Comment on essd-2021-201

The authors of this paper use machine learning techniques to calculate the median sediment particle size (D50) in U.S. streams. The scarcity of in situ measurements and continuous regional maps make this a worthwhile challenge. A total of 2577 D50 measurements and 76 predictive attributes were used to train a machine learning model and subsequently generate a D50 map for the contiguous U.S. The machine learning model used is a Gradient Boosting variant called XGBoost, its hyperparameters were optimised using the Optuna framework. The model is further improved by trimming the input features through iterative calculation of their feature importance scores. While the main contribution of this work seems to be the resulting National D50 map, I consider the clearly documented ML approach along with a couple of insightful comments on the use of said algorithms to be at least as valuable.

The article is well written and organised and for the most part seems methodically sound, at least from my machine learning point fo view.
My first mayor remark concerns the actual usefulness of the final data product when taking the model performance into consideration: although the KGE might be an established performance metric in the field of hydrology, the testing R2 metric does not point to great predictive accuracy. Ultimately, the model usefulness should be assessed by experts in the field of hydrology (which I am not) and an extended discussion on different performance metrics (including some that facilitate physical interpretability like RMSE) would help with this assessment.
My second mayor remark regards a possible sample bias and echoes that of the second reviewer albeit from a data focused point of view. The large disparity in counts shown in Fig. 1d) as well as the fact that over 10% of the samples were measured at the same source (USACE Mississippi River main stem) make the question of data representativeness an important one. Along the histogram suggested by the second reviewer, I would suggest a dedicated discussion on how XGBoost handles skewed datasets and their impact on prediction performance.

Beyond this two points, I think the articles makes a good use of the existing data and a well informed use of machine learning to produce a new data product, and I would like to see this research published.

Further minor comments:

Line 85: Is the averaging of samples over time the best way to handle multiple values?

Lines 122-123: This sentence is problematic: while gradient boosting does descend a gradient in some way, it does not make use of *the* Gradient Descent algorithm (Curry, Haskell B., 1944) most machine learning users associate with the concept. Might be worth reformulating or clarifying.

Lines 170-184: A flowchart would be a welcome addition and a good way convey this information at glance.

Lines 187-193: 4.1 would fit better in the Data section.

Lines 208-209: What is the mean annual flow velocity SHAP value and how does it compare to other predictions? Were there any other interesting predictive variables eliminated? A table analogous to Table 1 before feature selection might be useful for this discussion.

Line 235: A title like “Model Sensitivity Analysis” might be more suitable.

Line 285: Could also move to the Data section.

Figure 1a): Seems unnecessary, conveys very little information and could be better described in words (sentence in lines 188-189) or included in Figure with distinct markers.

Figure 1b): X-axis seems half empty and smaller counts are unreadable. Consider reformatting.

Figure 3: Consider flipping the X- or Y- axis order, so as to have a more natural “1.00 diagonal”.