We would like to thank the reviewer for the thoughtful comments and feedback on our manuscript. In this response document, we present the original comments along with our corresponding responses. For clarity, Our responses are in blue while the reviewer's comments are in white. Moreover, line numbers in the revised manuscript are provided where appropriate.

The authors present the generalization of a method for statistically analyzing the joint extremes of different environmental variables under non-stationary conditions.

The work is scientifically relevant and addresses a current problem, especially under the conditions of ongoing climate change.

The provision of the code for applying the analysis to other case studies is highly appreciated. I believe the paper could be published after some minor revisions.

We thank the reviewer for this positive feedback and for recognizing the relevance of our work. In the following, we have addressed all the reviewer's comments.

In Table 1, I would suggest adding an explanation of the meaning of the third column.

An explanation was added in between lines 125 – 127 that reads:

Line 125-127: "For Archimedean copulas (e.g., Gumbel and Frank) $\gamma_\theta(\chi)$ denotes the generator function used in constructing the copula $C(u, v|\theta) = \gamma_\theta^{-1}[\gamma_\theta(u) + \gamma_\theta(v)]$."

In the description of the first case study, it would be appropriate to add a justification for the choice of a 45-day time window between two extremes. At first glance, this may seem strange, given that it involves the analysis of joint flood probabilities in small basins and wave heights. The choice could make sense if one wishes to consider the joint probability of events with a view to recovery between one event and the next, but I believe this needs to be explained further.

The 45-day time window was selected based on the impact-based temporal compounding perspective for coastal compound flooding. Following the typology of compound events proposed by Zscheischler et al. (2020), temporally compounding events occur when "a succession of hazards leads to an impact," where the key consideration is that the initial event increases vulnerability to subsequent events even without direct hydrodynamic interaction.

In coastal systems, when an extreme sea level event occurs, it can compromise coastal defense infrastructure, cause residual inundation, saturate drainage systems, and reduce the capacity of the coastal zone to effectively discharge fluvial flows. If a subsequent extreme river discharge event occurs while the system remains in this compromised state, the compound impact can be significantly amplified compared to either event occurring in isolation—even in the absence of direct temporal coincidence or dynamic coupling between the two drivers. This represents a temporally compounding event where "one event increases the vulnerability or exposure to subsequent events" (Zscheischler et al., 2020).

The 45-day window reflects a realistic timeframe for recovery of coastal systems following marine extreme events, considering factors such as drainage system recovery, infrastructure capacity restoration, and residual saturation effects. The choice of 30 days for univariate peak separation follows established practice in extreme value analysis (Coles, 2001), ensuring statistical independence of univariate extremes. The bivariate window of 45 days (1.5× the univariate threshold) allows capturing compound events where temporal proximity amplifies impact through system vulnerability, while maintaining sufficient separation for statistical modeling validity.

The appropriateness of this choice is further supported by the statistically significant temporal patterns observed in the coupling parameter (Figure 2c), indicating that the selected window successfully captures physically meaningful temporally compounding behavior consistent with impact-based compound event theory.

The manuscript now reads (lines 324-329):

Line 324 – 329: "The choice of a 45-day maximum allowable lag among bivariate peaks reflects the time-lag among univariate peaks (30 days) and the impact-based temporal compounding perspective (Zscheischler et al., 2020): when an extreme sea level event compromises coastal system capacity, a subsequent extreme river discharge within 45 days can produce amplified impacts even without direct hydrodynamic interaction. Following sampling, the average temporal separation among bivariate peaks was approximately 15 days."

References:

- Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, London.

- Zscheischler, J., Martius, O., Westra, S., et al. (2020). A typology of compound weather and climate events. Nature Reviews Earth & Environment, 1(7), 333-347. https://doi.org/10.1038/s43017-020-0060-z

From the text, it appears that in some cases the analyses are based on model results rather than recorded data. This, due to a possible bias between the simulations and the actual data, could impact the quantification of return periods, and I believe it should be explicitly stated. Or, if it has been done, specify whether any form of bias correction was applied to the model data.

We thank the reviewer for this important comment. The reviewer is correct that using model-based datasets can introduce additional uncertainty in return level estimation compared to direct observations. The datasets used in this analysis—Mentaschi et al. (2023) for waves and Tilloy et al. (2025) for river discharge—have been extensively bias-corrected against observational data. However, residual model biases cannot be entirely eliminated and do contribute to uncertainty in the estimated return levels. This limitation and its implications have been explicitly acknowledged in the revised manuscript:

Lines 298 – 304: "All the data in the following examples were obtained from model results. Possible biases in the model data can also find its way into quantification of return periods. Nonetheless, these model outputs provided a good basis upon which the general methodology developed in this paper could be substantiated. For the wave dataset used in the first and second case study, we used a bias-corrected version of the results reported in Mentaschi, et al., 2023. The bias correction was based on Quantile mapping, and we focused on values above 50th percentile. The bias-correction was based on comparison with satellite measurements."

Lines 315 – 320: "Although residual model biases contribute to uncertainty in return level estimation, the use of model data for coastal hazard mapping is widely established practice, as models provide complete spatiotemporal coverage that cannot be achieved through sparse observational networks.The wave dataset employed in this study (Mentaschi et al., 2023) was post-processed to reduce biases through quantile mapping against satellite observations. Likewise, the river discharge dataset (Tilloy et al. 2025) was evaluated using 2448 river gauging stations to assess model skill."

References:

- Mentaschi, L., Vousdoukas, M. I., Mentaschi, L., García-Sánchez, G., Fernández-Montblanc, T., Roland, A., Voukouvalas, E., Federico, I., Abdolali, A., Zhang, Y. J., & Feyen, L. (2023). A global unstructured, coupled, high-resolution hindcast of waves and storm surge. Frontiers in Marine Science, 10. https://doi.org/10.3389/fmars.2023.1233679
- Tilloy, A., Paprotny, D., Grimaldi, S., Gomes, G., Bianchi, A., Lange, S., Beck, H., Mazzetti, C., & Feyen, L. (2025). HERA: a high-resolution pan-European hydrological reanalysis (1951–2020). Earth Syst. Sci. Data, 17(1), 293–316. https://doi.org/10.5194/essd-17-293-2025

We would like to thank the reviewer for her thoughtful comments and feedback on our manuscript. In this response document, we present the original comments along with our corresponding responses. For clarity, Our responses are in blue while the reviewer's comments are in white. Moreover, line numbers in the revised manuscript are provided where appropriate.

General comments

The paper generalizes an approach developed for the estimation of univariate extremes in the non-stationary context to multivariate extremes, through copulas. It is a very interesting approach which deals with a very important and yet not much studied issue, and it is worth being published. Since I have contributed to similar developments in the univariate context, I was excited to read this paper, and I have a couple of comments.

We thank the reviewer for this positive comment and for recognizing the originality and importance of our work. We are also glad that our work resonates with many of your earlier contributions in the univariate context and we made an effort to address all of your comments in the revised manuscript.

Specific comments

The idea of computing a stationary variable to estimate non-stationary extremes in the univariate context had also been proposed by Parey et al. (2010, 2013, 2019) for temperature and applied to rainfall in Acero et al. 2017. It could be added to the references,

We thank the reviewer for these valuable references. We have now included Parey et al. (2010, 2013, 2019) and Acero et al. (2017) in the revised manuscript, as they provide complementary perspectives on non-stationary extreme value analysis. These additions strengthen the contextualization of our methodology within the broader literature.

and I suggest some additions in the description of the section "Non-stationary marginals":

Lines 154-155: "where $y(t)$ is the non-stationary series, $x(t)$ is the assumed stationary series, $Ty(t)$ and $Cy(t)$ are generic terms representing the long-term variation in the mean and amplitude of $y(t)$, respectively." Stationarity is not easy to assess globally, however in Parey et al. 2013, a statistical test is proposed for the stationarity of the extremes of x(t) and its application to temperature shows that they can be considered as stationary (when Ty(t) is the trend in mean and Cy(t) the trend in standard-deviation, estimated by LOESS with an optimal smoothing parameter).

We thank the reviewer for this valuable methodological suggestion. Assessing post-transformation stationarity in large datasets is indeed challenging, and we will carefully consider implementing the test proposed by Parey et al. (2013) in future developments of the tsEVA framework. For the case studies presented in the revised manuscript, we verified the stationarity of each transformed signal by applying the Mann-Kendall test to the annual percentile series (both before and after transformation, as shown in Figures 2-4). This approach, while less formal than the bootstrap-based test suggested by the reviewer, provided adequate validation for the datasets considered here. This was explained in the manuscript.

Lines 262 – 265: "We used MK to assess non-stationarity of the univariate marginals by applying the test on the annual percentile series. To see the effectiveness of transformation, the test was applied on both stationary and non-stationary series (i.e., before and after application of transformation). This method was applied as a shortcut to other sophisticated tests for stationarity that rely on bootstrapping (e.g., Parey, et al., 2013)."

In the section "Joint sampling of the extremes", I agree with the stated limitations of GEV. However, I am not convinced that it can be applied to "relatively slow, seasonal phenomena, such as drought and heat waves" because in those cases, EVT application is not fully justified theoretically. Indeed, it is an asymptotic theory, in that the distribution of the maxima tend to a GEV distribution when the block length tends to infinity. With annual blocks, one assumes that 365 values is a sufficient number for assuming that the distribution of the maximum is a GEV, this can be called a "probabilistic assumption". This however requires that the phenomenon under study is independent, or slightly depended, so that the block length is not too low (too much lower than 365). The number of heat waves and droughts each year is not very large, so the probabilistic assumption may be too strong. In the case of low flows for example, Parey and Gailhard 2022 use stochastic generation.

We thank the reviewer for this important theoretical point regarding the application of EVT to temporally dependent phenomena. The reviewer is correct that the asymptotic justification for the GEV distribution relies on block independence (or weak dependence), and that slowly-varying seasonal phenomena such as droughts and heat waves exhibit temporal correlation structures that may challenge these assumptions.

We acknowledge this limitation explicitly in the revised manuscript (lines 513 - 515).

Lines 513 – 515: “It is true that the asymptotic justification for the GEV distribution relies on block independence (or weak dependence), and that slowly-varying seasonal phenomena such as droughts and heat waves exhibit temporal correlation structures that may challenge these assumptions.”

However, we note that despite these theoretical concerns, the GEV distribution has been shown to provide robust empirical fits for aggregated hydroclimatic indices in numerous studies. For example, Stagge et al. (2015) demonstrated that GEV outperformed alternative distributions for SPEI across Europe, suggesting that for moderately aggregated variables (e.g., monthly temperature, SPEI-6), the temporal dependence may be sufficiently reduced to permit reasonable GEV approximation within the practical range of return periods typically considered in risk assessment.

We recognize that alternative approaches, such as the stochastic generation methods employed by Parey and Gailhard (2022) for low flows, may provide more theoretically rigorous frameworks when temporal dependence is strong. However, for the application presented here—where we demonstrate the tsEVA methodology across diverse hazard types—the GEV approach offers a balance between theoretical defensibility and practical applicability. Future work could explore more sophisticated treatments of temporal dependence for specific applications where this becomes critical.

Lines 515 - 522: “Although SPEI-6 represents a slowly-developing process, the GEV distribution has been shown to provide adequate fits for such variables at appropriate aggregation scales. Stagge et al. (2015) demonstrated that GEV outperformed alternative distributions for SPEI across Europe for accumulation periods from 1 to 12 months. For the moderate temporal aggregations considered here (monthly temperature, SPEI-6), the GEV provides a practical and empirically supported approximation for extreme value analysis. Alternative approaches such as stochastic generation methods (Parey and Gailhard, 2022) may offer additional rigor for specific applications with stronger temporal dependence”

References

- Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., van Loon, A. F., & Stahl, K. (2015). Candidate Distributions for Climatological Drought Indices (SPI and SPEI). International Journal of Climatology, 35(13), 4027–4040. https://doi.org/https://doi.org/10.1002/joc.4267

Case studies:

Case studies 1 and 2:

1) Are 40 years sufficient to robustly fit the Copulas? How many joint extremes are used on average? Could the uncertainty in the fitting be quantified (uncertainty of the coupling parameter? Is a change from 0.65 to 0.79 significant for example?).

We thank the reviewer for these important questions regarding the robustness of the copula fitting procedure.

Sample size and joint extremes: The datasets span 71 years (Case Study 1) and 74 years (Case Study 2). The 40-year moving window was selected through prior sensitivity analysis to balance two competing requirements: smaller windows produce excessive noise in the coupling parameter, while larger windows provide insufficient temporal resolution to detect non-stationarity. On average, each 40-year window contains 66 joint extremes (Case Study 1) and 76 joint extremes (Case Study 2), which provides adequate sample sizes for stable copula parameter estimation. We added the following information in the revised manuscript:

Lines 331 - 333: On average, each 40-year window contains 66 joint extremes, which provides adequate sample sizes for stable copula parameter estimation.

Lines 341 - 342: Each time window contained 76 joint extremes on average.

Statistical significance: To assess whether the observed changes in the coupling parameter are statistically significant, we applied the modified Mann-Kendall test (Hamed and Rao 1998; Hamed 2008; Yue and Wang 2004) to the coupling parameter time series in the revised manuscript. This modified test takes account for auto-correlation in the input series, which arises due to the 1-year sliding time window and the resulting overlap between windows. After applying this method, the reported p-values in the revised manuscript changed. For example, in the first case study, the p-value shifted from 8.4e-11 to 4.7e-05. Despite these adjustments, the conclusions regarding the statistical significance of the observed changes in the dependency parameter remained essentially unchanged.

These clarifications have been added to the revised manuscript in lines 266 – 271:

Lines 266 – 271: "To assess the non-stationarity of the joint distribution, we applied the modified Mann–Kendall test (Hamed 2008; Hamed and Rao 1998; Yue and Wang 2004), which explicitly accounts for autocorrelation arising from the 1-year sliding window used to compute the temporal evolution of the coupling parameter. By considering this autocorrelation, the test provides a more accurate assessment of statistical significance, avoiding the overestimation of trend significance that can occur with overlapping windows."

References:

- Hamed, K. H. (2008). Trend detection in hydrologic data: The Mann–Kendall trend test under the scaling hypothesis. Journal of Hydrology, 349(3), 350–363. https://doi.org/https://doi.org/10.1016/j.jhydrol.2007.11.009
- Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology, 204(1), 182–196. https://doi.org/https://doi.org/10.1016/S0022-1694(97)00125-X
- Yue, S., & Wang, C. (2004). The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. Water Resources Management, 18(3), 201–218. https://doi.org/10.1023/B:WARM.0000043140.61082.60

2) The time evolution of the parameter of the copula and of the correlation coefficients is assessed through a Mann-Kendal test. However, they are computed from moving windows of 40 years, with a 1-year sliding, which makes them highly correlated. Doesn't it undermine the Mann Kendall test?

We thank the reviewer for raising this important point. We fully agree that 40-year moving windows with a 1-year step introduce strong serial dependence between successive estimates. This violates the independence assumption of the classical Mann–Kendall (MK) test and would indeed undermine its validity if applied in its standard form.

In response, we have revised the analysis and now apply a modified Mann–Kendall test that explicitly accounts for autocorrelation in the input series (Hamed and Rao, 1998; Hamed, 2008; Yue and Wang, 2004). This approach corrects the variance of the MK statistic to reflect the effective sample size under serial dependence, thereby avoiding inflated significance levels that may arise from overlapping windows.

Only the time-varying copula parameter is subjected to this test. The correlation coefficients derived from the same moving windows are not used for trend detection and therefore do not involve the MK test; they serve solely to assess the goodness of fit of the copulas within the Monte-Carlo evaluation.

Applying the modified MK test affects some p-values of the coupling-parameter trends, but the overall conclusions regarding non-stationarity remain unchanged.

The revised manuscript now states (Lines 266–271):

Line 266 – 271: “To assess non-stationarity in the joint distribution, we apply the modified Mann–Kendall test (Hamed 2008; Hamed and Rao 1998; Yue and Wang 2004), which adjusts for autocorrelation arising from the 1-year sliding window used to compute the time-evolving copula parameter. This correction provides a robust estimate of trend significance and prevents the artificial inflation of significance that may result from overlapping windows.”

3) How are the Return Levels computed? In the non-stationary context, there is not any standard definition of a Return Level, and different propositions have been made and compared in the univariate context (Yan et al. 2017). In Parey et al., the Return Level is computed for the stationary variable and back transformed into a Return Level of the variable under study through the changes in mean and standard-deviation; they are therefore representative of a targeted future climate period.

We thank the reviewer for this important conceptual question. The reviewer is correct that in non-stationary contexts, the interpretation of return periods requires careful definition, and different frameworks have been proposed (Yan et al., 2017; Parey et al., 2010).

In Parey et al. (2010), return levels are computed for the stationary transformed variable and then back-transformed to represent a specific target future climate period. The tsEVA framework adopts a similar transformation-based philosophy but with a different implementation and interpretation. Following the original tsEVA formulation (Mentaschi et al., 2016), we transform non-stationary series to stationarity, fit extreme value distributions, and back-transform to obtain time-varying marginal distributions (GEV or GPD parameters that vary with time).

For bivariate analysis, these time-varying marginal distributions are combined with time-varying copula parameters (estimated from moving time windows) to compute bivariate return periods using the AND and OR definitions (Equations 11 and 12). The resulting return periods are conditional on the statistical properties of each time window, allowing us to quantify how joint exceedance probabilities evolve as both marginals and dependence structures change over time.

We added further details to the revised manuscript as follows:

Lines 225 – 231: “In the non-stationary context, there is not any standard definition of a return period, and different propositions have been made (Yan et al. 2017). Therefore, the interpretation of return periods requires careful definition, and different frameworks have been proposed (Parey et al. 2010; Yan et al., 2017). In Parey et al. (2010), univariate return levels are computed for the stationary transformed variable and then back-transformed to represent a specific target future climate period. The tsEVA framework adopts a similar transformation-based philosophy but with a different implementation and interpretation."

Lines 247 – 253: "Based on the above definition, we restricted the use of joint return periods to bivariate cases only. In the non-stationary context, return periods are conditional on the statistical properties (marginals and dependence) of each time window. The computation procedure is as follows: For each time window, we estimate (i) the time-varying marginal distributions from the back-transformed non-stationary GEV or GPD parameters, and (ii) the copula function C(u,v) from the coupling parameter specific to that window. These are combined using the AND and OR definitions (Equations 11 and 12) to compute joint exceedance probabilities."

Case study 3: I am not sure that the application of EVT is fully justified in that case, as stated above, especially with monthly SPEI values.

We thank the reviewer. This comment was addressed under the specific comments section, 3rd comment.

References:

Acero F.J., Parey S., Hoang T.T.H., Dacunha-Castelle D., Garcia J.A. and Gallego M.C.: Non-stationary future Return Levels for extreme rainfall over Extremadura (SW Iberian Peninsula). Hydrological Sciences Journal, 2017, DOI: 10.1080/02626667.2017.1328559

Parey S., Dacunha-Castelle D, Hoang TTH.: Different ways to compute temperature return levels in the climate change context; Environmetrics, 2010, DOI 10.1002/env

Parey S., Hoang TTH, Dacunha-Castelle D.: The importance of mean and variance in predicting changes in temperature extremes, Journal of Geophysical Research: Atmospheres, Vol 118, 1-12, 2013, doi:10.1002/jgrd.50629

Parey S., Hoang T.T.H., Dacunha-Castelle D. : Future high temperature extremes and stationarity, Natural Hazards (2019) 98:1115–1134, https://doi.org/10.1007/s11069-018-3499-1

Parey, S.; Gailhard, J. Extreme Low Flow Estimation under Climate Change. Atmosphere 2022, 13, 164. https://doi.org/10.3390/atmos13020164

Stagge, J. H., Tallaksen, L. M., Gudmundsson, L., van Loon, A. F., & Stahl, K. (2015). Candidate Distributions for Climatological Drought Indices (SPI and SPEI). International Journal of Climatology, 35(13), 4027–4040. https://doi.org/https://doi.org/10.1002/joc.4267

Lei Yan, Lihua Xiong, Shenglian Guo, Chong-Yu Xu, Jun Xia, Tao Du, Comparison of four nonstationary hydrologic design methods for changing environment, Journal of Hydrology, Volume 551, 2017, Pages 132-150, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2017.06.001.