Title: Predicting Dynamics of Daily River Discharge in the Armley Region

Arvin Shahid

Abstract

The Armley Region in Northern Canada contains rivers prone to flooding, which has resulted in serious harm and monetary loss. To develop flood-loss-reduction initiatives, decision-makers need to be able to predict flood events more accurately. Classical statistical offer more accurate and affordable when it comes to modeling the intricate physical processes of floods through mathematical expressions. This work proposes efficient ways that combine a state-of-the-art Deep Learning method, a classical ML algorithm, and a classical statistical method to improve flood prediction for the Red River of the North. The techniques, Decision Tree (DT), Recurrent Neural Network (RNN) and Seasonal Autoregressive Integrated Moving Average (SARIMA). Utilizing hourly level information from three U.S. Geological Survey sites in Pembina, Drayton, and Grand Forks, we assessed the water levels for various periods. Unlike the other locations, Pembina, which is downstream, has a water level gauge but not a flow-gauging station. It is hypothesized that the findings of the floodwater-level forecast demonstrate that the RNN approach performs better than the SARIMAX or DT techniques. The resulting RNN Root Mean Squared Errors (RMSE) values from the study for Pembina, Drayton, and Grand Forks are 0.188, 0.150, and 0.109.

Key Words

Flood Prediction, Time Series, Armley, River, Discharge, Deep Learning, Water Levels

Introduction

For flood warning and water resource management, river and lake water level forecasting is essential. Researchers usually use time-series hydrological-prediction models to estimate future data because water-level data from hydrological stations usually have a time series structure. Utilizing historical data to forecast future water levels (future behavior) might uncover hidden information that is crucial for managing water resources, minimizing the consequences of flooding, and preventing or decreasing disasters (Elganiny 2018).

Canada's Armley Region contains the Red Basin. The yearly and seasonal discharge varies, and multiple variables, including as population expansion, economic development, and climate change, might cause the river's water demands to increase in the future. Variability in precipitation is reflected in patterns in the basin's seasonal and yearly streamflow. Floods in Armley region occur when water levels rise over the tops of riverbanks because of prolonged, heavy precipitation over the same region. This precipitation can take the form of rain, thunderstorms, snowfall mixed with spring melt, or ice jams. The mid-latitude regions of North America are more susceptible to spring-melt floods because of their flat terrain, poor permeability soil, long and harsh winters for snow buildup, and increased springtime temperatures (Kim 2015).

The Armley region experiences spring-melt floods frequently as it flows north. The Red Basin becomes hydrologically active in the spring thaw when the southern portion of the basin melts first; the northern portion of the basin is frequently frozen. In addition to the uniformly flat surface, river activity creates a meandering, slow-moving river that overflows into the northern portion of Canada's Red Basin, causing flooding. When there are large floods, these rivers overflow its weak banks due to surface runoff from snowmelt, flooding the whole valley and wreaking enormous devastation. The Red Basin watershed is seeing a sharp increase in the frequency of floods (Rice 2015).

Past methods to forecast hydrological events, such as storms, runoff or rainfall, shallow streamflow, hydraulic models, and more instances of global circulation, which includes the interaction between atmosphere, water, and floods, alternative approaches primarily rely on physically based models (Li 2015). Another method models the streamflow hydrodynamics using mathematical models. Issues are that physical models may anticipate a wide range of flooding scenarios, but they usually need multiple hydro-geomorphological-monitoring datasets, which makes short-term prediction impossible and requires expensive computers (Fernandez 2016). Furthermore, it highlighted that it may be more difficult to create physically based models as they often require in-depth knowledge and experience in hydrological variables (Borah 2011). Moreover, a variety of studies show that physical models' capacity for short-term prediction is lacking. Using ML techniques for short term prediction is a key area to discover.

Early flood forecasting can assist in giving communities early notice so they can safeguard their homes and property and lessen the effects of flooding. Since its inception, there has been a growing need to refine the identification and characterization of precursors, which impact the hydrological conditions responsible for spring-snowmelt floods, and to refine predictions to lessen the damage caused by Canadian floods.

Methods

Regional Background

The Red Basin is an international, multi-jurisdictional watershed covering 45,000 square miles. It is a unique basin that drains 45,000 mi2 and flows via Pembina River from the south of the region northward into Canada (De Loe 2009). The basin is 315 miles long and reaches a maximum width of around 60 miles. The very sinuous, low-sloping northern Red River canal, which spans 545 river miles serves as the boundary with northern America. Because of snowmelt, precipitation on the snowpack, or heavy rain on saturated soil, the majority of the streamflow happens in the spring and early summer of an average year. In the spring and early summer, flooding is more frequent, and during the rainy season, it is more severe (Biau 2016). In addition, the basin's level topography and the previously mentioned climate frequently result in significant floods in the Red Basin and its tributaries.

Data Collection and Pre-Processing

Along the major tributaries, USGS gauging devices and stations from OTT Hydromet (based in New Jersey), are developed in providing field-based estimates of river flow and river stage for the modeling system's validation. These three dataset’s water levels are gathered from the USGS's hourly gauge-height record. While preprocessing the data, we filled in the missing values using the interpolation approach if the number of consecutive missing values was less than twelve hours. We eliminated the time from our dataset where there were missing values for more than 12 hours. This approach evaluates for hourly water level forecasting using actual Red Basin information. While linear statistical models, like SARIMA, may not be ideal for representing the nonlinear interactions within the time series, they are adequate for representing the linear aspect. In the meanwhile, any nonlinear component (universal approximator) was modeled using non-parametric statistical machine learning models like RNN. Moreover, DT was chosen for the final approach since hydrological applications frequently employ it as an ML technique (Akar 2012). In the section that follows, each of these three chosen approaches are covered. The studied data involve 70% of the data as a training set, 15% as validation, and 15% as a testing set. I looked at years ranging from 2007-2019 for the analysis.

Model Building

For SARIMA we used R core package and 'River Flow' as the main time series variable representing the flow of the Red Basin (Azad 2022). AR_Comp_1, AR_Comp_2, etc. are autoregressive components capturing the linear dependence on past values. The 'Differenced_River_Flow' is the differenced time series variable to achieve stationarity (Integrated component). MA_Error_1, MA_Error_2 is moving average components modeling dependency on residual errors. We also have seasonal autoregressive components capturing seasonal patterns and seasonal moving average components capturing seasonal dependencies.

For RNN we used TensorFlow package, and again use variable 'River Flow' as the main time series variable representing the flow of the Red Basin. LSTM_Input_1, LSTM_Input_2, etc. are input features for the LSTM layer (Le 2019). Furthermore, we create hidden layers of the LSTM network. 'Output_Layer' is the output layer of the RNN and 'Dropout_Rate' is added to prevent overfitting.

For Decision Tree, we use Sci-Kit Learn package and implement relevant features used by the Decision Tree algorithm (Lin 2017). 'Splitting_Criterion' is the criterion for splitting, e.g., Gini impurity or entropy and 'Tree_Depth' is the maximum depth of the decision tree. 'Min_Samples_Leaf' is the minimum number of samples required to constitute a leaf node. We experiment for ensemble methods, with the weight assigned to each decision tree.

Results

The monthly and yearly statistics for these three chosen sites are shown in Figures 1a and 1b. With an average water level of 25.34 feet, Figure 1a clearly shows that April has the largest streamflow at the Pembina station. On April 15, 2009, the highest water level ever recorded at this station was 53.28 feet. This long-term dataset has yielded insightful information that has made it possible to analyze patterns in streamflow and water quality.

It is evident from Figure 1a that April has the greatest flow at the Drayton station, with an average water level of 20.56 feet. April 6, 2009, was the highest water level measured during the survey, at an average of 41.25 feet. Furthermore, Figure 1a shows that, for both the Pembina and Drayton stations, May has the second-highest streamflow, with average water levels of 25.12 feet and 19.01 feet, respectively. With an average water level of 19.35 feet, the Grand Forks station had the largest streamflow in May, as seen in Figure 1a. With an average water level of 50.36 feet, the greatest water level during the period of this research was recorded on April 6, 2009.

A bar chart representing the yearly water-level data for three hydrology stations along the Red River of the North is presented in Figure 1b. In 2019, the highest recorded average yearly water levels at Pembina, Drayton, and Grand Forks stations were 21.54 feet, 16.05 feet, and 18.84 feet, respectively.

Following the implementation of the algorithms on three distinct sample stations, the models were retrieved for additional analysis and tallied in Table 1. For the Pembina, Drayton, and Grand Forks datasets, the table provides information on the average forecast outcomes of all evaluated techniques at five distinct time intervals: six hours, twelve hours, one day, three days, and one week. Reduced RMSE values signify an increased prediction precision of the selected models. It was confirmed that the RNN is the most accurate model by identifying the structures of the SARIMA, DT, and RNN models.

The RMSE values of the RNN in the Pembina station are 76.25% and 79.67% lower, respectively, then those of the DT and SARIMA models. Additionally, employing RNN reduces the RMSE by 24.28% and 32.59%, respectively, between the DT and SARIMA models at Drayton station. Lastly, the Grand Forks station RNN RMSE values are 95.42% lower than the SARIMA model and 81.50% lower than the DT model.

The RNN produces the best results when forecasting one week ahead of time since it can accurately capture the trend of the real data. With an average difference of 0.624 ± 0.18 feet between the tested and forecasted water levels for three stations, the results demonstrate that the RNN outperformed the DT and SARIMA in predicting the water level. For DT and SARIMA, the mean discrepancy between the measured and expected water levels is 0.871 ± 0.52 feet and 1.948 ± 0.67 feet, respectively. For the Pembina station, the other two approaches are not as effective as RNN. The results of utilizing SARIMA, DT, and RNN to anticipate the water level at Drayton station one week ahead are shown in Figure 3. Figure 3c illustrates a similar outcome to the Drayton station scenario in that the peak may be predicted by RNN with a high degree of accuracy one week in advance. In projections one week out, it still accurately represents the pattern of the data, but the inaccuracies are substantial.

The RNN technique forecast was overstated for all water levels in all three sites, as seen in Figures 2, 3, and 4c. Figures 2, 3, and 4a show that SARIMA overestimated the water level at Grand Forks station but underestimated the water level at Pembina and Drayton stations. In conclusion, the DT approach overestimates the water level at Pembina station while underestimating the water level at Grand Forks and Drayton stations (Figures 2, 3, and 4b).

Discussion

Accurately forecasting time series is a difficult but crucial endeavor, particularly when it comes to water levels for flood warning systems. The early flood-warning system depends heavily on the water-level projections from the Red Basin flow-gauging stations, particularly for downstream sites like Pembina in our research that lack any discharge data. In this study, we have studied three methods: deep learning (RNN), classical learning (DT), and classical statistics (SARIMA). The RNN approach produced superior outcomes.

In contrast, a water-stage time series often exhibits both linear and nonlinear correlation features. The RMSE values for models fit to the series obtained at Pembina, Drayton, and Grand Forks are 0.188, 0.150, and 0.109, respectively, for one-week-ahead prediction, as shown in Table 1. These outcomes show how the Deep Learning algorithm is a dependable option for flood prediction due to its great precision. The Pembina, Drayton, and Grand Forks stations' experimental findings demonstrate that the RNN model performs better across all prediction times. At Drayton Station, the RMSE reductions between the DT and SARIMA models are 24.28% and 32.59%, respectively when using RNN. The RMSE values for RNN at the Grand Forks station are 95.42% lower than the SARIMA model and 81.50% lower than the DT model.

The research on short-form time series analysis for flood prediction in the Red Basin is crucial for enhancing public safety and resource management. Armley’s Rivers have unique hydrological characteristics, combined with its transboundary nature, necessitate precise and timely flood predictions to mitigate potential risks to communities and infrastructure. In response to the semi-arid climate and susceptibility to significant spring and early summer flooding in the Armley region, a focused approach to time series analysis is imperative.

Short-form time series, capturing temporal dynamics over a concise period, offer a pragmatic solution for flood prediction, particularly considering the rapid changes in environmental conditions leading to flood events, such as snowmelt, precipitation, and variations in soil saturation. Predicting future flood events is of paramount importance as it enables initiative-taking measures, including early warnings, evacuation planning, and resource allocation, thereby minimizing potential impacts on human settlements and agricultural areas.

Despite the recognized significance of accurate flood predictions, a noticeable dearth of comprehensive research exists on river predictions within the context of machine learning methodologies, particularly in the red basin. Traditional hydrological models, while valuable, struggle to capture the intricacies of dynamic and rapidly changing river systems. The incorporation of machine learning approaches offers a novel avenue for addressing this research gap, providing an opportunity to unlock a deeper understanding of the river's hydrological processes.

This study contributes to filling the existing void by exploring and implementing advanced machine learning models tailored for short-form time series analysis. The comparative lack of existing research underscores the novelty and timeliness of this investigation. Through the development and validation of predictive models, this research aims to bridge the gap between traditional hydrological methods and contemporary machine learning techniques, fostering a more holistic and accurate approach to flood prediction in the Armley Region.

Acknowledgements

I extend my sincere gratitude to the individuals and institutions who have played a pivotal role in the completion of this research project. I would like to acknowledge Professor Albert Ruhi and Instructor Kyle Leathers for valuable guidance, unwavering support, and scholarly insights.

Literature Cited

Akar, Ö.; Güngör, O. Classification of multispectral images using Random Forest algorithm. J. Geodesy Geoinf. 2012, 1, 105–112.

Arnold, J.G.; Srinivasan, R.; Muttiah, R.S.; Williams, J.R. Large area hydrologic modeling and assessment part I: Model development. JAWRA J. Am. Water Resour. Assoc. 1998, 34, 73–89.

Azad, A.S.; Sokkalingam, R.; Daud, H.; Adhikary, S.K.; Khurshid, H.; Mazlan, S.N.A.; Rabbani, M.B.A. Water Level Prediction through Hybrid SARIMA and ANN Models Based on Time Series Analysis: Red Hills Reservoir Case Study. Sustainability 2022, 14, 1843.

Biau, G.; Scornet, E. A random forest guided tour. TEST 2016, 25, 197–227.

Borah, D.K. Hydrologic procedures of storm event watershed models: A comprehensive review and comparison. Hydrol. Process. 2011, 25, 3472–3489.

Bui, D.T.; Pradhan, B.; Nampak, H.; Bui, Q.-T.; Tran, Q.-A.; Nguyen, Q.-P. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. J. Hydrol. 2016, 540, 317–330.

De Loë, R. Sharing the Waters of the Red River Basin: A Review of Options for Transboundary Water Governance; Prepared for International Red River Board, International Joint Commission; Rob de Loë Consulting Services: Guelph, ON, Canada, 2009.

Elganiny, M.A.; Eldwer, A.E. Enhancing the Forecasting of Monthly Streamflow in the Main Key Stations of the River Nile Basin. Water Resour. 2018, 45, 660–671.

Feldman, A. Hydrologic Modeling System HEC-HMS Technical Reference Manual: US Army Corps of Engineers; Hydrologic Engineering Center: Davis, CA, USA, 2000.

Fernández, C.; Vega, J.A.; Fonturbel, T.; Jiménez, E. Streamflow drought time series forecasting: A case study in a small watershed in North West Spain. Stoch. Hydrol. Hydraul. 2009, 23, 1063–1070.

Fernández-Pato, J.; Caviedes-Voullième, D.; García-Navarro, P. Rainfall/runoff simulation with 2D full shallow water equations: Sensitivity analysis and calibration of infiltration parameters. J. Hydrol. 2016, 536, 496–513

Kim, B.; Sanders, B.F.; Famiglietti, J.S.; Guinot, V. Urban flood modeling with porous shallow-water equations: A case study of model errors in the presence of anisotropic porosity. J. Hydrol. 2015, 523, 680–692.

Le, X.-H.; Ho, H.V.; Lee, G. River streamflow prediction using a deep neural network: A case study on the Red River, Vietnam. Korean J. Agric. Sci. 2019, 46, 843–856.

Li, L.; Simonovic, S.P. System dynamics model for predicting floods from snowmelt in North American prairie watersheds. Hydrol. Process. 2002, 16, 2645–2666.

Lim, Y.H.; Voeller, D.L. Regional flood estimations in Red River using L-moment-based index-flood and bulletin 17B procedures. J. Hydrol. Eng. 2009, 14, 1002–1016.

Lin, L.; Wang, F.; Xie, X.; Zhong, S. Random forests-based extreme learning machine ensemble for multi-regime time series prediction. Expert Syst. Appl. 2017, 83, 164–176.

Tables and Figures

Table 1. The table displays Root Mean Square Error (RMSE) values for diverse forecast horizons and prediction models at three distinct locations: Pembina, Drayton, and Grand Forks. The models evaluated include SARIMA, DT, and RNN. The table encompasses RMSE values for forecast horizons spanning 6 hours, 12 hours, 1 day, 3 days, and 1 week.

Table 1: RMSE for 3 River Time Series

		Forecast Horizon	SARIMA_RMSE	DT_RMSE	RNN_RMSE
Pembina	0	6 h	0.109	0.100	0.024
	1	12 h	0.205	0.161	0.030
	2	1 Day	0.506	0.268	0.040
	3	3 Days	1.862	0.863	0.074
	4	1 Week	2.267	2.285	0.188
Drayton	0	6 h	0.042	0.037	0.029
	1	12 h	0.073	0.097	0.034
	2	1 Day	0.153	0.182	0.043
	3	3 Days	0.532	0.704	0.066
	4	1 Week	1.488	1.816	0.150
Grand Forks	0	6 h	0.611	0.136	0.021
	1	12 h	0.654	0.244	0.029
	2	1 Day	0.757	1.056	0.050
	3	3 Days	1.196	1.637	0.085
	4	1 Week	2.030	2.670	0.109

Figure 1a. Three side-by-side bar charts, each representing the monthly water levels for various locations: Pembina, Drayton, and Grand Forks. The x-axis of each chart corresponds to the months from January to December, while the y-axis represents the water levels in feet.

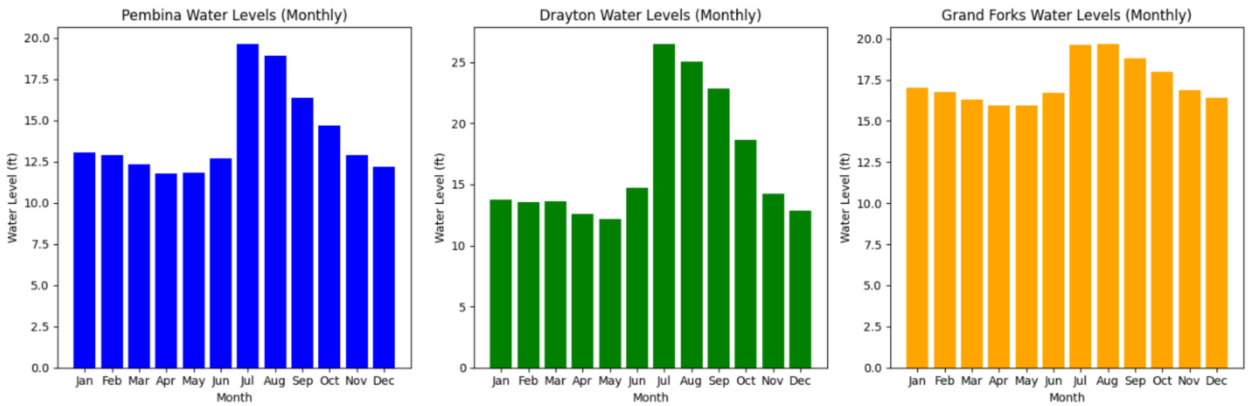


Figure 1b. Three side-by-side bar charts, each representing the yearly water levels from 2007-2019 for different locations: Pembina, Drayton, and Grand Forks. The x-axis of each chart corresponds to the Years, while the y-axis represents the water levels in feet.

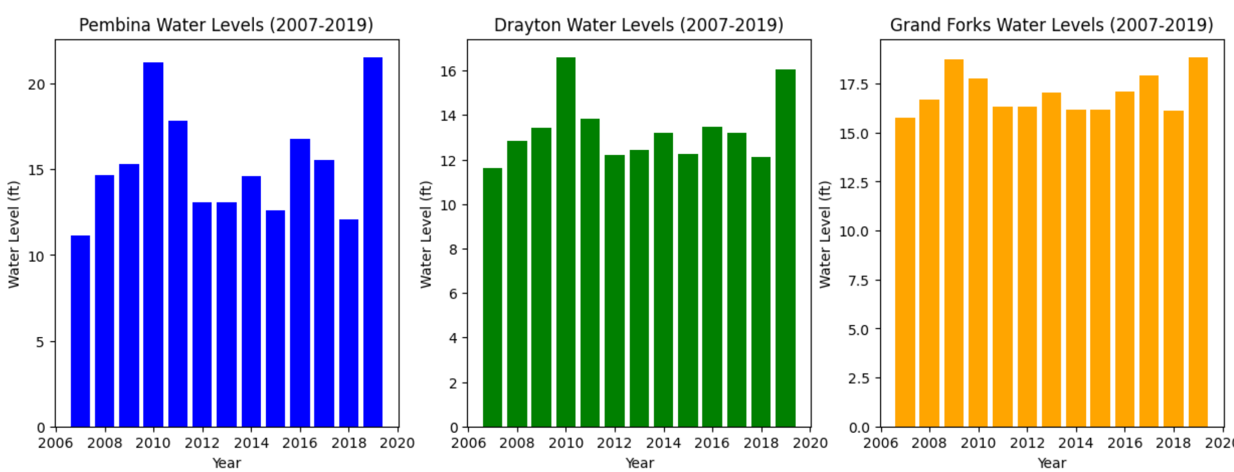


Figure 2a.

Figure 2 a: Pembina Water Level Prediction Using SARIMA

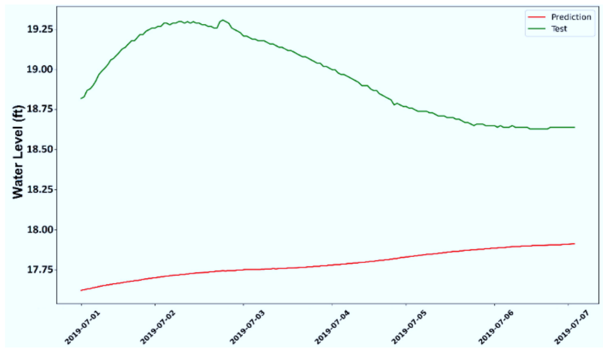


Figure 2b.

Figure 2 b: Pembina Water Level Prediction Using DT

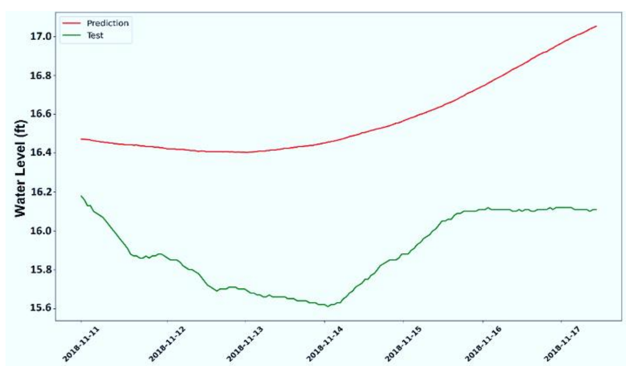
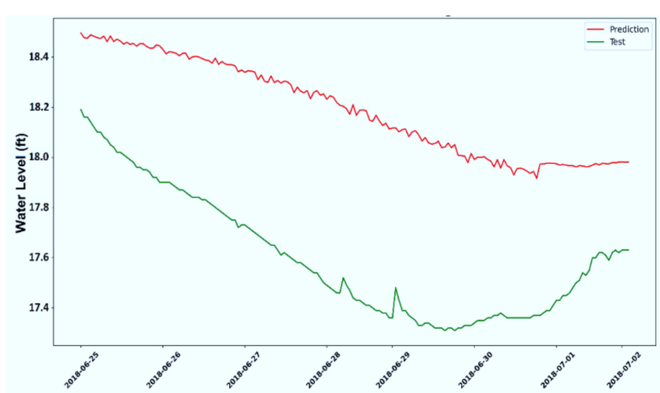


Figure 2C.

Figure 2 c: Pembina Water Level Prediction Using RNN



Pembina River predictions with the three models

Figure 3a.

Figure 3 a: Dayton Water Level Prediction Using SARIMA

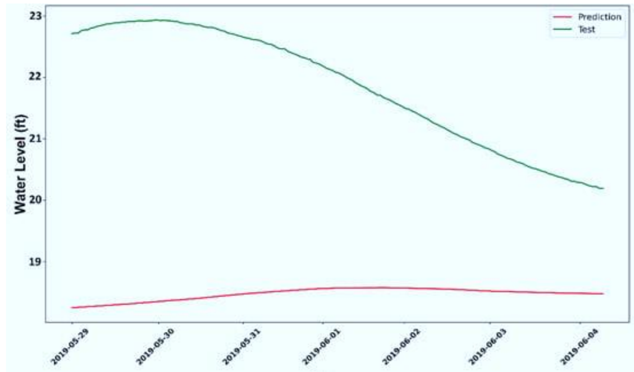


Figure 3b.

Figure 3 b: Dayton Water Level Prediction Using DT

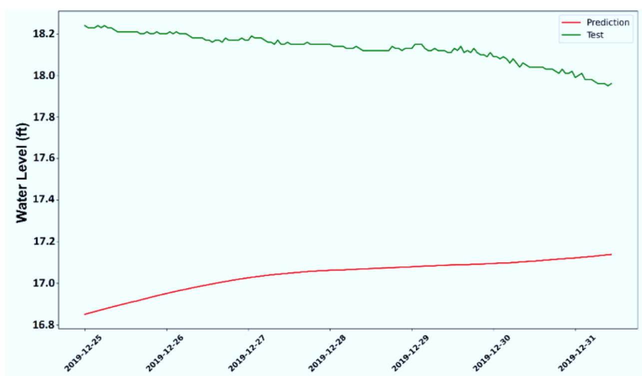
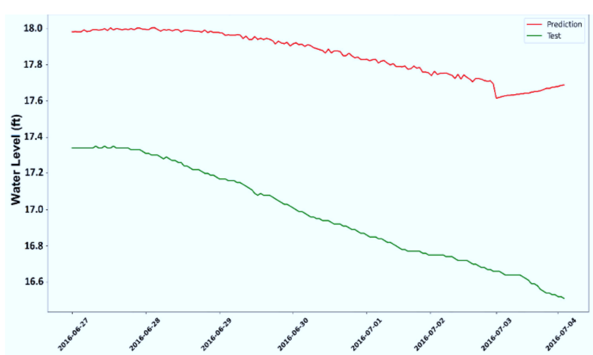


Figure 3c.

Figure 3 c: Dayton Water Level Prediction Using RNN



Dayton River predictions with the three models

Figure 4a.

Figure 4 a: Grand Forks Water Level Prediction Using SARIMA

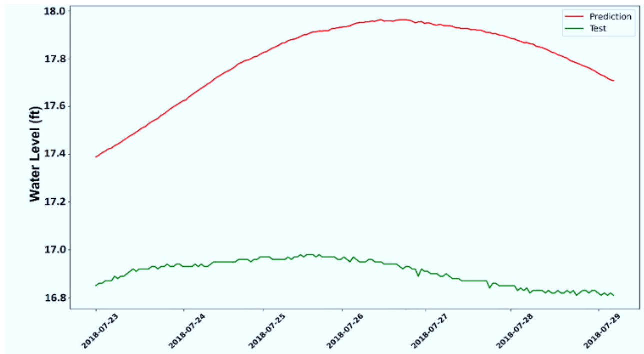


Figure 4b.

Figure 4 b: Grand Forks Water Level Prediction Using DT

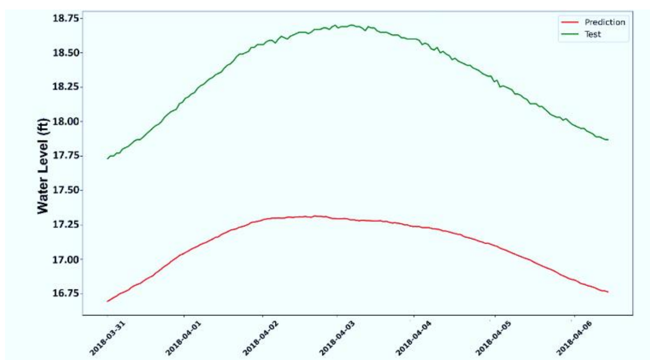
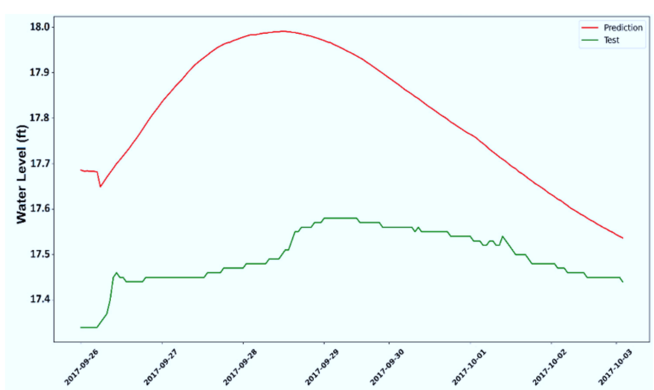


Figure 4c.

Figure 4 c: Grand Forks Water Level Prediction Using RNN



Grand Forks River predictions with the three models