Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2020-347-RC2, 2020

Natural Hazards and Earth System Sciences
Discussions
Open Access
EGU

Interactive comment on "Quantification of Continuous Flood Hazard using Random Forrest Classification and Flood Insurance Claims at Large Spatial Scales: A Pilot Study in Southeast Texas" *by* William Mobley et al.

Jerom Aerts (Referee)

j.p.m.aerts@tudelft.nl

Received and published: 17 December 2020

General Comments:

The authors developed a novel method for creating continues flood hazard maps using a random forest model trained on a flood insurance claim dataset and independent terrain and hydrological predictors. The necessity of such a method is clearly motivated and shows great potential as a quantitative alternative to the currently in place flood hazard maps. The novelty of the study lies in the use of the insurance claim dataset.

Printer-friendly version

Discussion paper

Overall, the manuscript is well written and shows interesting new insights into how random forest models can be used for flood prediction.

Although the majority of this study uses open datasets, the main dataset in question is not open to public due to privacy concerns. Therefore, it is not possible to reproduce the results and to assess the quality and or accuracy of this dataset. An anonymized version of this dataset would elevate the study and benefit the flood modelling community. In addition, the study would benefit from archiving the code used to create the model and analyze the results.

Specific Comments per chapter:

Introduction:

L40-41: The authors propose machine learning models as an alternative. The introduction would benefit from a broader overview of the different types of available flood hazard model. For example, probabilistic catastrophe models or physics-based models. The added value of ML models would then be more clear.

Sampson, C. C., Fewtrell, T. J., O'Loughlin, F., Pappenberger, F., Bates, P. B., Freer, J. E., and Cloke, H. L.: The impact of uncertain precipitation data on insurance loss estimates using a flood catastrophe model, Hydrol. Earth Syst. Sci., 18, 2305–2324, https://doi.org/10.5194/hess-18-2305-2014, 2014.

Wing, O. E. J., P. D. Bates, C. C. Sampson, A. M. Smith, K. A. Johnson, and T. A. Erickson (2017), Validation of a 30 m resolution flood hazard model of the conterminous United States, Water Resour. Res., 53, 7968–7986, doi:10.1002/ 2017WR020917.).

L46-47: This statement needs clarification. What sampling procedures were used by previous studies? How do they undermine model reliability? How is your method better than previous studies? Do previous studies refer to Woznicki et al. (2019) and Knighton et al. (2020)?

Background:

L71: What does “through this process” refer to?

L74-75: Is this the case for all flood hazard models or only for all ML flood hazard models?

Methods and Materials:

L115: Why did the authors choose the period 1976 to 2017? How often is the NFIP dataset updated? Please clarify.

L118-119: In Barbet-Massin et al., 2012 (referenced in this line) one of the conclusions is that pseudo-absences taken too far from the presence data would not be very informative. The spatial extent of the study would impact model performance. Please elaborate on, if and how the methodology constrained the selection of pseudo-random samples being taken to far from the presence data.

L121-125: Due to the importance of this section for the methodology please clarify what is meant by “one-on-one sample by watershed and year” and elaborate on why this reduces potential bias form an unbalanced dataset.

L129: Why and how was the decision made to resample to a 10-meter raster instead of a 30-meter raster? Table1 indicates that most variables are native at a 30-meter resolution.

L153-155, 160-162: The reasoning behind choosing the nearest date due to absent data is valid. However it could have implications (skewed behavior) on the results as the authors stated the importance of land use change. Please highlight this in the discussion chapter.

L202-203: The manuscript could benefit from an overview table or flow chart of method and model setup, e.g. number of trees, tree depth, amount of data points, inputs, outputs.

L212: Figure should be in the main text instead of the appendix. After the introduction

of each variable it should be placed here.

Results:

Are neutral water bodies excluded from the analyses? This would cause a high positive hit rate bias.

L231: As a reader I am interested in the results of the twelve watersheds individually as it gives an indication of spatial model performance. If possible, please add to the appendix and refer in the main text.

L252-258: The text expresses flood probability in percentages while the figure uses return periods.

L253: The text mentions that the plots appear visually similar but in fact differs, a table should be added to clarify the differences.

Discussion:

L260-261: “computationally efficient” although one might expect this from a random forest model there is no mention of model runtime in the manuscript making this statement not well founded.

L261-262: “Overall, the model creates an accurate representation of flood hazard for the study area and demonstrates strong discriminatory power.” This conclusion is difficult to make as the SFHA flood hazard maps are deemed outdated. Moreover, the results are not validated against actual flood events.

L267-268: Interesting, please add a range of values, or a plot of slope against model performance.

L272-275: This should be made more clear in the methods chapter.

Conclusions:

L291-293: The study area is sufficient in size that the assumption of homogeneous

precipitation is invalid. Given the importance of mesoscale convective systems for major flood events one can argue that the location and duration of such systems is a main driver for floods.

L300: Although the ease of using TWI is shown in the manuscript it is not used in the final model, please use another alternative driver as an example.

Figure corrections:

Figure 5: This figure could use a clearer description of what is depicted. Headers above each column enhances readability. In addition the figure would benefit from an overview map depicting the location of each column.

Figure A1: Font size is too small. Add a vertical line at 0.05 to indicate which variables are removed. This is only clear based on the text.

Technical corrections:

L28: NFIP abbreviation not written out, first occurrence in L49-50.

L38-39: Missing spaces.

L191: Start new paragraph.

L219: “A calibration plot shows”, please refer to the figure number.

L233: “was”, typo?

L235: The word “should” seems not necessary in this sentence.

L247-249: Some typesetting errors.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-347, 2020.