Review of: **Median bed-material sediment particle size across rivers in the contiguous U.S.**
by *Guta Wakbulcho Abeshu, Hong-Yi Li, Zhenduo Zhu, Zeli Tan, and L. Ruby Leung*

The authors have attempted to solve an important problem of increasing resolution of available sediment D50 in contiguous USA. The authors should be lauded for attempting to tackle this problem using a unique approach of predicting D50 for regions that lacked measurements using a Gradient-Boosting based machine learning method. The paper is well written, though the presentation can be improved. The approach taken by the authors is unique, though I don't think the resultant synthetic dataset generated meets the standards of the measured datasets usually published in this journal. I say this because, it is well known that a ML model's predictive capability is constrained by the range of values present within the training data-set. Even if a trained model shows decent prediction capability for the testing data-set, there is no guarantee that the model will be able to predict correctly for cases that have input parameter values beyond the range of the training data. When the authors synthetically generated the data for the whole US using a model trained using (extremely sparse) 2577 spatial locations, they didn't mention about the range of values for the input parameters in the 26000 flowlines, which were later used to generate the synthetic dataset. I would recommend the authors to submit this paper to a different journal, e.g. WRR, Advances in Water Resources, Geoscientific Model Development, JGRE, etc., and increase the discussion about why the 13 parameters that SHAP value indicate to be the ones most responsible for determining the D50 of sediment at a stream location.

As the above statement about the dataset suitability for the journal could be deemed subjective, I leave the final decision on suitability of the synthetically generated dataset for ESSD to the Editors. Going ahead, I will only comment on the technical and presentation issues of the paper.

1) In line 35 and later, the authors cite "Garcia, 1975". There is not citation in the references to match it.

2) Please re-write line 38 to make is clearer.

3) In line 61, the authors talk about how ML based approaches can allow establishment of successful predictive models without sufficient process-based knowledge. This has proven to be true in different fields; though in others, utilizing ML without a process based understanding has also led to erroneous models that lack generalizability. Often times the difference between success and failure of a ML model is based on the amount of data available for training the model. The authors in the current study have attempted to develop a generalized model for predicting sediment D50 in USA, based on different channel and catchment properties. It is hard to fathom that this could be achieved based on a dataset with only about 2600 data points, without any prior input about the processes involved.

4) In line 72, the authors mention that the dataset has some points with only a single sediment size value. The authors could try to see if they can utilize this extra data, even though D50 calculation is not possible. Maybe the data can be used for further validation of the model.

5) Please include more information in the caption of figure 1 a, e.g. things like how many locations are actually shown on the map. Also, data that is shown as histograms (1 b,c), could/should be represented spatially on the map. This would provide the readers additional information about the spatial variation of different aspects of the data.

6) In all the locations that has multiple values of D50 reported in time, what is the variability in the D50 value over time ? Even though the timescale across which the D50 data was collected is smaller than geomorphological timescales, it is important to check for the variability in order to be sure that the data was collected at stable stream-reaches. Also, what is the scientific basis for calculating a representative D50 by taking a mean ?

7) In line 109, the authors mention that if there are multiple sediment sampling locations for a flowline, they assigned a simple average to come up with the representative D50 for the flowline. This approach is simplistic, as this will work if all the sampling points on a flowline are equidistant. The authors should devise a method that accounts for the relative spatial location of each sampling location, else the representative D50 will be inaccurate.

8) Starting at line 115, the authors mention two studies, specifically Chen and Guestrin (2016) and Zheng et al. (2019), to argue that the ML method they have used is appropriate for the current study. It should be pointed out that even though the XGBoost method that the aforementioned studies used performed admirably, the model developed in those studies were for specific locations. Chen and Guestrin's used it for snowpack spatial patterns in the Sierra Nevada of California, and Zheng et al. used it for predicting water storage changes of a specific lake Inner Mongolia plateau. On the other hand the authors are trying to develop a general model for the whole of USA. Thus, the suitability of the adopted ML technique is debatable.

9) The use of KGE as the model performance parameter is interesting, especially as the KGE values for the testing dataset is relatively much better than the traditional R^2. Though, KGE itself is fraught with issues (Onyutha, 2020). So it would be informative if the authors also provide model performance quantification using Nash-Sutcliffe efficiency, CMA (Onyutha, 2020), etc.

10) Once the possible model input parameters has been reduced to 13 parameters (2 channel and 11 basin characteristics), the figures that show results for them (e.g. Fig. 3) should use names of the parameters that are intuitively understandable, rather than something that one has to look up a table (table 1) to recollect. So, please redo the figures.

11) The authors through this exercise of statistically trying to find the most relevant parameters are onto something very interesting and informative. Though, the study isn't complete without a detailed discussion about why or how the parameters that that the model zeros onto are physically connected to the process of sediment D50 formation. Doing this, the reader will have more confidence on the model's predictions and will be a step towards generalization.

12) In line 215, the authors mention that despite lack of any obvious one-on-one correlation between the 13 model input parameters and the D50, they believe the XGBoost model will be able to capture the "high-order interactions" among the input parameters. The authors do not provide any proof to indicate the accuracy of this statement. KGE > 0.5 for the testing dataset is encouraging, though on the other hand dismal R^2 (< 0.38) clearly indicate the large amount of dispersion in the model prediction. Thus there is no indication that the model has been able to accurately capture the general trends and processes that decide sediment D50 at a stream-reach. Thus, using this model to synthetically generate possible D50 values is USA, which can then be used model large-scale hydraulic and geomorphological processes is fraught with issues.

13) The authors suggest that the predicted D50 values can be used for producing a map of Manning's roughness coefficient for different streams and reaches in USA. This is hydraulically incorrect. Yes, there are certain stream reaches where D50 is a good indicator of the Manning's roughness coefficient, on the other hand there are many different scenarios under which this will fail. For example, if a stream has vegetation within its flood-plane, the Manning's roughness coefficient will be substantially higher than what the D50 of the channel will predict. Thus, the authors should either remove any mention of Manning's roughness coefficient calculation from the D50 map, or mention the circumstances under which the prediction of Manning's roughness coefficient will be inaccurate.