Extended Abstract: Estimating Reservoir Sedimentation using Deep Learning and the USACE RSI System Dataset

Deanna Meyer, Graduate Research Assistant, Saint Louis University, St. Louis, MO, deanna.meyer@slu.edu
Amanda Cox, Associate Professor, WATER Institute, Saint Louis University, St. Louis, MO, amanda.cox@slu.edu
Alejandra Botero-Acosta, Research Scientist, WATER Institute, Saint Louis University, St. Louis, MO, alejandra.boteroacosta@slu.edu
Vasit Sagan, Associate Professor, Taylor Geospatial Institute, Saint Louis University, St. Louis, MO, vasit.sagan@slu.edu
Ibrahim Demir, Associate Professor, University of Iowa, Iowa City, IA, ibrahim-demir@uiowa.edu
Marian Muste, Research Engineer, University of Iowa, Iowa City, IA, marian-muste@uiowa.edu
Chandra Pathak, Senior Engineer/Policy Advisor, US Army Corps of Engineers, Washington, D.C. chandra.s.pathak@usace.army.mil
Paul Boyd, Hydraulic Engineer, US Army Corps of Engineers, Omaha, NE, paul.m.boyd@usace.army.mil

Introduction

The current reservoir and dam infrastructure across the contiguous US provides water supply and hydroelectricity to localities, mitigates flood damage, supplies navigation, and provides recreational opportunities for communities. Maintenance of the collective system is pertinent to continued societal function. However, reservoirs throughout the nation are filling with sediment, which diminishes their life cycle and reduces their effectiveness, while increasing the cost of maintenance (Sholtes et al., 2018). Small-capacity reservoirs in high sediment yield regions are geologically prone to rapid loss of storage capacity and are at risk of sedimentation complications. The costs of remediating the accumulated sediment in these structures are exceedingly expensive, with dam removal providing the greatest expense in dam decommissioning (U.S. Bureau of Reclamation, 2006).

The USACE has implemented the Enhancing Reservoir Sedimentation Information for Climate Preparedness and Resilience (RSI) program to monitor reservoir aggradation and dam operation suitability for water-resource management. This unprecedented dataset contains information on approximately 400 dams (excluding navigation structures). However, given that over 90,000 dams exist in the US, the RSI dataset represents less than 1% of the US dams. Thus, there is a critical need to develop methods for estimating reservoir sedimentation at unmonitored sites. Existing reservoir sedimentation modeling methods have been unable to analyze large temporal or spatially scaled patterns of sedimentation, due to a lack of available data required for modeling. Previous sedimentation models utilized smaller and more local temporal and spatial scales that required daily to yearly hydrologic records, bathymetric reservoir details, and grain-size distribution of sediment (Ackers, 1988; Lajczak, 1996; Tarela and Menendez, 1999; Sundborg, 1992; Rowan et al., 2001). The RSI system provides reservoir data characterized by spatially diverse reservoirs across the contiguous United States, with some collection records spanning at least 100 years.

Additionally, due to the complex nonlinear behavior of natural sedimentation processes influenced by differing hydraulic flow factors, the utilization of machine learning introduces an ideal tool for constructing reservoir sedimentation estimations at unmonitored sites. (Abrahart et al., 2001; Zounemat-Kermani et al., 2019; Zounemat-Kermani et al., 2020). Thus, the objective of this research was to provide a tool for estimating reservoir sedimentation quantities using machine learning methodologies on the RSI system and other remotely gathered data. This tool could provide the USACE with a technique to monitor reservoir sedimentation and enable informed remediation efforts for RSI system reservoirs.

Methods

The USACE RSI data was combined with supplementary reservoir information corresponding to hydrologic and sedimentation processes to form a composite dataset utilized in this study. The supplementary dataset was compiled through the use of tools, such as GIS, MATLAB, and Pythonic Application Programming Interfaces (APIs). These APIs included Google Earth Engine and ArcGIS. Additionally, these tools exclusively use publicly available data, such as digital elevation models (DEMs), the National Landcover Database (NLD), USGS soil maps, monthly precipitation maps, the National Inventory of Dams (NID), as well as the Environmental Protection Agency's (EPA) ecoregion classification map, and the International Energy Conservation Code's (IECC) climate zone classification map. Table 1 distinguishes the types of parameters compiled through the use of these tools and public resources.

Table 1. Supplemental Data Compiled for the RSI System

Fixed Watershed Parameters	Time-Dependent Watershed Parameters	Reservoir Parameters
Basin Area	Total Upstream Normal Storage	Avg. Time Since Dam Completion
Mean Basin Slope	Total Upstream Max. Storage	Initial Trap Efficiency
Basin Relief	Total Upstream Dam Height	Initial Capacity
Hydraulic Length	Mean Monthly Precipitation	
Channel Slope	Cumulative Precipitation	
Mean Elevation	Duration of time between Surveys	
Max. Elevation	Maximum Monthly Precipitation	
Min. Elevation	Normalized Max. Monthly Precipitation	
Elevation Standard Deviation		
Composite Curve Number		
Percent Forested Area		
Mean Basin Latitude		
Mean Basin Longitude		
EPA Ecoregions Zones		
IECC Climate Zones		

A data anomaly detection was performed to reduce inclusion of erroneous data within the composite dataset. This included anomaly removals utilizing Autonomous Anomaly Detection (AAD) (Angelov et al., 2016; Gu and Angelov, 2017), which flagged 18 records corresponding to 15 reservoirs, and the Kolmogorov-Smirnov and Efron (KSE) outlier detection method (Jirachan and Priomsopa, 2015), which flagged 15 records corresponding to 10 reservoirs. Removal of anomalous data within the dataset used for model development varied by model based on individual performance, with the OLS model performing best with the full dataset and the machine learning models performing best with the removal of the KSE-identified anomalous data. A logarithmic transformation, as well as standard scaling or a minimum-maximum scaling of the datasets, was performed, depending on which anomalous removal method was conducted, to diminish the bias and skew of the variables' distribution. The following provides the equation for the log transformation:

$$\boldsymbol{x_{l_i} = \mathbf{sgn}[\mathbf{ln}(|x_i| + 1)]} \qquad \textbf{Eq. (1)}$$

where x_i is the original data value; x_{l_i} is the log-transformed value; i is the number of observations; and the **sgn** function multiplies the value by either a value of one if x_i is a positive value or a value of negative one if x_i is a negative value. Cox et al. (2022) provide additional details on the transformation and scaling of the variables.

A feature importance analysis was then conducted to analyze the sensitivity of variables in hindering statistical model performance. This resulted in the creation of a refined dataset with colinear features removed, known as the recursive feature eliminated (RFE) composite dataset. The RFE composite dataset was used to evaluate the modeled prediction of capacity loss within a reservoir, depending on the given predictor variables within the dataset. The data was examined in each iteration of a statistical or machine learning modeling method. For all models analyzed, a 70%/30% split of the datasets was applied for the training and testing of the models, respectively.

The first statistical analysis method was the Ordinary Least Squares (OLS) multilinear regression model. For this analysis, the full RFE dataset including anomalies, was utilized. The second analysis consisted of four supervised machine learning regression models: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Partial Least Squares (PLS). The third analysis used deep neural network (DNN) models. These models were modeled on the RFE dataset with the KSE anomalies removed. In the DNN model survey, four base DNN architectures were analyzed. For each DNN architecture, the hidden layer structure either progressively increased its decision node complexity from its initial input, or it progressively decreased. For each model studied, error and accuracy measures were analyzed with graphs of predicted versus observed measures of capacity loss, which is the change in capacity between two consecutive surveys.

Results and Conclusions

The recursive feature elimination process generated a dataset that contained twelve predictor variables, that were deemed most influential to providing modeling accuracy based on recursive model refinement through the use of a Random Forest regression model. Table 2 provides a list of the variables and their ranked order of significance.

For the OLS multilinear regression model, a training/testing analysis, as well as a fully calibrated OLS analysis were performed using the full dataset. This was done to provide the overall best-fit equation shown in Equation 2:

$$\begin{aligned} \boldsymbol{y_{OLS_l}} &\boldsymbol{= -10.3 + 0.553x_{l_1} + 0.476x_{l_2} + 0.383x_{l_3}} \\ &\boldsymbol{-0.181x_{l_4} + 0.561x_{l_5} + 1.63x_{l_6}} \\ &\boldsymbol{-0.0494\,x_{l_7} + 0.0250x_{l_8} - 0.0267x_{l_9}} \\ &\boldsymbol{+1.91x_{l_{10}} + 0.192x_{l_{11}} + 0.0589x_{l_{12}}} \end{aligned} \qquad \textbf{Eq. (2)}$$

where y_{OLS_l} is the log-transformed predicted capacity; x_{l_p} are the log-transformed predictor variables; and the numeric subscript p on the x_l terms denotes the variable index (Table 2). The calibrated OLS had an R^2 value of 0.40 and a mean absolute percentage error (MAPE) of 195%. Based on the model performance interpretation guidelines presented in Ayele et al., 2017, this OLS model performance is considered unsatisfactory (i.e., R^2 < 0.50). Table 2 defines the twelve predictor variables with their associated indices and coefficients. The unscaled OLS coefficients cannot be compared directly to determine the relative influence of each term. Therefore, standard scaling was used to further analyze the magnitude of influence each predictor variable had on the target variable (capacity loss) within the OLS equation.

Table 2. Recursive Feature Eliminated (RFE) ranked dataset variables.

Index	Variable	Units	Calibrated Standard Scaled Data - OLS Coefficients	Calibrated Unscaled Data - OLS Coefficients
1	Basin Area	mi^2	1.42	0.553
2	Initial Capacity	acre-ft	1.03	0.476
3	Cumulative Precipitation	in	0.323	0.383
4	Hydraulic Length	ft	-0.259	-0.181
5	Max Monthly Precipitation	in	0.234	0.561
6	Curve Number	n/a	0.144	1.63
7	Total Upstream Dam Height	ft	-0.119	-0.0494
8	Total Upstream Normal Storage	acre-ft	0.100	0.0250
9	Basin Relief	ft	-0.0369	-0.0267
10	Channel Slope	ft/ft	0.0226	1.91
11	Average Basin Latitude	°	0.0197	0.192
12	Mean Monthly Precipitation	in/mo.	0.0158	0.0589

Similar to the evaluation conducted on the OLS model, the best supervised machine learning model and DNN models were identified based on the highest R^2 values present in the untransformed training and testing datasets. A comparison between the supervised machine learning methods showed that the RFR had the highest accuracy in terms of predictive performance, when validated and calibrated on the respective data. With a training set R^2 of 0.61 and a testing set R^2 of 0.57, this model showed precision in terms of model fitness, when compared to the predicted versus observed values of capacity loss. Figure 1 shows the graphical testing results for this model, as well as the comparison to the testing performance of the other models studied. Notably, for the RFR model, there was a significant increase in MAPE on the testing dataset's forecasting accuracy. This signifies that the model's training results were overestimating the model's performance, regardless of the relatively high R^2 value present on the testing dataset.

In analyzing the DNNs, the complex DNNs performed significantly better in terms of accuracy based on the MAPE and R^2 values. The DNN_{PI1} was identified as the best DNN model variation based on maximizing R^2 and minimizing the relative root mean squared error (RRMSE). The DNN_{PI1} had training and testing R^2 values of 0.83 and 0.70, respectively. This makes the DNN_{PI1} the best fitting model in terms of performance. The RRMSE values of the DNN_{PI1} were the lowest RRMSE values compared across all analyzed machine learning models. However, the MAPE and RRMSE values showed a relatively large percentage increase between training and testing, meaning there may be underlying forecasting inaccuracies.

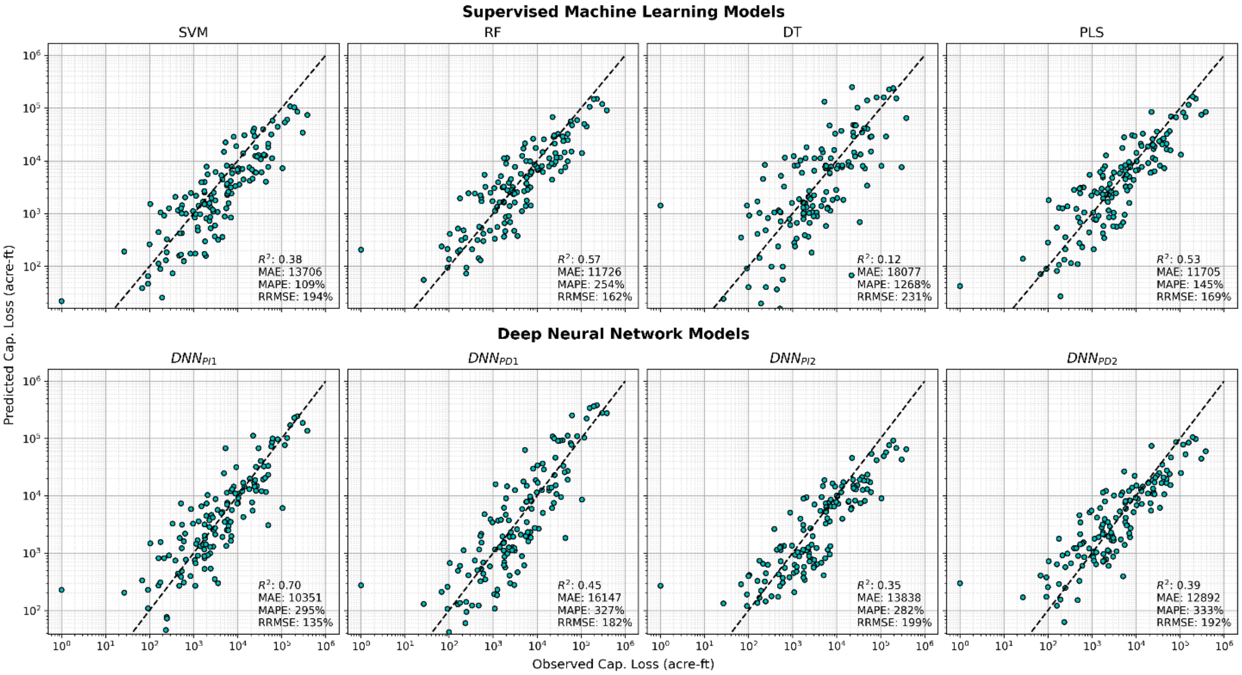

Figure 1. Comparison of Supervised ML and DNN predictive models

The model recommended for capacity loss prediction is a calibrated DNN_{PI1} model. The calibrated DNN_{PI1} was established through training the original, best-performing DNN_{PI1} model on the entire RFE dataset. This was conducted to overcome potential inaccuracies associated with the relatively smaller number of data points available, which is the case with the current RSI dataset. For this calibrated model, the R^2 increased to 0.81 and the MAPE value decreased to 38%. This shows a significant improvement in terms of forecasting accuracy, compared to all other models. Figure 2 illustrates the observed versus predicted capacity loss values for the calibrated DNN_{PI1}. Thus, this modeling method is interpreted as the most promising tool for the identification of vulnerable reservoirs within the RSI system.

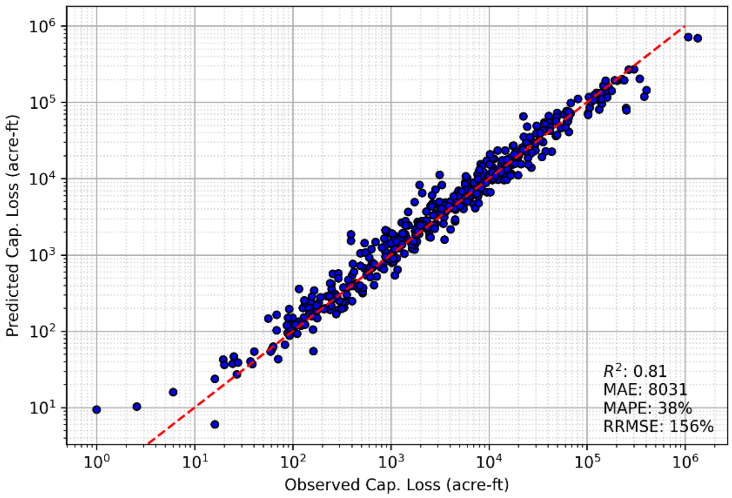

Figure 2. Observed versus predicted capacity loss values for the recommended machine learning model (calibrated DNN$_{PI1}$).

References

Abrahart, R. J., & White, S. M. (2001). Modelling sediment transfer in Malawi: comparing backpropagation neural network solutions against a multiple linear regression benchmark using small data sets. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *26*(1), 19-24.

Ackers, P. (1988). Alluvial channel hydraulics. *Journal of Hydrology*, *100*(1-3), 177-204.

Angelov, P., Gu, X., Kangin, D. and Principe, J. (2016). Empirical data analysis: A new tool for data analytics. 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 9-12, 000052-000059.

Ayele, G. T., Teshale, E. Z., Yu, B., Rutherfurd, I. D., and Jeong, J. (2017). Streamflow and sediment yield prediction for watershed prioritization in the Upper Blue Nile River Basin, Ethiopia. Water, 9(10), 782.

Cox, A.L., Meyer, D., Botero-Acosta, A., Sagan, V., Demir, I., Muste, M. (2022). "Data anomaly detection and sediment yield estimation in the US Army Corps of Engineers' Reservoir Sedimentation Information (RSI) database." Submitted to the US Army Corps of Engineers, Vicksburg, MS, July, 110 pp.Gu, X. and Angelov, P. (2017). Autonomous anomaly detection. 2017 Evolving and Adaptive Intelligent Systems (EAIS). 31 May-2 June 2017. 1-8.

Jirachan, T. and Piromsopa, K. (2015). Applying KSE-test and K-means clustering towards scalable unsupervised intrusion detection. 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). 22-24 July 2015. 82-87.

Lajczak, A. (1996). Modelling the long-term course of non-flushed reservoir sedimentation and estimating the life of dams. *Earth surface processes and landforms*, *21*(12), 1091-1107.

Rowan, J. S., Price, L. E., Fawcett, C. P., & Young, P. C. (2001). Reconstructing historic reservoir sedimentation rates using data-based mechanistic modelling. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *26*(1), 77-82.

Sholtes, J. S., Ubing, C., Randle, T. J., Fripp, J., Cenderelli, D., and Baird, D. C. (2018). Managing infrastructure in the stream environment. JAWRA Journal of the American Water Resources Association, 54(6), 1172-1184.

Sundborg, Å. (1992). Lake and reservoir sedimentation prediction and interpretation. *Geografiska Annaler: Series A, Physical Geography*, *74*(2-3), 93-100.

Tarela, P. A., & Menéndez, A. N. (1999). A model to predict reservoir sedimentation. *Lakes & Reservoirs: Research & Management*, *4*(3-4), 121-133.

U.S. Bureau of Reclamation (2006), Hydrology, hydraulics, and sedimentstudies for the Matilija Dam Ecosystem Restoration Project, Venture,CA—Draft report, 323 pp., Sediment. and River Hydraul. Group,Denver, Colo.

Zounemat-Kermani, M., Kisi, O., Piri, J., & Mahdavi-Meymand, A. (2019). Assessment of artificial intelligence–based models and metaheuristic algorithms in modeling evaporation. *Journal of Hydrologic Engineering*, *24*(10), 04019033.

Zounemat-Kermani, M., Mahdavi-Meymand, A., Alizamir, M., Adarsh, S., & 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.