Authors’ replies to Reviewer 1 comments for NHESS-2021-236

We would like to thank Dr. Philip Ward and the two anonymous Reviewers for their helpful comments and suggestions, which much improved our manuscript. We appreciate the constructive comments received, which are discussed in detail. In general, some of the comments in particular, dealing with the way comparisons between extreme value distributions are done, prompted a shift in the focus of the manuscript. As Reviewer 2 puts it, the manuscript provides a comparison between possible approaches to extreme water level analysis. We have thus proposed a more balanced title, which does not needlessly over-emphasize the most recently proposed method and more objectively reflects the findings and the newly introduced analyses.
Below we provide our discussion of Reviewer’s comments (in blue italic font) and describe the changes addressing them.

From the title of the manuscript, the reader expects to read a study about extreme storm surge. However, the study’s objectives (Lines 66-68) refer to extreme sea level. Later on, Line 150, the Authors say that they will investigate the variable h(t) being the sum of tide and storm surge, so sea-level without mean sea level. I would encourage the Authors to clearly state the variable of interest and the variable used when performing the analyses, see also other comments below.
Thank you for this suggestion. To avoid any confusion about the focus of our work, we have changed the title of the manuscript to “*Extreme coastal water level estimation and projection: A comparison of statistical methods*”. Throughout the manuscript we now use the term “coastal water level” when referring to the sum of the tide and surge components.

Information regarding MEVD, which is the main method investigated in the manuscript, is limited. The Authors say that this method guarantees “the least amount of a-priori assumption” (line 56). However, the following assumption must be made: F(x,θ) in Eq. 2, the threshold for the ordinary values, the estimation window for parameter estimation, the time-lag to ensure independence between ordinary values. How then is this method the one with the least amount of a-priori assumptions? I suggest clarifying further the advantages of the MEVD compared to the other two methods investigated.
In the revised manuscript, we have expanded the description of the MEVD and we better emphasize the differences between the MEVD and traditional approaches. In particular, we now clarify, at line 61, that: “Moreover, the MEVD framework (i) is a non-asymptotic extreme value distribution, which does not require the number of events/year to be large as in the traditional theory, and (ii) makes no a-priori assumptions on the properties of the event occurrence process (while, e.g. POT-GPD assumes a Poisson occurrence process)”.

Moreover, additional information should be discussed: how the threshold for the ordinary value was selected (line 121 says “as small as possible”).
We have moved the discussion about threshold selection, previously at lines 276-277, to line 121, to clarify. The revised manuscript now discusses here that: “The threshold is set to be large enough to filter out water level peaks that are likely to be associated to conditions without any storm contribution and sufficiently low to maximize the amount of information used. In addition to the above, we choose the threshold value that produces the minimum extreme-value estimation error under the MEVD framework”.

How the 5-year estimation window was selected.
The revised manuscript now includes, at line 125 a discussion of the optimal estimation window length: “In the present application, the optimal estimation window length was set to 5 years to obtain a more robust parameters estimation, especially when few values in each year are available”.

Why the 30-day lag time for the independence of the ordinary value is so different compared to the values found in the literature (lines 173-179).

We agree that the wording in the previous manuscript generated some confusion on this point. The previous literature introduces minimum time lags with different objectives. Here a minimum time lag separation is introduced to eliminate correlations between water level peaks induced by the deterministic tidal contribution, which has a long periodicity linked to the main lunar cycle of about 28 days. The existing literature, on the other hand, focuses on the storm-surge component only and thus uses shorter time lag values due to the shorter correlation of the surge component due to atmospheric drivers. The revised manuscript now clarifies, at line 173, that: "The existing literature, which focuses on the storm-surge component only, uses shorter time lag values due to the shorter correlation of the surge component due to atmospheric drivers. For example, the independence… ".

How F(x,θ), which turns out to be a GDP (Line 267), is different compared to the classical GDP.

We have renamed what was originally called GEV-POT as POT-GPD, to avoid the confusion pointed out by this Reviewer.

We interpret this question as asking how the MEVD is different from the POT-GPD methods, which also uses a GPD distribution. The difference is that the GPD used in the MEVD approach aims to capture the distribution of all the ordinary values (i.e. those in the main body of the probability distribution), obtained by imposing a "low threshold". On the contrary, the classic POT-GPD method adopts a very high threshold to select independent large events in the tail of the distribution. Additionally, the MEVD does not require one to assume that event arrival is Poisson-distributed. Zorzetto et al. (2016) highlighted that if one assumes (i) x to be the excess over a high threshold, (ii) $F(x;\,\theta)$ to be a Generalized Pareto Distribution (with fixed, deterministic parameters), and (iii) n to be generated by a Poisson distribution, then the GEV distribution is recovered as a particular case of the MEVD by means of the POT approach. We now briefly clarify this point in the revised manuscript as follows: "We highlight that the GPD used in the MEVD framework is obtained by imposing a small threshold (differently from the high threshold adopted in the POT-GPD approach) to capture the distribution of the main body of the probability distribution of the ordinary events and does require the event arrival process to be Poisson (Marani and Zorzetto, 2019)".

I do see the value in implementing the cross-validation procedure to assess the predictability power of the distribution selected as representative of the observations. At the same time, I see the cross-validation as an additional measure of goodness of fit rather than the main one.

We respectfully disagree on the statement regarding out-of-sample vs in-sample tests. Cross-validation is not a measure of goodness of fit. Comparisons of estimation outcomes against independent data (not used in calibration) quantify how a statistical model can predict the likelihood of the "next event" and gauge how inferences from a statistical method can capture the properties of the underlying physical process, as opposed to just describe the specific dataset on which it is calibrated.

The NDE only tests if the one quantile associated with the return period Tr of interest is well captured. What about the other quantiles? Is the distribution representative of the entire sample?

We are grateful to this reviewer for this useful comment. The NDE represents an average measure (average among the realizations $p = 1, \dots, Nr$ where $Nr = 1000$ in this application) of a standardized distance between estimated and empirical quantiles, and can be computed for any return period. We had focused on the highest quantile, estimated from the independent test dataset, because its estimation is the most uncertain and the most valuable in practice. However, we agree with the Reviewer that it is useful to ask the question "is the extreme value distribution representative of the entire range of return times of interest?". To this end, we have performed additional analyses to evaluate methods performance also for intermediate *Tr* values, greater than the calibration sample size (for Tr<S empirical quantiles can be used, with little need of distribution fitting). The results are

reported in the Figure below (obtained by estimating the probability distribution parameters on 30-year calibration sub-samples).

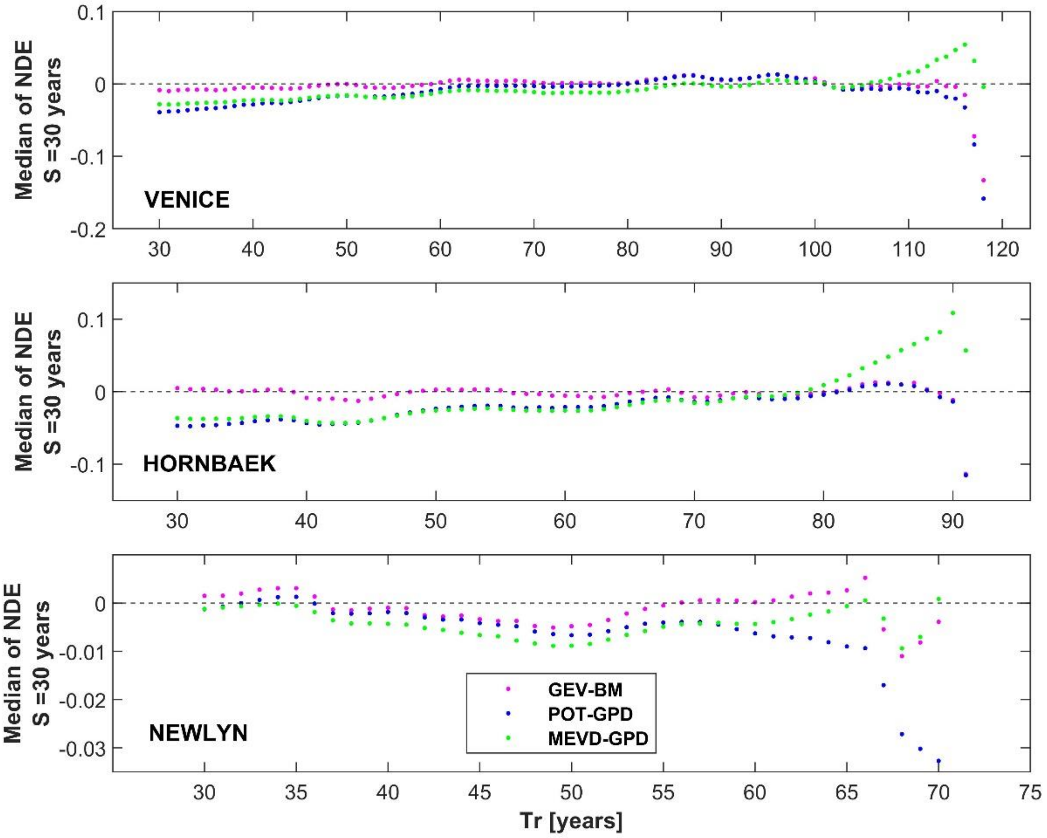

Figure 1. Median of non-dimensional error for any return period in test sub-sample obtained by estimating the distribution parameters on 30-year calibration sub-samples.

The new analyses suggest that when we focus on the median error associated with moderate values of the return period, GEV-BM displays an overall greater robustness (e.g., in the case of Venice and Hornbæk sites) with respect to POT-GPD and MEVD, which exhibit greater fluctuations. In particular, results show that MEVD is a good model for the highest values of the return period, but exhibit a greater absolute value of the estimation error for smaller *Tr*. The size of the available datasets does not allow to explore what happens for even greater values of *Tr*. Future work could investigate if estimation error can be reduced using different approaches to parameter estimation (e.g., by assuming “time-invariant” parameters in the ordinary distribution, whose estimation would thus be performed on the entire calibration dataset, rather than on relatively short sliding windows) and compute NDE for much greater values of *Tr* based on synthetic water level time series (which may be as long as is desired) produced by one of the several existing numerical models. We have now added this Figure to the manuscript, along with the above discussion. This addition, stemming from this Reviewer’s suggestion, also prompted a slight change of title, to reflect that the focus is a comparison of methods and to avoid suggesting that the MEVD is necessarily superior at all *Trs*.

Also, how the observed quantile h(obs,p) is calculated? Which sample (M, S, or V) is used? The Q-Q plots are mentioned only in the results section and they are only performed for the 30 years in-sample test. In my opinion, the Q-Q plots put the NDE into perspective and should be included as goodness-of-fit method. Also, it would be useful to have them in the main manuscript. I do understand that the space is limited, maybe the Authors could consider including in the main manuscript only the ones related to the MEDV.

We have added the following description: “As usual in frequency analysis, we associate to each observed yearly maximum, x_i, an empirical frequency value given by Weibull’s estimator $F_i = \frac{i}{(V+1)}$, where i is the rank of x_i in the list of yearly maxima sorted in ascending order, and $V = M - S$ is the sample size in the validation sub-sample. The return period Tr associated with each yearly maximum is then simply $Tr_i = \frac{1}{1-F_i}$."

We have now clarified that the test sub-sample (*V*) is used to extract the empirical quantiles to compare them with estimated ones. Regarding the Q-Q plots, we agree with the Reviewer and we have included in the main text the QQ-plots related to the MEVD.

In the section Return Period, the definition of Equation 4 needs to be further discussed. Even if the Authors replace (h) with (z-msl), Equation 4 is still the return period of (h), and not the return period of the (z), as indicated by the Authors. Mean sea level (msl) shows a clear linear trend and such trend is recognizable in (z). Similarly, in Equation 5, the distribution G is the distribution of the variable (h) and not the variable (z) as reported in line 341. This has an implication in Figure 5. I assume that the y-axis in Figure 5 "water level" refers to the variable (z). This variable (z) is time-dependent, while in Figure 5 it seems like the statistical properties of (z) are constant. I would have expected something similar to the effective return level plots, to show the effect of sea-level rise. How (msl), which is time-dependent, is added to (h), which is not time-dependent, to derive Figure 5? I suggest clarifying the transition from the analysis on the variable (h), a random variable, to (z), which presents a linear trend due to (msl). I also suggest being more precise with the notation and the terms used throughout the manuscript. It is very difficult to understand the variables the Authors refer to because are often called with many different terms, e.g., total sea level, water level, extreme sea level...

We apologize for the lack of clarity. In the revised manuscript, we have revised all the notation and terms used. In particular, we have indicated the variable *z* as "total water level" and *h* as "coastal water level". To avoid confusion, in Figure 5 we have replaced "water level" with "total water level" (i.e. the variable *z*).

Regarding the comments on $G(h)$, we see that some confusion may have arisen. We have now clarified how the exceedance distribution of variable H (coastal water level) is the same as the exceedance distribution of Z (total water level) by expanding the existing discussion (formerly lines 240-242), which now reads: "Because for a fixed value of mean sea level there is a one-to-one relation between the value of the sum of the astronomical and the storm surge contribution, h, and the total water level, $z = h + msl$, one can write $G_h(h) = P[H > h] = P[H > z - msl] = P[Z - msl > z - msl] = P[Z > z] = G_z(h)$, such that Eq. 3 can be used, once the cumulative distribution is known and for each value of *msl*, to determine the return period of the total water level: $Tr(z) = 1/1 - G_z(h)$".

This equality is independent of the fact that *msl* may change over time as the one-to-one relation between H and Z holds at all times. The possibility of projecting probability distributions and return periods into the future precisely depends on the fact that we need only substitute updated values of $msl(t)$ to infer the probability of future extreme total water levels.

The Authors say that "MEVD proves to be a good model for the extreme sea levels" (line 288) and that "MEVD-based estimates outperform the traditional approaches" (line 301). I do fail to see what the Authors describe. In the QQ-plots Figure S2-6, MEVD in the in-sample analysis has, in general, the highest variability, especially compared to the GEV. In the out-of-sample, MEVD looks better for lower quantiles, but it has quite a large variability for higher quantiles, compared to the other distributions. Overall, it is difficult to quantify which distribution performs best. This is also reflected in the NDE plots, Figure 3, where the differences between distributions are minimal.

We agree with this Reviewer that differences in performance are not large between MEVD and GEV, and the revised manuscript now does not draw a definitive conclusion on which approach is best independently from the return period of focus. However, small differences in the estimation accuracy are relevant for engineering applications when dealing with rare extreme events. Figure 3 shows a better performance by MEVD for the largest quantile for all sites except Marseille. We thus argue that the improved performance of MEVD for large Tr may have a significant impact on the effective

design of coastal defense structures (e.g., see Table 3 and Figure 4(a), (b) and (c)). The additional graph produced above to answer previous comments by this Reviewer shows small differences in the estimation accuracy of different approaches at different sites. In particular, the results suggest that no single approach is clearly superior at all values of *Tr*, due to a large variability in the estimates. For example, for the Venice site there is a decrease (in many cases an unbiased estimates) in MEVD NDE values for intermediate *Tr* (between 85 and 105 years) while for greater *Tr* values (but smaller than Tr_{max}) the error shows an overestimation of the actual quantile with respect to traditional approaches (which exhibit an underestimation tendency). To be more specific, if *Tr*> 105 years are considered, MEVD yields error estimates between zero and <10%, while errors associated with GEV-BM and POT-GPD lie between zero and <-20%.
The Hornbæk site shows similar results to the Venice site, while Newlyn's results exhibit more fluctuations for large *Tr* values with much reduced smaller amplitudes and values of the NDE.

Point by point comments:

Line 92. Please revise the notation. Pr(Mn<= x) = F(x)^n where Mn is the maximum of a sequence of independent random variable X. See also Coles 2001 (line 415)
Agreed, we have changed the notation accordingly.

Line 154. Additional discussion is needed concerning the fact that h(t) can be considered a stochastic variable even though a determinist component is included. Also, a literature review on indirect and direct methods (Line 149) for extreme sea level is missing.
Thanks for the suggestions. We have improved the discussion concerning the two aspects highlighted from the Reviewer.

Lines 133. The Authors discuss the negligibility of tide-surge interaction. Does this condition hold in the case of Punta della Salute which is located within the Venice Lagoon?
This statement may have generated some confusion and needs additional discussion. We now clarify that the tide-surge interaction is significant and needs to be taken into account when the surge and tide components are studied separately.
The revised manuscript now explains, at line 134, that: "However, this effect is significant and needs to be taken into account when the surge and tide components are studied separately. Since here we do not attempt to separate these contributions but we only analyze the sum given by the combination of the water level setup, induced by meteorological forcing, and the astronomical tide, hereafter we will neglect their non-linear interactions and we will consider the observed sea level as the sum of additive components".

How the GDP threshold is selected and tested?
As described from lines 272 to 275, the optimal GPD threshold value was determined by studying the stability of the GPD shape and modified scale parameters. To evaluate the goodness of fit of the distribution with respect to different threshold values, diagnostic graphical plots were constructed.
In the discussion to a previous Reviewer comment we clarify that we have moved the discussion about threshold selection, previously at lines 276-277, to line 121. The revised manuscript now discusses here that: "The threshold is set to be large enough to filter out water level peaks that are likely to be associated to conditions without any storm contribution and sufficiently low to maximize the amount of information used. In addition to the above, we choose the threshold value that produces the minimum extreme-value estimation error under the MEVD framework". Hence, the optimal threshold for the ordinary values selection is selected by testing different threshold values and evaluating the goodness of fit of the distribution by means diagnostic plots.

It would be very interesting and useful to appreciate the difference between the performance of the distribution functions to see the sample of maxima used for fitting the distributions.

We think the Reviewer is highlighting the need to include, in addition to Table 2, a comparative figure between the extreme time series used for fitting the distributions. The Supporting Information now includes this additional figure that displays the sample of maxima used to infer the distributions, i.e. annual maxima (GEV-BM), exceedances over the threshold (POT-GPD) and ordinary values (MEVD).

Lines 205-209. My suggestion is to revise this paragraph. The terminology is confusing. I believe the Authors here are discussing the variable (z), in which storm surge is a component.

Thank you, agreed. The revised manuscript now reports: “Future increases in the frequency of extreme sea levels due to climate change will have serious impacts on coastal regions. These impacts will vary temporally and regionally, depending on (i) the local mean se-level rise (including possible subsidence or uplift), (ii) current storm-surge intensity probability distributions, and (iii) changes in the dominant meteorological dynamics. In this particular application to extreme coastal water levels (i.e. the sum given by the combination of the water level setup, induced by meteorological forcing, and the astronomical tide), only the first two factors are considered”.

Lines 220-221. The Authors say that the tidal and storm components do not change over time as mean sea level. How did the Author check that no trend is detected in the variable h?

When we consider potential future changes in extreme high water level, our approach assumes to separate changes in mean sea level and atmospheric component. Here we focus only on the characterization of future changes in the statistics of mean sea level because it is the main driver of extreme sea level variations. Our statement, cited by this Reviewer, finds confirmation in previous studies of past and future changes in extreme high water levels (e.g., Zhang et al., 2000; Woodworth and Blackman, 2004; Menéndez and Woodworth, 2010; Lowe et al, 2010; Haigh et al., 2014; Wahl et al., 2017). According to this literature, it is reasonable to assume that increases in extreme high sea levels are primarily a result of the rise in mean sea level. This implies that variations in storm activity (e.g. magnitude, trajectories and frequency) are comparatively smaller than future rise in mean sea level at most locations. Our assumption is also confirmed in the IPCC AR5 report, which states that there is “low confidence” in region-specific projections of storminess and associated storm surges.

Future work could compare these two different entities, an analysis that is beyond the scope of our application.

Section 3: Was the trend test performed only on the annual maxima or also on the samples of maxima used to compute the GPD and the MEVD?

The trend test was performed only on the annual maxima. The revised manuscript now clarifies at line 251 that: “To answer this question, in this work we focus on the deviation of the yearly maxima from yearly mean sea level and test the presence of trend by the two-tail Mann-Kendall test”.

Line 281: Storm surge or storm surge and tide?

Thank you for pointing this potential lack of clarity. It is storm surge and tide. In the revised manuscript, we have revised all the notation and terms used. The variable defined as the sum between surge and tide is now indicated as “*coastal water level”*.

Line 285: what is L?

Thank you for catching this. We apologize for the mistake. “L” has been replaced with “M” which is the correct symbol used to indicate the time series length (as indicated in line 199).