Buffalo River Systems Neural Network

Riley Blasiak[1*]

[1]SUNY at Buffalo CIE-500

1. INTRODUCTION

Combined sewer systems (CSSs) are systems in which stormwater and municipal wastewater flow in a joint sewer infrastructure and are treated together before being discharged. While spatially conservative, CSSs are susceptible to combined sewer overflows (CSOs) when system capacity is exceeded and untreated wastewater is discharged into surrounding waterways. Untreated water from CSOs is likely to have high concentrations of nutrients and other contaminants that have adverse effects on the waterway into which they are discharged. CSOs contain contaminants such as nitrogen and phosphorous that can lead to algal blooms, in addition to high concentrations of total suspended solids (TSS) from both urban and industrial runoff and untreated sewage.

This study aims to quantify the impact of CSOs on local waterways, in Buffalo NY, through the development of a Neural Network (NN) model. The NN will predict water quality from independent variables that are consistent across the study area, with the goal of compiling a consistent dataset containing all the parameters of interest for each node to develop a Graphical Neural Network (GNN) for the system. The goal of this study is to utilize historical radar based precipitation data to create an accurate NN model in predicting water quality in Buffalo rivers, with a broader goal of creating a system-wide Graph Neural Network for CSO impact forecasting.

2. BACKGROUND

Machine learning advancements have significantly improved water quality prediction capabilities, specifically through artificial neural networks (ANNs) and graph neural networks (GNNs). ANNs, such as multiple layered models, have demonstrated strong performance in modeling non-linear relationships in river and water distribution systems, achieving high prediction accuracy in water quality parameters including pH, total dissolved solids (TDS), and electrical conductivity (EC) Ubah et al. (2021); Setshedi et al. (2021). These models excel at capturing temporal patterns even for networks with sparse data Li et al. (2024). GNN have additional capabilities by incorporating spatial dependencies, which allow nodes to make predictions based on the values at the previous nodes.

3. DATA ACQUISITION

3.1. Collection

This project utilizes two data sources. The first dataset contains river water quality for waterways within Erie county, collected by Buffalo Niagara Waterkeeper (BNW) and citizen scientists. Data collection started in April 2014 and is still on going for many locations in the Buffalo Waterway system as seen in Figure 1. The organization measures conductivity, dissolved oxygen, total dissolved solids (TDS), temperature, and turbidity along with sample location, date and time. Figure 2 illustrates how sporadic the data collection is, with sampling events happening about once a month during the summer months. Additionally, every site is not sampled during every sampling event which leaves some data gaps that need to be considered.

The second data source is radar data from NOAA. Historical NOAA radar data is available for Buffalo

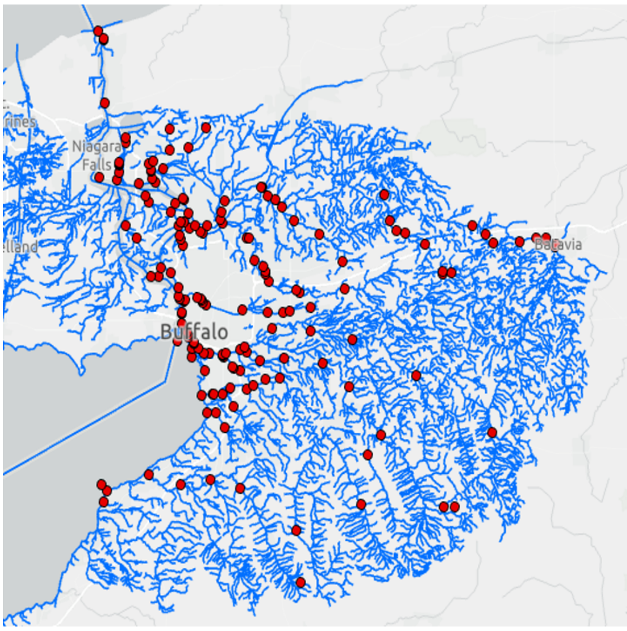


Figure 1. Water quality sample locations provided by Buffalo Niagara Waterkeeper

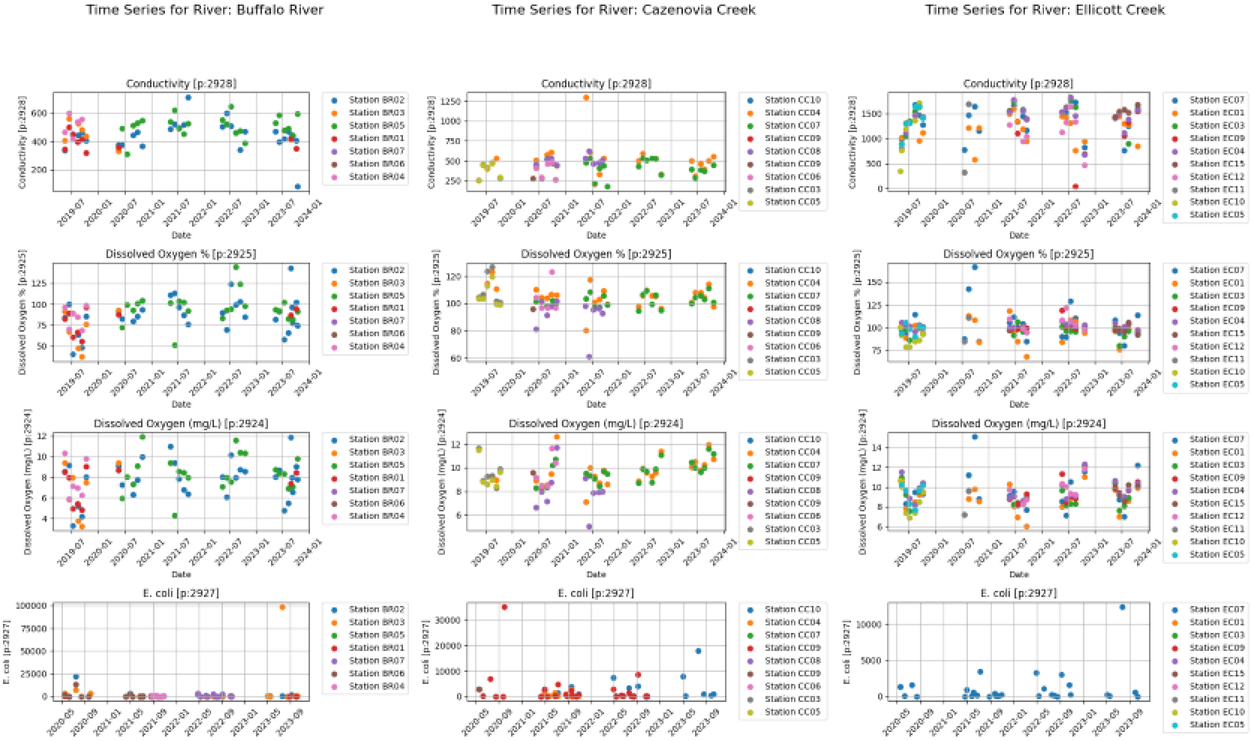


Figure 2. Buffalo Niagara Waterkeeper raw data snapshot

(KBUF) from 1992 to present. The Next Generation Weather Radar (NEXRAD) system contains 160 S-band Doppler weather radars operated by the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the US Air Force. NEXRAD Level III has over 75 products including: Base Reflectivity, One-Hour Precipitation, Storm total precipitation, and Tornadic Vortex Signature. Each product has a code associated with it, e.g. One-Hour Precipitation (DAA,170). This one-hour precipitation product shows precipitation accumulation on a 0.13 x 1-degree grid. A dual-polarization QPE algorithm is used allowing for 256 potential data levels National Centers for Environmental Information (2025).

The historical data was accessed through the NOAA Open Data Dissemination Program, from a Google cloud dataset Center (1992).

3.2. Processing

Metpy was used to read the radar files, since the data was in archive file format (.tar). All the data was cataloged in an excel spreadsheet to keep track of file names and locations, as well as which sampling dates had radar data available. Precipitation data close to the water quality sample time and location are separated from the rest of the data by downloading the radar file from the associated date and time, and pulling out a singular precipitation value for each sample location using a python script. An example radar image for one time point with the location for precipitation extraction is shown in Figure 3.

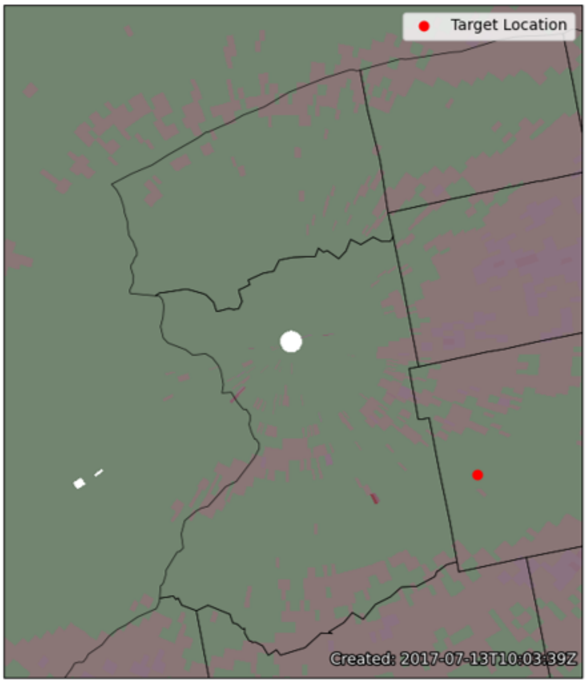


Figure 3. Radar imagery with target point for precipitation evaluation.

The final constructed dataset is called *BNW_sub_with_products.xlsx*. Before building the Neural Network (NN) the data was processed to remove null values and scale the data. Missing values were removed using listwise deletion to ensure consistent input dimensions for the model. To normalize the features and target variable, standardization was applied using scikit-learn's StandardScaler.

4. NEURAL NETWORK CONSTRUCTION METHODS

The cleaned and standardized dataset was divided into training and testing subsets using an 80/20 split via train_test_split. A feedforward fully connected neural network was constructed using PyTorch Paszke et al. (2019). The network architecture consisted of:

1. Input Layer: N neurons, corresponding to the N predictive features.
2. First Hidden Layer: 100 neurons, ReLU activation.
3. Second Hidden Layer: 60 neurons, ReLU activation.
4. Output Layer: 1 neuron, linear activation to output the continuous prediction of the target parameter.

This architecture as shown in Figure 4 was chosen through a trial and error approach, and resulted in a stable training procedure with diminishing losses.

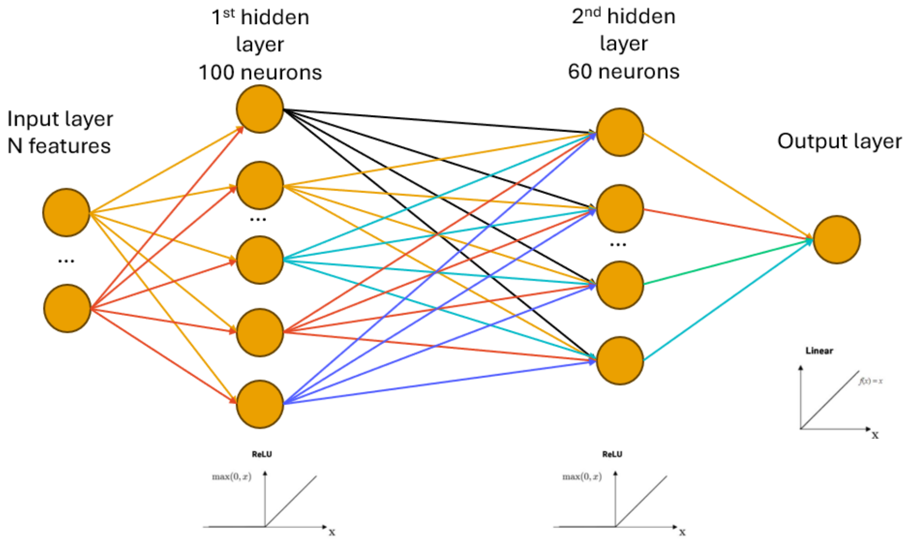


Figure 4. Neural network layout schematic

Column 1	Column 2
1	Precipitation
2	Turbidity
3	Dissolved Oxygen %
4	Conductivity
5	Total Dissolved Solids (TDS)
6	Latitude
7	Longitude
8	Temperature

Table 1. Parameters of interest for Neural Network construction

4.1. Training

The model was trained using the Adam optimizer with a learning rate of 0.001, and the Mean Squared Error (MSE) loss function. A batch size of 64 was selected for training. Training was conducted for 1000 epochs, with model performance evaluated for each epoch.

4.2. Testing and Evaluation

After training, the model was evaluated on the test set. The torch.no_grad was used to ensure that no gradients are computed during testing and to conserve memory usage for training mode that utilizes gradients. The predictions and actual data were scaled back to their original values to check the model performance on unscaled items.

To determine which parameters are best predicted by independent variables, a sensitivity analysis was performed as seen in Table 2. From the analysis it can be seen that conductivity is the most important predictor of TDS. If only the independent variables are to be considered (1,6,7), the most accurate prediction is with the target parameter TDS and yields a root mean squared error of 0.376.

5. RESULTS

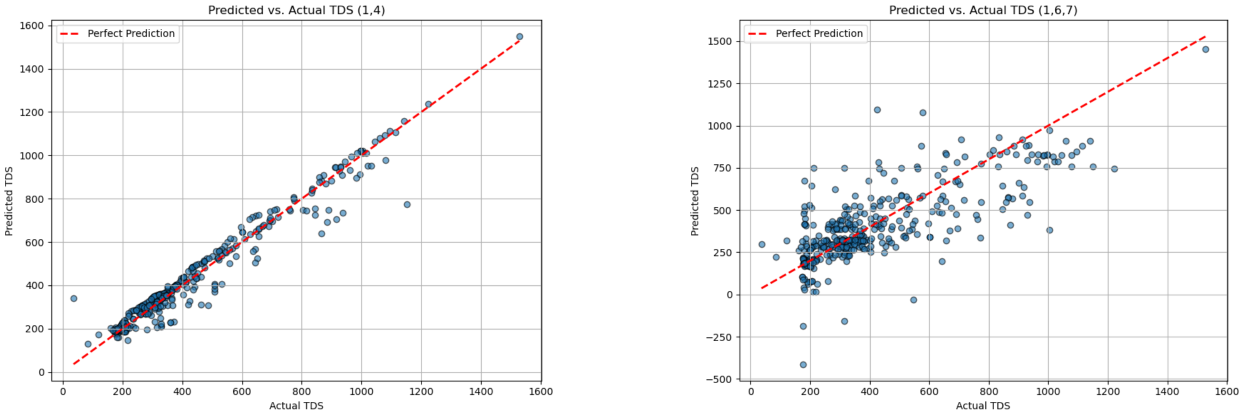
	Total Dissolved Solids	Conductivity	Dissolved Oxygen %	Turbidity
1,2,3,4,6,7	0.0744			
1,2,3,4,6	0.0837			
1,2,3,4,7	0.0661			
1,2,3,7	0.9389			
1,2,4,7	0.069			
1,3,4,7	0.0603			
1,4,7	0.0466			
1,4	0.0398			
4	0.0397			
1,6,7	0.3763	0.4343	0.953	0.8582
1,6,7,8	0.5237			
1,6	0.7195			
1,7	0.8224			

Table 2. Sensitivity analysis on model performance with various parameters as predictors. The value within the table is the test mean squared error for each parameter combination. The rows are features defined in Table 1 and the columns are Targets

From the results in Table 2 a comparison between the actual values of TDS to the predicted values of TDS from two simulations was conducted. The first case considering precipitation and conductivity (1,4) is shown in Figure 5a, the actual and predicted values of TDS are nearly perfect, with a slight under estimation from the model. The root mean squared error of the unscaled values was 52.3, meaning there is an average difference of 52.3 between the predicted and actual values. In the second case, considering only independent variables of precipitation and location (1,6,7), seen in Figure 5b, there are both under and over estimations of the predicted values. The root mean squared error of the unscaled values was 165.6, meaning there is an average difference of 165.6 between the predicted and actual values, which is about three times larger than the first case. If the model consistently had an overestimation, implementing a bias could improve the model, but with both over and under estimations that becomes difficult. There were not many significant outliers to remove to improve model training. If the training water quality data was collected more uniformly through space and time the model would perform better.

6. CONCLUSION

This study demonstrated the feasibility of using a NN to estimate water quality parameters from location and precipitation data. Extracting precipitation data from historical NEXRAD radar archives proved to be a complex and time-intensive process; however, the methods developed during this project may offer valuable guidance for similar efforts in future research. The sensitivity analysis revealed that although conductivity is the strongest predictor of TDS, a combination of other independent predictor values creates a reasonably accurate model. Once this NN accurately predicts water quality values across the network, a graphical NN can be created to model interactions between different parts of the system, as well as provide river water quality forecasts. A model like this could support real-time decision-making and proactive management of CSO events by predicting impacts of water quality based on forecast precipitation. Ultimately, this research



(a) Predicted vs Actual total dissolved solids with precipitation and conductivity as predicting variables.

(b) Predicted vs Actual total dissolved solids with precipitation, latitude, and longitude as predicting variables.

Figure 5. Comparison of predicted vs actual total dissolved solids using different feature sets.

utilized data driven tools to create a model, and this large river system may need a physics-based approach to improve accuracy.

7. FUTURE WORK

The NN model will be improved to better predict water quality parameters before moving into graphical NN development. More independent variables such as land use and other radar products will be used to improve the model. Perhaps a physics informed NN will be utilized to fill data gaps in the network through solving advection diffusion equations.

8. DATA AVAILABILITY

Some or all data, models, or code generated or used during the study are available in a repository online (https://github.com/blasiak2/CIE500).

REFERENCES

Center N. N. W. S. N. R. O., 1992, NOAA Next Generation Radar (NEXRAD) Level 3 Products, https://console.cloud.google.com/marketplace/details/noaa-public/nexrad-l3

Li Z., Liu H., Zhang C., Fu G., 2024, Water Research, 250, 121018

National Centers for Environmental Information 2025, Next Generation Weather Radar (NEXRAD), https://www.ncei.noaa.gov/products/radar/next-generation-weather-radar

Paszke A., et al., 2019, in Advances in Neural Information Processing Systems. pp 8024–8035, https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

Setshedi K. J., Mutingwende N., Ngqwala N. P., 2021, International Journal of Environmental Research and Public Health, 18, 5248

Ubah J., Orakwe L., Ogbu K., Awu J., Ahaneku I., Chukwuma E., 2021, Scientific Reports, 11, 24438