Comment on nhess-2022-221

Francesca Ribas (Referee)

Referee comment on "A predictive equation for wave setup using genetic programming"
by Charline Dalinghaus et al., Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2022-221-RC2, 2022

General comments

This article presents a new application of genetic algorithms to obtain empirical formulas for the maximum wave setup. The authors train and test a GP model with 9 different data sets of wave setup measurements. The pieces used by the GP model to build the equations are 6 variables and 6 operators, which are among those used in previous setup formulas. Then, they compare the obtained predictors with other 7 existing formulas for maximum wave setup. They obtain an extremely simple predictor that give wave setup with the same accuracy than the best of the previous existing formulas and a more complex expression that outperforms them.

Obtaining more accurate formulas for the maximum wave setup is important to increase our capacity to predict flooding since wave setup can contribute significantly to inundation episodes. This is especially crucial in the present framework of climate change and the urgent need of quantifying its potential effects. Beyond the fact that the two new obtained formulas can be directly used, the article also shows the potential of a new methodology (genetic algorithms) to capture the trends and provide accurate formulas for wave setup. This is very interesting since it can be applied to obtain better empirical formulas as soon as new better-quality data is available. The article is well written and their approach and results are of great interest for the coastal research community.

My overall impression is very positive but I still have a few comments and suggestions that might improve the manuscript: there are a few extra analyses that I missed in this study (specific comment 1), some parts of their methodology and results should be described in more detail (specific comment 2) and some parts of the text could be synthesized (specific comment 3). Below there is a longer description of these comments, together with a list of possible technical corrections. Overall, I recommend publication after these minor issues have been considered by the authors.

Specifics comments

1) Potential extra analysis (from more to less important) that could be added:

- To my opinion, it would be extremely interesting to apply the two new formulas together with the other 7 to a new data set not included in the training and test sets. This would allow quantifying if the new formulas maintain their better predictive capacity beyond the data that was used to train them.

- If available, it might be enlightening to show beach profiles corresponding to the 9 data sets used in this study. Maybe the sets that show less correlation (Duck 94 and SandyDuck) have a specificity. In particular, I would say that Duck beach has a mixture of fine and coarse sand and it often displays a double slope profile, with a larger slope at the foreshore corresponding to a coarser portion and a smaller slope in the rest of the profile linked to the finer portion.

- A variable that is important for nearshore processes that do not appear in the study is wave direction with respect to the shore normal. Do the authors have the wave direction corresponding to the different points in their data set? At least, they could plot the mean direction during the experiments and discuss if the sets that show less/more correlation show a specific direction.

- I am curious to know what would happen if Eq. (14) is applied using only the two first terms (so without the term related with grain size). Maybe this is out of the scope of this study but what would happen if this is plotted in Figure 5, too? That is, I wonder how sensitive is equation (14) to the last term? The text related to Figure 5 about the role of D50 (lines 240-244) might be expanded by including such analysis, if the authors find this interesting.

2) The following issues should be clarified in the text:

- Section 1: How are surf zones slopes and foreshore slopes defined? (e.g. which water depths?)

- Line 124: Iribarren number in the paper is computed with the foreshore slope, right? This is not what was used in the previous study where it first appear and thereby in the definition of line 58. Please,

- Section 2: The authors should justify the choice of the variables in their GP model. The 6 used variables sound completely reasonable to me given the previous predictors published in the literature but I think the paper would benefit from a sentence arguing this. Moreover, why don't they also include beta_s (surf zone slope)? Related to this, a comment about the introduction. Somewhere at the beginning of page 3, it would be nice to summarize what are the variables that have been used in the existing predictors (that are presented in the following paragraphs). The variables keep popping up each time a new predictor is presented and I think that anticipating potential variables at the beginning of page 3 would improve the text.

- Section 2: Also, the choice of operators in the GP model should be justified. In particular, what does it mean x^x? Why is this operator chosen instead of x^y or x^{const}? Also, why the parameter values are limited from -5 to +5?

- Line 174: Clarify what is a parsimony coefficient and how does it work.

- Section 3: How the GP model arrives at the factor 1625 in Eq. (14)? What is written in the Methods section is that constant range is from -5 to +5 (e.g. Table 2).

- Section 3: How robust is the result of the GP model? How the final formulas are chosen? Every time it is run, even with the same input parameter values of Table 2, it provides different formulas like Eq. (13) and (14)? Or it manages to always converge to these two selected ones?

- In line 233, the authors write that Duck94 showed less correlation than the rest. However, watching in detail at Fig. 4, my impression is that SandyDuck also stands out, at least in Eq. (13) plot.

- Line 254: The authors mention that the best of the previous models (Ji et al., 2018) has the disadvantage of having one coefficient more than Eq. (13). What do they mean?

- Line 270-271: Ji et al. (2018) formula also presents a good fit for dissipative and reflective conditions, right? This should be acknowledged.

3) The text is in general well written and to the point but, to my opinion, the following parts could be synthesized:

- Delete lines 18-24 in the Intro. This contains standard knowledge, included in any book on nearshore processes, it is not needed in research articles, in my opinion.

- Delete or synthesize lines 40-45 in the Intro. Is this adding something, given that most of the formulas below are purely empirical? At least, the formula could be deleted, maintaining only a summary of the main findings of Bowen et al. (1968). Alternatively, when introducing the work of Battjes (1974), the previous result of Bowen et al. (1968) could be acknowledged.

- Lines 250-258 could be more to the point. For example, I think it is unnecessary to write all the numbers of Table 3 (they can be seen on the Table).

Also, some of the suggestions of the list below are meant to synthesize the text.

Technical corrections

Line 15: According to the same authors, as -> As

Line 28: such as -> including [I understand the authors write here two examples of nearshore currents that are especially sensitive to setup, other currents being less sensitive to it.]

Line 29: in the flow circulation and so to sediment exchanges -> in the flow and sediment exchanges

Line 34: Delete "using Eq. (2) as the initial point" since many of the references use empirical approaches, right? [in fact, I suggest to delete Eq, 2]

Line 46: Define in one sentence \eta_M as (= maximum setup, which always occurs at the shoreline). Being the main character of your story, it deserves a careful definition, right?

Line 57: than only using Hs0 by relating setup and the surf zone similarity parameter -> by relating setup with the surf zone similarity parameter

Line 58: L0 should be defined here instead of after Eq. (6). Also, the authors could already introduce wave steepness here, instead of doing it before Eq. (7).

Lines 58-60: Since "foreshore slope" is not mentioned before in the text, I do not understand this sentence.

Line 63: I suggest to write Eq. (6) right after it is mentioned for the first time, so before the sentence that now starts "The equation (Eq. 6)…"

Line 70: was also -> had also been [I would say]

Lines 86-87: simplifications, uncertain or -> as well as simplifications, uncertainty or

Lines 113-114: This sentence belongs more to the Introduction, to my opinion.

Line 129: conditions with -> conditions, with

Table 1: Since every data set occupies two lines in the table, I suggest to add horizontal lines to separate between data sets and increase readability.

Figure 2: It should be enlarged significantly to make it readable. Probably it should be a horizontal 1-page figure.

Line 170: From the best solution, a new solution is created -> From the best solutions, a new set of solutions is created [What understand is that this step is done after a new generation has been created, in order to create the next generation. Thereby many

solutions are created, right? Also, to make a crossover you need to combine at least 2 parent equations…]

Line 181: The range of parsimony coefficients in the text do not include the value finally used (Table 2).

Lines 207-209: Order the citations by year, as in the rest of the article

Figures 3-6: They could be slightly enlarged to make them more readable (or maybe only enlarge numbers and letters font).

Line 212: predictor which -> predictor, which [I would say that which sentences must be between commas]

Line 213: interpretability is -> interpretability, is [I would say that which sentences must be between commas]

Line 217: This equation is -> It is

Figure 5: My impression is that showing the results of the two equations in the same panel would be more illustrative, since they could be more easily compared.

Line 279: presented in the equations are the same used-> present in the two obtained predictors are the same as those used

Line 284: was also introduced -> also appears

Line 298: The sentence starting with "The novel inclusion…" could start a new paragraph, given that this one and the following sentences cover a different topic, right?

Line 304: Equipments as -> Applied equipments include

Line 330: Currently, innovative data-driven approaches, such as genetic programming -> Innovative data-driven approaches, such as the genetic programming applied in this study

Line 336: We expect that the results of this work will -> The results of this work can