How Machine Learning Can Improve Predictions and Provide Insight into Fluvial Sediment Transport in Minnesota

J. William Lund, Hydrologist, USGS Upper Midwest Water Science Center, jlund@usgs.gov
Joel T. Groten, Hydrologist, USGS Upper Midwest Water Science Center, jgroten@usgs.gov
Diana L. Karwan, Associate Professor, University of Minnesota, dlkarwan@umn.edu
Chad Babcock, Assistant Professor, University of Minnesota, cbabcock@umn.edu

Extended Abstract

Understanding fluvial sediment transport is critical to addressing many environmental concerns such as exacerbated flooding, degradation of aquatic habitat, excess nutrients, and the economic challenges of restoring aquatic systems. However, fluvial sediment transport is difficult to understand because of the multitude of factors controlling the potential sources, delivery, mechanics, and storage of sediment in aquatic systems. While physical fluvial sediment samples are an integral part of developing solutions for these environmental concerns, samples cannot be collected at every river and time of interest. Therefore, accurate and cost-effective estimates of sediment loading are needed to manage riverine sediment transport at a multitude of scales (Ellison et al. 2016); also needed are methods to estimate sediment transport at sites where little or no physical samples have been collected (Gray & Simões 2008). The application of machine learning (ML) approaches to estimate sediment transport has grown over the past two decades (Afan et al. 2016). ML used in sediment transport research has shown multiple benefits over traditional approaches, such as increased prediction accuracy, the ability to learn complex linear and non-linear relations amongst the dataset and providing the ability to interpret these complex relations with important features used in the model (Cisty et al. 2021; Francke et al. 2008; Khan et al. 2021; Zounemat-Kermani et al. 2020; Cutler et al. 2007).

The main objectives of this study (Lund et al. 2022) were:

1) Organize representative physical sediment samples, streamflow, and publicly available geospatial datasets that describe watershed, catchment, near-channel, and channel features in Minnesota
2) Engineer new features from streamflow data to better account for bankfull streamflow and rising or falling hydrographs
3) Train extreme gradient boosting (XGBoost) ML models to provide estimates of total sediment transport at stream locations where little or no physical samples have been collected but streamflow and geospatial data is available (Chen & Guestrin 2016)
4) Evaluate the final ML model against the more simplified streamflow control ML model to show prediction accuracy gained by feature engineering
5) Compare cumulative loads from in-situ sediment surrogate models to ML models that were trained without any data from the surrogate site, highlighting the ability to

transfer knowledge of sediment transport process from sites with physical samples to sites without

6) Interpret the final ML models important features with Shapley additive explanations (SHAP) values to assess what the ML model learned and how predictions were made, while making connections to known processes controlling fluvial sediment transport (Lundberg & Lee 2017; Molnar 2019)

Separate XGBoost ML models were developed and trained to predict suspended-sediment concentration (SSC) and bedload (BL) from sampling data collected in Minnesota by the U.S. Geological Survey (USGS). A total of 1,382 SSC samples from 56 sites and 638 bedload samples from 43 sites were included in the final dataset (Lund & Groten 2022). Approximately 400 watershed (full upstream area), catchment (nearby landscape), near-channel, channel, and streamflow features were retrieved or developed from multiple sources, reduced to approximately 30 uncorrelated features, and used in the final ML models. The results from Table 1 indicate suspended sediment and bedload final ML models explain 69% and 78% percent of the variance in the respected datasets.

Table 1. Goodness-of-fit results from comparison of streamflow control machine learning models to final machine learning models [BL, bedload; bR^2, bias correlation coefficient; NSE, Nash-Sutcliffe efficiency; RMSE, root mean squared error; SSC, suspended-sediment concentration] (Lund et al. 2022).

Model	RMSE	NSE	bR^2
SSC—streamflow control	377.6	0.59	0.35
SSC—final	329.4	0.69	0.45
BL—streamflow control	178.9	0.68	0.69
BL—final	150.3	0.78	0.67

Normalizing streamflow by the 2-year recurrence interval (RI) helped to constrain the variability in streamflow across sites of varying sizes and indicates when streamflow was below, near, or above bankfull. Calculating the slope of this new dimensionless hydrograph in relation to 24 hours before and after the sample was collected quantified if the sample was collected during stable, slowly/quickly rising, or falling streamflow. These feature engineering steps to normalize streamflow and calculate the slope of the hydrograph were found to increase model variance by 10% for both the SSC and bedload models when compared to streamflow control models that used streamflow in cubic feet per second and a categorical value of 1 for rising and 0 for falling hydrographs.

By comparing ML model outputs to in-situ sediment surrogate model outputs at sites that were not included in the training or testing of the ML model provided an opportunity to validate the ML modeling approach. The site-specific ML cumulative daily suspended-sediment loads (SSLs) were within the sediment surrogates 90% prediction intervals at all four sites.

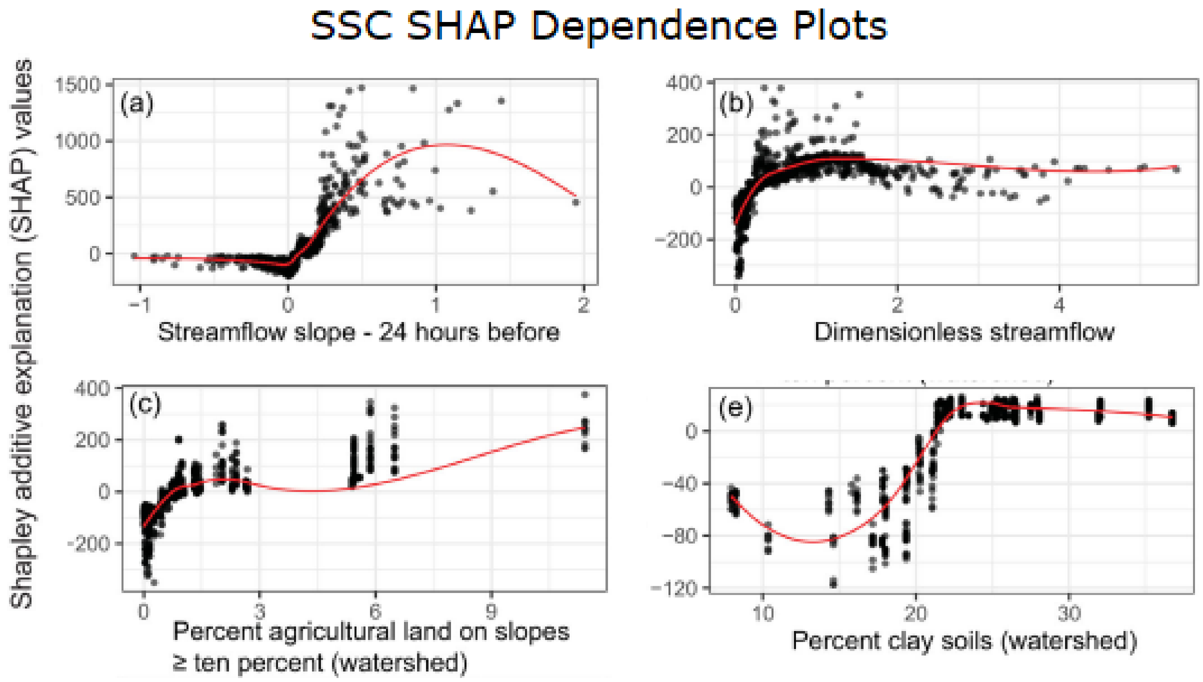

Figure 1. Selected Shapley additive explanation (SHAP) dependence plots from final suspended-sediment concentration (SSC) machine learning model. SHAP values on the y-axis and features observed values on the x-axis, each subplot has different scales. A positive SHAP value indicates the feature observation had a higher impact on predicting a target value greater than the mean of the observed values. A negative SHAP value indicates the feature observation impacted a prediction that was lower than the mean of observed values (Lund et al. 2022). A locally estimated scatterplot smoothing (LOESS) is presented as a red line.

SHAP values provided a quantitative way to support the model by displaying the relation and interaction of feature values and prediction output. Interpretation of SHAP values provided insight into how ML models made predictions and the processes controlling sediment transport. The dimensionless streamflow SHAP dependence plots showed the highest SHAP values were near the 2-year RI (x=1), which indicates higher sediment transport near bankfull streamflow (Figures 1b). These results are consistent with bankfull streamflow being the most geomorphically active (Biedenharn et al. 2008; Lane 1955). The results from the ML models suggest that the engineered streamflow features helped reduce uncertainty between streamflow and sediment transport across varying river sizes and regions in Minnesota. The streamflow and geospatial features are helping the ML models account for complex sediment source and transport processes which has been found to be difficult when using traditional approaches (Atieh et al. 2015; Ellison et al. 2016; Francke et al. 2008; Vaughan et al. 2017)

Advancements in data science and ML allowed for enhanced data driven sediment transport modelling, prediction accuracy, and interpretation techniques. Normalizing streamflow with the 2-year RI reduced variability and constrained the streamflow dataset around geomorphically active bankfull streamflow. Calculating streamflow slope features helped to better account for changing streamflow conditions. Geospatial datasets that account for local, near-channel, and watershed features helped improve predictions by allowing the model to learn complex processes related to sediment transport. Comparing ML model SSLs to modeled SSLs from in-situ sensors highlighted the utility of ML model's ability to learn and apply complex relations when making predictions at sites without physically collected samples. This study is a promising step forward in making fluvial sediment transport predictions using machine learning. Ongoing research is currently being completed by the USGS in other basins across the U.S. to make improvements to these methods by including time-series datasets like gridded precipitation and soil moisture to help capture complex antecedent conditions in the upstream watershed and local catchment while also using high-resolution digital elevation models to derive channel openness and slope-area indices to better describe the channel geomorphology.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Afan, H.A., El-shafie, A., Mohtar, W.H.M.W., and YaseenPast, Z.M. (2016). Present and prospect of an artificial intelligence (AI) based model for sediment transport prediction. Journal of Hydrology, 541B, 902–913. https://doi.org/10.1016/j.jhydrol.2016.07.048

Atieh, M., Mehltretter, S.L., Gharabaghi, B., and Rudra, R. (2015). Integrative neural networks model for prediction of sediment rating curve parameters for ungauged basins. Journal of Hydrology, 531, 1095–1107. https://doi.org/10.1016/j.jhydrol.2015.11.008

Biedenharn, D.S., Watson, C.C., and Thorne, C.R. (2008). In M. H. Garcia (Ed.), Chapter 6. Sedimentation engineering, processes measurement, modeling, and practice (pp. 355–386). Ph.D. American Society of Civil Engineers, Practice No. 110. ASCE. https://doi.org/10.1061/9780784408148.ch06

Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 13–17-August-2016 (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Cisty, M., Soldanova, V., Cyprich, F., Holubova, K., and Simor, V. (2021). Suspended sediment modelling with hydrological and climate input data. Journal of Hydroinformatics, 23(1), 192–210. https://doi.org/10.2166/hydro.2020.116

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K T., Gibson, J., and Lawler, J.J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Ellison, C.A., Groten, J.T., Lorenz, D.L., and Koller, K.S. (2016). Application of dimensionless sediment rating curves to predict suspended-sediment concentrations, bedload, and annual sediment loads for Rivers in Minnesota. U.S. Geological Survey Scientific Investigations Report 2016-5146 (68 p.). https://pubs.er.usgs.gov/publication/sir20165146

Francke, T., Lopez-Tarazón, J., and Schröder, B. (2008). Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. Hydrological Processes, 22, 4892–4904. https://doi.org/10.1002/hyp.7110

Gray, J.R., and Simões, F.J M. (2008). Estimating sediment discharge: Appendix D. In Sedimentation engineering: processes, measurements, modelling, and practice. American Society of Civil Engineers. (pp. 1067– 1088). https://doi.org/10.1061/9780784408148.apd

Lane, E.W. (1955). The importance of fluvial morphology in hydraulic engineering. Proceedings of the American Society of Civil Engineers, 81 (art.745), 1–17.

Lund, J.W., & Groten, J.T. (2022). Extreme gradient boosting machine learning models, suspended sediment, bedload, and geospatial data, Minnesota, 2007–2019. U.S. Geological Survey data release. https://doi.org/10.5066/P9VOPSEJ

Lund, J.W., Groten, J.T., Karwan, D.L., and Babcock, C. (2022). Using machine learning to improve predictions and provide insight into fluvial sediment transport. Hydrological Processes, 36(8), [e14648]. https://doi.org/10.1002/hyp.14648

Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4766–4775. https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Vaughan, A.A., Belmont, P., Hawkins, C.P., and Wilcock, P. (2017). Near-channel versus watershed controls on sediment rating curves. Journal of Geophysical Research: Earth Surface, 122(10), 1901–1923. https://doi.org/10.1002/2016JF004180

Zounemat-Kermani, M., Mahdavi-Meymand, A., Alizamir, M., Adarsh, S., and Yaseen, Z.M. (2020). On the complexities of sediment load modeling using integrative machine learning: Application of the great river of Loíza in Puerto Rico. Journal of Hydrology, 585, 124759. https://doi.org/10.1016/j.jhydrol.2020.124759