2020-01-01

A Comparison Of Data-Driven And Process-Based Modeling For Nutrient Estimation In A Eutrophic Reservoir

Yohtaro Kobayashi
University of Texas at El Paso

A COMPARISON OF DATA-DRIVEN AND PROCESS-BASED MODELING FOR NUTRIENT ESTIMATION IN A EUTROPHIC RESERVOIR

YOHTARO CHRISTOPHER KOBAYASHI

Master's Program in Environmental Engineering

APPROVED:

Ivonne Santiago, Ph.D., P.E. Chair

Saurav Kumar, Ph.D., P.E. Co-Chair

Shane Walker, Ph.D., P.E.

Deana Pennington, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

A COMPARISON OF DATA-DRIVEN AND PROCESS-BASED MODELING FOR NUTRIENT ESTIMATION IN A EUTROPHIC RESERVOIR

by

YOHTARO CHRISTOPHER KOBAYASHI, BSCE

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENVIRONMENTAL ENGINEERING

Department of Civil Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2020

ACKNOWLEDGEMENTS

I would like to thank my family, especially my parents. Without their support, I would not be in a position to be attending University. Their help had allowed me to go back to school and finish what I had started some years ago.

I would like to thank my friends I made while attending University. The study sessions and discussions we had helped in learning the material. Our comradery as students boosted our morale as exams neared.

I would like to thank the professors in the Department of Civil Engineering at UTEP. Their knowledge helped direct me towards the path I ultimately chose for my specialization in civil engineering.

I would like to thank Dr. Ivonne Santiago for all the support she has given me. Her help and advice while I was an undergraduate and as a graduate student has been greatly appreciated.

I would like to thank Dr. Saurav Kumar for his guidance. It is thanks to him that my interest is in environmental engineering. From when I was an undergraduate to the present, he has been looking out for me.

Finally, I would like to thank everyone here for their patience. Writing a thesis has been difficult and if not for everyone's patience, I would have had an even harder time finishing.

ABSTRACT

As land use around bodies of water changes, the need to model the body of water increases. Models help to educate, understand, and predict the state of water. Process-based models are commonly used in modelling bodies of water, but there are challenges with these kinds of models. They require data which can be difficult for certain communities to obtain due to logistics or cost, are computationally intensive, technically complicated, and require calibration. In contrast, a data-driven model simply connect relationships from the data, are not as computationally intensive nor technically complicated, and do not require calibration. This research compared a data-driven model with a process-based model to verify if a data-driven model is a viable alternative to process-based model using the same sets of data. The research also attempted to find a relationship between water quality data, hydrological data, meteorological data, and remote sensing data in the form of electromagnetic radiation obtained by satellites Landsat 5 and Landsat 7. The study area for this research was in Occoquan Reservoir over a five-year period (2008-2012). A long short-term memory neural network model was developed and fed with data. The results of the model were then compared with the results from a CE-QUAL-W2 analysis. The comparison suggested that a data-driven model cannot be used as an alternative to a process-based model. Further research is required as the data used had multiple gaps which affected the performance of the data-driven model. Optimal data for future research should have high frequency of sampling, less censored data, and electromagnetic radiation readings obtained from an unmanned aerial vehicle as opposed to a satellite.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Water quality is an important issue as water sustains life. As land use changes around bodies of water, the influx of nutrients gets affected. In order to understand the state of a body of water, computer modeling is required. A computer model can predict the changes in water quality occurring within a body of water. For example, an increase of nutrients, such as phosphates and nitrates, within the water can promote harmful algal blooms (HABs) Algal blooms can affect the taste of water as well as being toxic to both humans and animals (Falconer, 1989). Though water quality parameters are commonly modeled using a process-based model, there are difficulties that arise in that the underlying equations used in the models require extensive data, some of which can be difficult to obtain due to logistics, costs, or other reasons. Software that run the models require multiple inputs aside from the raw data, such as the area, depth, precipitation, and evaporation to name a few thus increasing both the complication of the software and the work needed to collect such data (Tong and Chen, 2002). They also require calibration which increases the amount of effort required to run the model. The benefit of these steps though is that they add real world boundaries to the model parameters increasing the accuracy and precision. Additionally, due to the number of processes the software runs concurrently, the computer that runs the software must have a high computing power (Cox, 2003). The lower the computing power, the longer it takes to run the software.

In contrast, data-driven models completely ignore the equations and simply bridge the input and output variables through statistical or machine learning methods (Orouji et al., 2013). Data-driven models are less technically complicated as it only requires a background of statistical knowledge. With data-driven models, while there are no strict rules in what kind of data can be used, there should at least be some degree of relationship between the data. This can be both an

advantage and a disadvantage as the model can detect relationships between data where none was thought possible, but also not able to detect the difference between actual data and erroneous data. Without calibration and additional inputs, the model isn't constrained by physical and scientific laws. Data-driven models allow for the combination of data from different sources that is normally not seen in process-based models (Shen et al., 2019). The source that is used in process-based models are the water itself as well as the weather conditions. The source not commonly used in process-based models comes from the electromagnetic radiation that's being sent by the sun and being reflected off the Earth's surface.

Remote sensing is the act of recording data at a place of interest without physically being at that location. Some examples of remote sensing are installing a sensor in a river to measure flow rate, using cameras to count the number of cars that pass through an intersection, and using satellites to measure the electromagnetic radiation. Satellites, such as the Landsat series, are used to record the electromagnetic radiation reflected off the Earth's surface. All particles on Earth reflect and absorb light, or electromagnetic radiation, to some degree. Clear water can be thought of as clean water whereas water with high turbidity, or cloudiness, can be thought of as dirty as it indicates a high amount of particles. Those particles can then be detected using RS which leads to further improving our understanding of bodies of water. Some examples of RS in water quality are using RS for monitoring purposes (Ritchie et al., 2003), quantification of shallow water quality parameters (Liu et al., 2003), and estimating coastal water quality (Brando and Dekker, 2003). There are some limitations in remote sensing for water quality though. RS can only capture the state of the water at the surface. As the depth increases, the amount of light that can penetrate the water decreases. There is also interference with the data recorded depending on the state of the

weather at the time the images are taken. Clouds can completely cover up a scene and atmospheric effects can affect the readings of the images.

The purpose of this project was to compare a data-driven model with a process-based model. Due to the reliance on computers and processing power, data-driven models and remote sensing are still relatively new technologies compared to how well-established process-based models are. What this research did was take the data from water quality sampling, hydrological data, and meteorological data and combine it with data obtained from remote sensing to predict the nutrients in the water. A long short-term memory neural network (a type of data-driven model) was developed with which the data was fed into and the results compared with a previous analysis done using CE-QUAL-W2 (a process-based model). The CE-QUAL-W2 analysis was used as a criterion to rank the performance of the data-driven model.

This project used Occoquan Reservoir, a highly eutrophic reservoir located in Virginia, as a testbed. It has a seasonal pattern of high nutrients in the summer which produces a large amount of HABs in the water. The reservoir is used for recreational purposes as well as a source of drinking water for the population living nearby. This is a cause for concern as, previously stated, algal blooms affect the taste of water and can be toxic to humans. This location was chosen as there has been a previous analysis using a CE-QUAL-W2 (a process-based model) by the Occoquan Watershed Monitoring Laboratory (OWML). The analysis was from 2008 to 2012.

Study Area

Occoquan Reservoir is a body of water in northern Virginia that sustains life for both residential and wildlife, as well as being used for recreational purposes. The reservoir is formed by a dam on Occoquan River and the river is fed by two sub-basins, Bull Run and Occoquan Creek

(*The Occoquan Watershed | OWML | Virginia Tech*). The reservoir occupies an area of 2100 acres and supplies 40% of clean drinking water for approximately 2 million people. The watershed which drains into the river covers 590 square miles and according to the 2000 census, contains a population of 363,000 residents (*Occoquan Reservoir par. 6*). The major land use is for agricultural with urban land use on the rise (Miller et al., 1997).

While being used as a fresh water supply, the reservoir is currently on Virginia's Impaired Waters – 303(d) list which is a list of a state's list of impaired and threatened waters. The reasons for being on the list are due to high levels of phosphorous, turbidity, low dissolved oxygen, the presence of copper sulfate, and growing presence of pharmaceuticals (*Occoquan Reservoir*). The reservoir is considered highly eutrophic during the summer which leads to algal blooms.

The Occoquan Watershed Monitoring Laboratory (OWML) collects and analyzes water samples from the Occoquan Reservoir on a weekly basis at various sampling stations located throughout the reservoir. Figure 1 shows the reservoir and the locations of the sampling stations. In this figure, the water is flowing towards the SE point.

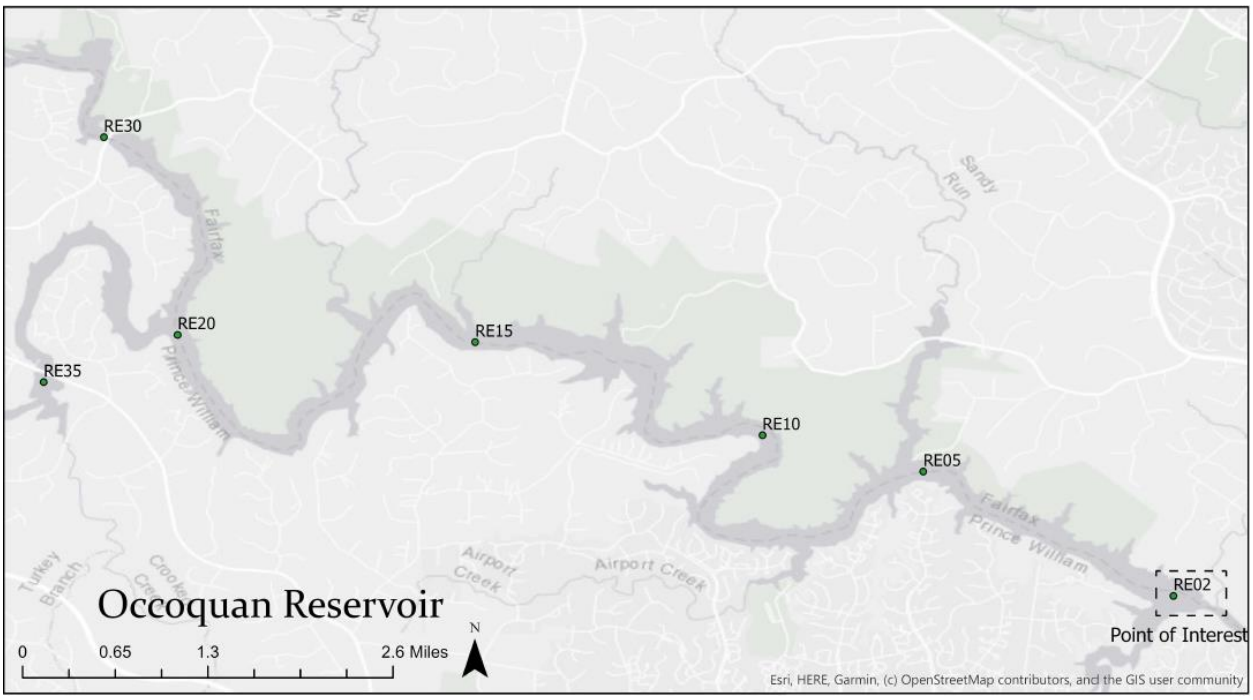


Figure 1 Occoquan Reservoir

While the samples taken are from the surface and bottom of the reservoir, this project focuses on just the surface as well as for just one station, RE02. A previous analysis was done by OWML on Occoquan Reservoir using a linked model of a watershed model, HSPF and a receiving water model, CE-QUAL-W2 for the years 2008-2012.

Figure 2 shows the complete watershed which ultimately flows to the reservoir.

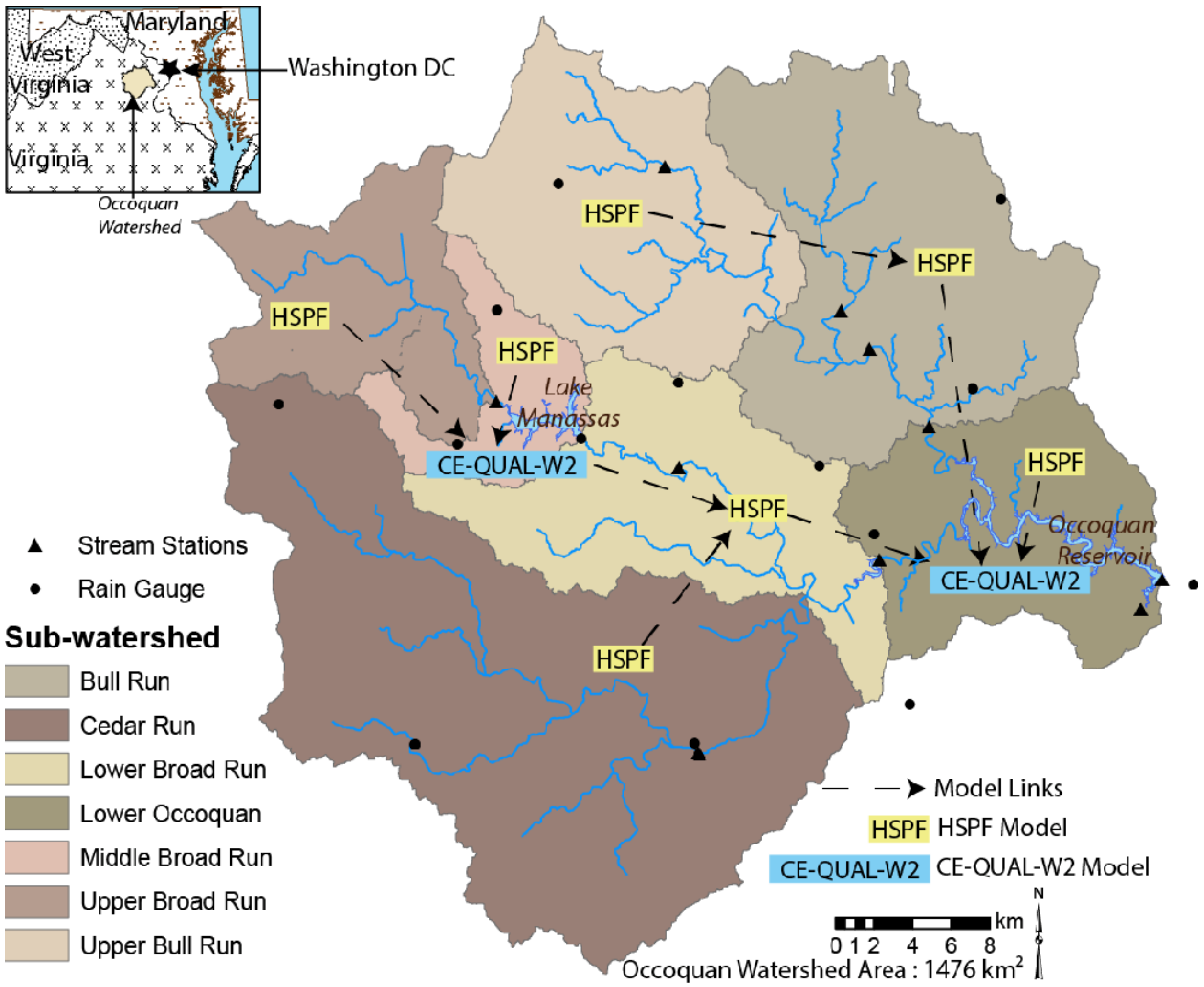


Figure 2 Occoquan Watershed Linked Model (Kumar et al., 2014)

The analysis uses the flow rate of the reservoir, dissolved oxygen (DO), temperature, alkalinity, organic phosphorous (OP), Total Nitrate [ammonium (NH_3-N) + Nitrate-Nitrite (Ox-N)], total suspended solids (TSS), chlorophyll a (Chla) as well as meteorological data which consists of air temperature, dew temperature, wind speed, and cloud cover. CE-QUAL-W2 is a process-based model in that it uses mathematical equations derived from mass, energy, and momentum conservation along with empirical observations to simulate conditions.

CHAPTER 2: GOALS AND OBJECTIVES

The main goal of this project is to predict concentrations of nutrients in water using machine learning combined with remote sensing data. Specific objectives include:

1) Develop a machine learning model using Long Short-Term Memory neural network to estimate concentration of nutrients in a eutrophic reservoir.
2) Compare the neural network model with CE-QUAL-W2, a process-based model.
3) Determine non-parametric relationships between remote sensing data, hydrologic data, and water quality data using machine learning.

CHAPTER 3: LITERATURE REVIEW

Water Quality Modeling

Managing water is an important task as water is required for all living organisms. To manage water, a model is required in order to simulate the system and see where, if any, problems are occurring. The change in regulations are the deciding factor in what the specific problem can be. Due to that, as regulation changes, so do models (Ambrose et al., 2004).

Though observed data is preferred in determining if a body of water is impaired, there are situations where models are used. Models are an option in areas where getting water samples would be difficult. Another situation would be to predict the change in water quality due to change in land use or from changes in water management (Loucks et al., 2017).

The models can simulate and predict the changes in chemical pollutant through three fundamental principles, the conservation of energy, mass, and momentum. Conservation in such that all three can neither be created nor destroyed but can be altered in some form. The amount going into a system should be equal to the amount coming out of the system. The applications of the conservation of energy in water modeling is to model evaporation of water through the change in temperature within the system and to find the interactions of water due to kinetic energy. Conservation of mass is the base of most water quality models. It is applied to water mass for hydrodynamics and the mass of matter to find the change in concentration of said matter. Newton's first law of motion, which is based off the law of conservation of momentum, states that an object at rest will stay at rest and an object in motion will remain in motion until acted upon by an external force. From the conservation of momentum, multiple equations for fluid motion are derived. (Martin and McCutcheon, 1999). Mass balance, for example, combines the conservation of mass and the conservation of momentum. Multiple flow rates going into a lake combine to equal the

flow rate coming out of the lake with each different inflow carrying different concentrations of matter. The matter can interact with one another once inside the lake to change form, but the total mass will remain the same which then comes out of the lake. The equation for a mass balance as well as conservation in general is shown below:

$$Output = Inputs \pm Change$$

These models which are derived from laws are considered a mechanistic model. We can apply mathematics to known relationships and calculate the changes. The other type of model is called an empirical model. Empirical models use statistics to find relationships within the data without using any established laws or background information of the data. Many modern software which can model water quality use a combination of mechanical and empirical models.

Models can range from simple to complex where some factors are omitted in the simple models and in a complex model, multiple factors are included in the calculation. The simple models model an ideal world where the properties stay constant and as the model gets more complex, it starts to get closer to reality. No model can accurately capture reality though as reality always has some degree of randomness. Simple models can also be done by hand as the calculations are relatively quick.

While this project focuses on the CE-QUAL-W2 model, there are numerous other models being used in water quality. Watershed models that are commonly used are Soil and Water Assessment Tool (SWAT) (Neitsch et al., 2002), Storm Water Management Model (SWMM) (Huber and Barnwell, 1988), and Hydrologic Simulation Program-FORTRAN (HSPF) (Bicknell John C Imhoff John L Kittle et al, 2005.). OWML uses a linked model for their analysis of Occoquan Reservoir in which the outputs of an HSPF model are the used as the inputs for the CE-QUAL-W2 model.

CE-QUAL-W2

CE-QUAL-W2 (W2) is a two-dimensional hydrodynamic and water quality model that is commonly used in long and narrow waterbodies. The model was developed by Edinger and Buchak in 1975 and is still currently being updated. The most recent version of the model is version 4.2.1 which was released on July 21, 2020 (ce.pdx.edu/w2). It has been used in various waterbodies worldwide since as early as 1979. The W2 model was first used for Occoquan Reservoir in 1994 by the Northern Virginia Planning District Commission (NVPDC 1994). W2 models basic eutrophication processes such as algae/nutrient/DO dynamics. It can simulate the water surface elevation, velocity, temperature, and other water quality constituents, some of which are DO, alkalinity, NH_3-N, Ox-N, TSS, and CHLa. The model assumes lateral homogeneity which explains the longitudinal and vertical water quality gradients. Some of the capabilities of the model are long term simulations, multiple branches, waterbodies, inflows, and outflows which allow complex water systems to be modeled, variable grid spacing so that higher resolution can be used where needed, and customization in the output. (Cole and Wells, 2003). The W2 model uses five governing equations, the x and z momentum, continuity equation state equation, and free surface equation. Table 1.1 shows the data required to run the model as well as a brief description of the data.

Table 1.Required Data for W2 Model

Data Required	**Brief Description of Data**
Geometric Data	Defines the finite difference in the watershed.
Initial Conditions	The condition of waterbody when model first starts
Boundary Conditions	Inflows, Outflows, Head and Surface Boundary Conditions.
Hydraulic Parameters	Dispersion and diffusion coefficients.
Kinetic Parameters	Coefficients that affect the constituent kinetics.
Calibration Data	Provides initial and boundary conditions and assesses performance.

Data-Driven Modeling (Machine Learning)

Data-driven, or machine learning, models are a relatively new concept in water quality modeling. The rise of technology in computational power allows us the capability to feed data into an algorithm and produce results formed from the hidden relationships in the data as well as capturing the underlying physics and chemistry. Some of the methods used are neural networks (Kuo et al., 2007), fuzzy inference methods (Orouji et al., 2013), support vector machines (He et al., 2014), and k-nearest neighbors (Towler et al., 2009). As this project uses a neural network, this section will cover the neural network concept as well as previous literature on water quality that uses neural networks. Before neural networks are discussed though, statistical modeling needs to be reviewed as it is the basis to data-driven models.

STATISTICAL MODELING

Statistical modeling is the process applying mathematical equations to raw data and reaching a conclusion. The models can be used to find the relationship between variables and non-variables and measuring the correlation. The purposes of a statistical model are to predict, estimate, and describe (Friendly and Meyer, 2015).

Linear regression is one such method used to model the relationship between single or multiple variables. The model has a dependent variable, the variable that you're interested in, and the independent variables, which can explain the dependent variable. The equation for a univariate linear regression is:

$$Y = a + bX$$

Where Y is the dependent variable, X is the independent variable, b is the slope of the line, and a is the intercept of the line.

For multivariate linear regression:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Where n is the number of variables that are being used to explain the dependent variable.

An error term is commonly added to the equation to account for the difference between the model and observed values. A scatterplot is useful when looking at raw data as a trend can be inferred by visual inspection. A regression line can then help identify the outliers and influential data points. The closer the points are to the line, the more those points influence the direction of the line. A coefficient of correlation, R value, can be calculated with values ranging from -1 to 1 to determine the degree of relationship between the variables. At 1, as one variable increases, so does the other; and at -1, as one variable increases, the other decreases. In situations with multiple variables, the coefficient of determination, R^2, is a better term. The values for the coefficient of determination range from 0 to 1 and explain the percentage variation in y explained by all the x variables. A value close to 1 shows that the variables are highly correlated.

The phrase, "correlation does not imply causation" appears in statistical literature. It indicates that even if variables have a high correlation, it is not the direct cause of the variable that you're interested in.

NEURAL NETWORK

Neural networks get its name from the neural network of an animal's brain. A neural network can be described as a network of nodes with weights in-between that adjust as the learning progresses. Neural networks are typically made up of three layers, the input layer, the hidden layer, and the output layer. The input layer, in statistical terms, can be thought of as the independent variable. In machine learning terms, it is the features of the dataset. The output layer, in statistical

terms, can be thought of as the dependent variable. In machine learning terms, it is the label of the dataset. The hidden layers contain the activation functions. The activation function are monotonic differential functions that assigns an output value from an input. Depending on the function used, it can range from 0 to 1, 0 to ∞, or -1 to 1 (Fausett, 1994). Figure 3 shows a basic neural network with three layers. The input layer has three neurons, the hidden layer has six neurons, and the output layer has one neuron.

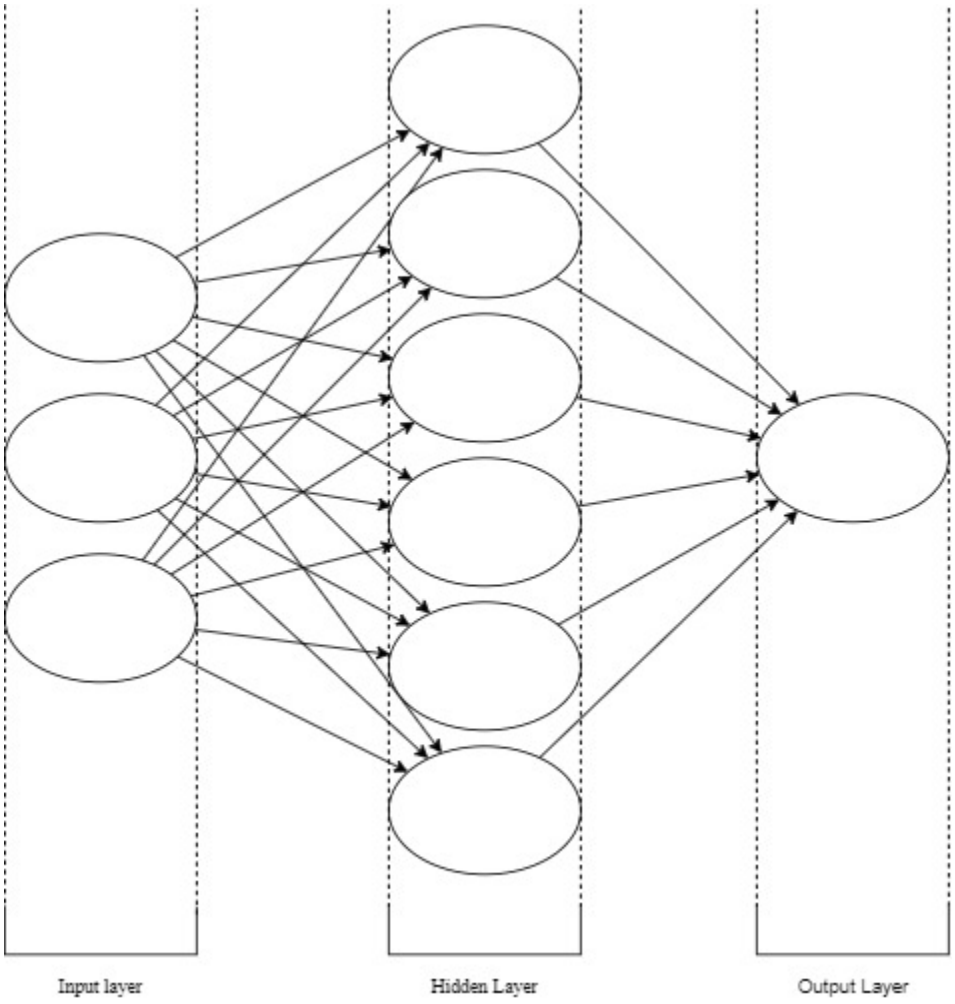


Figure 3. Basic Neural Network with three inputs and one output

A neural network is considered deep learning if the number of hidden layers is larger than one or the total layers, including input and output, are greater than three. Deep learning can be

beneficial for more complex datasets that involve time-series or computer vision (Hinton and Osindero, 2006).

Some examples of previous literature that have used neural networks is the prediction of monthly water quality parameters in Axios River in Northern Greece (Diamantopoulou et al., 2005), estimating the biochemical oxygen demand (BOD) of an inlet to a wastewater treatment plant (Dogan et al., 2008), computing the dissolved oxygen (DO) and BOD using eleven water quality input variables (Singh et al., 2009), multi objective optimization of water quality management (Wen and Lee, 1998), estimating water quality index in the Langat River Basin, Malaysia (Juahir et al., 2004).

NEURON

A neuron is basis for other neural networks. The equations for other forms of neural networks are essentially made up of neurons. A neural network is made up of inputs (x), hidden states (h), and outputs (y). To get both the hidden states and the output, a sigmoid function, or an activation function, is applied to a linear equation with the inputs and a weight applied to it and a bias term to squash the values between 0 and 1.

$$h = \sigma\left(W_h^T x + b_h\right)$$

$$\hat{y} = \sigma(W_o^T h + b_o)$$

Ultimately, the goal of neural networks is to optimize the weights of each neuron with an optimization function.

HYPERPARAMETERS

The hyperparameters of a neural network determine how complex the model is. It allows the model to be more flexible depending on your inputs and the output you're interested in. The activation function is considered a hyperparameter, and it for example, can be changed to better suit if the output you're looking for is categorical or quantitative.

Another hyperparameter, the optimization function, determines how the model is learning and how quick it's learning. Gradient descent is one example of an optimizer. Optimization functions, like activation functions, have advantages and disadvantages to different datatypes. Knowing when to use which function is learned either through experience or with the use of a hyperparameter tuner.

RECURRENT NEURAL NETWORK

While an artificial neural network can handle time series data, a recurrent neural network (RNN) is better suited to handle time series data as it can use its internal state (memory) to process sequences of inputs. A standard artificial neural network is a feed forward model in that the data only moves forward. While it's still learning, it's not learning using previous memory. An RNN uses back propagation to take what it has learned and apply it to new data. Due to this, it is becoming the common choice for sequential, or time series prediction.

The figure below is a diagram for how data travels through an RNN.

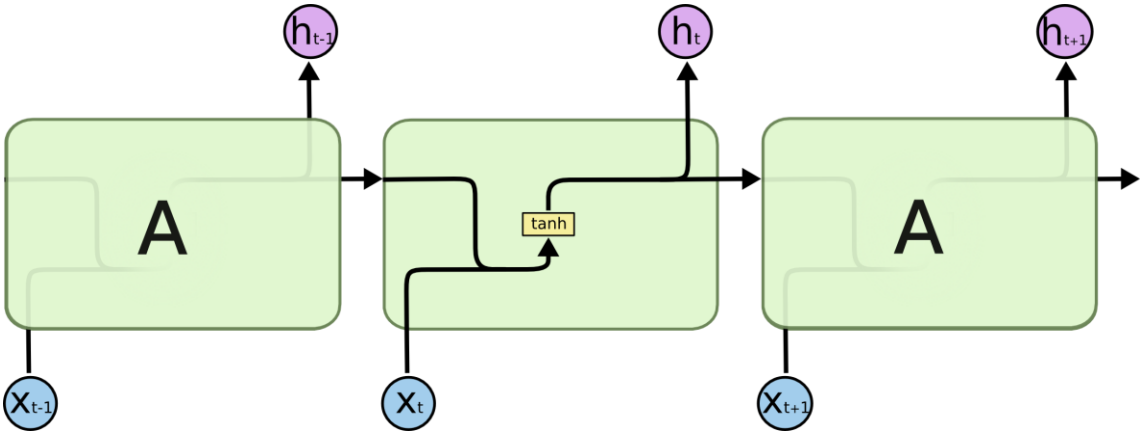


Figure 4: RNN Diagram Olah, Christopher "Understanding LSTM Networks" August 2015 https://colah.github.io/posts/2015-08-Understanding-LSTMs/

The diagram can be explained through the following equations:

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

Where h_t is the hidden state, or output vector, W_{hh} is the weight for the hidden state, h_{t-1} is the previous hidden state, W_{xh} is the weight of the input, x_t is the input, W_{hy} is the weight of the output, and y_t is the output. In this case, the activation function is a tanh. Essentially, both the previous state and the current input along with the corresponding weights applied are summed and squashed to a value between -1 and 1 through the tanh function. That value is then fed as the next hidden state into the next cell along with the input at time step +1. This is repeated until all time steps are used, then the final output is produced.

One of the drawbacks of an RNN is that as the gap increases between two points in time, the RNN has difficulty in learning to connect the information (Hochreiter, 1991). Long Short Term

Memory (LSTM) is a type of RNN which can handle long-term dependencies (Hochreiter and Urgen Schmidhuber, 1997).

LONG SHORT-TERM MEMORY (LSTM)

The cell of an LSTM contains multiple gates that allow or prevent data from moving to the next cell using a range of 0 to 1 where 0 stops data from passing through and 1 completely opens the gate to allow data through. The figure below shows the diagram of an LSTM.

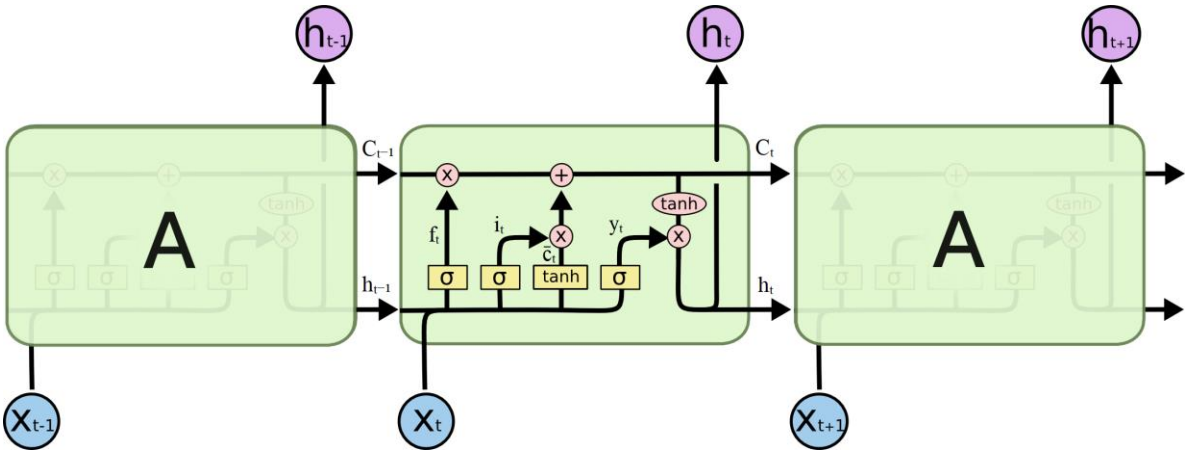


Figure 5: LSTM Diagram Olah, Christopher "Understanding LSTM Networks" August 2015 https://colah.github.io/posts/2015-08-Understanding-LSTMs/

The first gate, called the forget gate (f_t), determines if the data should be kept or thrown away with the equation:

$$f_t = \sigma(w_f x_t + R_f h_{t-1} + b_f)$$

The next gate, the input gate (i_t), determines which value will get updated. This is done through two equations:

$$i_t = \sigma(w_i x_t + R_i h_{t-1} + b_i)$$

$$\bar{c}_t = \sigma(w_c x_t + R_c h_{t-1} + b_c)$$

Then, the cell state is updated using the equation:

$$c_t = f_t c_{t-1} + i_t \bar{c}_t$$

The final gate, the output gate (y_t), then determines what exactly is going to be outputted using the equations:

$$y_t = \sigma(w_y x_t + R_y h_{t-1} + b_y)$$

$$h_t = y_t \sigma(c_t)$$

Where x_t is the input vector, w_i, w_f, and w_y are weights matrix for the input, forget, and output gates to the input, R_i, R_f, and R_y define the weights matrix for the input, forget, and output gates to the input, b_i, b_f, and b_y are the input, forget, and output gate bias, c_{t-1} and h_{t-1} are the previous cells output vector and h_t is the output vector (Barzegar et al., 2020).

Remote Sensing

Remote sensing, as the name implies, is the process of obtaining data from an object without physically being there. Generally, the term is applied when the object of interest is the Earth. Sensors are attached to either a satellite or an aircraft and records images when flown over the point of interest. The type of remote sensing can be either passive, which relies on the radiation emitted from the Sun, or active, which the sensor itself emits energy. The data that's being recorded is the reflection and absorption of electromagnetic radiation. The data is then preprocessed to

account for the atmospheric effects on the radiation. Further processing of the images involves removing the cloud cover, changing the resolution of the images, or to stich the images into a mosaic.

Electromagnetic radiation is a type of energy that can propagate through space. All matter that are above absolute zero emit some electromagnetic energy. Electromagnetic energy is modeled either with a wave model or a particle model. In a wave model, the electromagnetic radiation is modeled as harmonic waves which are characterized by the wavelength and frequency. The formula used to calculate the wavelength is:

$$c = \lambda v$$

Where c is the speed of light ($c=3x10^8$ m/s), λ is the wavelength (measured in meters), and v is the frequency (measured in hertz).

The electromagnetic spectrum, as seen in Figure 6, is a representation of the range of the electromagnetic energy.

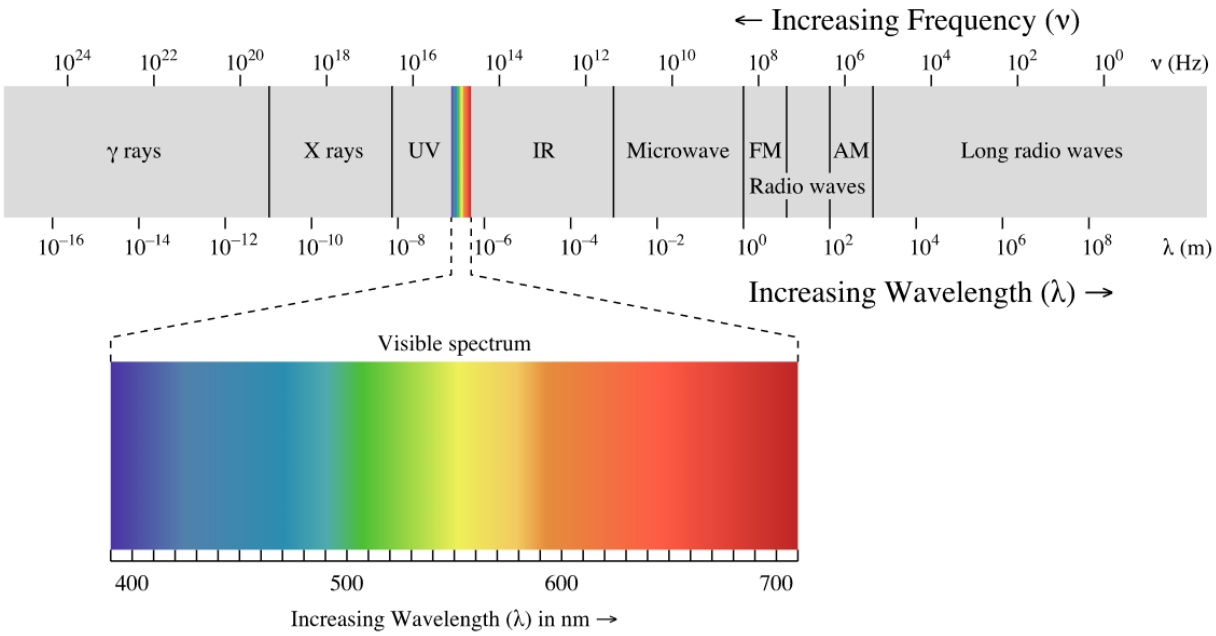


Figure 6. Electromagnetic Spectrum Ronan, Philip "EM spectrum" August 2007

https://en.wikipedia.org/wiki/File:EM_spectrum.svg Licensed under the Creative Commons Attribution SA 3.0

In remote sensing, the visible spectrum, infrared (IR), and microwave are commonly used. The visible spectrum is what the human eye perceives as visible light. It ranges from approximately 400 nanometers to 700 nanometers. Infrared can be broken up into three sections, near infrared, mid infrared, and thermal infrared. Both the visible and infrared range can be used to detect composition of the area of interest and vegetation. The values within the electromagnetic spectrum are what is measured with sensors. As the sun emits electromagnetic radiation, the sensors pick up what is reflected from the Earth's surface. Since the air in the atmosphere is made up of multiple gases, there is an atmospheric interference on the radiation. As the electromagnetic radiation travels through the atmosphere, the molecules of in the air cause the radiation to get absorbed or scatter.

The factors that affect the quality of the data taken by sensors are the spatial resolution, temporal resolution, spectral resolution, and the radiometric resolution. Digital images are made

up of pixels, essentially a square made of varying color intensities. The higher the resolution an image has, the more pixels there are. The spatial resolution is the ratio of a pixel to an area within the image taken. When the area is smaller, it is easier to identify features within the image, such as buildings, trees, or plants for example. With a larger area, the multiple features within the pixel are obscured by performing an algorithm on the electromagnetic data of all the features in the pixel. A sensor attached to a satellite has a larger field of view than a sensor attached to an aerial vehicle on Earth. The tradeoff though is that the images taken from the satellite have a much larger pixel area.

The temporal resolution of a dataset is the time between when the image is taken of the point of interest. Sensors are commonly attached to satellites which revolve around the Earth taking images. Different satellites revolve around the Earth at various velocities and orbits. When the satellites take an image at a certain location, it takes time for the satellite to get back to the same spot over the Earth to take another image. Sensors on an aerial vehicle aren't restricted by time as they can be flown when needed. Which satellite to use and if an aerial vehicle can be used is highly dependent on user interest.

Sensors cannot detect the complete range of the electromagnetic spectrum. Sensors have multiple bands where each band has a specific range of values of the electromagnetic spectrum which can be captured. The more bands a sensor has, the more information can be obtained within a single pixel. Sensors can be divided into two categories, hyperspectral and multispectral. While both include multiple bands, the main difference being that the bands in hyperspectral are contiguous. Since different components on Earth will reflect and absorb electromagnetic radiation, choosing a sensor that has a higher number of bands isn't necessarily the correct choice. If the

electromagnetic characteristics of the substance of interest is known, the sensors that can detect those values can be chosen.

Radiometric resolution describes the brightness or the shade of each pixel which relates to the amount of detail in an image. Sensors take images with a certain amount of bits per pixel. As the bits increase, the higher shades of a color can be recorded. For example, a 1-bit image will only show black or white whereas a 2-bit image will show black, dark grey, light grey, and white. The range of shading grows exponentially as the bit size increases.

Landsat 5 and Landsat 7 are two satellites that have been used for remote sensing. Landsat 5 was launched into orbit on March 1st, 1984 and was deactivated on June 5th, 2013. Landsat 7 was launched into orbit on April 15, 1999 and is still in used today. Both satellites orbit the Earth every 16 days at an altitude of 705 kilometers. Landsat 5 carries two sensors, a multispectral scanner (MSS) and a thematic mapper (TM). Landsat 7 carries only one sensor, an enhanced thematic mapper plus (ETM+). The two tables below show the bands and the spectral ranges of each band. For this research, TM bands 1-7 of Landsat 5 and ETM+ bands 1-7 of Landsat 7 were used.

Table 2: Landsat 5 Bands

Bands	**Spectral Range**
Band 4 Visible Green (MSS)	(0.5 to 0.6 µm)
Band 5 Visible Red (MSS)	(0.6 to 0.7 µm)
Band 6 Near-Infrared (MSS)	(0.7 to 0.8 µm)
Band 7 Near-Infrared (MSS)	(0.8 to 1.1 µm)
Band 1 Visible (TM)	(0.45 - 0.52 µm)
Band 2 Visible (TM)	(0.52 - 0.60 µm)
Band 3 Visible (TM)	(0.63 - 0.69 µm)
Band 4 Near-Infrared (TM)	(0.76 - 0.90 µm)
Band 5 Near-Infrared (TM)	(1.55 - 1.75 µm)
Band 6 Thermal (TM)	(10.40 - 12.50 µm)
Band 7 Mid-Infrared (TM)	(2.08 - 2.35 µm)

Table 3: Landsat 7 Bands

Bands	**Spectral Range**
Band 1 Visible (ETM+)	(0.45 - 0.52 µm)
Band 2 Visible (ETM+)	(0.52 - 0.60 µm)
Band 3 Visible (ETM+)	(0.63 - 0.69 µm)
Band 4 Near-Infrared (ETM+)	(0.77 - 0.90 µm)
Band 5 Near-Infrared (ETM+)	(1.55 - 1.75 µm)
Band 6 Thermal (ETM+)	(10.40 - 12.50 µm)
Band 7 Mid-Infrared (ETM+)	(2.08 - 2.35 µm)
Band 8 Panchromatic (ETM+)	(0.52 - 0.90 µm)

CHAPTER 4: METHODOLOGY

Data Collection

Two types of data were provided by Occoquan Watershed Monitoring Laboratory (OWML) for this research: water quality data and hydrological and meteorological data.

WATER QUALITY AND HYDROLOGICAL AND METEOROLOGICAL DATA

The data consisted of weekly readings (with the occasional missing data for a week) of dissolved oxygen (DO), temperature, alkalinity, ortho-phosphate phosphorus (OP), ammonia nitrogen (NH_3-N), oxidized nitrogen (Ox-N), total suspended solids (TSS), and chlorophyll a trichromatic (CHLA) and flow rate at various sampling stations from the year 2008 to 2012. As seen in Figure 1, RE02 was the station chosen. This station is located at the very end of the reservoir so that all the tributary streams and rivers are included in the analysis. In addition to the water quality data, the meteorological data was also provided. The meteorological data consisted of the air temperature (T_{AIR}), dew temperature (T_{DEW}), wind speed, and cloud cover. The data was obtained from the weather station at Dulles International Airport. In addition, the flow rate (Q), given by OWML, was included into the meteorological dataset. OWML used the data to predict nutrients to do an analysis using CE-QUAL-W2 which will be used as a comparison to the model developed in this project. The data that CE-QUAL-W2 and other process-based models don't utilize is electromagnetic data obtained from satellites.

The table below shows the distribution of total weekly samples provided by OWML for each year for a total of 193 entries.

Table 4: Description of Water Quality Data

Total Weekly Samples	Year	Comments
44	2008	2 weeks were skipped in Spring (dd),1 week skipped in Summer, 5 weeks skipped in Winter.
36	2009	4 weeks skipped in Spring, 4 weeks skipped in Fall, 8 weeks skipped in Winter.
36	2010	5 weeks skipped in Spring, 4 weeks skipped in Fall, 9 weeks skipped in Winter.
34	2011	6 weeks skipped in Spring, 4 weeks skipped in Fall, 8 weeks skipped in Winter.
40	2012	5 weeks skipped in Spring, 1 week skipped in Summer, 1 week skipped in Fall, 5 weeks skipped in Winter.

The following tables give a summary of the water quality data and the hydrological and meteorological data.

Table 5: Summary of Water Quality Data

	DO (mg/L)	TEMP (ºC)	TALK (mg/L as $CaCO_3$)	OP (mg/L as P)	NH3-N (mg/L as N)	OX-N (mg/L as N)	TSS (mg/L)	CHLA (µg/L)
count	193	193	193	193	193	193	193	193
mean	8.70	18.77	48.31	0.01	0.05	1.32	4.26	12.56
std	2.88	8.19	12.03	0.01	0.05	0.81	3.16	10.14
min	1.63	2.5	20.4	0.005	0.005	0.05	0.5	1
25%	7.2	11.7	39.1	0.005	0.02	0.83	2.7	5.3
50%	8.93	19.6	47.5	0.005	0.04	1.1	4	12
75%	10.98	26	56.1	0.01	0.07	1.61	4.8	17
max	13.83	31.7	74.3	0.07	0.26	4.35	25	79

Table 6: Summary of Hydrological and Meteorological Data

	T_{AIR} (ºC)	T_{DEW} (ºC)	WIND SPEED (m/sec)	CLOUD COVER	Q (m^3/s)
count	193	193	193	193	193
mean	17.6818	9.5328	2.5032	6.0811	15.5843
std	8.9038	9.6401	1.3582	3.2285	13.6130
min	-3.8545	-15.3157	0	1.25	1.35
25%	10.972	2.8867	1.545	2.8125	5.6
50%	18.4795	10.528	2.54	5.9375	12.32
75%	24.693	17.637	3.245	9.1667	21.55
max	33.1125	27.6855	6.19	10	86.34

SATELLITE DATA

The satellite data was obtained using a Python Script that pulled pixel data from the Landsat 7 and Landsat 5 satellite images in the Google Earth Engine archive. Specifically, the script was fed the coordinate data of the RE02 sampling station and a circular area with a radius of 100 meters centered around that point. The script then takes the average of the pixel data within the area. The radius of 100 meters ensures that multiple pixels are selected, and that no foliage is within the area.

There were two problems encountered with the satellite data. The first is that both Landsat 7 and Landsat 5 takes images of an area every 16 days. Most of the days that images were taken do not match up with the days that water samples were taken. The other problem, which contributed to the previous, is that a large chunk of satellite data was missing during the time span of 2008 to 2012 as can be shown in Table 7. A device which corrects the swaying of the satellite failed on May 31st, 2003. As a result, there are gaps in the images obtained by Landsat 7 leaving 78% of the pixels remaining.

Table 7: Description of Remote Sensing Data

Images Obtained	Images Missing	Year	Comments
7	28	2008	Images obtained were clustered around June to September.
9	24	2009	Images obtained were clustered around May to August, one image in October, and one image in November.
15	17	2010	Images obtained were clustered around April to September, one image in October, and one image in December.
8	19	2011	Images obtained were clustered around May to August, one image in February, and one image in October
4	15	2012	Two images obtained for May, one in July, and one in August.

A summary of the dataset is shown in Table 8.

Table 8: Summary of Remote Sensing Data

	B1	B2	B3	B4	B5	B6	B7	$B3^2$	B3/B2	B2/B3
count	146	146	146	146	146	146	146	146	146	146
mean	110.69	127.45	106.57	187.15	78.79	860.93	48.21	43819.46	0.25	0.36
std	191.52	213.21	180.79	312.70	143.89	1337.39	90.03	93719.75	0.39	0.56
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
75%	246.30	296.75	240.04	452.49	152.92	2885.85	78.57	57674.10	0.74	1.08
max	963.64	925.63	755.16	1400.41	615.29	2995.00	438.29	570270.52	1.15	1.45

Preprocessing the Data

Before the data can be fed into a machine learning algorithm, it was preprocessed using a Python script. For this project, as the data covers a span of five years, it was split so that the model will train with the first four years (2008 to 2011) and then tested on the fifth year (2012).

So that the datasets can be merged properly and leave the minimal amount of gaps, the dates were converted to week per year.

The features range from values of thousandth for concentrations and up to tens of thousands for the spectrum. So that the data is closer together in range, scaling was applied. Since water quality data can have outliers which can be attributed to rare events like storms, robust scaling was chosen as it is highly resistant to the effects of outliers (Spence and Lewandowsky, 1989).

$$z = \frac{X - Mdn}{Q_3 - Q_1}$$

X is the value of interest, Mdn is the median, Q_3 is the 75^{th} percentile and Q_1 is the 25^{th} percentile.

Grouping the Data

The data was separated into three main groups:

1) Water Quality Data (W)
2) Hydrological and Meteorological Data (H)
3) Remote Sensing Data (R)

Data sets were then created for all possible combinations for a total of seven different data sets as shown in Table 9.

Table 9: Dataset Combinations

Combination 1	W
Combination 2	H
Combination 3	R
Combination 4	**WH**
Combination 5	WR
Combination 6	HR
Combination 7	WHR

The idea behind using several combinations is to compare the accuracy of the results to see how well nutrients can be modeled while omitting certain set(s) data and if having remote sensing

data can improve the results of current modeling methods. CE-QUAL-W2, the process-based model, used the dataset of combination 4, water quality data and hydrological and meteorological data.

Creating the Neural Network Model

The neural network model chosen for this project was a Long Short Term Memory (LSTM) RNN which can handle long-term dependencies (Hochreiter and Urgen Schmidhuber, 1997). While an artificial neural network can be used for water quality data, RNN is better suited for time series data like the water quality data that was used in this project.

The code that was used for the LSTM model was written in Python coding language. The Python libraries used are listed in Table 10 along with a description of what they are used for.

Table 10: Libraries Used in Code.

Library	**Description**
Numpy	Numerical computing.
Pandas	Data manipulation and analysis.
Matplotlib	Data visualization.
Statsmodels	Statistical analysis.
Seaborn	Data visualization.
Tensorflow	Machine learning platform.
Geextract	Extract satellite imagery data from Google Earth Engine.
Keras Tuner	Optimizes the hyperparameters for neural network.

The code was written using a Jupyter Notebook environment. The benefit of using a Jupyer Notebook to run the code is the non-linearity of its workflow. It allows the changing of the models hyperparameters and to run the snippet of code for the model rather than having to run through the whole script.

Three models were created to predict three different parameters in water: chlorophyll a, total nitrogen, and total suspended solids. For each model, the datasets created from the combination of data groups were used as input.

Hyperparameter Tuning

A library for Python called Keras Tuner was utilized to tune the hyperparameters. In the code, the options for the number of neurons per layer were given, the number of layers, the activation function, the optimizer, and the learning rate for the optimizer. Table 11 shows the hyperparameters and the options that were entered.

Table 11: Hyperparameter Options

Hyperparameter	**Options**
Number of Neurons	32, 64, 96, 128, 160, 192, 224, 256
Number of Layers	1, 2, 3, 4
Activation Function	ReLU, Sigmoid, Tanh.
Optimizer	Adam, SGD
Learning Rate	0.0001, 0.001, 0.01, 0.1

Table 12 lists the optimal hyperparameters found using Keras Tuner for the seven combinations.

Table 12: Optimal Hyperparameters Found

Combination	**Number of Neurons**	**Number of Layers**	**Activation Function**	**Optimizer**	**Learning Rate**
1 (W)	256,128,224	3	sigmoid	adam	0.01
2 (H)	192,192,160	3	sigmoid	adam	0.01
3 (R)	128,128,224,256	4	sigmoid	adam	0.01
4 (WH)	224,192,256	3	sigmoid	adam	0.01
5 (WR)	128,128,224,224	4	sigmoid	adam	0.01
6 (HR)	192,224,256,224	4	sigmoid	adam	0.01
7 (WHR)	192,192,224	3	sigmoid	adam	0.01

In a sigmoid function, for values less than -5, it returns a value close to 0. For values greater than 5, it returns a value close to 1. The equation used is as follows

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

CHAPTER 5: RESULTS AND DISCUSSIONS

Processed data

A summary of the of the processed data for the water quality data, hydrological and meteorological data, and remote sensing data is shown in the table below. Merging all the datasets together to a week per year entry introduced additional zero values into the datasets. The additional zero values come from weeks where there is no water quality data, but there is remote sensing data and vice versa.

Table 13: Summary of Processed Water Quality Data

	DO	**TEMP**	**TALK**	**OP**	**NH3-N**	**OX-N**	**TSS**	**CHLA**
count	219	219	219	219	219	219	219	219
mean	-0.1508	-0.0949	-0.1731	0.9886	0.2972	0.1792	0.1704	0.1048
std	0.7057	0.5339	1.0729	2.3278	0.8429	1.0241	1.1678	0.7596
min	-1.4885	-0.9651	-2.4378	-1	-0.5455	-1.1561	-1.1429	-0.7410
25%	-0.5982	-0.5871	-0.5135	0	-0.3636	-0.3757	-0.4286	-0.4980
50%	0	0	0	0	0	0	0	0
75%	0.4018	0.4129	0.4865	1	0.6364	0.6243	0.5714	0.5020
max	0.9593	0.7346	1.5784	13	4.1818	3.8728	7.7857	4.9163

Table 14: Summary of Processed Hydrological and Meteorological Data

	T_{AIR}	**T_{DEW}**	**WIND SPEED**	**CLOUD COVER**	**Q**
count	219	219	219	219	219
mean	-0.0703	-0.0239	-0.0277	-0.0208	0.2332
std	0.5920	0.5685	0.7579	0.4839	0.8490
min	-1.1753	-1.4261	-1.1107	-0.7222	-0.6032
25%	-0.5791	-0.5121	-0.5826	-0.5556	-0.3822
50%	0.0000	0.0000	0.0000	0.0000	0.0000
75%	0.4209	0.4879	0.4174	0.4444	0.6178
max	0.96032	1.140054	1.958678	0.611111	4.771242

Table 15: Summary of Processed Remote Sensing Data

	B1	B2	B3	B4	B5	B6	B7	B3Squared	B3/B2	B2/B3
count	219	219	219	219	219	219	219	219	219	219
mean	73.7911	84.9667	71.0456	124.7697	52.5279	573.9541	32.1385	29212.9729	0.1638	0.2390
std	164.7156	184.0150	155.8087	269.9188	123.1167	1164.1020	76.8736	79188.5148	0.3356	0.4887
min	0	0	0	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0	0	0
75%	0	0	0	0	0	0	0	0	0	0
max	963.6442	925.6350	755.1626	1400.4135	615.2867	2995	438.2931	570270.5172	1.1533	1.4455

Due to the large amount of zero values in the remote sensing data, the scaling method had little effect.

Neural Network Model

The following figure shows the flow of the neural network for just the water quality data by itself.

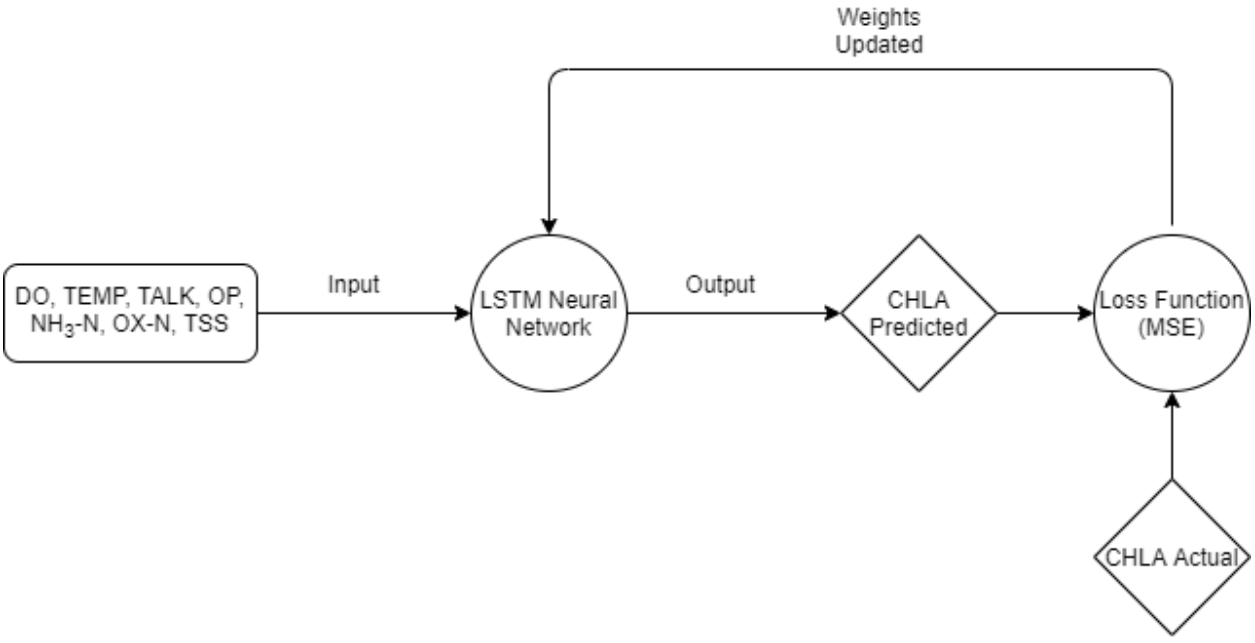


Figure 7: Neural Network Process

The steps are as follows:

1. Data is fed into the neural network.
2. Chlorophyll-a is predicted.
3. A loss function, mean squared error, is applied using the predicted chlorophyll-a and the actual chlorophyll-a.

$$MSE = \frac{1}{n}\Sigma(y - \hat{y})^2$$

   n is the number of data points, y is the actual value, and ŷ is the predicted value.

4. The calculated error is then backpropagated through the neural network to update the weights.
5. This process is repeated to minimize the loss function until it starts to plateau and the network stops learning.

The model uses the first 177 entries as the train data as it represents data taken from 2008 to 2011. The first entry is the first week of 2008 and the 177th entry is the 51st week of 2011 (52nd week is missing data). Since the window of time steps used is 5, there are 37 entries available to test on. The 37 entries represent 2012 starting with the 178th entry as the first week in 2012 and the 214th entry as the 47th week in 2012. In the following figures, know that the time steps are weekly for a year, but with missing gaps in-between.

The R2 score used to evaluate the performance of the models uses the following equation.

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})}{\Sigma(y - \bar{y})}$$

Because the score is being evaluated on unseen data, the results can fall outside of the range of 0 to 1 that is commonly seen in statistical analysis. A negative result means that the model is performing worse with the new unseen dataset than on the data that the neural network trained on. This suggests that for the fifth year which the model tested on, it's receiving a combination of values which the model had not seen for the first four years that the model trained on.

CHLOROPHYLL A

Figure 8 through Figure 14 represent the combinations 1-7 as previously seen in Table 9. The table that follows the figures lists the R^2 values for all the models.

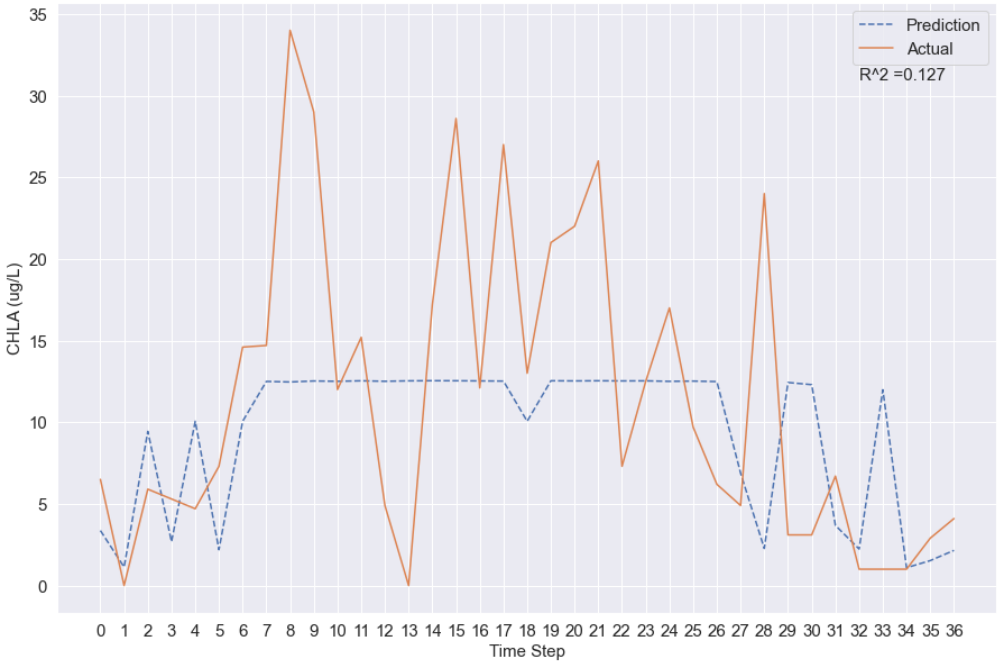


Figure 8: Chlorophyll-a - Combination 1 (W)

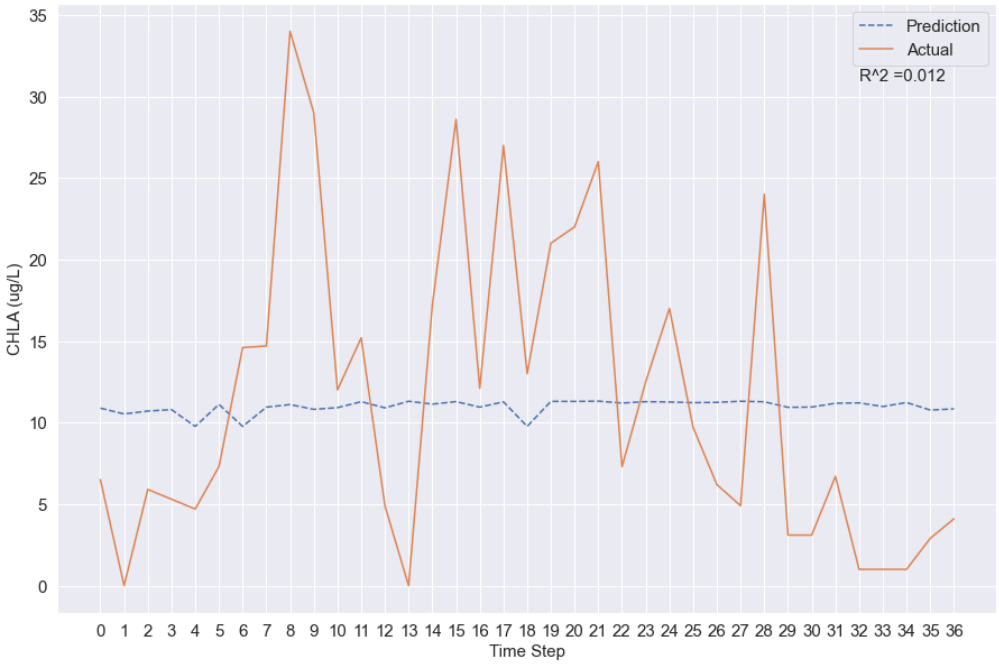


Figure 9: Chlorophyll-a - Combination 2 (H)

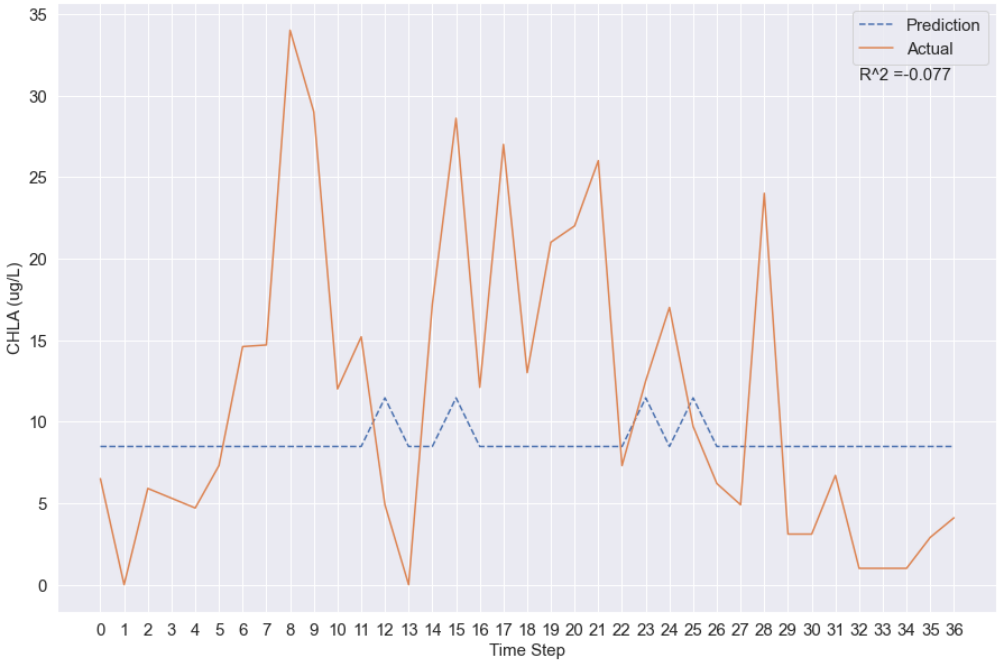


Figure 10: Chlorophyll-a - Combination 3 (R)

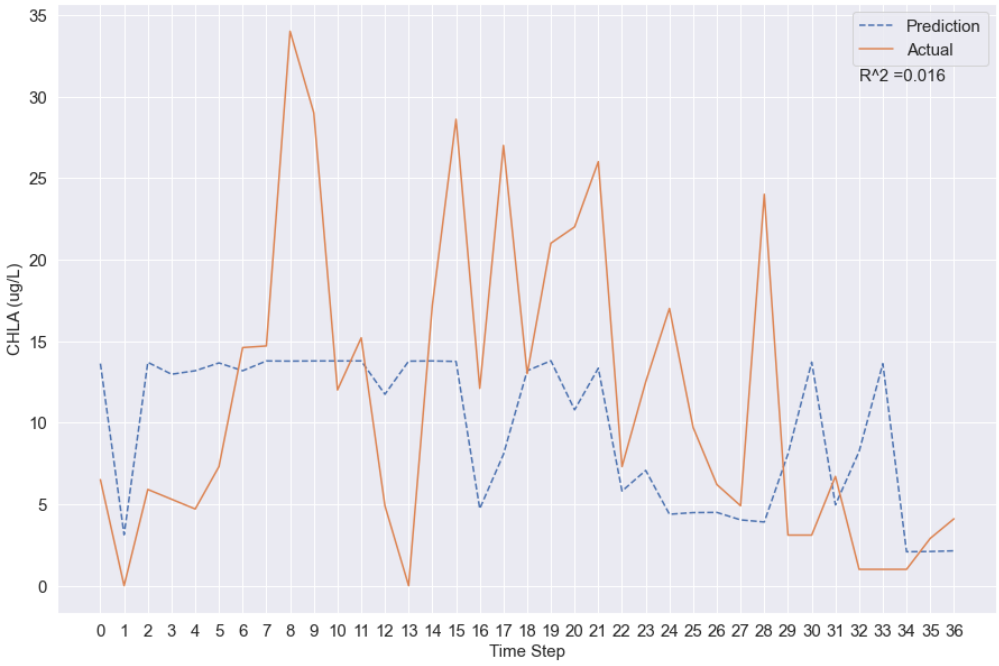


Figure 11: Chlorophyll-a - Combination 4 (WH)

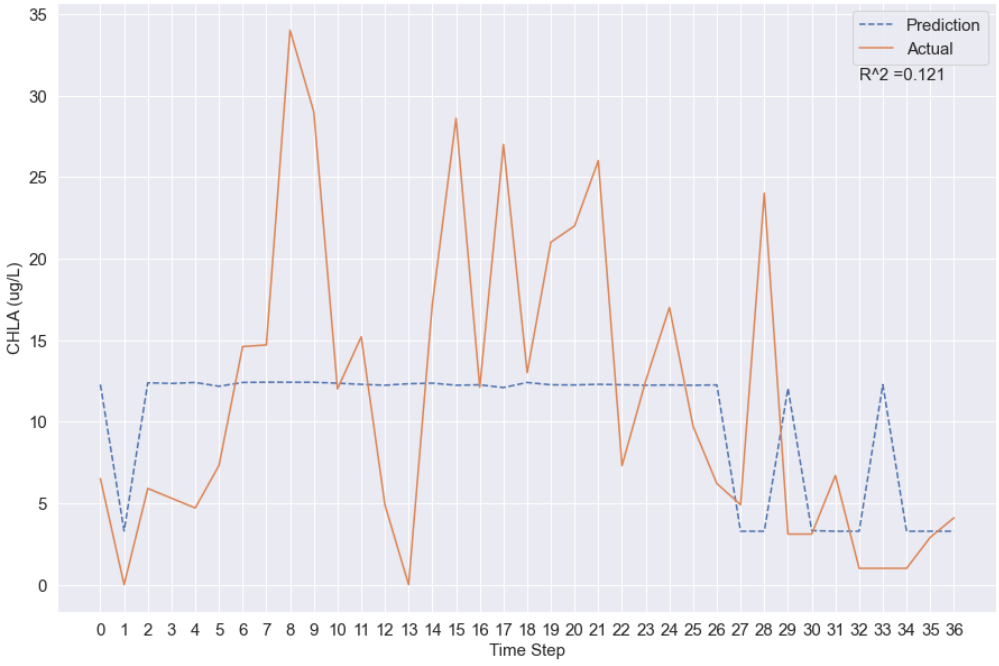


Figure 12: Chlorophyll-a - Combination 5 (WR)

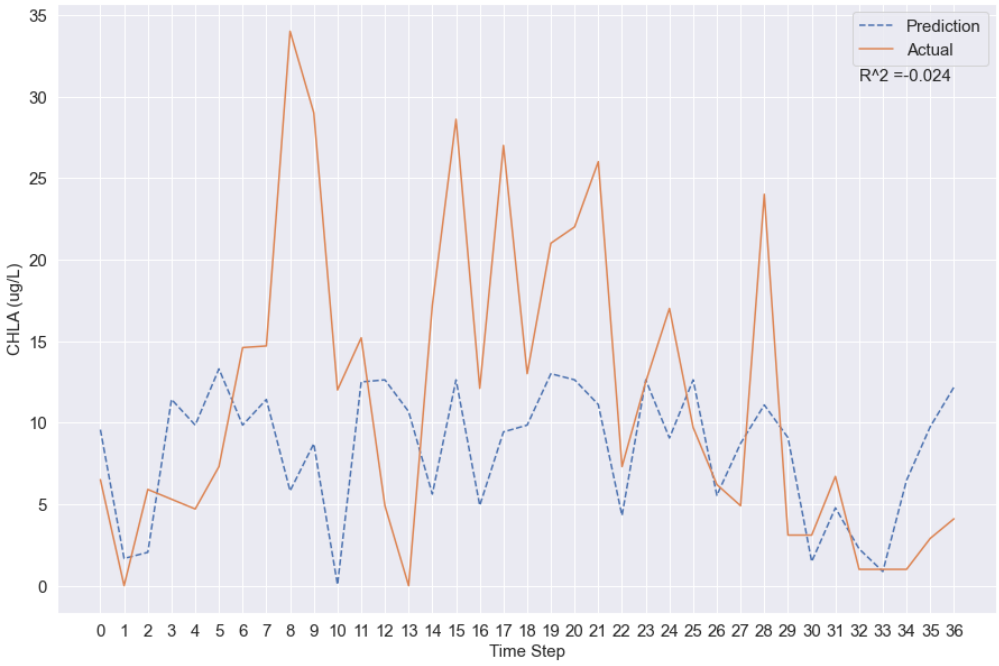


Figure 13: Chlorophyll-a - Combination 6 (HR)

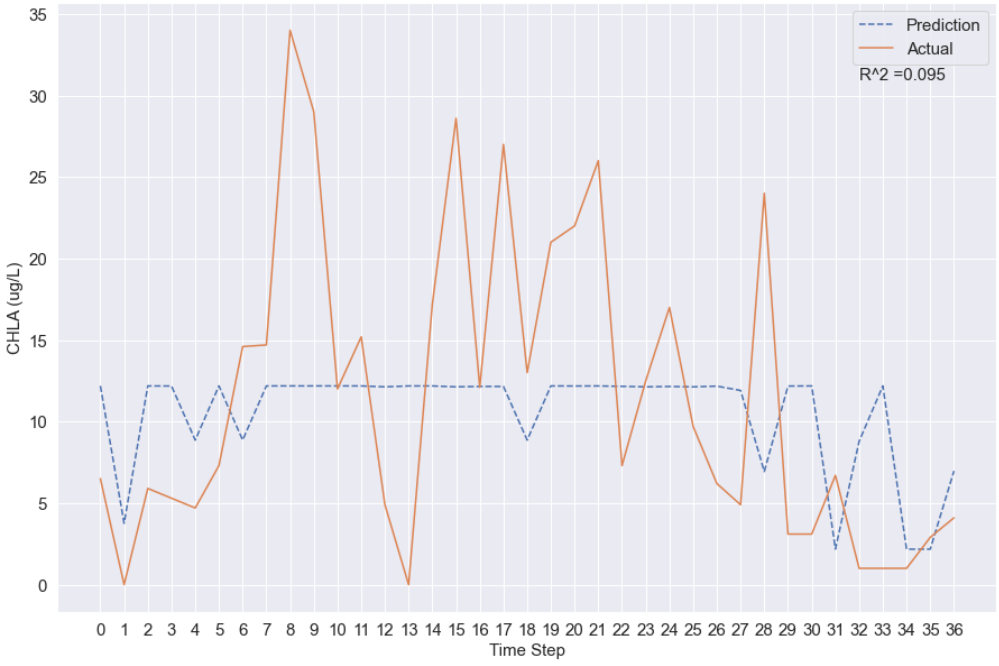


Figure 14: Chlorophyll-a - Combination 7 (WHR)

Table 16: Chlorophyll a Coefficient of Determination

Combination	**R^2**
1 (W)	0.127
2 (H)	0.012
3 (R)	-0.077
4 (WH)	0.016
5 (WR)	0.121
6 (HR)	-0.024
7 (WHR)	0.095

The model that has the highest R^2 value is combination 1, water quality data by itself. The model that has the lowest R^2 value is combination 3, remote sensing data by itself. The models are underfitting and fail to estimate the sudden peaks of the actual values. The CE-QUAL-W2 model results, which are in the appendix, has an R^2 value of 0.23. While the R^2 is still low, the CE-QUAL-W2 model does estimate spikes in the data.

TOTAL NITROGEN

Figure 15 through Figure 21 represent the combinations 1-7 as previously seen in Table 9. The table that follows the figures lists the R^2 value for all the models.

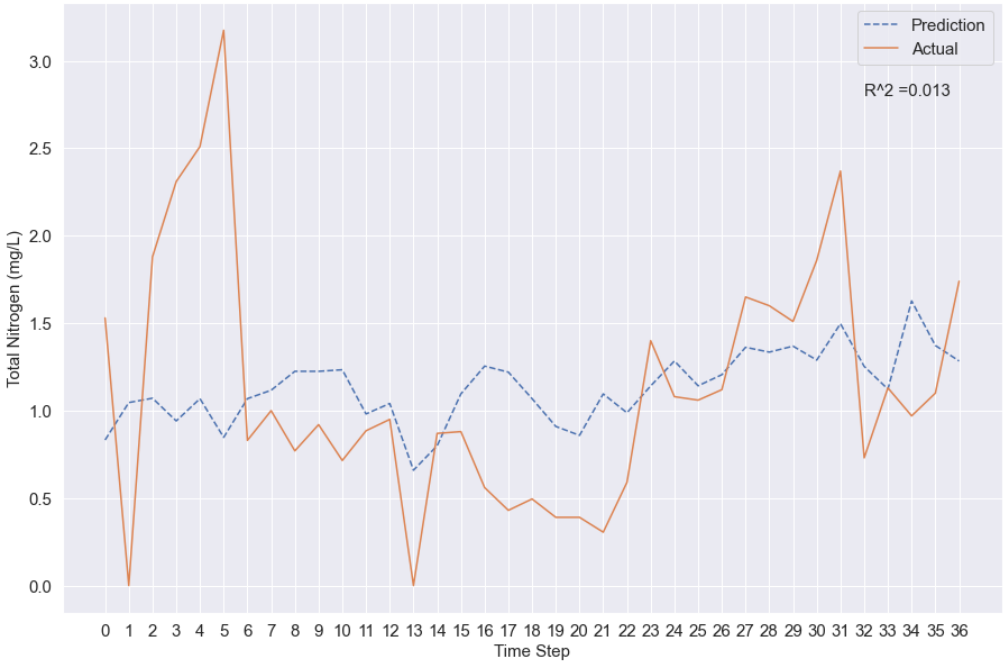


Figure 15: Total Nitrogen - Combination 1 (W)

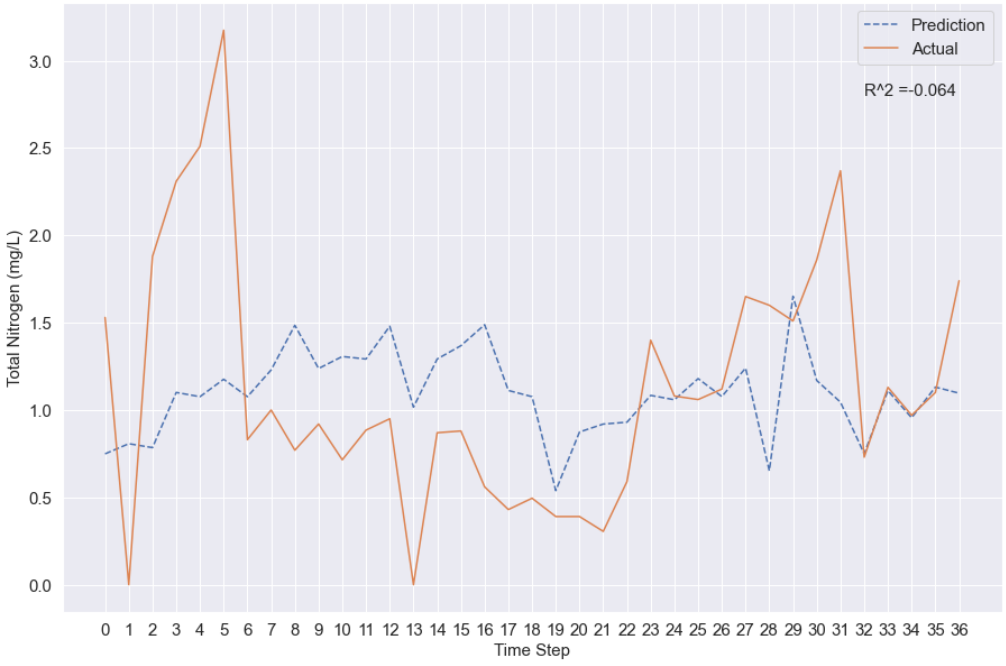


Figure 16: Total Nitrogen - Combination 2 (H)

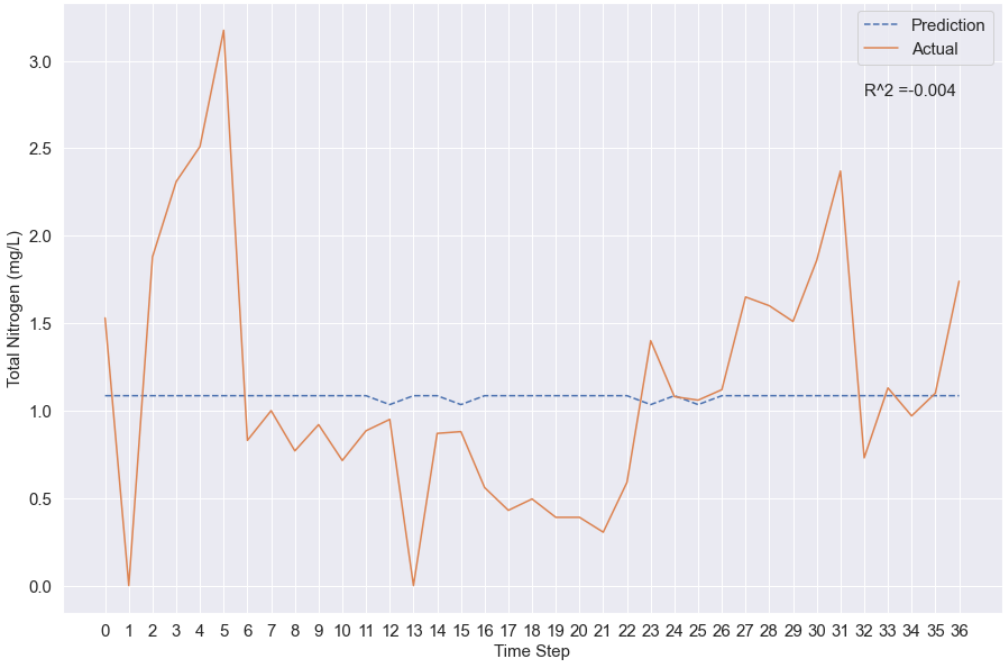


Figure 17: Total Nitrogen - Combination 3 (R)

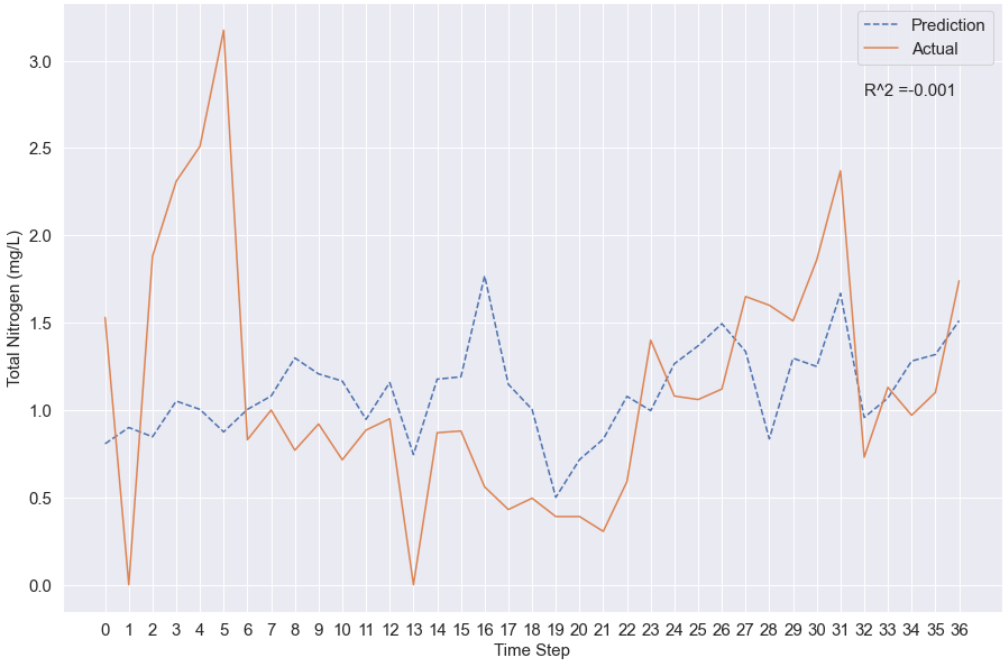


Figure 18: Total Nitrogen - Combination 4 (WH)

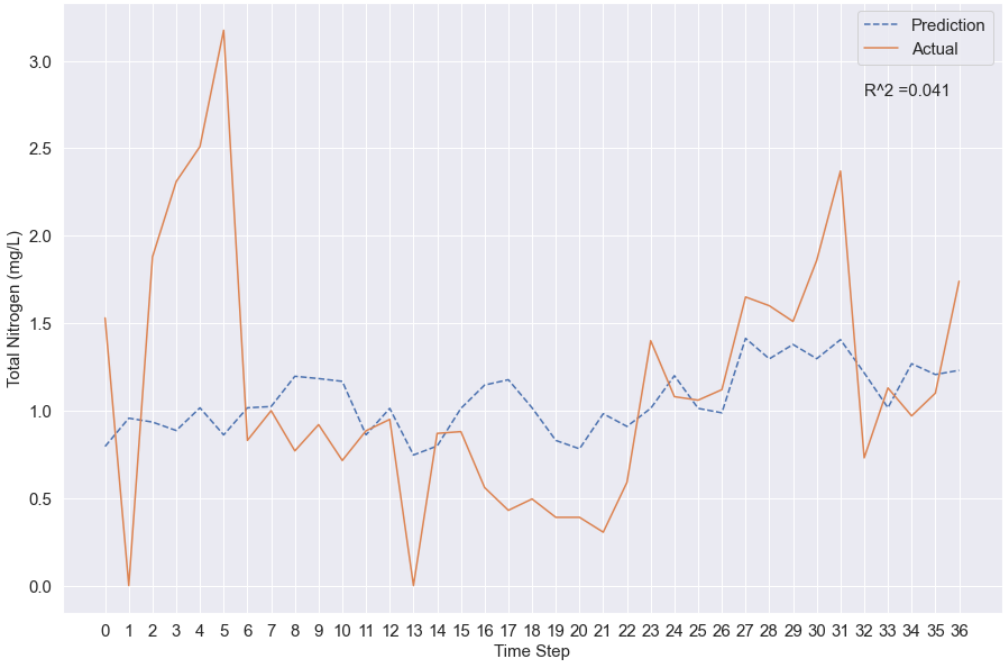


Figure 19: Total Nitrogen - Combination 5 (WR)

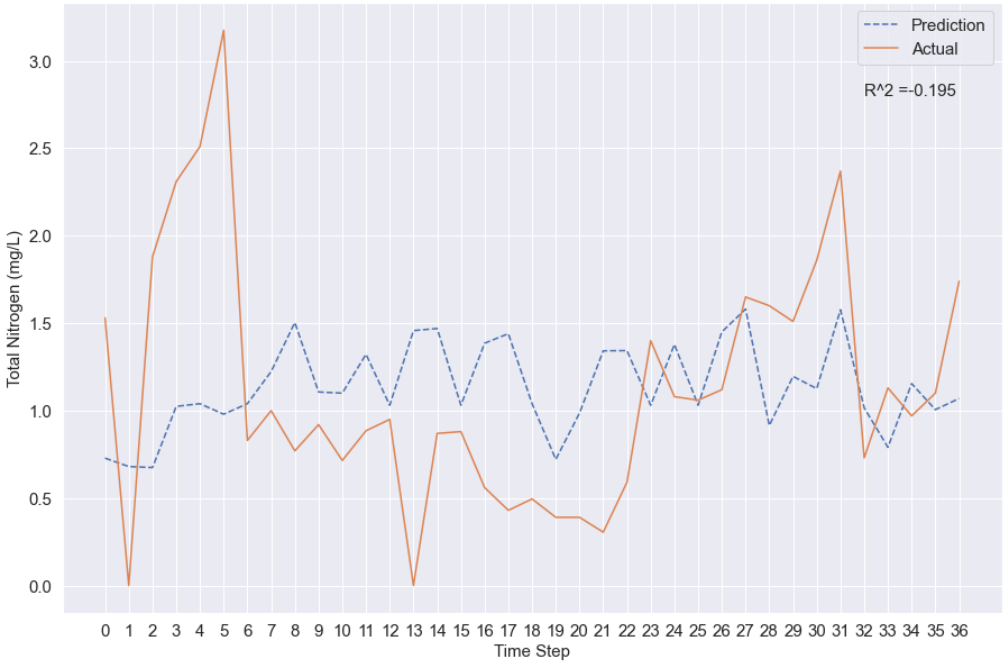


Figure 20: Total Nitrogen - Combination 6 (HR)

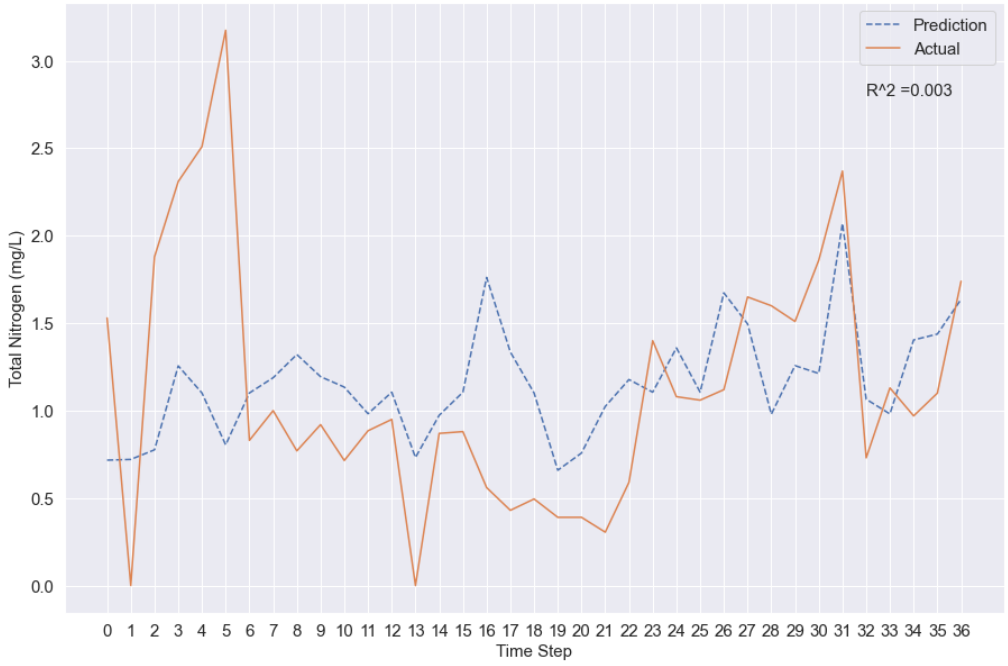


Figure 21: Total Nitrogen - Combination 7 (WHR)

Table 17: Total Nitrogen Coefficient of Determination

Combination	**R^2**
1 (W)	0.013
2 (H)	-0.064
3 (R)	-0.004
4 (WH)	-0.001
5 (WR)	0.041
6 (HR)	-0.195
7 (WHR)	0.003

The model that has the highest R^2 value is combination 5, water quality data and remote sensing data together. The model that has the lowest R^2 value is combination 6, hydrological and meteorological data and remote sensing data. The overall performance for total nitrogen is lower than the models for chlorophyll a. The CE-QUAL-W2 model has an R^2 value of 0.168.

TOTAL SUSPENDED SOLIDS

Figure 22 through Figure 28 represent the combinations 1-7 as previously seen in Table 9. The table that follows the figures lists the R^2 value for all the models.

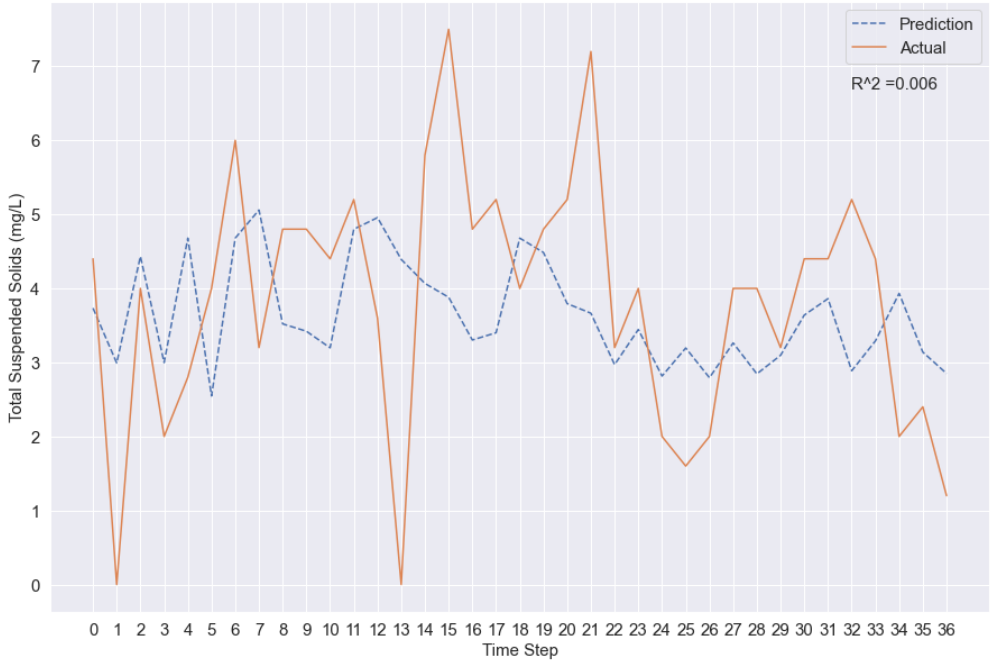


Figure 22: Total Suspended Solids - Combination 1 (W)

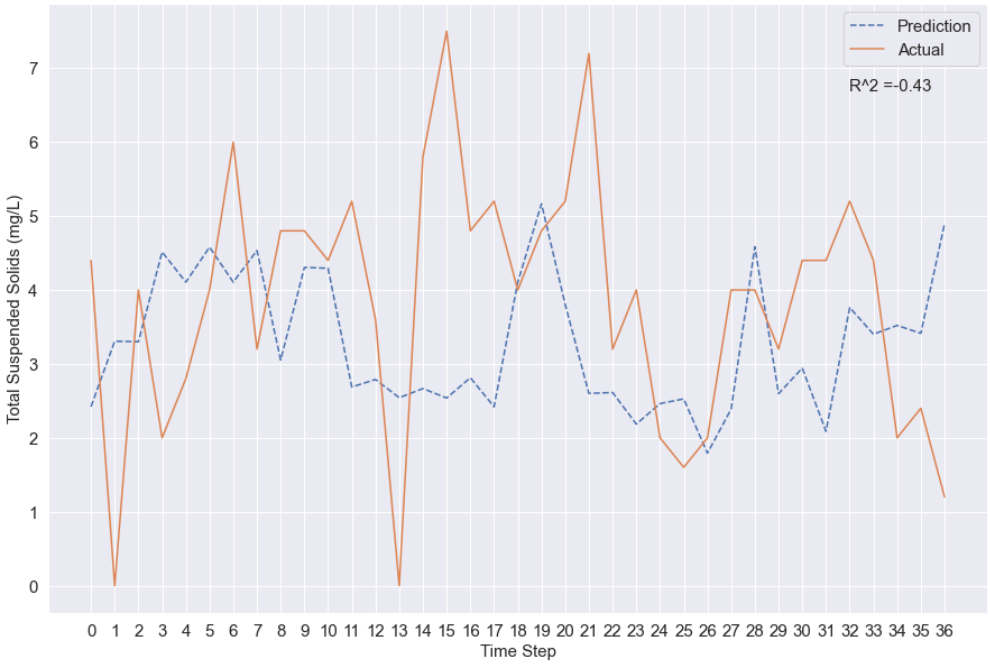


Figure 23: Total Suspended Solids - Combination 2 (H)

Figure 24: Total Suspended Solids - Combination 3 (R)

Figure 25: Total Suspended Solids - Combination 4 (WH)

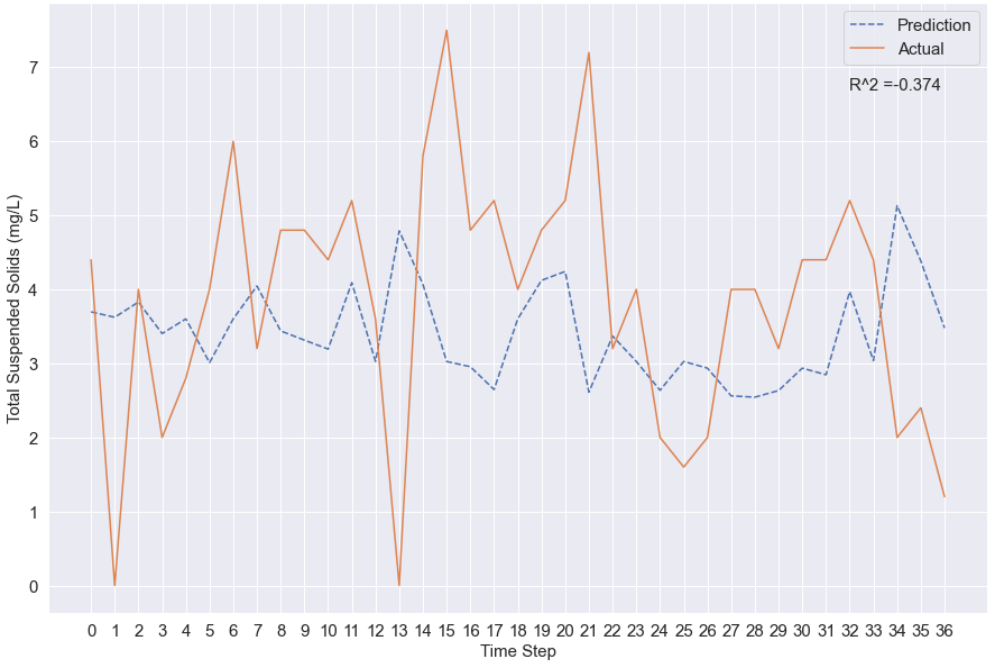


Figure 26: Total Suspended Solids - Combination 5 (WR)

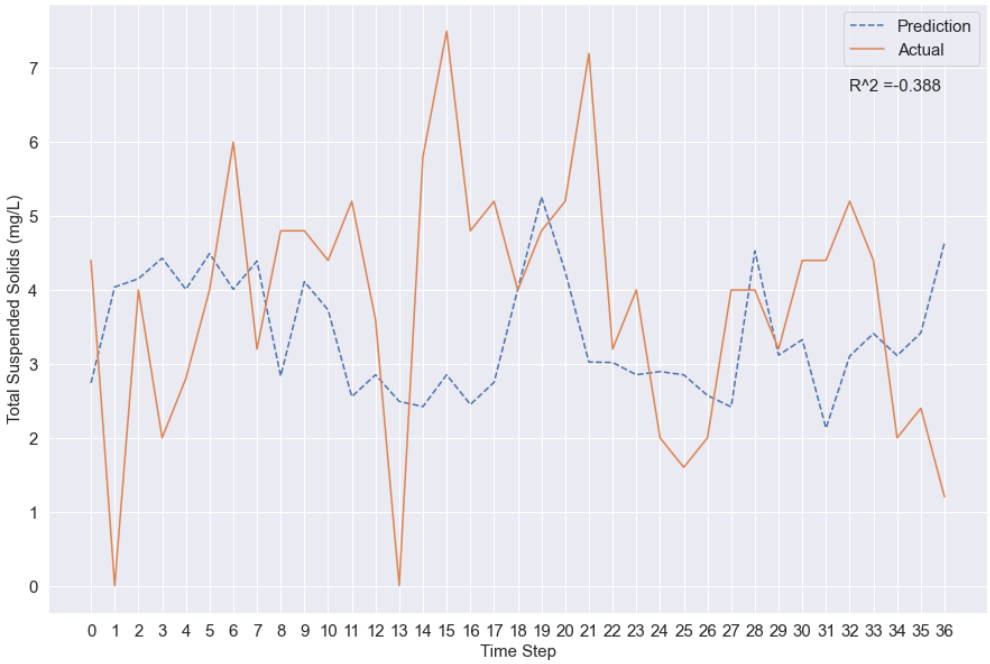


Figure 27: Total Suspended Solids - Combination 6 (HR)

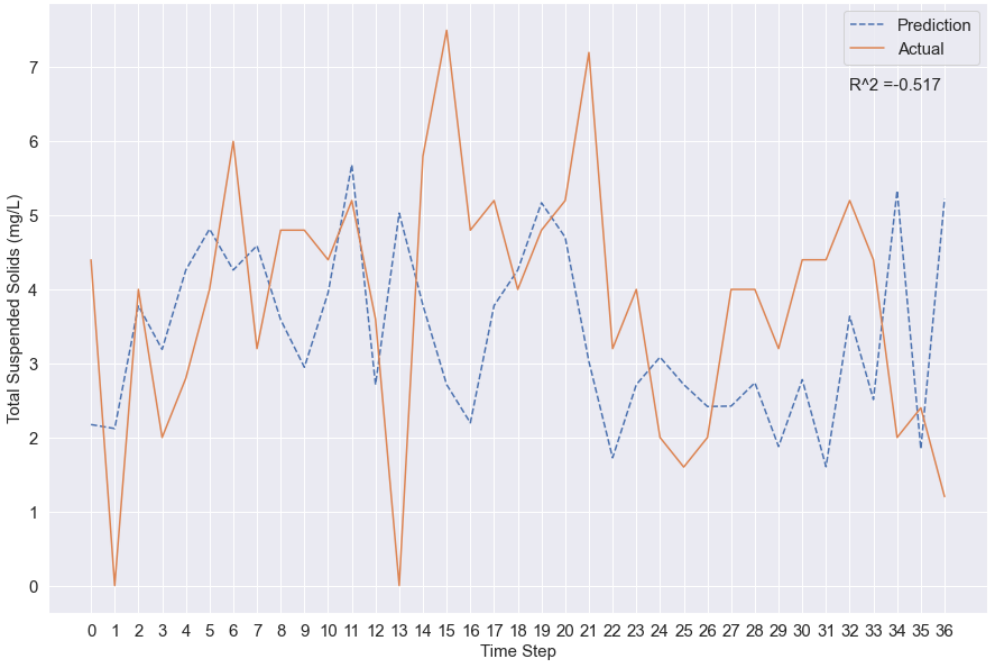


Figure 28: Total Suspended Solids - Combination 7 (WHR)

Table 18: Total Suspended Solids Coefficient of Determination

Combination	**R^2**
1 (W)	0.006

2 (H)	-0.43
3 (R)	-0.156
4 (WH)	-0.299
5 (WR)	-0.374
6 (HR)	-0.388
7 (WHR)	-0.517

The model that has the highest R^2 value is combination 1, water quality data by itself. The model that has the lowest R^2 value is combination 7, all three datasets combined. The overall performance of total suspended solids is lower than chlorophyll a and total nitrogen. The CE-QUAL-W2 model has an R2 value of 0.197.

One thing to note though about the comparison between the neural network and CE-QUAL-W2 is that CE-QUAL-W2 has used the dataset as calibration so it's modelling from data that it has already seen. The neural network is created by learning the weights from the data for the first four years, but is being tested on the fifth year, which it has never seen.

Discussion

All three neural network models performed worse than the CE-QUAL-W2 model. While all three don't do a good job predicting the spikes in the data, chlorophyll-a appears to perform the worst out of the three nutrients. In chlorophyll-a, the remote sensing data by itself has the lowest R^2 value but is improved when coupled with water quality data. Water quality and remote sensing together perform better than water quality and hydrological and meteorological data. In total nitrogen, the model is improved according to the R^2 when remote sensing data is added to water quality data, but not with the hydrological and meteorological data. In total suspended solids, the R^2 value suggests that the remote sensing data does not improve the performance.

There are multiple factors that could be affecting the performance of these models. The factors can be separated into two categories, external factors and internal factors.

EXTERNAL FACTORS

The glaring issue that contributes the most to the performance is the remote sensing data. Aside from the large amount of missing data that had been previously mentioned, there are issues with the temporal resolution, spatial resolution, and the spectral resolution. Since images are taken from Landsat 5 and Landsat 7 every 16 days, the readings do not match up with the dates at the time the water samples are taken. At best, they are within a few days of the sampling date. Landsat 5 and Landsat 7 has a pixel resolution of 30 meters by 30 meters. There is a loss of data when 30 square meters is condensed into one pixel. Especially when the model is predicting nutrients which are magnitudes smaller in comparison. The two satellites can only detect seven bands of spectral data. There could be still be a relationship occurring in the electromagnetic spectrum that isn't being detected by the Landsat satellites. The peak reflectance of chlorophyll-a is at about 700 nm wavelength (Jiao et al., 2006) which is between the detectable range of the bands of the Landsat satellites.

The water quality data itself contributes to the performance. A true value of the concentration of a nutrient is difficult to detect. If the concentration falls below a certain threshold, a sensor detects a large amount of noise so common practice is to set the concentration to the threshold limit regardless of what's being read or simply state that the value is less than the limit. The term used for this practice is censored data. As a result, the dataset will have deceiving repeated values which introduce errors in analysis and models. This is apparent in Table 5 for Organic Phosphorous as 50% of the data has a value of 0.005. The method of collecting water samples is another possible contributor. The current method collects water samples once every week, but not at a set time step. For example, on one week the sample is collected on a Monday, but then on the following week, the sample is collected on a Wednesday, then the week after that,

the sample is taken on a Tuesday. Some weeks are skipped as well due to the reservoir being frozen at the sampling station. (Tate et al., 1999) suggest that not only is frequency important for modeling, but also timing. Samples should be taken before, during, and after storms.

Some papers have attempted to resolve the issue of censored data. (El-Shaarawi and Dolan, 1989) estimate water quality concentrations using maximum likelihood. (Gilliom and Helsel, 1986) found that the best method to minimize errors is the log-probability regression method.

The CE-QUAL-W2 analysis provided to me by OWML is for 2008 to 2012, but the complete data they have collected goes from 1973 to the present. Since this research was to compare the results of the neural network model and the CE-QUAL-W2 model, it was limited to the amount of samples used. Next step in the research would be to remove the limits and feed all the data available into the neural network model.

INTERNAL FACTORS

While it is possible that a LSTM neural network would not work for water quality prediction, the results of other research papers say otherwise. A hybrid CNN/LSTM model successfully predicts dissolved oxygen and chlorophyll-a with data samples every 15 minutes (Barzegar et al., 2020). They used electrical conductivity, oxidation-reduced potential, pH, and water temperature as inputs. Another LSTM model predicts nutrients with a high degree of accuracy with samples every 6 hours and a total of 2188 entries (Xu et al., 2019). The two papers suggest that for a neural network to predict successfully, a large data set and high frequency of sampling are required. In comparison, the dataset for this project could be considered sparse. The data being sparse is what could be contributing to the model performance. Something else that the sparse data affects is the scaling of the data. While robust scaling can be used to reduce the impact

of outliers, the addition of multiple zero values into the dataset reduces the utility of the scaling. And in the case of remote sensing data where most of the dataset is missing, it does not work.

The hyperparameters in a neural network allow for multiple combinations of parameters. As each parameter becomes a choice, the total amount of simulations required become multiplicative. The total amount of simulations requires a longer time to run a search algorithm for the optimal hyperparameters so the choices for this project had to be culled to a reasonable length. Future attempts could attempt a brute force method on a dedicated computer and largely increase the choices and run the search.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

When comparing the R^2 values of a data-driven model and a process-based model, the data-driven model performs less than the process-based model. The performance of the data-driven model could be improved with a high frequency dataset, but the same could be said for a process-based model. The comparison between the two with high frequency data would be for future research. What this research revealed though is that a data-driven model, specifically a long short-term memory recurrent neural network, fails at predicting with low frequency, small dataset compared to the process-based model.

The remote sensing data from Landsat 5 and Landsat 7 at the particular location and for the particular time span is unfortunately not usable. It is inconclusive if remote sensing data could be used to enhance the accuracy of a neural network for nutrient prediction. A suggestion for future research would be to use unmanned aerial vehicles (UAV). UAV's are small enough that they can be carried to the field on the same day sampling occurs. Instead of the data being several days apart, it can be reduced to within hours of when the water sample was taken. Since the UAV is closer to the surface of the water than a satellite, the spatial resolution is higher detailed. The UAV also bypasses any atmospheric interference that occurs as well as cloud coverage. Different sensors can be attached to the UAV for a wider spectral range allowing for further inputs into the neural network model.

Other types of satellites could also be used. Various satellites that are currently in orbit have different resolutions than the resolutions mentioned for Landsat. While Landsat images are freely available to the public, some satellite images have a cost attached to them. The costs for obtaining images from those satellites may be cheaper than the cost-of-entry for UAVs.

Future research would benefit from using an alternate form of assessing the performance of the model rather than using an R^2 score that this research did. The values were low for this research, but had the R^2 been higher, an additional measure would've been utilized such as ANOVA. The negative values of the R^2 can also be misleading for those not familiar with neural networks.

Future research should investigate hybrid models which combine both data-driven models and process-based models. (Mekonnen et al., 2015) fused a data-driven model, an artificial neural network, with a process-based model, Soil and Water Assessment Tool, to predict the runoff generation from prairie landscapes. Their results suggested that the fused model can improve modelling capabilities.

REFERENCES

Ambrose, Robert B. †., et al. "Development of Water Quality Modeling in the United States." *Eng. Res*, vol. 14, no. 4, 2004, pp. 200–10, doi:10.4491/eer.2009.14.4.200.

Barzegar, Rahim, et al. "Short-Term Water Quality Variable Prediction Using a Hybrid CNN-LSTM Deep Learning Model." *Stochastic Environmental Research and Risk Assessment*, vol. 34, 2020, doi:10.1007/s00477-020-01776-2.

Bicknell John C Imhoff John L Kittle, Brian R., et al. *HYDROLOGICAL SIMULATION PROGRAM-FORTRAN USER'S MANUAL FOR RELEASE 11*. 1996.

Brando, Vittorio E., and Arnold G. Dekker. "Satellite Hyperspectral Remote Sensing for Estimating Estuarine and Coastal Water Quality." *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 6 PART I, June 2003, pp. 1378–87, doi:10.1109/TGRS.2003.812907.

Cole, Thomas M., and Scott A. Wells. *PDXScholar CE-QUAL-W2: A Two-Dimensional, Laterally Averaged, Hydrodynamic and Water Quality Model, Version 3.1 Part of the Hydrology Commons, and the Water Resource Management Commons*.

Cox, B. A. "A Review of Currently Available In-Stream Water-Quality Models and Their Applicability for Simulating Dissolved Oxygen in Lowland Rivers." *Science of the Total Environment*, vol. 314–316, Elsevier, 2003, pp. 335–77, doi:10.1016/S0048-9697(03)00063-9.

Diamantopoulou, M. J., et al. "The Use of a Neural Network Technique for the Prediction of Water Quality Parameters of Axios River in Northern Greece." *European Water*, vol. 11, 2005.

Dogan, Emrah, et al. “Application of Artificial Neural Networks to Estimate Wastewater Treatment Plant Inlet Biochemical Oxygen Demand.” *Environmental Progress*, vol. 27, no. 4, Dec. 2008, pp. 439–46, doi:10.1002/ep.10295.

El-Shaarawi, A. H., and D. M. Dolan. “Maximum Likelihood Estimation of Water Quality Concentrations from Censored Data.” *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 46, no. 6, NRC Research Press Ottawa, Canada , June 1989, pp. 1033–39, doi:10.1139/f89-134.

Falconer, Ian R. “Effects on Human Health of Some Toxic Cyanobacteria (Blue-Green Algae) in Reservoirs, Lakes, and Rivers.” *Toxicity Assessment*, vol. 4, no. 2, John Wiley & Sons, Ltd, May 1989, pp. 175–84, doi:10.1002/tox.2540040206.

Fausett, Laurene. *Fundamentals Neruals Networks*. Edited by Don Fowley, First, prentice Hall, 1994.

Friendly, Michael, and David Meyer. *Discrete Data Analysis with R Visualization and Modeling Techniques for Categorical and Count Data*. 2015.

Gilliom, Robert J., and Dennis R. Helsel. “Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques.” *Water Resources Research*, vol. 22, no. 2, John Wiley & Sons, Ltd, Feb. 1986, pp. 135–46, doi:10.1029/WR022i002p00135.

He, Zhibin, et al. “A Comparative Study of Artificial Neural Network, Adaptive Neuro Fuzzy Inference System and Support Vector Machine for Forecasting River Flow in the Semiarid Mountain Region.” *Journal of Hydrology*, vol. 509, Elsevier, Feb. 2014, pp. 379–86, doi:10.1016/j.jhydrol.2013.11.054.

Hinton, Geoffrey E., and Simon Osindero. *A Fast Learning Algorithm for Deep Belief Nets Yee-Whye Teh*.

Hochreiter, Josef. *DIPLOMARBEIT IM FACH INFORMATIK Untersuchungen Zu Dynamischen Neuronalen Netzen*. 1991.

Hochreiter, Sepp, and J. J. Urgen Schmidhuber. "(No Title)." *MEMORY Neural Computation*, vol. 9, no. 8, 1997.

Huber, Wayne C., and Thomas O. Barnwell. *STORM WATER MANAGEMENT MODEL, VERSION 4: USER'S MANUAL*. 1988.

Jiao, H. B., et al. "Estimation of Chlorophyll - A Concentration in Lake Tai, China Using Situ Hyperspectral Data." *International Journal of Remote Sensing*, vol. 27, no. 19, Taylor and Francis Ltd., 2006, pp. 4267–76, doi:10.1080/01431160600702434.

Juahir, Hafizan, et al. "APPLICATION OF ARTIFICIAL NEURAL NETWORK MODELS FOR PREDICTING WATER QUALITY INDEX." *Jurnal Kejuruteraan Awam*, vol. 16, no. 2, 2004.

Kumar, Saurav, et al. *Extending Occoquan Reservoir Water Quality Model For Stakeholder Involvement Part of the Water Resource Management Commons*.

Kuo, Jan Tai, et al. "Using Artificial Neural Network for Reservoir Eutrophication Prediction." *Ecological Modelling*, vol. 200, no. 1–2, Elsevier, Jan. 2007, pp. 171–77, doi:10.1016/j.ecolmodel.2006.06.018.

Liu, Yansui, et al. "Quantification of Shallow Water Quality Parameters by Means of Remote Sensing." *Progress in Physical Geography: Earth and Environment*, vol. 27, no. 1, Sage

PublicationsSage CA: Thousand Oaks, CA, Mar. 2003, pp. 24–43, doi:10.1191/0309133303pp357ra.

Loucks, Daniel P., et al. "Water Quality Modeling and Prediction." *Water Resource Systems Planning and Management*, Springer International Publishing, 2017, pp. 417–67, doi:10.1007/978-3-319-44234-1_10.

Martin, James L., and Steven C. McCutcheon. *Hydrodynamics and Transport for Water Quality Modeling*. Lewis Publishers, 1999.

Mekonnen, Balew A., et al. "Hybrid Modelling Approach to Prairie Hydrology: Fusing Data-Driven and Process-Based Hydrological Models." *Hydrological Sciences Journal*, vol. 60, no. 9, Taylor and Francis Ltd., Sept. 2015, pp. 1473–89, doi:10.1080/02626667.2014.935778.

Miller, Cherie V., et al. "Nutrients in Streams during Baseflow in Selected Environmental Settings of the Potomac River Basin." *Journal of the American Water Resources Association*, vol. 33, no. 6, American Water Resources Assoc, 1997, pp. 1155–71, doi:10.1111/j.1752-1688.1997.tb03543.x.

Neitsch, S. L., et al. *VERSION 2000*.

NVPDC, Northern Virginia Planning District Commission. *Mission Statement for the Occoquan Basin Model Upgrade*. 1994.

Occoquan Reservoir. http://www.pwconserve.org/issues/occoquan/index.html. Accessed 11 Nov. 2019.

Orouji, H., et al. "Modeling of Water Quality Parameters Using Data-Driven Models." *Journal of Environmental Engineering*, vol. 139, no. 7, American Society of Civil Engineers, July 2013,

pp. 947–57, doi:10.1061/(ASCE)EE.1943-7870.0000706.

---. "Modeling of Water Quality Parameters Using Data-Driven Models." *Journal of Environmental Engineering*, vol. 139, no. 7, July 2013, pp. 947–57, doi:10.1061/(ASCE)EE.1943-7870.0000706.

Ritchie, Jerry C., et al. "Remote Sensing Techniques to Assess Water Quality." *Photogrammetric Engineering and Remote Sensing*, vol. 69, no. 6, American Society for Photogrammetry and Remote Sensing, 1 June 2003, pp. 695–704, doi:10.14358/PERS.69.6.695.

Shen, Jian, et al. "A Data-Driven Modeling Approach for Simulating Algal Blooms in the Tidal Freshwater of James River in Response to Riverine Nutrient Loading." *Ecological Modelling*, vol. 398, Elsevier B.V., Apr. 2019, pp. 44–54, doi:10.1016/j.ecolmodel.2019.02.005.

Singh, Kunwar P., et al. "Artificial Neural Network Modeling of the River Water Quality-A Case Study." *Ecological Modelling*, vol. 220, no. 6, Mar. 2009, pp. 888–95, doi:10.1016/j.ecolmodel.2009.01.004.

Spence, Ian, and Stephan Lewandowsky. *ROBUST MULTIDIMENSIONAL SCALING*. no. 3, 1989.

Tate, Kenneth W., et al. "Timing, Frequency of Sampling Affect Accuracy of Water-Quality Monitoring." *California Agriculture*, vol. 53, no. 6, University of California Agriculture and Natural Resources (UC ANR), Nov. 1999, pp. 44–48, doi:10.3733/ca.v053n06p44.

The Occoquan Watershed | OWML | Virginia Tech. http://www.owml.vt.edu/aboutowml/about.html. Accessed 11 Nov. 2019.

Tong, Susanna T. Y., and Wenli Chen. "Modeling the Relationship between Land Use and Surface

Water Quality." *Journal of Environmental Management*, vol. 66, 2002, pp. 377–93, doi:10.1006/jema.2002.0593.

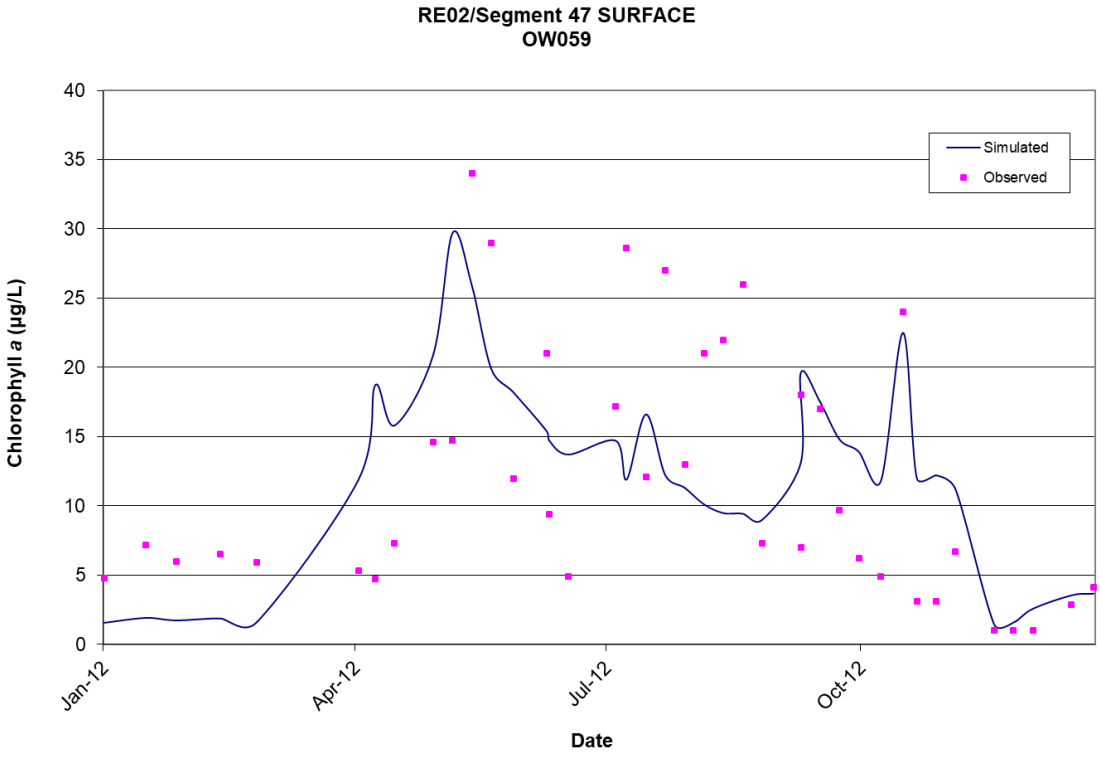
Towler, Erin, et al. "Simulating Ensembles of Source Water Quality Using a K-Nearest Neighbor Resampling Approach." *Environmental Science and Technology*, vol. 43, no. 5, American Chemical Society, Mar. 2009, pp. 1407–11, doi:10.1021/es8021182.

Wen, Ching-Gung, and Chih-Sheng Lee. "A Neural Network Approach to Multiobjective Optimization for Water Quality Management in a River Basin." *Water Resources Research*, vol. 34, no. 3, Mar. 1998, pp. 427–36, doi:10.1029/97WR02943.

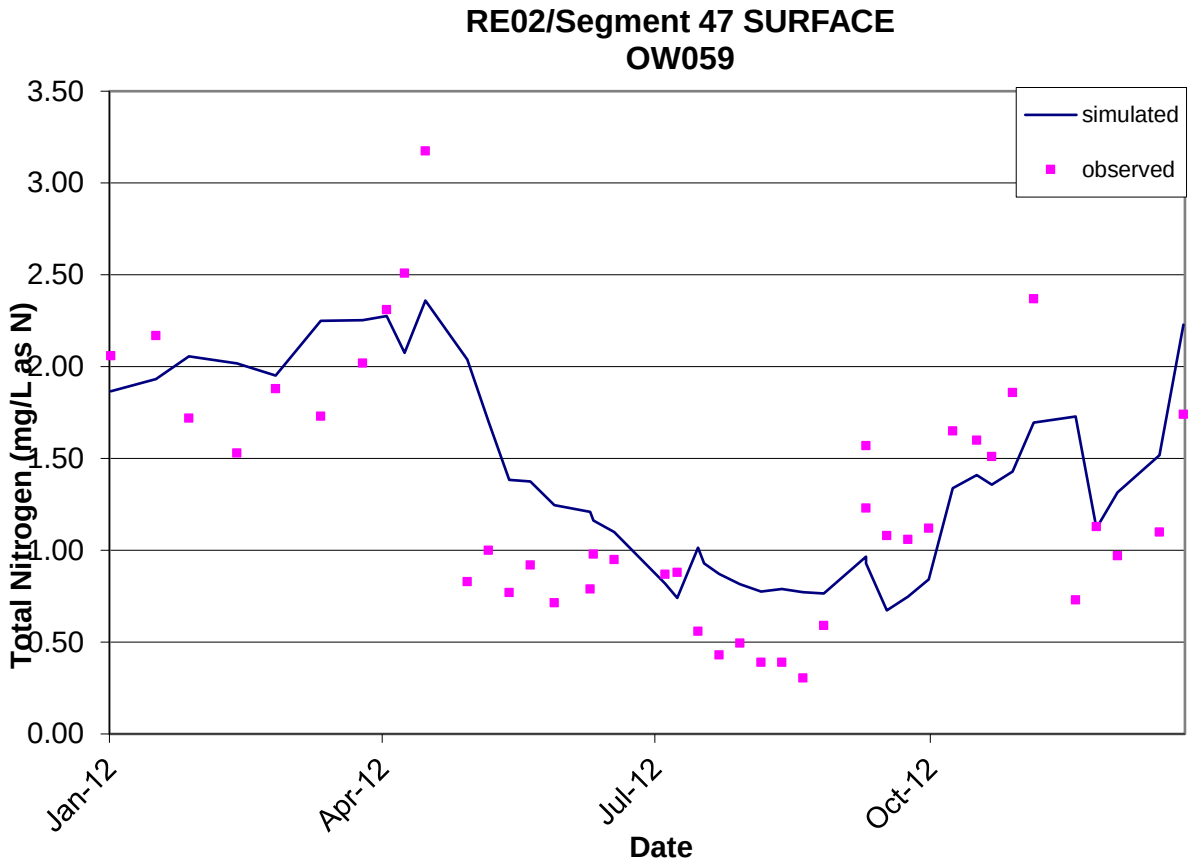
XU, Lin, and Ke LIU. "Analysis of Water Quality Monitoring Data Based on LSTM." *DEStech Transactions on Environment, Energy and Earth Sciences*, no. iccis, 2019, pp. 478–84, doi:10.12783/dteees/iccis2019/31698.

APPENDIX

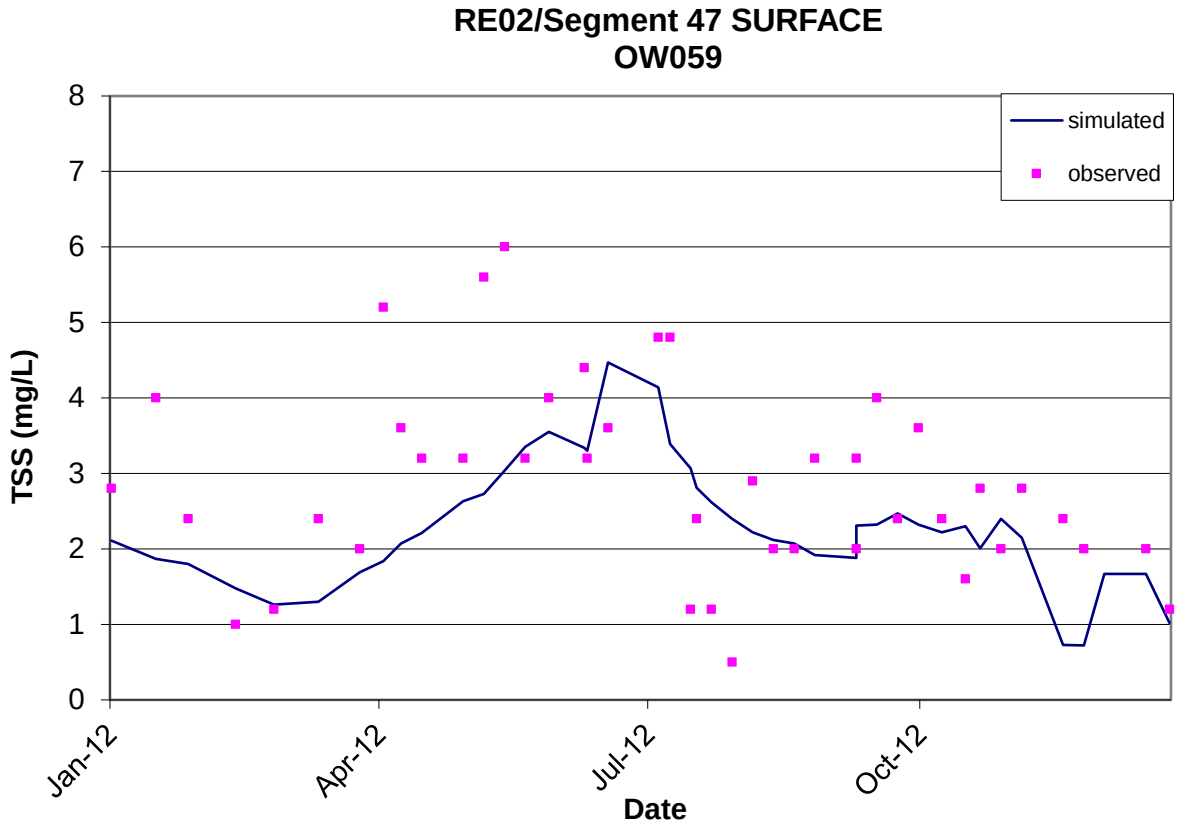
The figure shows a plot of the estimation of chlorophyll a from CE-QUAL-W2. The calculated R^2 value is 0.23.



The figure shows a plot of the estimation of total nitrogen from CE-QUAL-W2. The R2 value is 0.168.



The figure shows a plot of the estimation of total suspended solids from CE-QUAL-W2. The R2 value is 0.197.



VITA

Yohtaro Kobayashi transferred to the University of Texas at El Paso in the year of 2016. While attending school, he joined the American Society of Civil Engineers and the Chi Epsilon Honor Society of Civil Engineers. He also became an undergraduate research assistant for Dr. Saurav Kumar and for the Center of Environmental Engineering. While as an undergraduate, he attended and presented to two conferences, EWRI and iEMSs. He then graduated with a Bachelor of Science in Civil Engineering in the year of 2018.

Afterwards, he continued his education at UTEP as a graduate student in pursuing a Master of Science in Environmental Engineering and will have graduated in August 2020.

Yohtaro Kobayashi is currently looking for a job in water quality or water management.

Contact Information: yckobayashi@miners.utep.edu