REPLY TO REVIEWER ZHEJUN HUANG

Dear Dr. Zhejun Huang,

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.

We propose to address your comments as we illustrate below. For easing the reading of our rebuttal, original comments are reported in italics after the tag "***Reviewer:***", while our reply is flagged using the tag "**Authors**".

Reviewer: *In the present study, an approach that blends several geomorphic descriptors together using the Decision tree method is proposed and is compared with a univariate approach, Geomorphic Flood Index (GFI). Only DEM-based geomorphic descriptors from pre-existing flood hazard maps were used. The decision tree method was employed to establish two types of models: one for classifying flood-prone areas and one for predicting water depth. Seven distinct geomorphic descriptors (GDs) were considered and blended in the multivariate approach. They demonstrated that the performance of the proposed multivariate approach was better than the univariate approach using GFI only.*

The article is well written and easy to read. However, tables need to be improved to be more readable.

Authors: we thankfully acknowledge the overall positive evaluation of our manuscript, and take advantage of this further opportunity to thank the Reviewer for his useful comments. We will improve the readability of all tables.

Reviewer: *In the present study, a new approach that blends several GDs together is proposed. This work indeed fills some gaps in the field of flood hazard assessment, the techniques and methods are feasible. I consider this work a contribution in this field and this manuscript can be accepted for publication after revisions.*

Major comments.

1. *Only one univariate using the geomorphic descriptor GFI is used as the comparison method. Is the performance of the univariate approach using GFI better than the other ones using the other six GDs?*

   Authors: Thank you for focusing on this point, that is very important for the evaluation of the results. We compared the multivariate approach with the GFI-univariate only based on: (1) several contributes in the literature indicating GFI as one of the most informative morphometric indexes for floodplain delineations (e.g., Samela et al., 2017, https://doi.org/10.1016/j.advwatres.2017.01.007), and (2) our preliminary analyses, all showing that GFI has better performances than any other index considered in our study. We will explain

with more detail this point in the next version of the manuscript.

2. ***Reviewer:*** *Does the proposed approach have advantages over the existing multivariate approaches? I would suggest comparing the proposed multivariate approach with one or two existing multivariate approaches to make this study more comprehensive and convincing.*

   Authors: The point highlighted in this comment is relevant, and we are pleased to have the occasion to elaborate further on it, while revising our manuscript. Our study does not aim specifically to propose a more accurate multivariate approach. We focused on an alternative way to approach multivariate DEM-based flood hazard assessment that differs significantly from other multivariate approaches presented in the literature (see e.g., Janizadeh et al, 2019, https://doi.org/10.3390/su11195426; Costache et al., 2020, https://doi.org/10.1016/j.jenvman.2020.110485); we tried to summarize the innovative aspects in the introduction (mostly) and conclusion sections of the manuscript, and we will revise these sections to make our message as clear as possible.
   Our approach is associated with two main advantages relative to existing methods. The first advantage is the feasibility and repeatability, as we only used descriptors that can be easily retrieved from DEM processing, while other studies exploited additional information (e.g., about geology, soil type, precipitation) that needs more extensive research and could be in some cases unavailable or unreliable. The second main advantage is the reliability, as our study does not use as target information records of historical events, instead, our model is trained and tested against previously published flood hazard maps, which consist of ensembles of hydraulic model output and represent scenarios with a certain return period.
   Due to the second point, it is quite difficult to compare performance metrics of our model with the ones of other studies due to the intrinsic differences among them; also, our performance metrics are computed pixel-based on a continuous domain, while other multivariate flood hazard assessments simply refer to sparse geographical locations that were inundated or not by specific historical flood events.
   Undoubtfully, the comparison between our approach and other multivariate models is very interesting. Nevertheless, our aim is not to improve the accuracy of multivariate DEM-based approaches, instead, we want to evaluate the benefits derived from the combination of different DEM-based descriptors compared to a univariate approach. Indeed, the Reviewer's suggestion is a good topic for further studies.

3. ***Reviewer:*** *How to determine the ratio of training and testing set and why? The authors should clarify it since the ratio is quite crucial for the learning model and thus affects the performance of the approaches. Usually, the performance in the test set is accepted rather than that in the training set. However, the manuscript used the results in the training area in the abstract (as shown in Table 1).*

   Authors: We sincerely appreciate this comment, as it allows us to give more details about some important aspects of the study. The ratio between training and testing followed two different rules during two consecutive phases of the study. First, we divided the entire dataset (i.e., the calibration area, consisting of the floodable zone close to rivers with buffer) into 85% for

training and 15% for testing, based on established proportion adopted for machine learning algorithms. This produces two datasets that contain millions of pixels, enough to compute reliable metrics. Second, during the extrapolation experiments, we divided the study area as showed in figure 7; training and test datasets amount to millions of pixels in this case as well. We reported performance metrics for both training area and test area to show the readers that our models are not overfitting on the training set, which is a frequent problem of machine learning algorithms.
Indeed, use of the training metrics instead of the test ones in the abstract results is a mistake and we will correct it.

4. ***Reviewer:*** *Tables 1, 2, and 3 are not very clear. I would suggest bolding the important values in these tables.*
   Authors: This suggestion is appreciated, we will make more readable the tables of the manuscript.

5. ***Reviewer:*** *The investigated area was divided into four parts: A, B, C, and D. It seems that B and C divided the area into left and right parts. However, the choice of A and D is not very clear, and most areas of A and D overlap.*

   Authors: Exactly as stated in the point raised by the Reviewer, B and C divide the area into left and right parts, to examine model performance if the training considers just the upstream (area B for training, C for testing) or the downstream (area C for training, B for testing) portions of the Po river basin. Area A entirely encompasses the Alps and the streamline of the Po river, and part of the Po Plain; after training the models in A, we test them in the resting portion of the study area, which contains lower mountain range (the Apennines) and smaller river catchments. In this way, we check if our approach is sensitive to these hydrological conditions. Area D comprehends most of the Alps and the Po Plain, and entirely the Apennines and the streamline of the Po river. After using B for the training, we test the model over the resting part of the Alps and Po valley, checking if the approach is capable to estimate flood hazard in rivers with similar hydrological conditions to the ones of the training. We will improve the descriptions of these regions and why we selected them in our study.

6. ***Reviewer:*** *The authors used Gini importance (GI) to measure the importance of each factor. Is that possible to use the information to give different weights to the GDs to build up learning models?*

   Authors: Many thanks for raising this point on Gini Importance, which is a fundamental part of the description of our study outcome. We believe that GI is an extremely useful metric for guiding future studies on multivariate DEM-based flood hazard mapping. One might use in principle the GI values we obtained in our study to set initial values of weights for the input descriptors for training multivariate models in other study regions. Nevertheless, this represents a critical aspect as GI values resulted to be overly sensitive to the training area being used. We will include this comment in the revised manuscript.

To conclude, we want to state also that we are grateful for the other minor comments, that we will incorporate and address while revising our manuscript.

Thank you again for your precious help,

With kind regards,

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