Response to Referee #1

Referee comments are in plain text
Author comments are in BOLD
NHESSD Manuscript text is in *italics*
Added Text is in *Bold Italics*
(Line numbers refer to the marked manuscript)

We thank the Referee for the comments and address each point below.

I have read the manuscript with pleasure and a lot of interest and I think it is an important contribution. The manuscript is also well written and the presentation of the results is good. Therefore I am very favorable to its publication in your journal I only have some minor comments/suggestions which I feel would improve the manuscript. The authors compiled a new wave run up dataset, by extending the already broad Stockdon et al dataset with other measurements. Following they apply machine learning to conclude that the latter performs better in predicting wave run up heights. Machine learning in general have been already shown to be very capable predictors of wave runup. One of the earliest examples can be found here (https://link.springer.com/article/10.1007/s10236-011-0440-5) and maybe the most recent before the present work is Abolfathi et al 2016, with all studies reporting very good results.

Thank you for pointing us toward the manuscript of Vousdoukas et al (2011). We now cite this work on Line 97-99:

"Previous Machine Learning work has focused on predicting runup and swash, but only for engineered structures, impermeable slopes, and/or for laboratory experiments (e.g., Kazeminezhad and Etemad-Shahidi, 2015; Bonakdar and Etemad-Shahidi, 2011; Bakhtyar et al., 2008; Abolfathi et al., 2016) and not on natural beaches ***apart from Vousdoukas et al., (2011) which used artificial neural network (ANNs) for shoreline contour elevation (which includes the wave runup), on a natural beach in Portugal****."*

Since empirical run up formulas are simplifications of the actual processes deriving site specific 'recalibrations' of existing wave run up formulas has been considered a recommended practice. That way at least the effect of some unknown parameters is reduced. I would recommend the authors to mention that somewhere when they discuss previous studies, since among the recent ones mostly Stockdon 2006 aimed to propose a universal parameterization.

This is a good point, and we now added it to the discussion section.

Line 404-406:

"Generally the results from machine learning technique are strictly related to the range of the training and validation datasets (original dataset in Fig. 1). This work demonstrated that the applicability of the predictors can sometimes be used beyond the range of the testing dataset (new dataset in Fig. 1). However it is unknown how predictors will perform in settings beyond those in the present work — future tests on new field data are therefore recommended. ***Furthermore, parameterizations always work better when free parameters are optimized to a given site by using existing data and it should be considered when proposing universal parameterizations."***

Wave setup is part of run up and is driven by wave breaking. The latter is controlled by the nearshore beach slope (and not of the beachface only) a parameter which most times remains unknown, among others.

We added a comment that nearshore slope, which is a controlling parameter, is excluded from this predictor

Line 393-397

*"Looking at the limitation of the proposed models, the variables taken into account (H_0, T_p, L_0, β) are easily accessible but also oversimplify the processes that affect swash. For instance, we do not include the influence of the wave directional spread (Guza and Feddersen, 2012), the cross-shore wind component and the tidal range (Vousdoukas et al., 2012). However, in order to include these and other aspects (e.g., role of underwater vegetation, **nearshore bathymetry**) it is necessary to perform more field experiments that record swash, runup and other relevant variables. An additional limitation is that the swash formulas obtained in this study approaches a nonzero value as wave height approaches zero."*

and we clarify the concept at:

Line 411-413

*"This result is in line with studies such as Ruggiero et al., (2001 and 2004) and in contrast with Stockdon et al., (2006), Senechal et al. (2011) and Ruessink et al., (1998). **Although difficult to quantify and extremely simplified (this parameter together with sediment diameter should integrate the effect of the entire cross-shore profile), our results suggests that some parameter involving the beach profile should be considered when predicting runup characteristics."***

Moreover, infragravity motions and wave setup are not the same thing but they could be confused in some field measurements. The authors could elaborate on these aspects when they discuss infragravity parameterizations.

We distinguishing wave setup from infragravity motions, we state at the beginning of the manuscript:

Line 2-5

"The height reached by waves can be defined from water level elevation time series at the shoreline $\eta(t)$ as the sum of two distinguished components: the wave set up (the temporal mean of the time series $\langle\eta\rangle$ relative to the still water level) and the swash $\eta'(t)$ (the vertical fluctuation of the water level around the wave set up)."

and we have added (on line 74-75):

*"**In addition the similarity in the temporal scales of wave setup and infragravity motions could also be a confounding factor in measurements.** Finally, a number of other studies have also proposed other predictors that introduce other parameters to account for the cross-shore wind component and the tidal range "*

The weak point of machine learning techniques is that their predictive skill is limited to the conditions covered by the parameter space of the training dataset. GP is superior in that aspect to ANNs, since the final product is a relationship that is based on parameterizations which were derived considering the physical processes. At the same time it is not meant that the coefficients estimated will result in reasonable results beyond the range of the training dataset. In this case the training dataset is quite extensive but given that most of the global coastline is not included, it is not for granted that the solution could fail in other parts of the world.

First, we would like to point out that this GP routine is not 'aware' of physical processes (some previous GP work by other researchers have been forced to conform to physical laws). Only parameters and coefficients to combine these variables were given to the GP (along with the data).

We have clarified some aspects of our work in the discussion, and how it relates to generalization and extrapolation beyond the range of input data.

Line 366-367: "*In this work we use data compiled by Stockdon et al., (2006) to build new predictors, by the use of GP, for both total and infragravity swash elevations. We then test the generalizability of these new predictors using new data (including some extreme conditions).* ***This is different from previous applications of ML in coastal settings in two ways: First, we are testing the ML-derived predictor on data that is collected from a different setting (compared to the training data)— three beaches not included in the training data. Second, the testing data includes events that are outside the data range of the training data — we are extrapolating the ML-derived predictor as a test of its generalizability.***

Line 400-406

"*… the data used in the analysis does not include the limit condition of 'no waves-no swash'. Consequently, even if the GP formulas obtained do not correctly predict the limit condition corresponding to a no wave scenario, the prediction for both datasets has smaller errors compared to commonly used formulas.* ***Generally the results from machine learning technique are strictly related to the range of the training and validation datasets (original dataset in Fig. 1). This work demonstrated that the applicability of the predictors can sometimes be used beyond the range of the testing dataset (new dataset in Fig. 1). However it is unknown how predictors will perform in settings beyond those in the present work — future tests on new field data are therefore recommended."***

All this is not criticism, I just think the authors should discuss the above points. In addition I believe that it will be helpful for the reader to provide information about the range of input parameters for which the formulas are valid (maybe in form of a table).

We have included the information regarding the range of parameters on which the formulas have been developed on Figure 1, where the blue circles represent the training and validation dataset and the red crosses the testing one. Figure 2 includes the entire range of input and target parameters for which these predictors have been evaluated and tested (and therefore for which their validity has been assessed in the present work).

Moreover details on parameters intervals divided for experiment are reported in Table 1. For highlighting the aspects you suggested we included the following sentences at line 401-407

Line 401-407

"*... the data used in the analysis does not include the limit condition of 'no waves-no swash'. Consequently, even if the GP formulas obtained do not correctly predict the limit condition corresponding to a no wave scenario, the prediction for both datasets has smaller errors compared to commonly used formulas.* ***Generally the results from machine learning technique are strictly related to the range of the training and validation datasets (original dataset in Fig. 1). This work demonstrated that the applicability of the predictors can sometimes be used beyond the range of the testing dataset (new dataset in Fig. 1). However it is unknown how predictors will perform in settings beyond those in the present work — future tests on new field data are therefore recommended."***

Thank you for considering the revised version of this manuscript for publication in NHESSD