Article

Demystifying the Relationship Between River Discharge and Suspended Sediment Using Exploratory Analysis and Deep Neural Network Algorithms

Wisdom Akatu [1], Marzieh Khosravi [1,*], Md Abdullah Al Mehedi [1], Jones Mantey [2], Hamed Tohidi [3] and Hanieh Shabanian [4]

[1] College of Engineering, Villanova University, Villanova, PA 19085, USA; wakatu@villanova.edu, mkhosrav@villanova.edu, mmehedi@villanova.edu,
[2] Gold Fields Ghana Limited, Environmental Department, P.O Box 26, Tarkwa – Ghana; jones.mantey@gold-fields.com
[3] Department of Civil Engineering, University of Memphis, TN 38152, USA; htohidi@memphis.edu
[4] Department of Computer Science, Northern Kentucky University, KY 41099, USA; shabanianh1@nku.edu
* Correspondence: mkhosrav@villanova.edu

Abstract: The dynamics of suspended sediment involves inherent non-linearity and complexity as a result of the presence of both spatial variability of the basin characteristics and temporal climatic patterns. As a result of this complexity, the conventional sediment rating curve (SRC) and other empirical methods produce inaccurate predictions. Deep neural networks (DNNs) have emerged as one of the advanced modelling techniques capable of addressing inherent non-linearity in hydrological processes over the last few decades. DNN algorithms are used to perform predictive analysis and investigate the interdependencies among the most pivotal water quantity and quality parameters i.e., discharge, suspended sediment concentration (SSC), and turbidity. In this study, the Long short-term memory (LSTM) algorithm of DNNs is used to model the discharge-suspended sediment relationship for the Stony Clove Creek. The simulations were run using primary data on discharge, SSC and turbidity. For the development of the DNN models and examining the effects of input vectors, combinations of different input vectors (namely discharge, and SSC) for the current and previous days are considered. Furthermore, a suitable modeling approach with an appropriate model input structure is suggested based on model performance indices for the training and testing phases. The performance of developed models is assessed using statistical indices such as root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Statistically, the performance of DNN-based models in simulating the daily SSC performed well with observed sediment concentration series data. The study demonstrates the suitability of the DNN approach for simulation and estimation of daily SSC, opening up new research avenues for applying hybrid soft computing models in hydrology.

Keywords: deep neural network; long short-term memory; suspended sediment; discharge

1. Introduction

Sediment-laden rivers and streams pose significant environmental and economic challenges [1]. The delivery of suspended sediment concentration (SSC) by a river has significant implications for its channel morphology, material fluxes, geochemical cycling, water quality, and the biotic and aquatic ecosystems that rely on the river. Furthermore, fine sediment, which has been identified as an important vector for transporting nutrients and contaminants, has a significant influence on river geomorphological and biological processes [2–4]. Because of the potential effects on biotic and aquatic habitats, as well as other land and water management processes, accurate estimation and long short-term forecasting of sediment transported by rivers are critical [5]. Streamflow has been identified as the primary explanatory variable for SSC although it is not always directly related

 doi:10.20944/preprints202211.0437.v2

to SSC and that the relationship between the two is known to vary greatly [6–8]. As a result, this variation is caused primarily by the complexity of the sediment concentration transported hydrological phenomenon as a result of several ambiguous parameters such as spatial variability of basin characteristics, river discharge patterns, and inherent non-linearity in hydro-meteorological parameters [9,10]. Other considerations include sediment availability, seasonality, and the location of the source within the watershed [11]. Significant variation in SSC may also result from a hysteresis effect with streamflow. The variation in streamflow, which provides vital information for the timing and changes in sediment concentrations, has been widely used in the development of SSC prediction models. Turbidity, SSC, and TSS are inextricably linked because they are all measures of suspended sediment in streams [1,12,13].

Physical models are built on simplified partial differential equations of flow and sediment. The models also rely on some unrealistic simplifying assumptions for flow and empirical relationships for rainfall and flow erosive effects [14]. These are extremely complex and sophisticated models, with some components representing physical processes. According to previous studies these models theoretically consider the effects of spatial variation in catchment properties as well as the uneven distribution of precipitation and evapotranspiration [15,16]. Process-oriented distributed models are impractical because most of the variables (for example, precipitation) are not currently measurable for much of the world. They have numerous drawbacks [17]. Because analytical methods make simple assumptions, numerical modeling is more applicable than analytical modeling [18,19]. However, in some cases, numerical studies of river-suspended sediment transport in actual conditions are impossible [20]. As a result, the feasibility of data-driven models as alternative methodologies for predicting SSC is commonly used. Data-driven (DD) models have an advantage over deterministic models in that they require fewer data and are better suited for forecasting. Kisi, 2012 argued that the ambiguity surrounding the internal structure of DD models was analogous to a 'black box' approach developed through a trial-and-error process. Despite their black-box nature, DD models are adaptable when it comes to capturing the nonlinearity of streamflow-sediment yield processes [14].

Deep Neural Network (DNN) has emerged as a popular DD modeling technique due to its ability to self-adapt, recognize patterns, and capture complex non-linear behavior between input and output parameters [21–23]. There are numerous examples of successful Deep Neural Networks (DNN) applications in the field of water resources engineering. These include the prediction of variables such as sediment load [24–30], river flows [31–36], runoff [37–41], flood frequency analysis [42,43], flood forecasting [44,45] and streamflow data infilling procedures [12,46].

Joshi et al., 2016 used different combinations of input vectors (i.e., discharge, sediment, and stage) to develop a robust network model structure for simulating the discharge-sediment and stage-discharge-sediment relationships provide a more accurate estimate of SSC for the Bhagirathi River [5]. The results of developed DNN model simulations were compared with the traditional conventional sediment rating curve (SRC) method in their study, and performance was tested using statistical parameters. In terms of the selected performance indices, DNN models outperformed the SRC method in simulating the daily SSC. These findings corroborate researchers' claims, such as Cigizoglu and Kisi, 2005, about using the feed-forward back propagation (FFBP) algorithm, sediment concentration, and streamflow data as inputs for daily or monthly SSC estimation and forecasting [47–49]. However, it has been reported that not enough studies have been conducted to evaluate the best performance of training algorithms within DNN modeling techniques [21]. This justification emphasizes the significance of training algorithm performance evaluation, as selecting an appropriate algorithm is as important as network architecture and geometry. The use of the neural network regression model in this study will aid in the introduction of a mechanism for integrating predictive models into the cloud platform, ultimately benefiting the research community.

This study aims to demystify the relationships between SSC and river discharge for the Stony Clove Creek using deep neural network algorithms in line with the above

 doi:10.20944/preprints202211.0437.v2

submissions. The specific objectives of the study are to analyze the hidden patterns in the distribution of river discharge, SSC, and turbidity series; to evaluate the performance of long short-term memory (LSTM) algorithms in predicting SSC in river flow; and perform sensitivity analysis and construct predictive models to forecast the discharge, SSC/turbidity relationship.

2. Materials and Methods

2.1. Study Location and Data Source

The Stony Clove Creek watershed is located in the central Catskill Mountain region of southeast New York State (**Figure 1**). The latitude and longitude of the study site are 42°07'40.6" N, 74°15'47.7" W with the coordinate system "North American Datum of 1983", Greene County, NY, Hydrologic Unit 02020006, on the right bank at the downstream side of bridge on Janssen Road in Lanesville. It has a 23.96 km^2 drainage area and the datum of gage is 390.14 m above NAVD of 1988, from the topographic map. The Stony Clove Creek flows from its headwaters at Notch Lake to its confluence with the Esopus Creek in the village of Phoenicia. Annual stream discharge within this part of North America generally peaks during spring snowmelt, but large rainstorms can cause rapid increases in discharge of local streams throughout the year due to the steep slopes and thin soils of the Catskill Mountains.

The Stony Clove Creek is a tributary of the Esopus Creek, eventually emptying into the Ashokan Reservoir, which supplies approximately 10% of NYC's drinking water (GCSWCD) [50]. The geology of the Catskill Mountains exerts a clear influence on the landscape, valley and stream channel morphology. The U.S. Geological Survey maintains a streamflow-gauging station on Stony Clove Creek (Gauge 1362336) that provides a continuous discharge record, SSC, and turbidity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

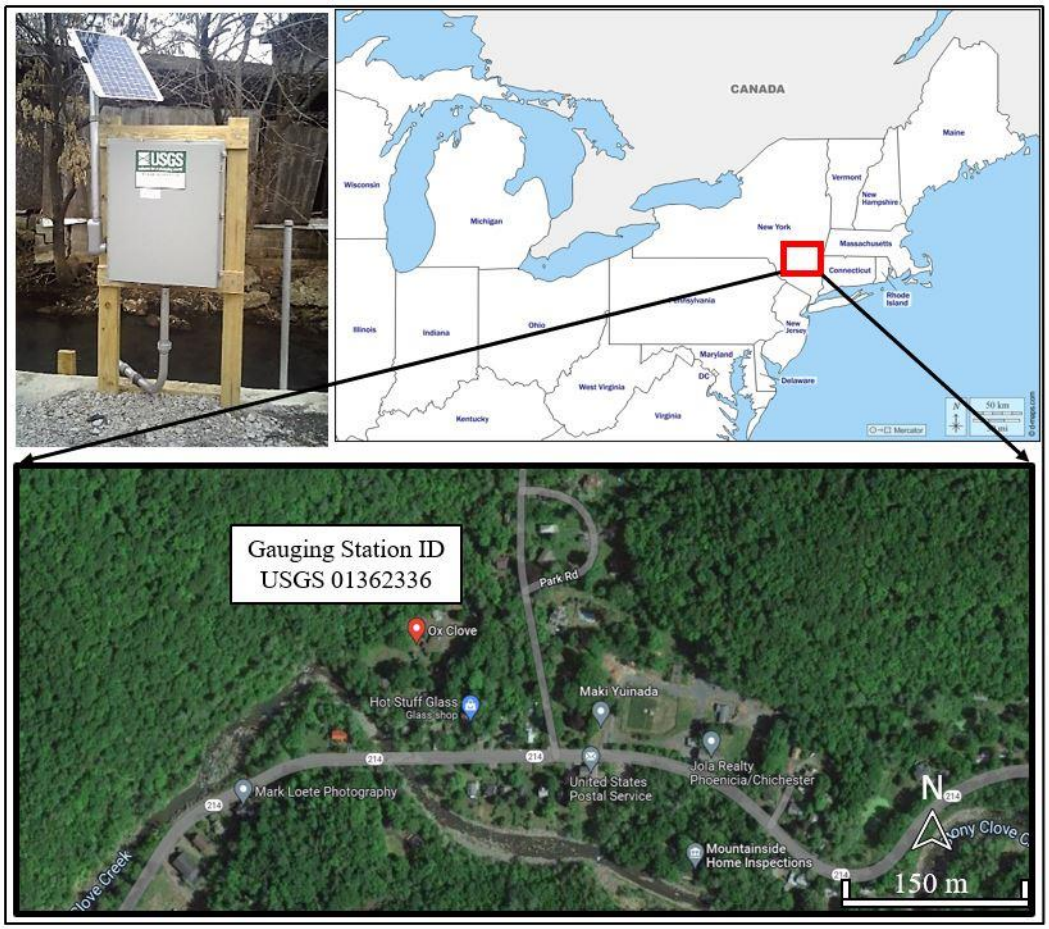


Figure 1. Study Location showing streamflow-gauging station on Stony Clove Creek.

The entire workflow of the Exploratory Data Analysis (EDA) and LSTM prediction tasks are divided into three distinct stages. In the first step, data collection from the USGS web portal, an exploratory analysis of the water quality variables, and feature engineering to transform the data for training/testing the LSTM algorithm is conducted. Variables are listed in **Table 1**. Activities in the first step are categorized as the transformer. After investigating the dataset and performing data transformation on the variables, LSTM neural network is trained using the data prepared in the first step to perform predictive analysis. LSTM neural networks regression model is assessed using several error matrices (e.g., Root Mean Square Error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2)). These activities are categorized as an estimator. The LSTM algorithm is tuned and optimized by altering the hyperparameters to reduce the errors in the prediction and achieve satisfactory performance. In the third step namely the evaluator, the model is deployed to predict the SSC/Turbidity for a new set of target variables. Model performance is further improved through iterative incorporation and validation of the input variables. The LSTM workflow of predicting the GSI performance indicator is illustrated in **Figure 2**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

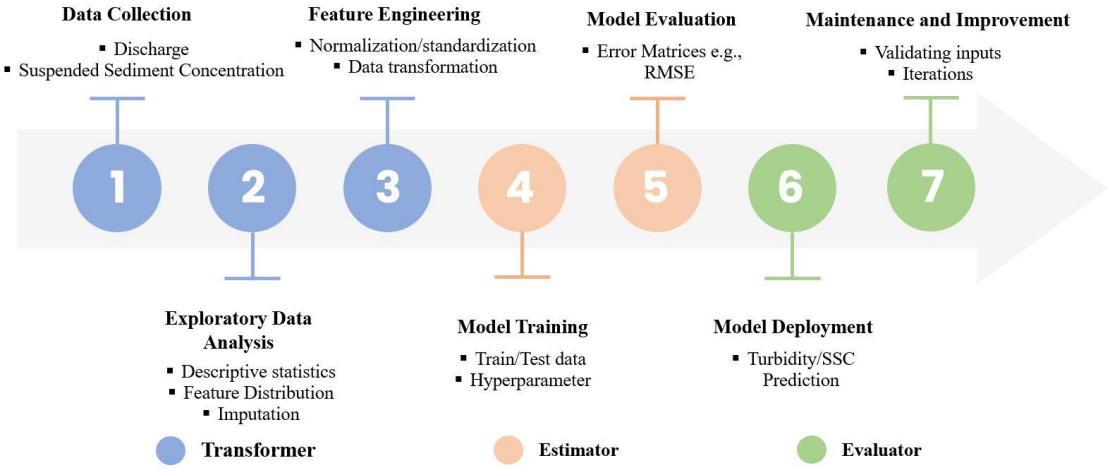


Figure 2. Pipeline of EDA and LSTM prediction tasks shows how the activities are linked from data pre-processing to model deployment. The steps are further classified into three groups: transformer, estimator, and evaluator.

In the first step of the workflow in **Figure 2** the time series of the discharge, SSC, and turbidity variables are retrieved from the USGS National Water Information System: Web Interface [51,52]. The range of the time series data for all the variables was different due to the various recorded duration. The range of the data used in this research is from 2/3/2020 to 9/30/2021, with observed data of about 19 months. The time-series data has 15 minutes interval between each recording. Historically, the extreme values for maximum and minimum discharges for the recorded period are equal to 48.14 m^3/s on Dec. 25, 2020, and 0.045 m^3/s on Oct. 22, 23, 2017, Sept. 23, 25, 28, 29, 30, Oct. 1, 2, 2019, respectively.

Table 1. List of the variables used for EDA and predictive analysis with LSTM model.

SW Parameters	Unit	Descriptions
Suspended Sediment Concentration	mg/L	The ratio of the dry weight of the sediment in a water-sediment mixture (obtained from a stream or other body of water) to the total weight of the mixture.
Turbidity	NTU	Water quality parameter that refers to how clear the water is
Discharge	m^3/s	Quantity of stream flow

2.2. Multivariate Exploratory Data Analysis

A detailed Exploratory Data Analysis (EDA) is performed in **Figure 2** activity 2 to understand the attributes and characteristics of the multivariate dataset. Exploratory Data Analysis (EDA) is a critical step in performing preliminary data investigations to obtain satisfactory LSTM model performance. Multiple visual techniques and numerical indices are used to investigate the internal temporal distribution of all three variables of discharge, SSC, and turbidity. EDA is the process of performing an initial investigation of input variables to understand the hidden pattern of the variables' distribution. EDA is further subdivided into activities. They are known as descriptive statistics, detecting outliers/extreme values, and normality check. Descriptive statistics is an excellent method for determining the distribution of the values of input variables based on the number of data points, mean, standard deviation, percentiles, interquartile range, and range. **Table 2** displays full multivariate descriptive statistics. Histograms with density plots are used as a visual representation of normality in the variables, and Pearson's coefficient of skewness (PCS) is used as a numerical indicator of skewness.

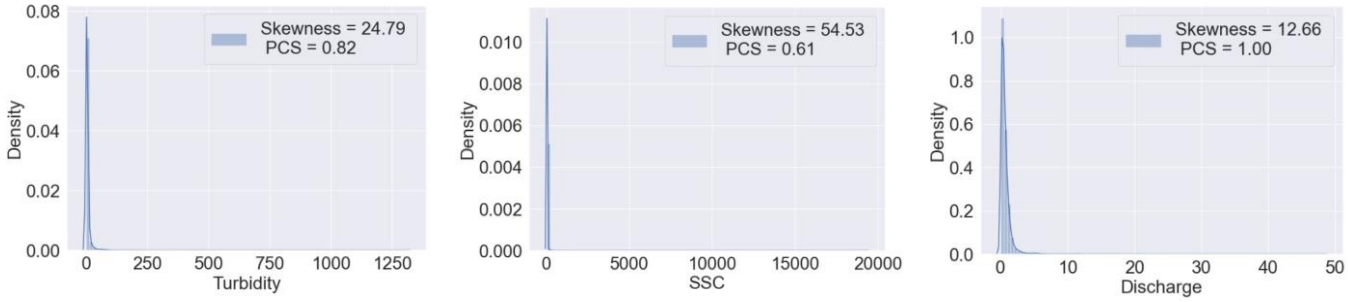
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

Table 2. Descriptive Statistics of the water quantity and quality variables.

	Count	Mean	Std	Min	25%	50%	75%	Max
Turbidity	57201	6.04	26.19	0.00	1.40	2.40	4.20	1310
SSC	57230	15.52	212.93	0.00	2.90	4.80	8.20	19400
Discharge	58051	0.77	1.59	0.00	0.24	0.46	0.83	48.14

Numerical interpolation is used to make the dataset consistent by filling missing values with the linear method between the first and last available observation for each non-value duration in the dataset. From the total of about 57000 historical observations, 850, 821, and 959 missing values are replaced with interpolations for turbidity, SSC, and discharge, respectively.

Figure 3 depicts a visual representation of the distribution of the input variables, indicating that the overall non-normality is high. All variables exhibit significant non-normality and skewness. Pearson's Correlation Coefficient (PCS) values of turbidity and SSC related to discharge and skewness of all three variables are indicated in **Figure 3** as a numerical measure of non-normality/skewness.



(a) (b) (c)

Figure 3. Distribution of the values of input variables, turbidity (a), SSC (b), and discharge (c)using histogram and density plot. The legend of each plot shows the skewness values.

Figure 4 illustrates that the linear connection between turbidity and discharge variables is higher than the linear relationship between SSC and discharge values. Lower values of the linear coefficients to delineate the overall non-linearity among several variables is high. The direction of the linear relationships is found to be positive in both turbidity and SSC with more weight on turbidity relative to discharge.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

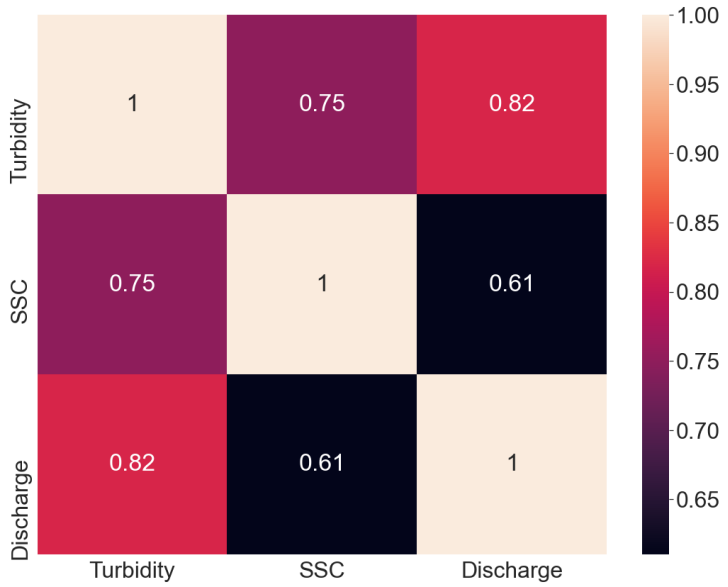


Figure 4. Bivariate correlation coefficients among the input variables represented by the correlation heatmap.

2.3. Feature Engineering

Following a successful preliminary investigation of the dataset using EDA, Feature Engineering (FE) is performed. Without a successful FE, the LSTM method may not produce satisfactory results with minimal error. Without a successful examination of the dataset, adequate optimization via iterative gradient descent cannot be achieved. As a result, extensive feature engineering is performed to transform the variables best suited for the LSTM learning algorithm. In this study, FE is used to perform imputation, data transformation, data standardization, and dataset segmentation into training, testing, and validation sets. Imputation is used to fill in the null values to make the entire dataset consistent. Sensor malfunctions resulted in null values or observations in every series.

These cells in the dataset are imputed with the values of the blank cell's nearest neighbors. However, due to the reduction in dataset size caused by the exclusion of observations, it is used in this study. Following successful imputation with variable median values, the distribution of the variable series is visually and numerically checked to confirm normality. The Pearson Coefficient of Skewness (PCR) is used to assess the normality of the variables. Because the distribution of discharge and water level values is highly skewed to the left, indicating a significant non-normality, neural network regression algorithms without appropriate data transformation do not contribute to satisfactory outcomes with good optimization.

Data normalization is a process of preparing the values of each variable for further analysis and implementing the DNN algorithm. The normalization process includes rescaling all the variables to the values that have the mean of zero and variance one**.** All the input variables' distribution are then converted to the bell-shaped normal distribution curve. As the variable considered in this study is the continuous independent variable, the normalization of the variable is crucial for training/testing the neural network algorithm.

All the values in the discharge series are normalized to prepare the training dataset for the LSTM model (**Equation** Error! Reference source not found.).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{(Equation Error! Reference}$$

 doi:10.20944/preprints202211.0437.v2

source not found.)

X denotes the variable of interest (e.g., discharge, turbidity, and SSC) and subscript norm, max and min represent the normalized variable, maximum and minimum value of the values of the variable. The entire normalized variable series is divided into two sections i.e., a training set that is used to train the model and a testing set that is used to evaluate the model. In this study the training and testing sets consist of 70 percent and 30 percent of the dataset, which is 58021, 38681, and 19340 number of observations, training observations, testing observations points.

2.4. Long Short-term Memory (LSTM) Recurrent Neural

In deep learning forecasting, where variables depend on prior historical knowledge throughout the data series, LSTM has proven to be a particularly popular technique for handling time series data. The long-term correlations and relationships between the variables can be recognized by utilizing LSTM. Because of the decaying error backflow, recurrent backpropagation requires a significant amount of computational time and effort to learn to store long-term information. Hence, the concept of the vanishing gradient problem in recognizing long-term dependency of Recurrent Neural Network (RNN) was introduced [49,53]. The main element of processing and retaining long-term information is LSTM feedback connections, and this characteristic distinguishes it from the conventional feedforward neural network. Both long-term memory (c[t-1]) and short-term memory (h[t-1]) are processed in a typical LSTM algorithm through the utilization of multiple gates to filter the information. For an unchanged flow of gradients, forget and update gates update the memory cell state. Three gates i.e., input gate i_g, forgot gate f_g, output gate o_g and cell state handle the information flow by writing, deleting, preserving past information, and reading respectively (**Figure 5**). As a result, LSTM can memorize data at various lead times, making it suited for time series prediction inside a specific window. In forget gate, long-term information enters and passes through a filtration where unnecessary information is discarded. The forget gate filters out unnecessary data by using the sigmoid activation function where the range of the function is 0 (gate close) and 1 (gate open). Input gate filter and quantify the significance of new data coming as input to the cell. Similar to the forget fate, input gates use binary activation functions to filter out information and control the flow of both long- and short-term information. The output gates regulate the value of the upcoming hidden state which is a function of the information on previous inputs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

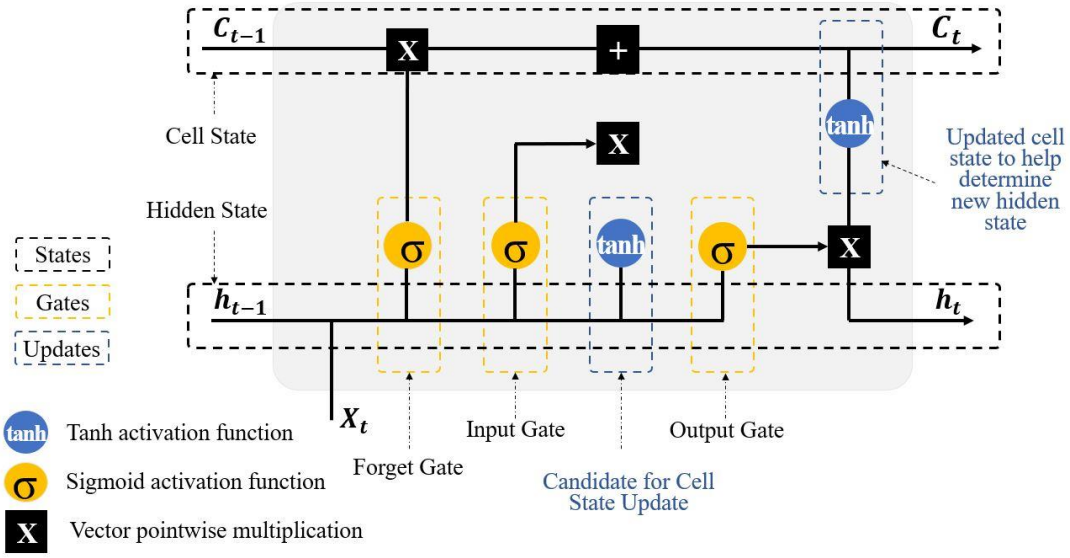


Figure 5. Schematic representation of an LSTM architecture.

In this research, a neural network with one LSTM hidden unit accompanied by a dense layer connecting the LSTM target output at the last time-step (t-1) to a single output neuron with non-linear activation function. The LSTM model was trained using the deep learning library, Keras in Python, the ReLU activation function, and the RMSE, MAE, and R^2 function [54–57]. To predict the discharge variable of a time-step in the future e.g., daily/weekly, values of the variables at the previous time-steps are used. Hyperparameters are tuned to maximize the performance of the LSTM model through an iterative trial-and-error approach. In this study, Keras, a python library that offers a space search for machine learning algorithms is used to find the best combination of the hyperparameters [58]. Considered hyperparameter of the LSTM algorithm in this study is the size of epoch and batch and number of neurons.

2.5. Time-series dataset

The initial dataset has 15-minutes interval for all the variables and the time-series analysis by LSTM conducted on 15-minutes for prediction. Among three variables in the dataset, turbidity and SSC were utilized in discharge prediction. To study the sensitivity of discharge prediction to SSC alone, in the first round of analysis (scenario one), only SSC was considered in discharge prediction as well as observed discharge values. In the second round of analysis (scenario two) both turbidity and SSC were considered in the prediction of discharge variable. The resultant prediction can indicate the importance and sensitivity of prediction based on the resolution of the input data. The discharge values were predicted by only SSC variables for the first round of models. The second round of model consists of SSC and turbidity as the features utilized to predict discharge. The comparison between these two scenarios demystifies the improvement of discharge prediction considering turbidity as a second feature added to SSC in further section.

2.6. Model Evaluation using Loss Functions

In the model evaluation step in the activity 5 in **Figure 2**, the performance of the LSTM model is evaluated using the top three standard error matrices e.g., RMSE, MAE, and R^2. Error matrices provide numeric values as the model performance indicator by comparing the observed and predicted values. The RMSE value is used to evaluate the LSTM model in showing the model performance improvement. The RMSE as a function is more sensitive to significant errors as the squared term multiplies greater errors

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

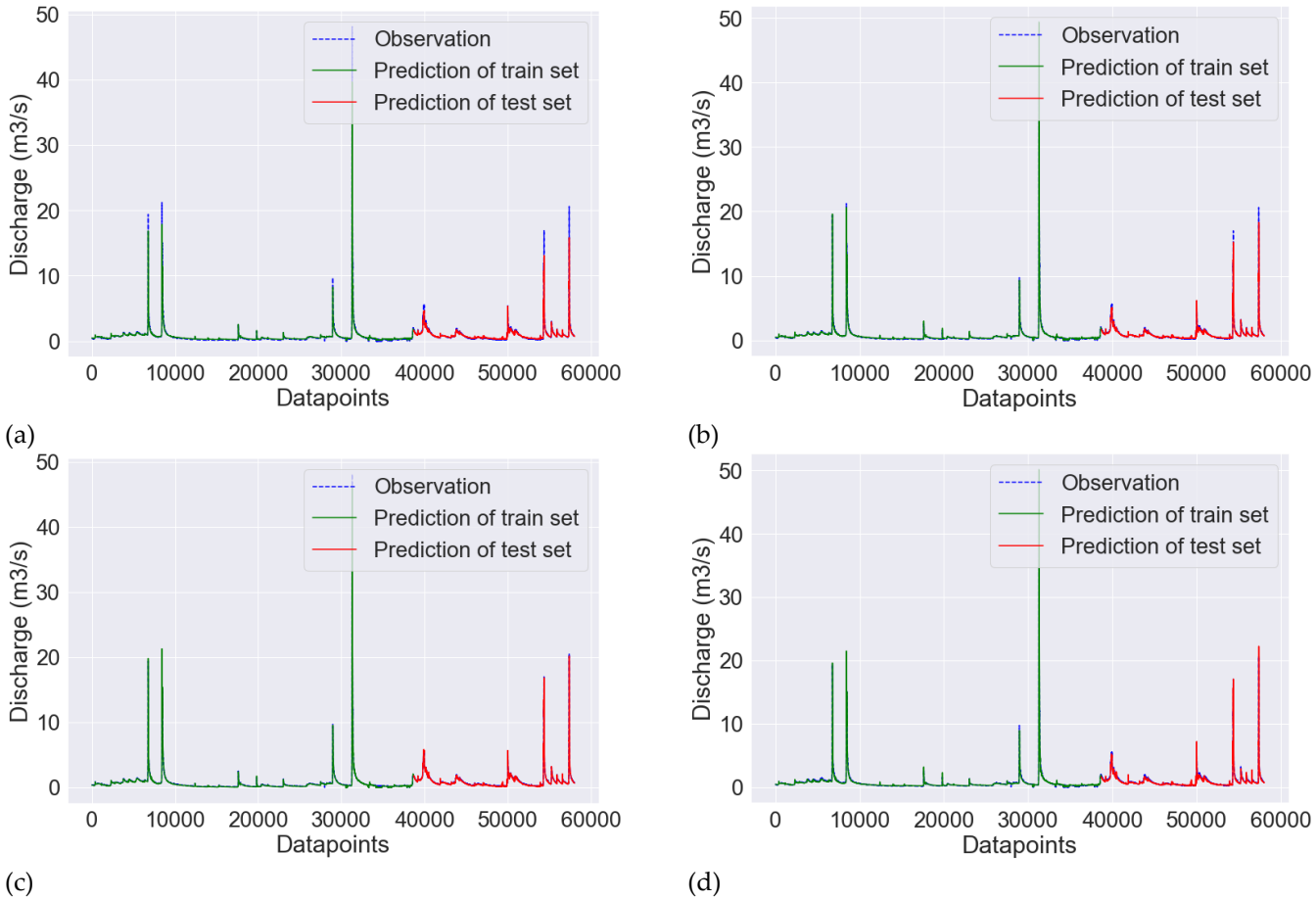
exponentially more than smaller ones. The lowest RMSE score corresponds to the best predictive accuracy. The better the model fits the data, the closer the R^2 value is to 1.

3. Results and Discussion

This LSTM neural network is used to predict the multivariate turbidity, SSC, and discharge variables based on the previous time series data. Several lead time durations selected for both time-series dataset (15-minutes and daily dataset). The prediction values of 1, 2, 3, 4, 5, 6, and 7 lead time chosen for both scenarios with the fixed time lag of 30 previous steps to be consistent with various LSTM models. Predicted values are compared to the observed dataset to quantify the error matrices. The RMSE, MAE, and R^2 functions are used to estimate the error from the predicted discharge, SSC, and turbidity variables. The model performance is improved by an increase in the number of epochs. Error matrices are obtained through multiple models runs to demonstrate the linkage between the model performance and the lead times.

3.1. Predicted and Observed Water Quantity and Quality variables

The output from the LSTM algorithm is compared to the observed values of the discharge, SCC, and turbidity variables through visual representation in **Figure 6**. Both the observed and predicted values of the discharge variable are plotted for the entire time series against the number of observations (data points). The overall distribution of the predicted values of discharge variable is approximately identical to the observed data providing a satisfactory performance of the LSTM algorithm. Three error metrics recorded for all variables and full-time series showed that LSTM performed well in case of both train/test sets. In **Figure 6**, the green portion of the plot illustrates the training portion of the dataset whereas the red portion shows the testing portion, and dashed lines with blue color show the observed data.



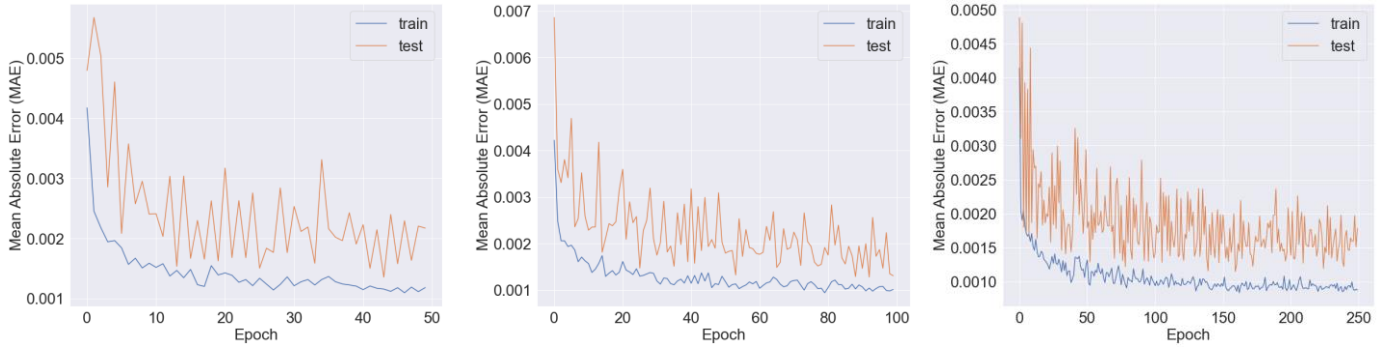
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

Figure 6. Distribution of observed value from the gage records (dashed blue lines) and predicted values from LSTM model for the discharge variable, considering SSC, lead time of 1 (a), considering SSC, lead time of 7 (b), considering turbidity and SSC, lead time of 1 (c), considering turbidity and SSC, lead time of 7 (d), with train/test split (green and red).

3.2. Model Evaluation Matrices and Improvement

The performance of the LSTM neural network is evaluated using three error matrices e.g., RMSE, MAE, and R^2. The performance of the model was also evaluated and improved by increasing the number of iterations i.e., epoch in the neural network. The value of MAE is observed with an increase in the number of epochs in **Figure 7**, where MAE value is the indicator of the model performance. The number of epochs is increased up to 250 to evaluate the performance variation. The RMSE and MAE values are found to decrease from 120.25 and 42.14 to 110.48 and 25.40, respectively by increasing the epoch number from 50 to 100 for discharge prediction based on only SSC variable (scenario one), which indicates satisfactory performance in the LSTM algorithm. Furthermore, the RMSE and MAE value is found to decrease from 9.40 and 3.44 to 8.94 and 3.17, respectively by increasing the epoch number from 50 to 100 for discharge prediction based on both turbidity and SSC variable (scenario two), which indicates satisfactory performance in the LSTM algorithm. On the other hand, the increase in epoch number from 100 to 250 revealed the opposite trend, when discharge predictions for both scenarios are considered. In the case of the first scenario, the RMSE and MAE values increased to 110.81 and 34.67 respectively. Taking into consideration scenario two, there was an increase in RMSE value to 9.18 and a decrease in MAE value to 2.62.

The model performance increases significantly from the very beginning of the iteration for both the train and test scenarios. The trend of a decrease in the RMSE values reaches a near-steady state after 20 epochs. A local decrease in the performance i.e., an increase in the RMSE value can be seen after 20 epochs. Furthermore, the performance variation with the change in the lead time considering the error metrics for both scenarios studied. The comparative study results for RMSE, MAE, and R^2 are represented in **Figure 8** and **Figure 9**.



(a) (b) (c)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 November 2022 doi:10.20944/preprints202211.0437.v2

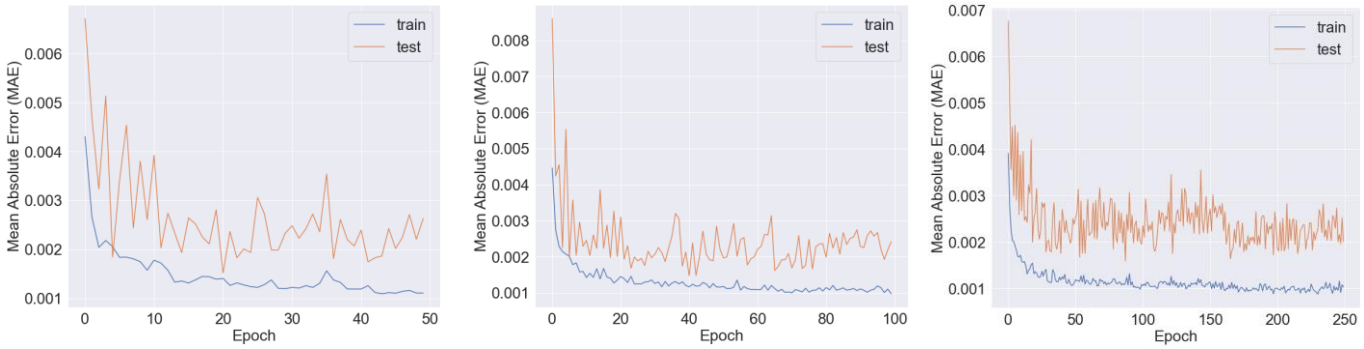


Figure 7. The increase in the number of epochs for the train and test set, discharge prediction considering SSC for, 50 (a), 100 (b), and 250 (c) epoch numbers, and discharge prediction considering both turbidity and SSC for, 50 (d), 100 (e), and 250 (f) epoch numbers.

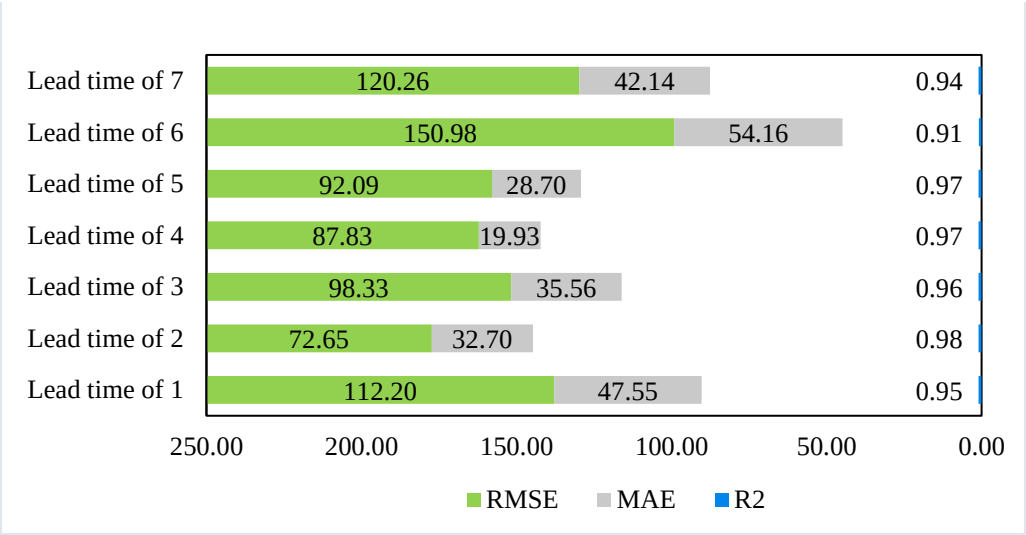


Figure 8. Model performance evaluation for the entire time series of scenario one.

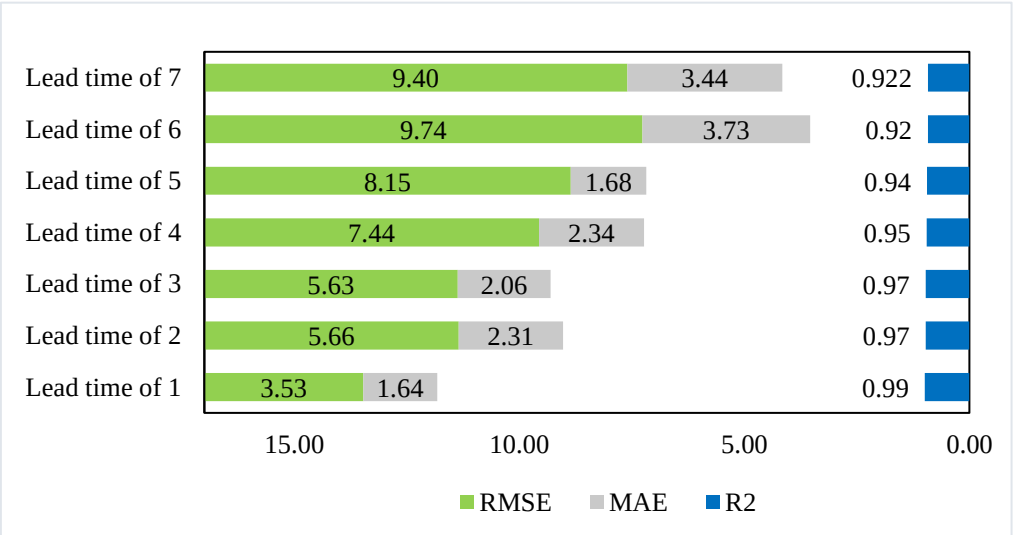


Figure 9. Model performance evaluation for the entire time series of scenario two.

 doi:10.20944/preprints202211.0437.v2

Error matrices e.g., RMSE, MAE, and R^2 are documented for several lead times. Lead times are pivotal parameters of LSTM algorithm towards model performance. Lead time values are for each step forward in time series. So, the lead time of 1 equal to 15-minutes and the lead time of 4 equals to 1-hour. The values of RMSE and MAE increases with the increase in the lead times whereas the R^2 decreases showing the degradation in the model performance with an increase in the lead times for the second scenario. There was much variation in the increasing and decreasing trend of the RMSE, MAE, and R^2 in the first scenario which made it harder to distinguish the lead time importance in the prediction of discharge variable only based on SSC. Therefore, the selection of the lead times should be based on the model performance and necessity. An R^2 value of 1 denotes a perfect fit between the observed and predicted values, with no error in LSTM prediction. The R^2 value for all models is within the range of 0.91-0.99, establishing an overall satisfactory performance from LSTM model prediction. The best prediction with minimum error is associated with a discharge prediction based on both turbidity and SSC at a lead time of 1, with a corresponding R^2 value of 0.99.

4. Conclusion

Many rivers are potential sources of SSC; therefore, an accurate estimate of sediment concentration is critical in understanding its relationship with water discharge for water resource planning and management. In this study, DNN algorithms are used to perform predictive analysis and investigate the interdependencies among the most pivotal water quantity and quality parameters i.e., discharge, SSC, and turbidity for the Stony Clove Creek. For simulation, primary data of discharge, SSC, and turbidity for the period from 2/3/2020 to 9/30/2021 are used. Combinations of various input vectors (namely SSC, discharge, and turbidity) are considered for model development and examining the effects of input vectors.

Consequently, a viable modelling strategy with acceptable model input structure is proposed for simulation of SSC based on several model performance indices. The LSTM technique was used to execute different model experiments in this study with two main scenarios to study the feature importance of SSC and turbidity. 15-minutes times-series values of discharge, SSC, and turbidity were utilized as input, and model simulations were run using a distinct set of input combinations. RMSE, MAE, and R^2 were used to test the performance of the LSTM model simulations.

In general, hysteresis generated by LSTM simulations captures nonlinear dynamics, generalize the structure of the entire data set, and perform well with observed sediment concentration series data. As a result, the study suggests that using soft computing techniques such as LSTM technique can provide precise estimates of SSC. The study shows that the LSTM neural network regression techniques are highly effective for simulating and predicting multivariate time series.

As a result, the study opens up new research avenues for enhancing other DNN algorithms and hybrid soft computing techniques for short-/long-term prediction of SSCs in water discharge. This research will be especially useful in managing water resources projects in both upstream and downstream areas of the southeast New York region in terms of sediment flux.

Author Contributions: Conceptualization, W.A, M.K. and M.A.A.M; methodology, W.A, M.K. and M.A.A.M; software, M.K. and M.A.A.M; validation, W.A, M.K. and H.T.; formal analysis, M.K. and M.A.A.M; investigation, W.A. H.T.; resources, W.A, J.M. and M.K.; data curation, W.A and M.A.A.M; writing—original draft preparation, W.A, M.K., J.M. and M.A.A.M.; writing—review and editing, W.A, M.K., H.T. and H.S.; visualization, H.T. M.K.; supervision, H.T. and H.S.; project

administration, W.A. H.S.; funding acquisition, H.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The code and data to reproduce our results is available at [59], https://github.com/MShivaKh/SuspendedSediment.git.

Acknowledgments: This study was supported by the Computer Science program in the School of Computing and Analytics at Northern Kentucky University and the Civil Engineering department of University of Memphis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. USGS Surface-Water Data for Minnesota Available online: https://waterdata.usgs.gov/mn/nwis/sw (accessed on 17 November 2022).
2. Caroni, E.; Singh, V.P.; Ubertini, L. Rainfall-Runoff-Sediment Yield Relation by Stochastic Modelling. *Hydrol. Sci. J.* **1984**, *29*, 203–218, doi:10.1080/02626668409490934.
3. Mantey, J.; Nyarko, K.B.; Owusu-Nimo, F.; Awua, K.A.; Bempah, C.K.; Amankwah, R.K.; Akatu, W.E.; Appiah-Effah, E. Mercury Contamination of Soil and Water Media from Different Illegal Artisanal Small-Scale Gold Mining Operations (Galamsey). *Heliyon* **2020**, *6*, e04312, doi:10.1016/j.heliyon.2020.e04312.
4. Mantey, J.; Nyarko, K.B.; Owusu-Nimo, F.; Awua, K.A.; Bempah, C.K.; Amankwah, R.K.; Akatu, W.E.; Appiah-Effah, E. Influence of Illegal Artisanal Small-Scale Gold Mining Operations (Galamsey) on Oil and Grease (O/G) Concentrations in Three Hotspot Assemblies of Western Region, Ghana. *Environ. Pollut. Barking Essex 1987* **2020**, *263*, 114251, doi:10.1016/j.envpol.2020.114251.
5. Joshi, R.; Kumar, K.; Adhikari, V.P.S. Modelling Suspended Sediment Concentration Using Artificial Neural Networks for Gangotri Glacier. *Hydrol. Process.* **2016**, *30*, 1354–1366, doi:10.1002/hyp.10723.
6. Tornes, L.H. *Suspended Sediment in Minnesota Streams*; Water-Resources Investigations Report; U.S. Geological Survey: St. Paul, MN, 1986; Vol. 85–4312;.
7. Tornes, L.H.; Brigham, M.E.; Lorenz, D.L. *Nutrients, Suspended Sediment, and Pesticides in Streams in the Red River of the North Basin, Minnesota, North Dakota, and South Dakota, 1993-95*; Water-Resources Investigations Report; U.S. Geological Survey: Mounds View, MN, 1997; Vol. 97–4053;.
8. Blanchard, R.A.; Ellison, C.A.; Galloway, J.M. Sediment Concentrations, Loads, and Particle-Size Distributions in the Red River of the North and Selected Tributaries near Fargo, North Dakota, during the 2010 Spring High-Flow Event. 36.
9. Fleming, G.; Harrison, A.; Fleming, J.; Kite, G.; Chitale, S.; Herbertson, J.; Collins, M. Discussion. Design Curves for Suspended Load Estimation. *Proc. Inst. Civ. Eng.* **1970**, *46*, 81–92, doi:10.1680/iicep.1970.6984.
10. Kumar, R.; Yazdan, M.M.S.; Mehedi, M.A.A. Demystifying the Preventive Measures for Flooding from Groundwater Triggered by the Rise in Adjacent River Stage. *Preprints.* **2022**, 2022090452. doi: 10.20944/preprints202209.0452.v1.
11. Knighton, D. *Fluvial Forms and Processes: A New Perspective*; 2nd ed.; Routledge: London, 1998; ISBN 978-0-203-78466-2.

12. Mehedi, M.A.A.; Yazdan, M.M.S.; Ahad, M.T.; Akatu, W.; Kumar, R.; Rahman, A. Quantifying Small-Scale Hyporheic Streamlines and Resident Time under Gravel-Sand Streambed Using a Coupled HEC-RAS and MIN3P Model. *Eng* **2022**, *3*, 276–300, doi:10.3390/eng3020021.
13. Abdullah Al Mehedi, M.; Reichert, N.; Molkenthin, F. Sensitivity Analysis of Hyporheic Exchange to Small Scale Changes In Gravel-Sand Flumebed Using A Coupled Groundwater-Surface Water Model. **2020**, 20319, doi:10.5194/egusphere-egu2020-20319.
14. Kisi, O. Modeling Discharge-Suspended Sediment Relationship Using Least Square Support Vector Machine. *J. Hydrol.* **2012**, *456–457*, 110–120, doi:10.1016/j.jhydrol.2012.06.019.
15. Wicks, J.M.; Bathurst, J.C. SHESED: A Physically Based, Distributed Erosion and Sediment Yield Component for the SHE Hydrological Modelling System. *J. Hydrol.* **1996**, *175*, 213–238, doi:10.1016/S0022-1694(96)80012-6.
16. Refsgaard, J.C. Parameterisation, Calibration and Validation of Distributed Hydrological Models. *J. Hydrol.* **1997**, *198*, 69–97, doi:10.1016/S0022-1694(96)03329-X.
17. Tayfur, G.; Guldal, V. Artificial Neural Networks for Estimating Daily Total Suspended Sediment in Natural Streams. *Hydrol. Res.* **2006**, *37*, 69–79, doi:10.2166/nh.2006.0006.
18. Bor, A. Numerical Modeling of Unsteady and Non-Equilibrium Sediment Transport in Rivers | PDF | Fluid Dynamics | Water Resources Available online: https://www.scribd.com/document/143503869/t-000755 (accessed on 17 November 2022).
19. Khosravi, M.; Arellano, D. Selection of Adequate EPS-Block Geofoam for Use in Embankments Subjected to Seismic Loads 2022.
20. Salih, S.Q.; Sharafati, A.; Khosravi, K.; Faris, H.; Kisi, O.; Tao, H.; Ali, M.; Yaseen, Z.M. River Suspended Sediment Load Prediction Based on River Discharge Information: Application of Newly Developed Data Mining Models. *Hydrol. Sci. J.* **2020**, *65*, 624–637, doi:10.1080/02626667.2019.1703186.
21. Mustafa, M.R.; Rezaur, R.B.; Saiedi, S.; Isa, M.H. River Suspended Sediment Prediction Using Various Multilayer Perceptron Neural Network Training Algorithms—A Case Study in Malaysia. *Water Resour. Manag.* **2012**, *7*, 1879–1897, doi:10.1007/s11269-012-9992-5.
22. Ahmad, M.; Al Mehedi, M.A.; Yazdan, M.M.S.; Kumar, R. Development of Machine Learning Flood Model Using Artificial Neural Network (ANN) at Var River. *Liquids* **2022**, *2*, 147–160, doi:10.3390/liquids2030010.
23. Mehedi, M.A.A.; Yazdan, M.M.S. Automated Particle Tracing & Sensitivity Analysis for Residence Time in a Saturated Subsurface Media. *Liquids* **2022**, *2*, 72–84, doi:10.3390/liquids2030006.
24. Kerem Cigizoglu, H.; Kisi, Ö. Methods to Improve the Neural Network Performance in Suspended Sediment Estimation. *J. Hydrol.* **2006**, *317*, 221–238, doi:10.1016/j.jhydrol.2005.05.019.
25. Alp, M.; Cigizoglu, H.K. Suspended Sediment Load Simulation by Two Artificial Neural Network Methods Using Hydrometeorological Data. *Environ. Model. Softw.* **2007**, *22*, 2–13, doi:10.1016/j.envsoft.2005.09.009.
26. Kisi, Ö. Constructing Neural Network Sediment Estimation Models Using a Data-Driven Algorithm. *Math. Comput. Simul.* **2008**, *79*, 94–103, doi:10.1016/j.matcom.2007.10.005.
27. Jothiprakash, V.; Garg, V. Reservoir Sedimentation Estimation Using Artificial Neural Network. *J. Hydrol. Eng.* **2009**, *14*, 1035–1040, doi:10.1061/(ASCE)HE.1943-5584.0000075.
28. Kisi, O.; Haktanir, T.; Ardiclioglu, M.; Ozturk, O.; Yalcin, E.; Uludag, S. Adaptive Neuro-Fuzzy Computing Technique for Suspended Sediment Estimation. *Adv. Eng. Softw.* **2009**, *40*, 438–444, doi:10.1016/j.advengsoft.2008.06.004.
29. Chou, W.-C. Modelling Watershed Scale Soil Loss Prediction and Sediment Yield Estimation. *Water Resour. Manag.* **2010**, *24*, 2075–2090, doi:10.1007/s11269-009-9539-6.
30. Guven, A.; Kişi, Ö. Estimation of Suspended Sediment Yield in Natural Rivers Using Machine-Coded Linear Genetic Programming. *Water Resour. Manag.* **2011**, *25*, 691–704, doi:10.1007/s11269-010-9721-x.

31. Smith, J.; Eli, R.N. Neural-Network Models of Rainfall-Runoff Process. *J. Water Resour. Plan. Manag.* **1995**, *121*, 499–508, doi:10.1061/(ASCE)0733-9496(1995)121:6(499).
32. Tawfik, M.; Ibrahim, A.; Fahmy, H. Hysteresis Sensitive Neural Network for Modeling Rating Curves. *J. Comput. Civ. Eng.* **1997**, *11*, 206–211, doi:10.1061/(ASCE)0887-3801(1997)11:3(206).
33. PANAGOULIA, D. Artificial Neural Networks and High and Low Flows in Various Climate Regimes. *Hydrol. Sci. J.* **2006**, *51*, 563–587, doi:10.1623/hysj.51.4.563.
34. Fernando, D.A.; Shamseldin, A.Y. Investigation of Internal Functioning of the Radial-Basis-Function Neural Network River Flow Forecasting Models. *J. Hydrol. Eng.* **2009**, *14*, 286–292, doi:10.1061/(ASCE)1084-0699(2009)14:3(286).
35. Edossa, D.C.; Babel, M.S. Application of ANN-Based Streamflow Forecasting Model for Agricultural Water Management in the Awash River Basin, Ethiopia. *Water Resour. Manag.* **2011**, *25*, 1759–1773, doi:10.1007/s11269-010-9773-y.
36. Kisi, O.; Nia, A.; Gosheh, M.; Tajabadi, M.; Ahmadi, A. Intermittent Streamflow Forecasting by Using Several Data Driven Techniques. *Water Resour. Manag. Int. J. Publ. Eur. Water Resour. Assoc. EWRA* **2012**, *26*, 457–474.
37. Shamseldin, A.Y. Application of a Neural Network Technique to Rainfall-Runoff Modelling. *J. Hydrol.* **1997**, *199*, 272–294, doi:10.1016/S0022-1694(96)03330-6.
38. Jayawardena, A.W.; Fernando, D.A.K. Use of Radial Basis Function Type Artificial Neural Networks for Runoff Simulation. *Comput.-Aided Civ. Infrastruct. Eng.* **1998**, *13*, 91–99, doi:10.1111/0885-9507.00089.
39. Ju, Q.; Yu, Z.; Hao, Z.; Ou, G.; Zhao, J.; Liu, D. Division-Based Rainfall-Runoff Simulations with BP Neural Networks and Xinanjiang Model. *Neurocomputing* **2009**, *72*, 2873–2883, doi:10.1016/j.neucom.2008.12.032.
40. Bhadra, A.; Bandyopadhyay, A.; Singh, R.; Raghuwanshi, N.S. Rainfall-Runoff Modeling: Comparison of Two Approaches with Different Data Requirements. *Water Resour. Manag.* **2010**, *24*, 37–62, doi:10.1007/s11269-009-9436-z.
41. Evsukoff, A.G.; Lima, B.S.L.P. de; Ebecken, N.F.F. Long-Term Runoff Modeling Using Rainfall Forecasts with Application to the Iguaçu River Basin. *Water Resour. Manag.* **2011**, *25*, 963–985, doi:10.1007/s11269-010-9736-3.
42. Kordrostami, S.; Alim, M.A.; Karim, F.; Rahman, A. Regional Flood Frequency Analysis Using an Artificial Neural Network Model. *Geosciences* **2020**, *10*, 127, doi:10.3390/geosciences10040127.
43. Hall: Regional Flood Frequency Analysis Using Artificial... - Google Scholar Available online: https://scholar.google.com/scholar_lookup?title=Regional%20Flood%20Frequency%20Analysis%20Using%20Artificial%20Neural%20Networks&publication_year=1998&author=M.J.%20Hall&author=A.W.%20Minns (accessed on 17 November 2022).
44. Karl, A.K.; Lohani, A.K. Development of Flood Forecasting System Using Statistical and ANN Techniques in the Downstream Catchment of Mahanadi Basin, India. *J. Water Resour. Prot.* **2010**, *2*, 880–887, doi:10.4236/jwarp.2010.210105.
45. Kar, A.K.; Winn, L.L.; Lohani, A.K.; Goel, N.K. Soft Computing–Based Workable Flood Forecasting Model for Ayeyarwady River Basin of Myanmar. *J. Hydrol. Eng.* **2012**, *17*, 807–822, doi:10.1061/(ASCE)HE.1943-5584.0000505.
46. Khalil, M.; Panu, U.S.; Lennox, W.C. Groups and Neural Networks Based Streamflow Data Infilling Procedures. *J. Hydrol.* **2001**, *241*, 153–176, doi:10.1016/S0022-1694(00)00332-2.
47. Cigizoglu, H.K.; Kişi, Ö. Flow Prediction by Three Back Propagation Techniques Using K-Fold Partitioning of Neural Network Training Data. *Hydrol. Res.* **2005**, *36*, 49–64, doi:10.2166/nh.2005.0005.
48. Mehedi, M.A.A.; Khosravi, M.; Yazdan, M.M.S.; Shabanian, H. Exploring Temporal Dynamics of River Discharge Using Univariate Long Short-Term Memory (LSTM) Recurrent Neural Network at East Branch of Delaware River. *Hydrology* **2022**, *9*, 202, doi:10.3390/hydrology9110202.

49. Yazdan, M.M.S.; Khosravia, M.; Saki, S.; Mehedi, M.A.A. Forecasting Energy Consumption Time Series Using Recurrent Neural Network in Tensorflow. *Preprints*. **2022**, 20220904041. doi: 10.20944/preprints202209.0404.v1.
50. Home - Greene County Soil & Water Conservation District Available online: https://www.gcswcd.com/ (accessed on 17 November 2022).
51. STONY CLOVE CREEK BLW OX CLOVE AT CHICHESTER NY Available online: https://waterdata.usgs.gov/monitoring-location/01362370/ (accessed on 17 November 2022).
52. USGS Surface-Water Historical Instantaneous Data for the Nation: Build Time Series Available online: https://waterdata.usgs.gov/nwis/uv?referred_module=sw&search_criteria=search_station_nm&search_criteria=search_site_no&search_criteria=site_tp_cd&submitted_form=introduction (accessed on 17 November 2022).
53. Khosravi, M.; Arif, S.B.; Ghaseminejad, A.; Tohidi, H.; Shabanian, H. Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices. *Preprints*. **2022**, 2022090341. doi: 10.20944/preprints202209.0341.v1.
54. Khosravi, M.; Tabasi, S.; Hossam Eldien, H.; Motahari, M.R.; Alizadeh, S.M. Evaluation and Prediction of the Rock Static and Dynamic Parameters. *J. Appl. Geophys.* **2022**, *199*, 104581, doi:10.1016/j.jappgeo.2022.104581.
55. Abdollahzadeh, M.; Khosravi, M.; Hajipour Khire Masjidi, B.; Samimi Behbahan, A.; Bagherzadeh, A.; Shahkar, A.; Tat Shahdost, F. Estimating the Density of Deep Eutectic Solvents Applying Supervised Machine Learning Techniques. *Sci. Rep.* **2022**, *12*, 4954, doi:10.1038/s41598-022-08842-5.
56. Karimi, M.; Khosravi, M.; Fathollahi, R.; Khandakar, A.; Vaferi, B. Determination of the Heat Capacity of Cellulosic Biosamples Employing Diverse Machine Learning Approaches. *Energy Sci. Eng. n/a*, doi:10.1002/ese3.1155.
57. Zhu, X.; Khosravi, M.; Vaferi, B.; Nait Amar, M.; Ghriga, M.A.; Mohammed, A.H. Application of Machine Learning Methods for Estimating and Comparing the Sulfur Dioxide Absorption Capacity of a Variety of Deep Eutectic Solvents. *J. Clean. Prod.* **2022**, *363*, 132465, doi:10.1016/j.jclepro.2022.132465.
58. Gupta, H.V.; Kling, H. On Typical Range, Sensitivity, and Normalization of Mean Squared Error and Nash-Sutcliffe Efficiency Type Metrics. *Water Resour. Res.* **2011**, *47*, doi:10.1029/2011WR010962.
59. SuspendedSediment/README.Md at Main · MShivaKh/SuspendedSediment Available online: https://github.com/MShivaKh/SuspendedSediment (accessed on 21 November 2022).