We thank the review for the thoughtful comments on our updated manuscript. We respond to the general comments below. We have updated all the typographical errors as identified by the reviewer in their attached PDF within the revised text.

From "Suggestions for revision" Section

The revised manuscript is much improved. There are, however, some lingering issues from the 1st round of reviews and several new (mostly minor) issues.

We first, thank the reviewer for reviewing a revised version of the MS.

See the attached annotated pdf for comments and revisions. Some of the comments within the document are fairly substantial.

We respond to these comments below in the "attached PDF" section.

The authors all but ignored serveral of the comments in my 1st review (Reviewer #4), most notebly:

2. The removal of outliers is overall acceptable but can be very problematic when using a fluvial dataset as the 'extreme' values are often just the few large rivers in a dataset. The authors warn the reader to only use/interpret the results within the range of the variables but they should more carefully examine the outliers and try to include realistic observations and maximize the dataset (and thus model) representation of large rivers.

Respectfully to the reviewer, we did include additional acknowledgement and analysis of these outliers in response to this comment in the first round of reviews. We apologize if we misunderstood the comment and the additional look at the data and the confirmation that we do not preferentially remove any sites from our training dataset did not fully respond to the reviewer's comment. To reiterate and as is stated in the paper- no sites are removed during this screening process.

To emphasize this point more thoroughly, we have added additional text acknowledging the influence of this screening process on the training dataset. We also looked at the range of sample sizes for the largest rivers in our dataset and confirmed that while the average number of samples is reduced for larger rivers, we still have a number of large river sites where the sample size exceeds the median number of samples of the entire dataset in Lines 135-145 (also pasted below in response to a PDF comment).

3. The metrics selected for representing the models' accuracy are reasonable but need some justification. Why MSE and not RMSE or PBIAS or R2?

We addressed the choice of MSE relative to RMSE in the response document to the reviewer's original comments. MSE is advantageous over the coefficient of determination (R2) for similar reasons, so we have added this to the existing text on line 179. There are additional reasons why R2 is less preferable than MSE, including that it is more commonly used to describe the explanatory power of a model, not the model's overall accuracy. The aim of model training is to optimize model performance around accuracy, so MSE is a natural choice. PBIAS is also less preferable to MSE because it does not necessarily capture the precision of the model, but rather the bias, or systematic distortion, of model predictions relative to data. It is our aim

during the training process to capture the overall accuracy and precision of the model, not the explanatory power of any variable or a specific distortion effect.

While we have added a reference to why MSE is preferable to R2, consistent to our description of why it is preferable to RMSE, we feel that further review of each possible type of error that we could use and its pros and cons is beyond the scope of this contribution. Given this, we chose to not add additional discussion of PBIAS which would require us to define this type of error only to say it is unsuitable. MSE is a standard choice for ML model optimization.

Regarding the new literature I listed in my 1st review - the authors added these references as an add-on at the end of the discussion but all but ignored the research and advances they presented.

We have added additional acknowledgement of the findings of Cohen et al. (2022) and Lammers and Bledsoe (2018) to the discussion of the model sensitivity analysis (Lines 341-342) and Lines 352-354. We have gone through the papers the authors suggested and feel as though the clearest incorporation of these works is in the discussion of potential applications of a model like the ANN presented here, given global-scale data availability. Thus, we have referenced the papers with this in mind in Lines 400-405. Discussion of the specific findings of these models regarding spatial or temporal variability in sediment flux from a global-scale modeling approach are beyond the scope of our paper's aims.

WBMsed and the MOSART-sediment model are a different class of model than the ANN presented here. The bedload transport module each model is one component of a larger modeling framework. The authors of these works have demonstrated that their models perform reliably against field observations and capture temporal and spatial variability in sediment transport well. However, referencing or explaining the specific findings of spatial or temporally variability in sediment flux on a global scale is well beyond the scope of our manuscript, which aims to present a new ANN model for bedload flux, but does not aim to explore temporal or spatial trends in ANN predictions (though certainly that could represent interesting future work!). Given this, we feel the way in which we have referenced these works is appropriate within the scope of our manuscript.

From Attached PDF

Is this just removing flooding events in the large sites? What is the consequence for the resulting model?

As is stated in the paper, removal of the upper and lowermost percentiles of the training dataset is a common technique to enhance model training and subsequent performance. We acknowledge that this step removes the largest flooding events of the database. However, these largest events are also likely to be the least commonly occurring, and while these more exceptional events may be interesting, this does not warrant a degradation in model performance to include them. Further, it is worth noting that the discharge range over which the trained model is reliably applied still spans many orders of magnitude and may still capture more exceptional flood or sediment transport events in smaller rivers. We have added the following to the text between Lines 135-145 to acknowledge this point more thoroughly for the reader.

While this removal of more extreme values is an important step to ensure model quality, we acknowledge that this step preferentially removes the most extreme flow and sediment transport events from the dataset. While there is significant interest in predicting sediment transport rates for extreme flow events, these largest events are the least frequently occurring in the dataset and more data would be needed to train an ANN model to reliably predict bedload flux under these conditions. Following this screening, we maintain 134 distinct datasets, emphasizing that the training data do encompass more frequently occurring small and intermediate floods across all available sites in the database. Thus, while the trained model presented here may not be appropriate to predict bedload flux in response to exceptional events, it can still be applied over many orders of magnitude of discharge, as described above. Following this screening process, the median number of samples across all sites is n=50. For larger rivers with maximum discharges exceeding 300 m^3/s (n=17), the median number of samples is reduced, n=23. However, five of these larger sites do have sample sizes exceeding the median sample size n=50, with a maximum sample size of n=146 for the Mondego River (1.8% of the full database). *Thus, following the screening process, large rivers remain adequately represented in the training dataset.*

Since MAE is used throughout the paper, this should be explained beforehand (in the Methodology)
We have added a new section to the methods detailing this procedure. See Lines 230-245 (described below).

Is the MAE for all the results calculated as abs(log(O)-log(P)) ? If so, it is quite misleading.

To clarify, we only use the log-transformed MAE when comparing the performance of each of the bedload transport models to the observed dataset. This is because model predictions can differ from observations by many orders of magnitude.

We respectfully disagree with the characterization of this method as misleading. The observational data is not normally distributed and spans multiple orders of magnitude, thus a log-transform more equally weights all orders of magnitude and portions of the existing distribution. In the calculation of MAE, MAE = sum(abs(O-P))/n where N is the Observed value, P is the predicated value, and n is the number of samples. If the predicted and observed values are within the same order of magnitude, MAE captures both over- and under-predication in an equivalent way. If the prediction over or underpredicts by over an order of magnitude, MAE becomes asymmetric and more greatly penalizes overpredictions. Based on the model performances in this contribution, this would result in model underpredictions by many many orders of magnitude to appear to perform better than models that equally over and underpredict the data by a single order of magnitude in each direction. We have added a section to the methods between Lines 230-246 to more thoroughly describing how and why we chose to calculate MAE on the log-transformed dataset. We have included this section below.

2.3.5 Quantitative comparison of ANN performance and bedload models
In order to evaluate the performance of the ANN relative to these existing models, we calculated MAE for the four previous bedload transport models and the ANN model

based on the direct measurements of bedload flux from the BedloadWeb database within the portion of the dataset reserved for the test (n = 1,624). MAE is calculated as:

$$MAE = \frac{\sum|observed - predicted|}{number\ of\ samples}. \quad (3)$$

We selected MAE as the primary criteria to assess the average model performance because it is less sensitive to extreme values (Willmott & Matsuura, 2005). To better compare under and overprediction of each model across multiple orders of magnitude, we log-transformed all bedload transport observations and predictions. This is because, based on Eq. 3, predicted values that fall multiple orders of magnitude below observed values will result in very small differences between predicted and observed, which, result, by definition, in very small MAE values. In extreme cases, MAE values computed for models that, on average, underpredict the observed data by multiple orders of magnitude (e.g. Fig. 2A) can be less than MAE values for models that equally over- and underpredict the observed data within the same order of magnitude (e.g. Fig. 2d). In this case, computing MAE on log-transformed observations and model predictions more equally weights underpredictions of each model relative to model overpredictions. Further, given that the observations of bedload transport span four orders of magnitude, this procedure helps to more equally account for model errors across the full range of observations and associated predictions.