The authors use observations from tide and precipitation gauges at three sites along the West coast of the United States to calculate various joint and conditional probabilities based on different sampling methods. They compare different scenarios of bivariate return periods as defined in Salvadori et al. (2016) and Serinaldi (2015) using a wide range of copulas and marginal distributions. Based on this, they conclude that annual maxima does not produce "worst-case" events and recommend instead wet season coinciding sampling to characterize compound flooding in semi-arid regions.

The authors have done a substantial amount of interesting work: they have tested many different scenarios and possibilities of copulas and marginal distributions. However, I find the manuscript difficult to follow, which as a result, makes the relevance of the findings unclear. More particularly, I find the abstract quite disconnected to the structure and the findings presented in the manuscript. For example, I am still uncertain which research gaps are being filled. I do think this might be resolved by carefully restructuring the manuscript and their findings to focus on the main points the authors are trying to convey.

- As pointed out by the authors, "Serinaldi (2015) suggests inter-comparing univariate, multivariate, and conditional probabilities and return periods is misleading as each probability type describes its associated event." (line 337-338). Yet, after reading the manuscript it gives the impression that this is exactly what the authors did by listing the results obtained from all possible hazard scenarios. In the methodology, the authors mention the type of question each hazard scenario might be answering, but this is not reflected in the abstract. Instead, only a broad mention of "compound events" is mentioned.
- The authors also frequently refer to the FEMA (2016) methodology which seems to imply that they want to compare their hazard scenarios with FEMA but this is not done at any point in the manuscript. FEMA (2016) provides a methodology to derive the 100 (500)- year return period of the water level while this study does not model the obtained water levels from the event pairs obtained. Furthermore, the FEMA flood maps should represent, in theory, the flood depths levels happening once every 100 (or 500) years on average. This probability cannot be compared with probabilities from the hazard scenarios and conditional probabilities as each of them only contains a subset of all potential flood events.
- It is unclear what is meant by "worst-case" in the manuscript which makes the abstract and conclusions very difficult to understand. Based on which results do the authors conclude that AM does not result in worst case scenarios? Is this for each hazard scenario?
- To my understanding, this study does not assess flood risk but derives the pairs of different joint/conditional return periods, except for the structural scenario where a mention on the flood extent is given. As such, I find the title a bit misleading in its current form. Also, the *implications* of the work could be better highlighted in the abstract. I find parts of the discussion interesting (lines 339-348) where the authors mention the value of testing different hazard scenarios to capture compound interactions.
- Line 9-10: "Although annual maximum sampling is commonly recommended for characterizing compound events, …". The current emphasis in the abstract about the annual

maximum sampling is not straightforward to me. This could also be because I am not sure what the authors mean by "worst case" event pairs. But I would actually tend to disagree with this statement. There is a wide variety of sampling methods commonly used to characterize compound floods that include both peaks-over-threshold and annual maxima, with no clear scientific consensus for one or the other method. Compound flood studies that have looked at the impact from these events, such as Santos et al. (2020) found AM methods to produce a more balanced datasets. However, without looking at the impact from your events, one cannot make a strong conclusion about this.

- I find the comparison of different sampling methods particularly confusing (section 4.3 and 4.4). As mentioned by reviewer 2, these methods sample different sets. Therefore, it is expected to lead to different values for the evaluation of the '100-year' probability. When presenting the results, this should be clearly discussed, otherwise this seems to imply that '100-year' event from the marginal, conditional and joint probabilities are the same, which is not true. Maybe this could already be clearly stated in the Methods and reiterated in the Results. In the end, an important message from your study is specifically that these hazard scenarios do not represent the same probability space and thus for the same probability level, they lead to different values. This means that extreme caution should be used for users/modelers who want to use a given hazard scenario to derive the "100-year" flood event.
- The authors did not show formal goodness-of-fit test (but only informal ones like the BIC and AIC). The lowest BIC or AIC value could still denote a poor fit from the model. It is therefore usually recommended to perform the Cramer-von Mises blanket test, see for example:
    - Ward et al. (2018). *Dependence between high sea-level and high river discharge increases flood hazard in global deltas and estuaries*. Doi: 10.1088/1748-9326/aad400
    - Genest et al. (2009). Goodness-of-fit tests for copulas: A review and a power study. Doi: 10.1016/j.insmatheco.2007.10.005
    - Couasnon et al. (2018). A Copula-based bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment scale: An application to the Houston ship channel, Texas. Doi: 10.3390/w10091190

Minor comments

The line numbers below refer to the track changes manuscript

- Line 1: Your study assumes stationary conditions so I am not sure the sentence on sea level rise is needed.
- Line 38: "and then adopting the more severe flooding result". From the reference cited (FEMA 2016c), it is shown that FEMA does not uses the most severe flood result but a combination of both fluvial and coastal flood depth probabilities to estimate the 100-year water level in transitional areas (see Figure 4.1 in the *FEMA Guidance for Flood Risk Analysis and Mapping: Coastal Flood Frequency and Extreme Value Analysis*). So I would rephrase this to better highlight the limitations from their methods.

- Line 50-51: Maybe rephrase to mention that *initial* studies were mainly focusing on hazard scenarios. I would then advise to change the title of Table 1 to a non-exhaustive list of studies that used hazard scenarios (if this is what the authors want to emphasize). Otherwise, the body of literature that used copulas to study compound flooding is much broader that what is stated there. Also, there is a lot of recent research that moved beyond single hazard scenarios in order to better capture all potential flood events using copulas. See for example:
    - Bevacqua et al. (2017). Multivariate Statistical Modelling of Compound Events via Pair-Copula Constructions: Analysis of Floods in Ravenna. Doi: 10.5194/hess-21-2701-2017
    - Santos et al. (2021). Assessing compound flooding potential with multivariate statistical models in a complex estuarine system under data constraints. Doi: 10.1111/jfr3.12749
    - Jane et al. (2022). Assessing the Potential for Compound Storm Surge and Extreme River Discharge Events at the Catchment Scale with Statistical Models: Sensitivity Analysis and Recommendations for Best Practice
    - Couasnon et al. (2022). A flood risk framework capturing the seasonality of and dependence between rainfall and sea levels – an application to Ho Chi Minh City, Vietnam. Doi 10.1029/2021WR030002
- Line 67: “humid climatic conditions”: This sounds very broad
- Line 75 : “most likely”. At this point, this term has not been introduced in the manuscript and without context, this may lead to the wrong understanding of this expression.
- Line 92: “December 19, 2023” Typo?
- Line 93-94: I am not sure of the added value of this sentence. Can you make a clearer link with your study design?
- Line 103: “by the event time”. Do you mean duration? Or time step?
- Line 114-116: “.. since the co-occurrence of precipitation and water levels follows the FEMA guidance for considering a “worst case scenario" approach .” I do not understand this. Can you elaborate?
- The authors use a lot of acronyms. I can understand why but the references to the acronym is not always mention in the Figures or titles. So the reader sometimes has to scroll to the text to find those acronyms again. For example, I couldn’t find what M refers to (I think Marginal distribution?). I would carefully review all Figures and Tables to make sure any acronym is mentioned to be able to interpret the figure or table independently of the text.
- Table 3 : The word “marginal” or “univariate” when describing the probabilities/return period is missing in the title.
- Line 194: “easily”. I would remove this term and rename this section “Univariate return periods”. In multivariate space, I think to interpret a return period is not easy at all!
- Line 320: “save the Cubic”: typo?
- Uncertainties are not quantified in your study which limits the interpretation of the implications. I am not asking to add this but at least mention this point somewhere in the discussion.

- Similarly, it would be interesting to mention the limitation of this method in locations with multiple pluvial or coastal flood seasons.