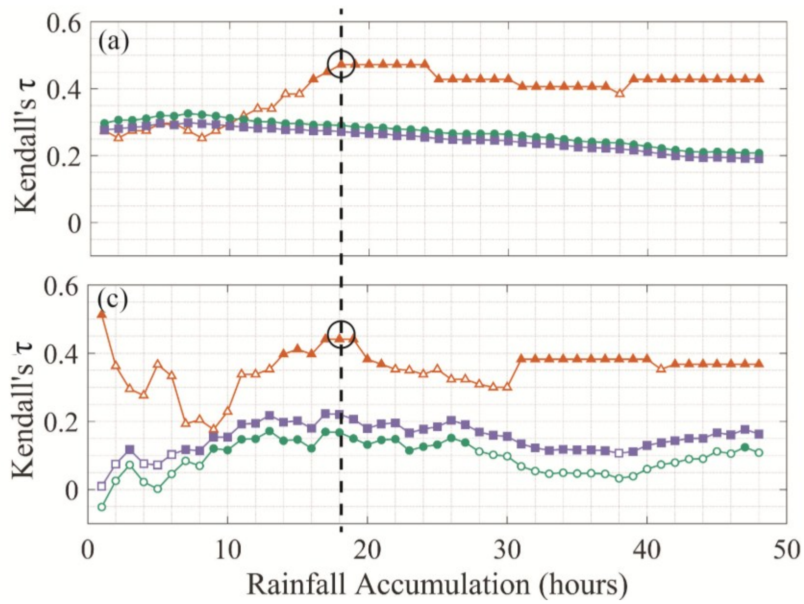

Figure R1: Kendall's τ between NTR and RF for different RF accumulation times for all events (purple), TCs (orange), and non-TCs (green) for samples conditioned on NTR (a and b) and RF (c and d). Filled markers indicate values that are significant at the 5% level. The black circles with vertical dashed lines show the selected RF accumulation for each location (Adapted from Maduwantha et al., 2024).

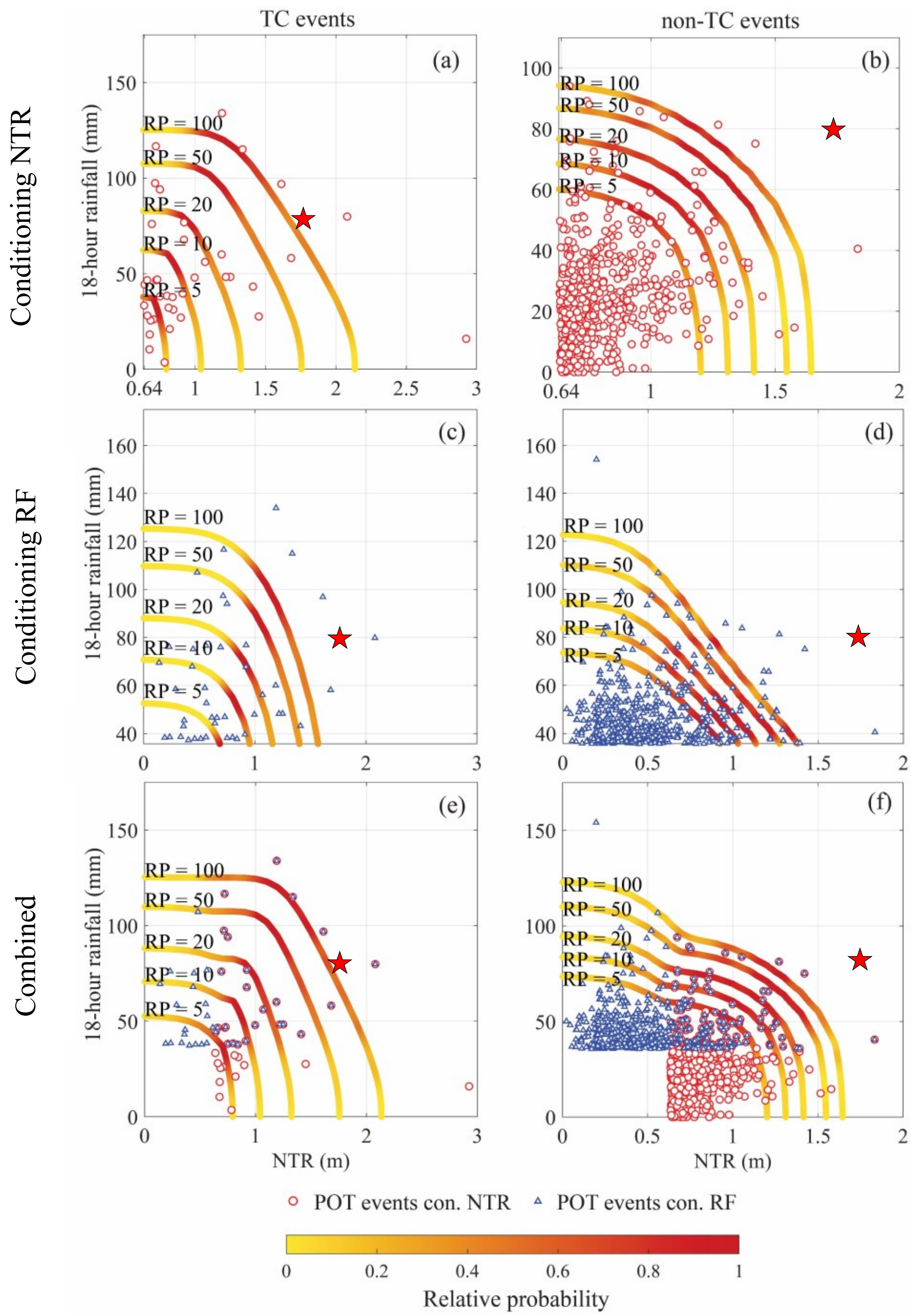

Figure R2: Results of bivariate statistical analysis for Gloucester City; TC events (left) and non-TC events (right): (a), (b) when conditioning on NTR; (c), (d) when conditioning on RF; and (e), (f) when two conditioning samples are combined. Quantile isolines of the 5, 10, 20, 50, and 100-year joint return periods are shown where the color scale indicates the relative probability of events along the isolines. Note the different x- and y-axis scales for better clarity (Adapted from Maduwantha et al., 2024). The red star indicates the selected synthetic event for the demonstration of events generation.

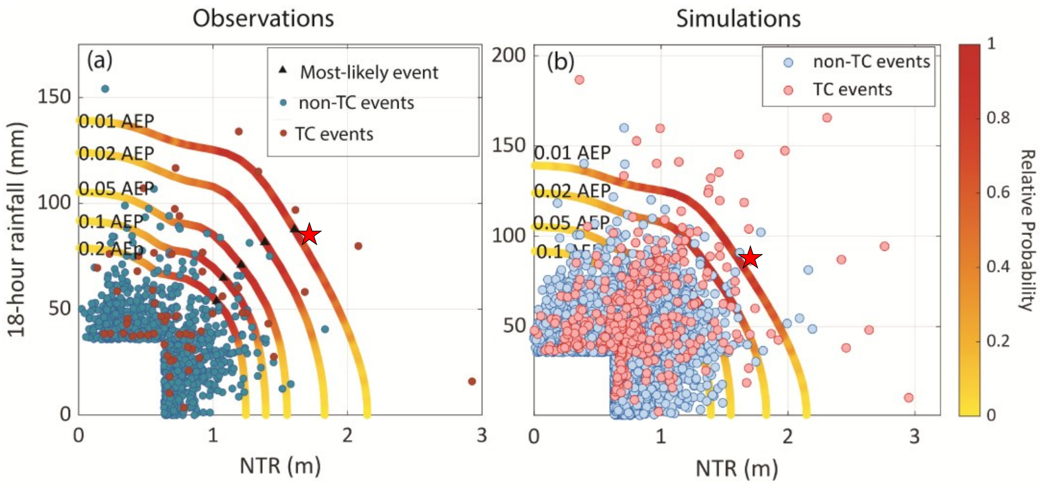

Figure R3: Modified Fig. 3 of the original manuscript. The red star indicates the selected synthetic event for the demonstration of events generation.