REPLY TO REVIEWER CATERINA SAMELA

Dear Dr. Caterina Samela,

We would like to thank you very much for your comments and suggestions. Your help is strongly appreciated, and we believe it will significantly contribute to the improvement of our manuscript.

In the following points, we address your major comments:

1. ***Reviewer:*** *Terminology: "flood hazard" maps is a terminology more appropriate to maps derived by hydrologic/hydraulic simulations. Topography-based (hydrogeomorphic) maps are generally termed in literature as flood-susceptibility maps, or flood-prone areas map or floodplain maps (see e.g. indersson, S., Brandimarte, L., Mård, J., and Di Baldassarre, G.: Global riverine flood risk – how do hydrogeomorphic floodplain maps compare to flood hazard maps?, Nat. Hazards Earth Syst. Sci., 21, 2921–2948, https://doi.org/10.5194/nhess-21-2921-2021, 2021.)*

   Authors: Thank you for pointing out our misleading terminology. We will revise our manuscript thoroughly according to your useful suggestion.

2. ***Reviewer:*** *One of the most important issues addressed in this work is the estimate of the water depth, a parameter of fundamental importance especially for estimating expected flood damage. Compared to the large number of published studies on the delineation of the areal extent of flood hazard areas, in the literature there are fewer studies concerning the estimation of water inundation depth with simplified methods. This is an added value of this work. However, since DEM-based methods find their primary purpose in applications in data-scarce environments (although not exclusively), it is perhaps worth because while reference data to calibrate the classification problem are often available also in these contexts, on the opposite flood hazard map providing water depth values (to use for calibrating the regression problem) are more difficult to find. In addition, this data should be characterized by good accuracy to train a simpler but reliable model based on it. I think a consideration on this aspect can find a place in the manuscript.*

   Authors: This observation focuses on a particularly important aspect of our work that we did not address in detail. Indeed, predicting the expected maximum water depth is more complex than delineating flood-prone areas, and this requires more accurate data to effectively train machine learning algorithms. A possible solution to the application of our approach to data-scarce environments is to use the outcome of continental/global studies, when they are available (see e.g. the dataset provided by the Joint Research Centre at 25 m resolution at global scale; this dataset, that is the one we used to train our models, should be sufficiently reliable to train and validate machine learning algorithms, even if it does not consider –as in our case study- watersheds with lower extension than 500 km^2), or the outcomes of detailed hydraulic studies in specific portions of the study region, again, where and when available. As suggested by the reviewer, we will include some considerations on this truly relevant issue, by revising the

introduction and discussion sections.

3. ***Reviewer:*** *I wonder about the choice of identifying the calibration area by setting a constant-radius buffer. In this study, testing the performance outside of calibration areas is part of the application, so it was possible for authors to perform a sensitivity analysis on the accuracy obtained with different buffers. However, readers who want to apply the methodology with no possibility to validate the results (e.g. in poor data environments) are left without guidance on how to set this constant buffer. Here, in the same work for the same study area, two different buffer values are considered the best for the two reference maps (2 km for the 500-years PGRA flood hazard map, and 5 km for the JRC 100-years flood map). Instead, a topographical-hydrological criterion (e.g. the one used by Degiorgis et al., 2012) offers the possibility of being adopted and re-applied in any context, responding at the same time to the characteristics of the study area and of the available reference map. This consideration does not influence the relevance of the investigation and the interesting results obtained, but is made thinking about how to replicate the study in different case studies.*

   Authors: The point highlighted in this comment is truly relevant, and we are pleased to have the occasion to elaborate further on it, while revising our manuscript. During our research, several experiments have been performed trying different calibration areas, that can be divided into two groups: the first consists of the merger of all the elementary basins (i.e. hydrological units that drain directly into an elementary stretch of the river-network) that are entirely or partially included in the target flood-prone areas; the second consists of the areas within a fixed buffer around the target flood-prone areas. Testing these alternatives, i.e. training our classification models on these two alternative calibration areas, we observed the same performances of the trained models, as opposed to an enormous difference in terms of computational effort for defining the calibration areas associated with the two techniques (i.e., buffering the target map is computationally more effective). For this reason, further analysis has been conducted with a fixed buffer calibration area.
   The same approach to define a specific area to calibrate the models has been proposed and used by Tavares Da Costa et al. (2019), who used a fixed buffer of about 1 km around the flood-prone areas of the target map, and named this as “classification area”. The approach by Degiorgis et al. (2012) is remarkably similar, as they considered for the training just the pixels within flood-prone target areas and their conterminous.
   During our research, we observed that the choice of the radius for the calibration buffer has some influence on the results, and, it needs to be large enough to enable the model to recognize non-flood-susceptible pixels. Also, as stated in the manuscript, different radius can perform differently depending on the target map being considered.
   Our manuscript does not provide the interested reader with enough detail on the choice of the right buffer radius for the calibration area, which requires some sensitivity analysis. We will better detail this part and suggest to resort to Tavares da Costa et al. (2019) or Degiorgis et al. (2012) when a sensitivity analysis is not viable (i.e. possibility to validate the model).

4. ***Reviewer:*** *The analyses are made up of a series of steps and sub-steps (and further sub-steps), not always easy to follow along, that are listed in the first lines of Section 4 “Framework of the*

analysis". Then, subsections 4.x do not follow any of the previous subdivision. Did you consider that the methodology would be easier to read, follow and reproduce if a subsection is dedicated to each of the major 4 steps?

Authors: We appreciate this suggestion, as it highlights a key section of the manuscript that can be significantly improved. We agree with this comment, and we think that the structure of Section 4 should be redesigned to be more linear and representative of our methodology.

5. ***Reviewer:*** *In tables 1, 2, 3 can be unclear the difference among the results of the first two rows and the other rows. Section 4.2 reports that the models have been applied a first time using the entire domain of the calibration areas, and then the models were applied again four other times after selecting four subdomains of the calibration area (to test extrapolation performances). I believe this should be better clarified and the section 4.2, in general, could be reorganized. For example, it first describes what happens in phase (3), then in phase (4), and toward the end of the section is nominated phase (2). Is there a possibility to simplify and re-order this description?*

   Authors: Many thanks, the difference of what we did in phase (3) and phase (4) is a fundamental aspect of our study. In phase (3) all the pixels in the calibration area were randomly divided into two groups: 85% for to train the models, and 15% for the validation. As the split was random, the datasets for the training and the validation have the same statistical distribution of the different values for the seven input indexes and for the target values. In phase (4), the pixels in the calibration area have been divided into training and validation sets with a geographical meaning. This leads to applications of the same approach to real world applications, where the input dataset has some hydrological and morphological characteristics depending on its morphological and hydrological features, and the validation set has different characteristics. We are extremely glad that this has been pointed out; we will rewrite Section 4 in a more effective and clear way.

Finally, we want to state also that we are very grateful for the other minor comments, that we will address while revising our manuscript.

Thank you again for your appreciated help,

Kind regards

Andrea Magnini, Michele Lombardi, Simone Persiano, Antonio Tirri, Francesco Lo Conti, Attilio Castellarin