In this short and well-written manuscript, the authors present an ANN model for predicting bedload flux based on a published dataset. Machine learning is increasingly used for modeling and predicting natural dynamics, with known strengths and limitations. Bedload is perhaps one of the more challenging processes to model given its strong dependency on highly dynamic and local variables. A number of models have recently been published that attempt to predict bedload over large scales (continental and global; see below). This paper is therefore quite timely and adds to the broader communities' efforts to better predict fluvial dynamics. The following issues should be addressed before it is accepted for publication. These are not very major issues but will likely require additional analysis.

We thank the reviewer for their thoughtful review of this manuscript. We respond to individual comments below.

1. The observational dataset includes an unequal number of observations for each river - if the spatial variability is larger than the temporal variability this may lead to overfitting. The authors addressed that to a degree, but need to better discuss this issue. As it stands the model predicts temporal dynamics using observations from different rivers. ANN may be flexible enough to deal with this but, again, needs more discussion and maybe an additional analysis using some sort of average value for each site (regression may be more suitable in this case given the small sample size).

We thank the reviewer for the comment. To explore whether or not particular variables had an outsized influence on the result, we performed a sensitivity analysis (as discussed in the manuscript) and computed the model error on both the training dataset and a validation dataset both for the ANN trained on all model inputs but also for each sensitivity test that we conducted. We found that the training and validation loss errors are consistent, indicating that the model is not overfit to the training data (See Lines 157-160). These relative values of MSE for training and validation remain consistent across all sensitivity tests of the model as well, again, indicating that the model is not likely to be overfit to the training data.

We compared site-specific MAE values for the ANN model to both the IQR and the full range of observed bedload transport rates at each site. We found that, on average, MAE values are less than both the IQR and the full range of qs values. We found 11 instances where MAE > IQR and only one instance where MAE exceeded the full range of observed values at a site, comprising less than 10% of sites in the database. However, the median number of samples in these cases was 17, relative to a median of 50 samples across all sites.

In addition to this, we looked at functional relationships between the site-specific model MAE for the test data versus the total number of samples at each site. We did this to ascertain whether or not the model was biased towards differences in sample size. We did not find any systematic or significant relationships between the sample size at any individual site and the computed errors between the ANN output and our test data. We

can include this in addition to the existing supplementary material in Figure S4 (referenced below).
Across 134 distinct datasets, the median number of samples is n=50. The 25th percentile for sample size is n=18 and the 75th percentile is n=83, with 82% of the data within one order of magnitude. Only 22 sites have more than 100 samples. The largest dataset is from Goodwin Creek, has 307 samples and comprises <4% of the full database. Given this, we do not expect that any individual dataset should dominate model training. This is further confirmed by the lack of any systematic relationship between model errors and sample size.

Because some of the input parameters to the ANN are dynamic (e.g. discharge, width), we also explored the absolute error between every individual observation in our database and the model input parameters. This is included in the supplement in Figure S4. We find that there is no systematic or significant relationship between the absolute error across all data points and any particular input parameter. We do find that the lowest measured transport rates result in increased errors at some sites, which is consistent with most bedload flux models in the partial or intermittent transport regime very close to the threshold for motion. (Figure S4, top-left)
In the revised version of the manuscript, we will make these details more explicit in the assessment of model performance.

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.

We note that our screening of extreme values does not reduce the number of field sites, rather it excludes the most extreme values across the field sites. This keeps the larger rivers within the dataset. Within a revised manuscript we will highlight the parameter space where the model is applicable and clarify that this screening process does not reduce the number of large rivers from the original dataset.

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

It is worth clarifying to the reviewer that we are using different error metrics to explore different things. MSE is used specifically in model training as the model iterates towards the optimized weighting of all input parameters. MSE is the most commonly used error metric for loss functions such as those shown in Figure 1B because it penalizes larger errors moreso than RMSE, which is the square root of MSE, or MAE, which reduces the impact of these outliers. This penalization of large errors by MSE is particularly helpful in the optimization of the ANN across multiple epochs. We can include this detail in a revised version of the manuscript in the methods section as it pertains to the training of the algorithm.

In contrast, to compare the average performance of both the ANN test data (unseen data) and the additional 4 bedload transport models, we chose to compute the Mean Absolute Error instead of MAE. We choose MAE in this case precisely because it is less sensitive to any individual outlier or large error. Because bedload transport is particularly noisy, we deemed MAE to be the most appropriate error metric to assess the average performance of each model. Additionally, as indicated by Figure 2, the existing models for bedload flux that we compare the trained ANN sometimes result in large differences between the estimated and observed values for qs. Given this, we felt that the MAE served as the most conservative comparison between the ANN and these models. We applied the same reasoning when looking at site-specific errors between the ANN and observations. We also computed the RMSE, which is provided in the supplement, and while the values differ, the relative performance of the models is the same using the RMSE metric versus the MAE (see supplementary figures). We can add a sentence about this in the main text.

We did note our choice of using MAE for the model/observation comparisons in the first paragraph of the discussion (See Lines 175-176). We feel as though this acknowledgement is sufficient for what is a fairly standard metric for quantification of errors.

4. The paper falls short in providing tools and guidelines for applying its outcomes. The paper's main outcome is to demonstrate the potential usefulness of ANN for modeling bedload flux. How can the reader use this knowledge moving forward? Will they have to develop their own ANN based on the dataset? How can it be used for other locations (as the authors suggested)? This is a common issue with ML modeling, but the authors can mitigate it with additional descriptions and tools (e.g. scripts).

We are happy to provide a documented Jupyter notebook with the ANN-model and a short user-guide in an updated supplement.

5. The authors are encouraged to explore recently published papers such as:
Cohen, S., Syvitski, J., Ashely, T., Lammers, R., Fekete, B., & Li, H. Y. (2022). Spatial Trends and Drivers of Bedload and Suspended Sediment Fluxes in Global Rivers. Water Resources Research, e2021WR031583.
Gomez, B., & Soar, P. J. (2022). Bedload transport: beyond intractability. Royal Society Open Science, 9(3), 211932.

Lammers, R. W., & Bledsoe, B. P. (2018). Parsimonious sediment transport equations based on Bagnold's stream power approach. Earth Surface Processes and Landforms, 43(1), 242-258.

Li, H. Y., Tan, Z., Ma, H., Zhu, Z., Abeshu, G. W., Zhu, S., ... & Leung, L. R. (2022). A new large-scale suspended sediment model and its application over the United States. Hydrology and Earth System Sciences, 26(3), 665-688.

Tan, Z., Leung, L. R., Li, H. Y., & Cohen, S. (2022). Representing global soil erosion and sediment flux in Earth System Models.

> Thank you for the suggestion. We will review and include these papers appropriately in the revision of our manuscript.