Dear Editor,

We are thankful to the two anonymous reviewers for the constructive comments and the appreciation of our work.

Based on the comments of reviewer #1, we revised the paper in the following aspects. (i) We analized the sea level data using the revised joint probability method suggested by the reviewer and we included it in the comparisons with other methods. (ii) We analyzed data from a different tide gauge (namely: Marseille). This tide gauge was chosen because of the long temporal coverage (>100 years), a tide regime different from Venice, and the lowest degree of non stationarity due to the absence of local subsidence. (iii) Where pointed out by reviewer #1, we expanded the text to improve the reproducibility of our analyses.

Where needed, we included in the text the references suggested by both reviewers, and corrected the terminology according to the suggestion from reviewer #2

All the points raised by the two reviewers were addressed. Detailed, point-to-point changes are listed in the attached file.

We hope that these changes will increase the impact of the study and are happy to hear back from you in due time.

Sincerely,
Damiano Baldan,
on behalf of all co-authors

Answer to reviewer #1

Reviewer's comment	Action taken	Line
General comments		
The evaluation of return levels obtained from this set of methods in particular (GEV, GPD, PP, and JPM) has not been performed before, to the best of my knowledge. However, previous studies compared the implications of using different extreme value distributions when accounting for non-stationarity (e.g., Razmi et al 2017; Menendez et al 2010). Thus, the paper will benefit from the analysis of a wider set of methodologies, particularly from including the revised joint probability method (RJPM), which is one of the most widely applied indirect methods for estimating non-stationary return levels.	Thank you for the comment. We included the Revised Joint Probability Method (RJPM) among the methods. We added a short description of the RJPM in the methods section and included it in the results (figures 4 and 5). See detailed answer below.	215-235
Finally, the uncertainties in estimating the parameters of the different extreme value distributions could be included in the analysis, as done in previous studies (e.g., Cheng et al 2014) to offer a more exhaustive analysis.	Thank you for the comment. Following the detailed comments, we included the parameters uncertainties in figure 3. See detailed answer below.	-
As is, the manuscript conclusions are appropriate for the study case (Punta Della Salute) only. In order to provide a quasi-standardized method for non-stationary analysis, the paper will benefit from applying the analysis to a larger set of tide gauge records, so the authors will be able to assess whether the conclusions can be extrapolated to other areas of study. Likewise, readers will be able to decide which method should be used according to the conditions of each case: tide range, relative relevance of surge vs tide in extremes – as in Dixon et al 1999-, location, record length, etc. In this sense, Haigh et al 2010 showed that differences between return levels estimated using different methods highly depend on the record length and on the presence of outliers (when using direct methods).	Thank you for the comment. We included data from a different tide gauge (namely: Marseille). This tide gauge was chosen because of the long temporal coverage (>100 years), a tide regime different from Venice, and because of the lower impact of non stationarity due to the absence of local subsidence. We mention the inclusion in the introduction: "We compare the implemented methods also for sea level data from the tide gauge in Marseille (Southern France)." We added a short rationale for the inclusion in the methods section (2.1): "On the contrary, area where the Marseille tide gauge is located has a lower tidal range (around 10 cm), and is located on a stable background, with a relative sea level rise of + 1.1 mm y^{-1} in the last 150 years (Letetrel et al., 2010; Wöppelmann et al., 2014)". We included the data source and quality in section 2.2: "Hourly sea level data recorded at Marseille are available for the time period 1849 –	73-74 & 94-96 & 113-117 & 351-353

	2017. Measurements were performed with a float-operated tide gauge until 1988, with an acoustic sensor for 1989 – 2008, and with a radar sensor from 2009 onwards. measurements were recorded mechanically until 1988 and electronically from 1989 onwards (Wöppelmann et al., 2014). A total record length of 77 years (spanning 1903 - 2017) was used to fit the model, since some years were discarded due to incomplete records". We included results from Marseille in figures 5 and 6. We analyzed both tide gauge data using a smaller (30 years) subset of the data and we present the results in figure S2, and briefly in the results section: "A consistent behaviour was observed when stationary models fitted to data covering 30 years were compared with JPM and RJPM (Figure S2)."	
In addition to the above comments, some concerns about the reproducibility of the work arise, including (i) the use of phrases such as "*We used our own script to generate the empirical sea level distribution function*" (line 151); (ii) the authors do not provide information on the additional tide gauges used to calculate mean sea level prior to 1924; (iii) they do not provide information on how they calculated the uncertainties shown for the stationary analysis in 30-year time windows (Figure 3 and Figure 4).	See answers to the specific comments below.	-
Specific comments		
Line 58: How do you define suitability? What do you use to compare the different distributions and decide which of them performs best?	Thanks for the comment. We added a short explanation of what we mean by 'suitable', including a reference to the manuscript section where we explain how we compared different models: "It is expected that different methods might be more or less suited (in terms of explained variance, see section 2.3.6) to accommodate non stationary data, and might lead to different estimates of extreme sea level	60

	probabilities (Wahl et al., 2017; Razmi et al., 2017)."	
Tide gauge: Is the sea level record at this particular tide gauge affected by the activation of the MOSE? Can that influence the measured extreme sea levels?	The MOSE system was first operated in fall 2020. We added the information to the text: "Data from 2020 onwards are affected by the activation of a storm surge barrier system that prevents ESLs to propagate inside the Venice lagoon (MOSE)."	99-100
- Tide gauges used to calculate the mean sea level prior to 1924: (i) in order to ensure reproducibility, the authors should provide the name and location of those tide gauges. (ii) Are these extra tide gauges affected by the same subsidence rate as Punta Della Salute? As a double-check, did you compare the mean sea level of the tide gauges and the Punta Della Salute one during the overlapping period?	Thank you for the comment. We included the requested information in the text, with a reference to current work on sea level rise in Venice: "To calculate long-term mean sea level before 1924, we used yearly mean sea level data from other tide gauge stations active in the city of Venice (and thus affected by the same subsidence rate as *Punta della Salute-Canal Grande*) whose records cover the period 1885 – 1922 (namely: *Campo Santo Stefano*, *Arsenale, and Punta della Salute*; for details see Zanchettin et al., 2021). "	110-112
Joint probability method: instead of using JPM, the authors might consider the revised JPM (RJPM) or the skew surge revised JPM (SSRJPM). These two methods allow the fitting of extreme value distribution (thus obtaining extreme probabilities) and apply an extreme index that accounts for dependence between extreme events.	Thank you for the comment. We included in the text a short description of the JRPM in the methods section: "The revised Joint Probability Method (RJPM) improves the JPM by fitting the surge distribution with a probability distribution function, to allow for the smoothing of the empirical distributions, and for projections beyond the highest measured surge (Tawn et al., 1989)."	217-219
Line 151: to ensure reproducibility you might want to clarify which kind of empirical distribution you have used instead of stating sentences such as "*we use our own script to generate the empirical sea level distribution*".	Thanks for the comment. We extended the text in section 2.3.4: "For the JPM, we used all the tide and surge data from the sea level decomposition to generate the empirical frequency distribution over classes of width 10 cm. The maximum theoretical sea level (sum of maximum tide and maximum surge) falls within the highest class. Then, we calculated the discrete convolution between the two histograms. For the RJPM, and fitted a Gumbel distribution function to the annual maxima of the surge (following Tawn et al., 1989). Then we calculated the probability for each histogram class as the integral of the distribution. The quantiles of the obtained histogram representing the empirical sea level distribution were calculated based on the frequency of each class, and used for the	228-235

	estimation of return levels (Pugh and Vassie, 1978; Marcos et al., 2009).“	
- Comparing different model configurations: Could you please expand the explanation of the meaning of M_0 and M_1?	We added the explanation to the text: “Two nested competing models $M_0 \subset M_1$ can be compared using the deviance statistic (Coles et al., 2001). For example, M_1 can be a model whose parameters depend on covariates, while M_0 a model whose parameters do not depend from covariates.”	265-266
Figure 3: I don’t think you explained how you calculated the gray envelope in the text. The analysis will benefit from the comparison of uncertainties resulting from the different model configurations. Also, why isn’t the scale parameter (yellow line) shown for GEV, GPD (MSL), and PP? doesn’t the scale parameter vary with time (equation 7)? Same for the location parameter (green line) for PP (NDT, MSL, MSL_L)?	Thank you for the comment. We modified figure 3 and added the confidence intervals on the parameters. The reason why not all the models are shown in figure 3 is to avoid a clogged figure. Only parameters from models that are significantly better than the models without covariates are shown. We added an explanation why fewer parameters are displayed in figure 3 in the caption. We explained the gray envelope and the reason: “Figure 3: Comparison between the parameters estimated in the time window analysis (full thin line; the grey envelope represents the uncertainty of the parameters from the time window analysis) and the parameters estimated by different models configurations over the full data length. Parameters from all the configurations of GEV, GPD and PP that do not include covariates are showed. Parameters from models with covariates are showed only if models improve significantly the fit (see Table 2 for the likelihood test) […]”	Fig 3 caption