We would like to thank dr. Víctor Malagón-Santos (Reviewer 1) for his thorough review and constructive suggestions. In this response we are addressing his comments and feedback. We have numbered the reviewer's comments (**R1.1, R1.2** etc.) in order to facilitate referencing to each comment. We have added here all the changes in manuscript (shown with green). The pages and line numbers in our responses refer to the revised manuscript with track changes.

Reviewer 1 (Víctor Malagón-Santos)

The manuscript by Athanasiou et al. proposes a surrogate model of a time-consuming numerical model (XBeach) for the prediction of coastal erosion under extreme conditions at the regional scale (the Netherlands' coast). The main benefit of the approach is speeding up the prediction process without significantly decreasing the accuracy of erosion estimates. This is particularly useful in the application of early warning systems. Similar methodologies have been proposed before, but the novelty here is the application on a regional scale and the inclusion of the beach profile as an input. The latter makes the model capable of reproducing coastal erosion for a wide range of typological coast profiles, enabling accurate and fast predictions regardless of the initial state of the beach. The framework is presented as a flexible and transferable tool, so similar applications to other processes should be viable.

The manuscript is well presented, it reads well, and figures are of scientific rigor. The proposed methodology is thoroughly explained and easy to interpret, facilitating its future application to similar issues. There are some instances where the quality of the manuscript would benefit from clarifications. For example, a few (but important) details of the statistical model used here were not explained in detail. Correcting this will not require great effort, hence I suggest accepting the manuscript after minor revisions.

We would like to thank the Reviewer for the positive comments.

Please find below a more detailed review, providing line numbers where some corrections or clarifications might help improve the quality of manuscript. I hope you find my suggestions useful.

R1.1:
Line 72. I would expect a sentence here highlighting the novelty and relevance of the proposed model as compared to the ones mentioned just before.

We agree with the reviewer and have now added this part*:*

[Line 72]: "The novelty of the proposed meta-model relative to previous similar methodologies is the inclusion of the pre-storm beach profile as input, which allows for large scale applications."

R1.2:
Line 125. What was the criteria for determining highly dynamic areas?

The highly dynamic areas included areas like the sand spits formed at the tips of the Wadden islands, were either way no dune features are present. This included some areas on the back of the Wadden islands (which are protected from North Sea storms). We have now changed the text to:

[Line 126]: "We excluded some of the 1,430 profiles from Athanasiou et al. (2021), which were found to be at highly dynamic areas (e.g., transect at the tips of the Wadden islands, where no clear dune features were identified) or transects at the non-exposed side of the Wadden islands, leaving 1,368 transect for our study."

R1.3:
Line 142. There could be Hs extreme events coinciding with a non-extreme SSL (and vice versa). What if, for instance, SSL (even though it does not exceed a threshold) is high enough to cause erosion together with an extreme Hs. Wouldn't it be more appropriate to include extreme Hs AND/OR extreme SSL?

While swell waves can occur at the Dutch coast even when SSL is not extreme, we don't expect these events to be impactful due to the high dune elevation at the Dutch coast. Furthermore, we picked these thresholds per station, based on the morphological characteristics of each region (see Lines 147-155), which means that they are representative of what constitute an event (i.e., collision regime) based on the local dune toe elevation.

R1.4:
Line 145. Remove (2016)

We have now removed it.

R1.5:
Line 172. I wonder if a GPD is appropriate to describe the marginal of D. D is not defined based on threshold exceedances here, and although a GPD can be used to describe other extremes (such as annual maxima), I am not convinced this variable may have a heavy-tail behaviour.

We based this choice on the results of Li et al. (2014), where a GPD was used to model the tail of event duration as well. To validate this, we performed a Kolmogorov–Smirnov test for all GPD fits and they were passed at the 95% confidence interval.

Li, F., Van Gelder, P. H. A. J. M. , Vrijling, J. K., Callaghan, D. P., Jongejan, R. B., & Ranasinghe, R. (2014). Probabilistic estimation of coastal dune erosion and recession by statistical simulation of storm events. Applied Ocean Research, 47, 53–62. https://doi.org/10.1016/j.apor.2014.01.002

R1.6:
Line 176. A Gaussian copula is mentioned here, so I assume it is a Multivariate Gaussian copula what is used here to model the dependence structure between the four hydrodynamic variables. This copula is chosen based on earlier work by Li et al. (2014b) as it was deemed suitable at the Ijmuiden-06 station. This may seem a reasonable choice, as long as the data used here does not differ substantially from Li et al. (2014b). But I wonder if the Multivariate Gaussian copula would also be appropriate for the other stations. I can see a validation at the Euro platform station (Figure 3), but not for the other stations.

Indeed, this is refereeing to a multivariate Gaussian copula. We have now added this to the text to make it more clear:

[Line 181]: "...and we then fit a multivariate Gaussian copula..."

We had decided not to include the pair-plots for each station in the initial manuscript for the sake of space. But the validation was done for all stations. We have changed this part of the text:

[Line 186]: "The relative differences between the observed and simulated dependency statistics

were on average smaller than 5%, 1% and 2% for ρ, τ, and TDC respectively, for all four stations (Figure 3 and Supplementary Figures S1-S3), verifying that the simulator was able to capture the dependency structure."

Additionally, now we have added a supplement presenting the event simulation and validation statistics for the other stations:

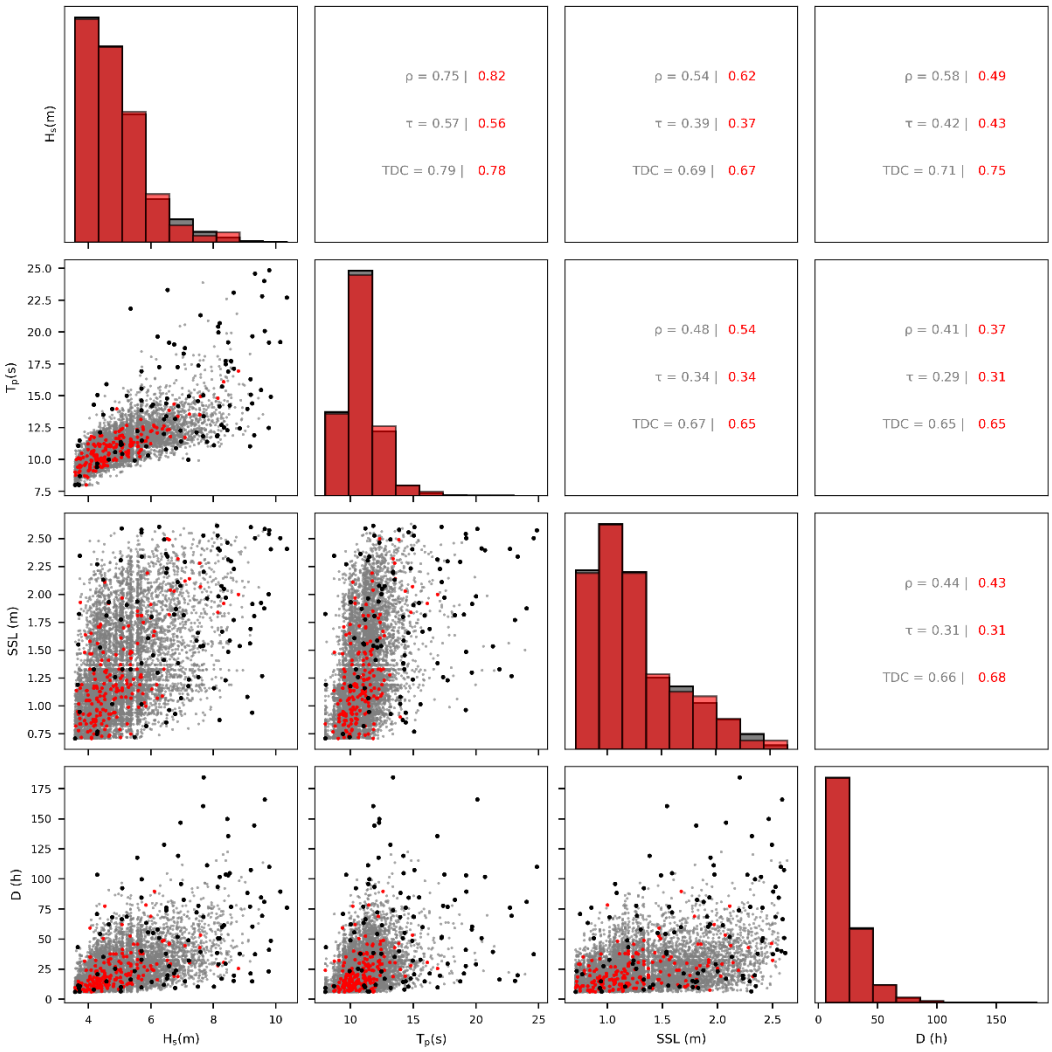

Figure S1: Copula-based events simulator for the Schiermonnikoog Noord station location (see Figure 2). Red, grey and black dots indicate observed, simulated and the 100 MDA selected events, respectively. The black dots are a subset of the grey dots (simulated events) that are selected with the MDA. Histograms of each storm parameter (*Hs, Tp, SSL* and *D*) for both observed and simulated events can be seen in the diagonal graphs. Below the diagonal, scatter-plots for each pair are plotted. Above the diagonal, 3 different dependency coefficients (ρ: Pearson correlation, τ: Kendall's rank correlation, TDC: non-parametric tail dependence) are shown for each pair for the observed and simulated events.

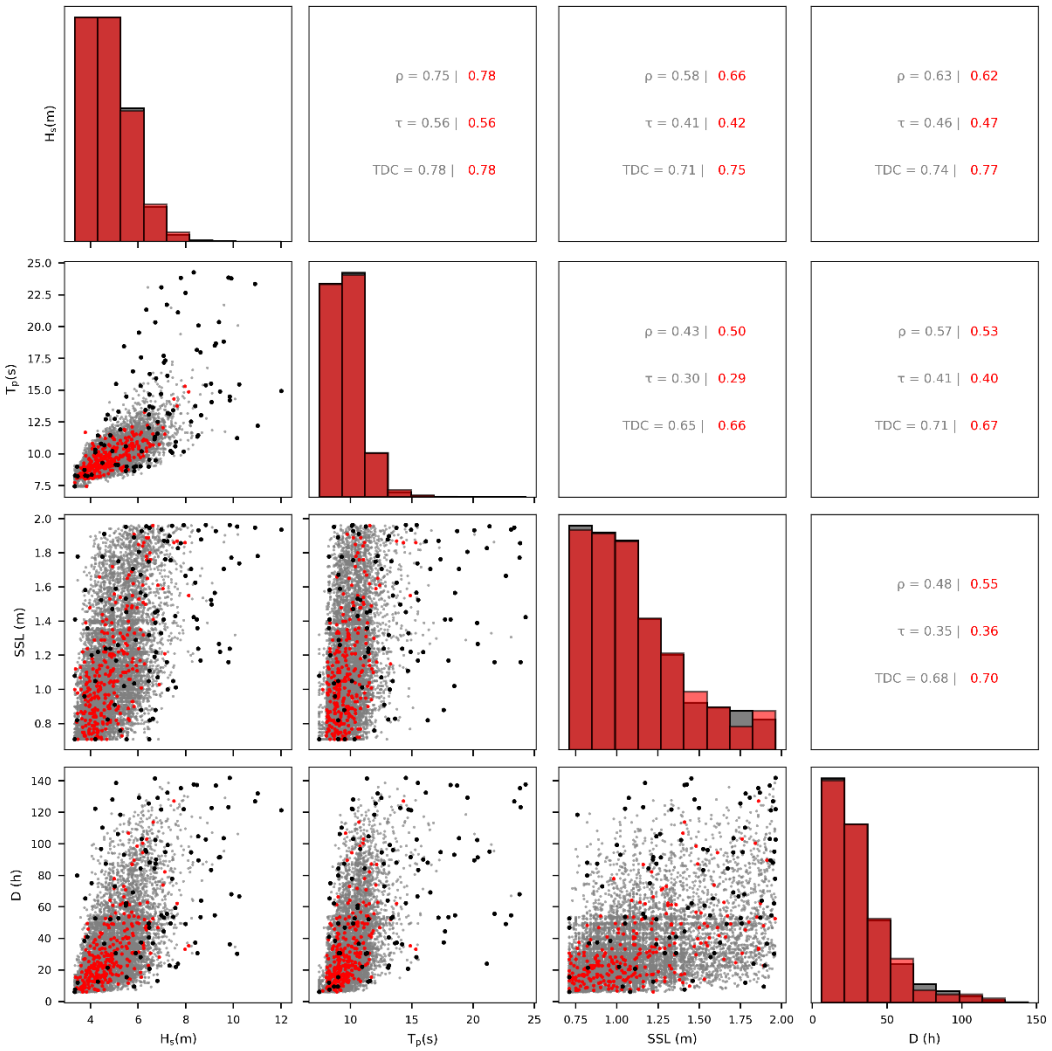

Figure S2: Copula-based events simulator for the Eierlandse Gat station location (see Figure 2). Red, grey and black dots indicate observed, simulated and the 100 MDA selected events, respectively. The black dots are a subset of the grey dots (simulated events) that are selected with the MDA. Histograms of each storm parameter (*Hs, Tp, SSL* and *D*) for both observed and simulated events can be seen in the diagonal graphs. Below the diagonal, scatter-plots for each pair are plotted. Above the diagonal, 3 different dependency coefficients (ρ: Pearson correlation, τ: Kendall's rank correlation, TDC: non-parametric tail dependence) are shown for each pair for the observed and simulated events.

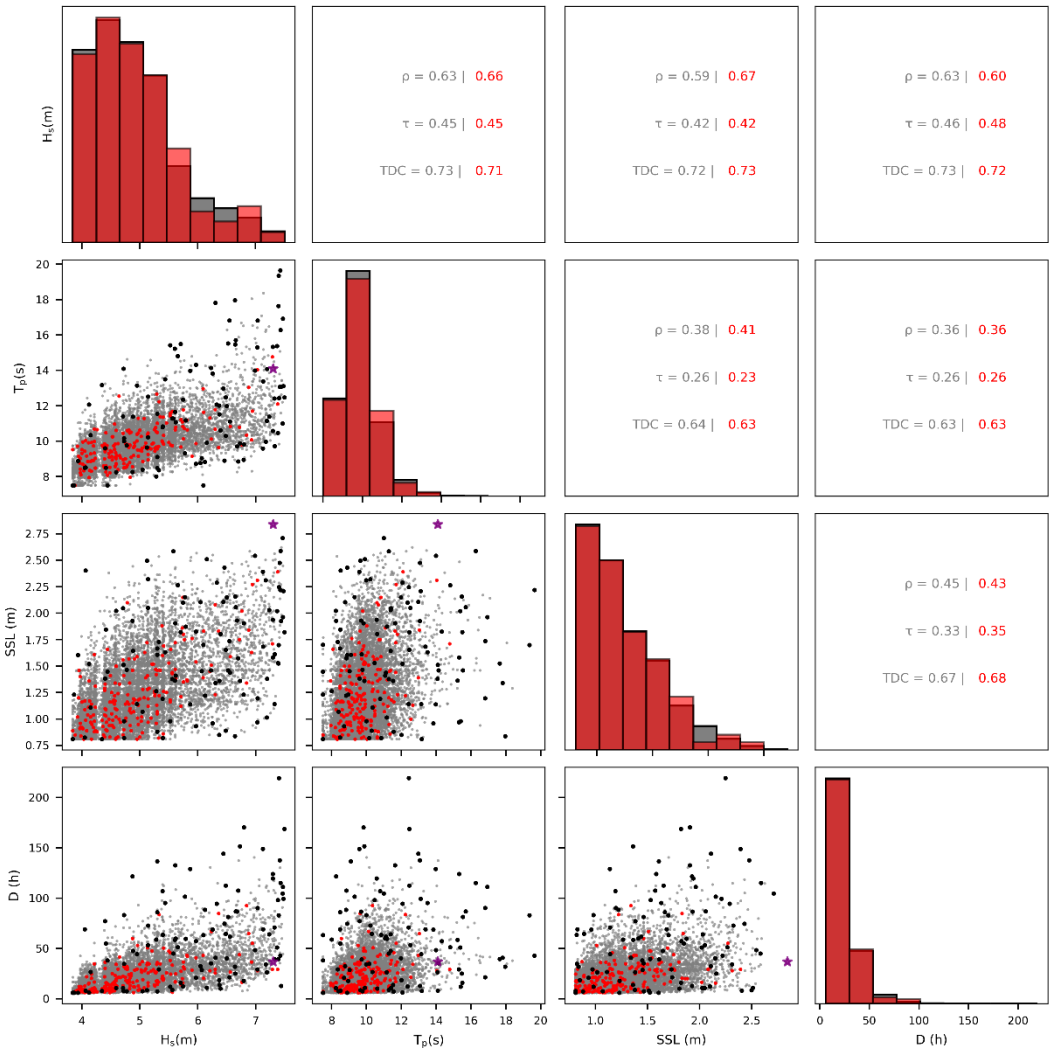

Figure S3: Copula-based events simulator for the IJmuiden-06 station location (see Figure 2). Red, grey and black dots indicate observed, simulated and the 100 MDA selected events, respectively. The black dots are a subset of the grey dots (simulated events) that are selected with the MDA. The purple star indicates the boundary conditions during the 1953 storm. Histograms of each storm parameter (*Hs, Tp, SSL* and *D*) for both observed and simulated events can be seen in the diagonal graphs. Below the diagonal, scatter-plots for each pair are plotted. Above the diagonal, 3 different dependency coefficients (ρ: Pearson correlation, τ: Kendall's rank correlation, TDC: non-parametric tail dependence) are shown for each pair for the observed and simulated events.

R1.7:

Line 180. 100,000 synthetic events are sampled based on Monte-Carlo. How does this number translate into length of data? What is the rate of events per year?

Since the objective of this step for this research was to create a representative set of training storms and not derive probabilistic risk estimates, the number of simulated events was chosen high enough to make sure that extremes are sampled well. But we did not model the frequency of occurrence of the events in a year. From the observed events though, the average number of events per year was ~6-7. Which would mean that 100,000 synthetic events would represent ~14,000 years, but as mentioned before we don't use this in our analysis.

R1.8:
Line 185. Related to the previous comment in line 176, it is stated here that dependency statistics were smaller than 5%, but I have the feeling this only refers to the data presented in Figure 3. Could you also report metrics for the other stations?

Please see our response to R1.6. We have now explicitly mention that this refers to all stations and have now added the figures of the other stations in a supplement.

R1.9:
Figure 3. The resolution of this figure should be increased. Also, how do the copula-based most extreme simulations presented here compare with estimates of most extreme historic events not included in the copula analysis as observations, such as the 1953 one? This could also give an indication of how realistic the most extreme synthetic events are, especially for simulations far more extreme than the observed ones. Perhaps a return period/value comparison between extreme historical events not included as observations (e.g., 1953) and synthetic events from the copula analysis would be insightful and reinforce your message about the suitability of the statistical model.

Any issues with the current resolution probably relate to the pdf conversion during the submission. During the final submission all figures will be provided in high resolution.

Following the observation from the Reviewer, we re-evaluated available data on the boundary conditions of the 1953 storm, and found that for the closest station of Hoek van Holland, a max TWL of ~ 3.85 m was observed. Using the local MHW from the Jarkus dataset this leads to a SSL of 2.84 instead of the 3 m we previously used. We have now plotted this event in the pair plots of the IJmuiden-06 station. It can be seen that for the SSL the 1953 is slightly larger than the largest SSL simulated. This can be connected with the difference in the way we calculated SSL for the observed events (we did a tidal analysis to calculate the historic tides from the TWL record), while for the 1953 storm we simply subtract the local MHW from the maximum TWL record. For the other parameters, (Hs, Tp and D) the 1953 storm fall well inside the simulated events cloud.

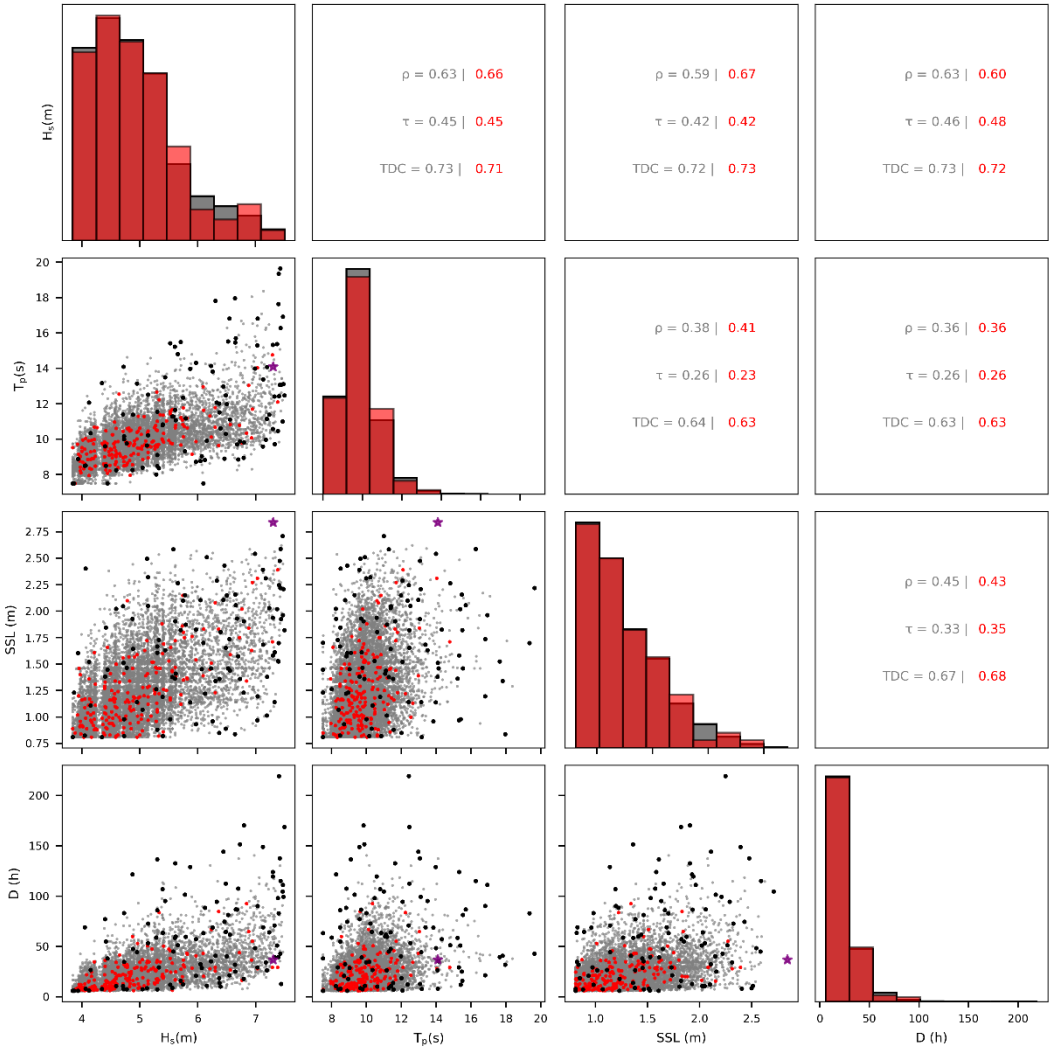

Figure S4: Copula-based events simulator for the IJmuiden-06 station location (see Figure 2). Red, grey and black dots indicate observed, simulated and the 100 MDA selected events, respectively. The black dots are a subset of the grey dots (simulated events) that are selected with the MDA. The purple star indicates the boundary conditions during the 1953 storm. Histograms of each storm parameter (*Hs, Tp, SSL* and *D*) for both observed and simulated events can be seen in the diagonal graphs. Below the diagonal, scatter-plots for each pair are plotted. Above the diagonal, 3 different dependency coefficients (ρ: Pearson correlation, τ: Kendall's rank correlation, TDC: non-parametric tail dependence) are shown for each pair for the observed and simulated events.

Using this updated boundary condition for SSL, we redone the validation of section 3.3. We the updated values, the predicted DEV for the 1953 is actually closer to the observed values than before.

Table 2:

Storm	SSL (m)	Hs (m)	Tp (s)	D (h)	Number of transects	Remarks
February 1953	2.84	7.3	14.1	37	1	A typical schematized profile of the Holland coast is used as the pre-storm profile, while for the observed dune erosion, reported values of 90±26 m^3/m are used (Van Thiel de Vries, 2009).

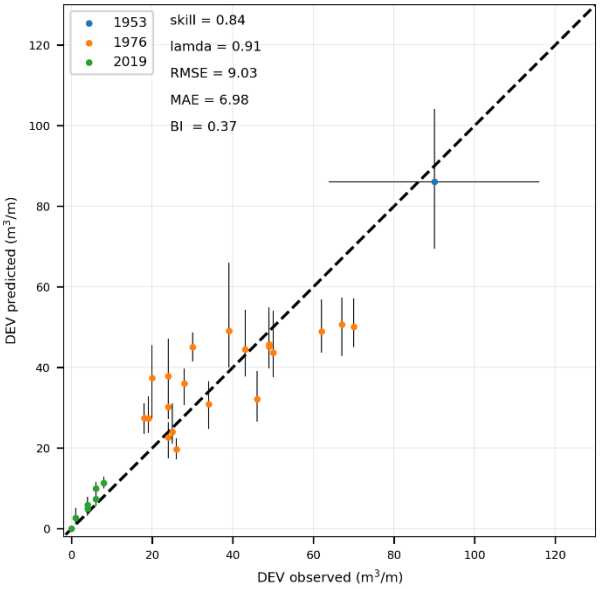

Figure 5: Scatter plot between the observed (x-axis) and ANN-predicted (y-axis) DEV, for three historic storms with variable number of transects per storm. The dots indicate the average predictions of the ANN, while the vertical lines show the min and max values as given by the ensemble of the ANN output. For the storm of 1953 the horizontal line gives the range of the values reported.

R1.10:

Line 275. You may mention this later, but how was this division done? 50/50? Was this a k-fold validation? How did you determine a suitable calibration/dataset division to ensure that was a good way of selecting the architecture of the ANNs?

Indeed, it was not a k-fold validation, since the training dataset was already created in a "clever" way in order to capture representative profile and storm conditions. The benchmark dataset was sampled out of the training dataset to ensure that these cases are unseen to the model. In order to ensure that the results were not biased to a single split, we repeated the split 10 times with different randomizations seeds. We explain this in the text:

[Line 285]: "The division of the benchmark dataset to a calibration and validation dataset was performed 10 times with different randomization seeds and the final mean error statistics were used, to ensure that any bias of the individual divisions was minimized. This meant that 10 different ANNs were produced (with the same architecture but different weights and biases), which will give an ensemble of DEV predictions, that can work as an uncertainty range."

R1.11:

Figure 10. Encouraging that the model performs best for the most complete pre- and post-storm profiles available (2019), but it is also true that this event was not particularly erosive. Would the model perform as well for more erosive events if we had complete pre- post/storm profiles as in 2019? I seem to remember the winter season of 2013-14 was particularly extreme (waves and surge) in northern Europe (this especially applied to the UK, but I imagine the Netherlands was also impacted by these series of storms). Are there records of complete pre- post-storm transects for that particularly extreme winter? It could be more insightful to show how the model performs for more erosive events while being validated with complete transects (if they are available).

This is indeed a great observation. Sadly, there are no pre- and post-storm measurements available for the winter of 2013-2014 to perform this kind of validation. Only a small part (~250 m) of the

coast at Egmond aan Zee was measured before the storm season (Ruessink et al. 2019), and the post-storm measurements have the effects of two consecutive storms (October and December 2013), which do not allow for validation.

Nevertheless, even with the incomplete measurements of the storms of 1976 and 2019, we could see that the expected variability between the storms, and between the profiles was captured in an acceptable manner.

Ruessink, G., Schwarz, C. S., Price, T. D. and Donker, J. J. A.: A multi-year data set of beach-foredune topography and environmental forcing conditions at Egmond aan Zee, the Netherlands, Data, 4(2), http://doi.org/10.3390/data4020073, 2019.

R1.12:
Figure 12. This is a nice and interesting figure. I wonder what's the effect of altering the number of TCPs included in the training process.

While this is a nice suggestion, it is not straight forward to compare this since in contradiction to the MDA algorithm that was used for the storms, the K-Means algorithm which was used for the TCPs does not maintain the order or the actual TCPs that are chosen for each cluster. This would mean that we would have to repeat all XBeach simulations for each scenario with a different number of TCPs, which was deemed out of the scope of the present study.

Nevertheless, we have already studied this effect in a previous paper focusing on clustering elevation profiles at the Dutch coasts (Athanasiou et al. 2021). There we tested different numbers of TCPs and saw that 100 TCPs was optimum to balance the computational effort and the good representation of the coast.

R1.13:
Line 552. There could be also problems in tropical-storm prone areas. Especially the fitting of marginal distributions and the copula approach, given the rarity of those events.

We have now added this part in the text:

[Line 554]: "Additionally, special care should be taken in applying these methods at tropical-storm prone areas, where it should be ensured that the training forcing conditions are representative of the extremes induced by tropical cyclones (Bloemendaal et al., 2022)."