Analysis, Prediction and Comparison Algorithms For Water Quality Variables

Reem Elmahdi[1]; Willie Brink[1]; Josefine Wilms[2]
[1]Department of Mathematical Sciences, Stellenbosch University, [2]CSIR

Introduction

Water is an important resource for several activities. Although water is found in many forms, this study will focus on surface water. Water contains many measurable variables**[1]:** pH, specific conductance, temperature, etc. These variables are affected by pollutants which can result in decreased quality, in turn determining the water's usefulness. Hence, it is important to keep track of the values of these variables.

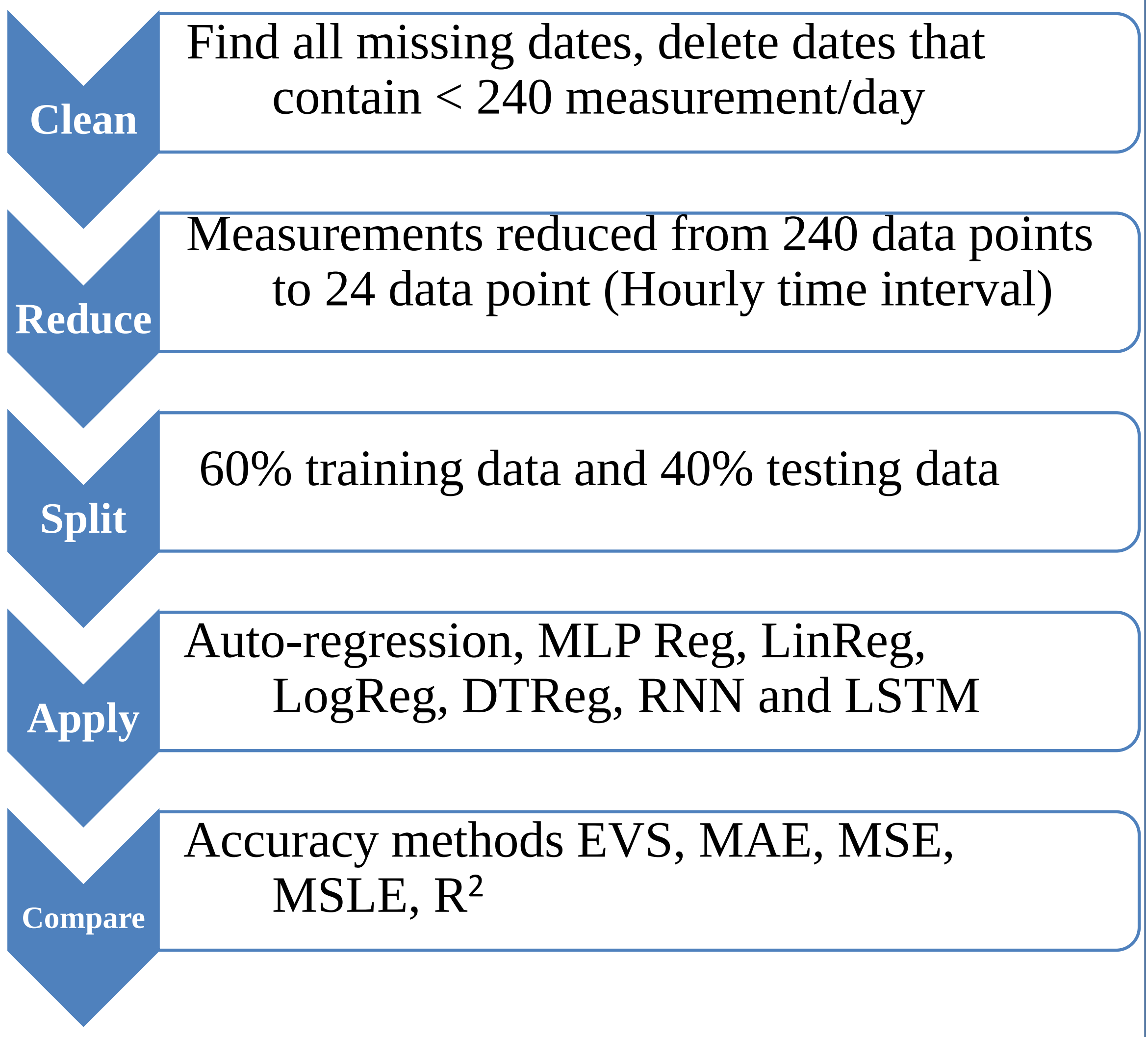
Measuring all the variables can be labour intensive, costly and time consuming. However, machine learning can potentially be used to predict the future values of the variables from recorded historical data.

Objectives

This study aims to analyse water quality variables. The analysis process involves an investigation into cross-correlation between different variables and auto-correlation, predicting future values of the variables using machine learning, and also comparing results of the prediction models to determine optimal models with respect to the dataset.

The dataset used in this study was obtained from the United States Geological Survey (USGS), specifically measurements collected at the Hog Island channel monitoring station. The data of 12 variables from August 2010, with 6-minutes time intervals, are used **[2]**.

Methodology


Clean
Find all missing dates, delete dates that contain < 240 measurement/day
Reduce
Measurements reduced from 240 data points to 24 data point (Hourly time interval)
Split
60% training data and 40% testing data
Apply
Auto-regression, MLP Reg, LinReg, LogReg, DTReg, RNN and LSTM
Compare
Accuracy methods EVS, MAE, MSE, MSLE, R²

All possible combinations of MLP solver and activation function were applied. The model with the highest accuracy is presented here.

Results

Analysis

Table 1. Statistics of four sample variables.

Var	Min	Max	Mean	Std	Days	Unit
pH	-2.55	3.47	0.44	1.42	79	-
DO	-1.8	30.7	13.24	7.88	2406	mg/L
SC	32200	57700	45774. 1	2242.1	2283	S/m
Chl	2.5	18.7	8.75	2.35	2374	mg/L

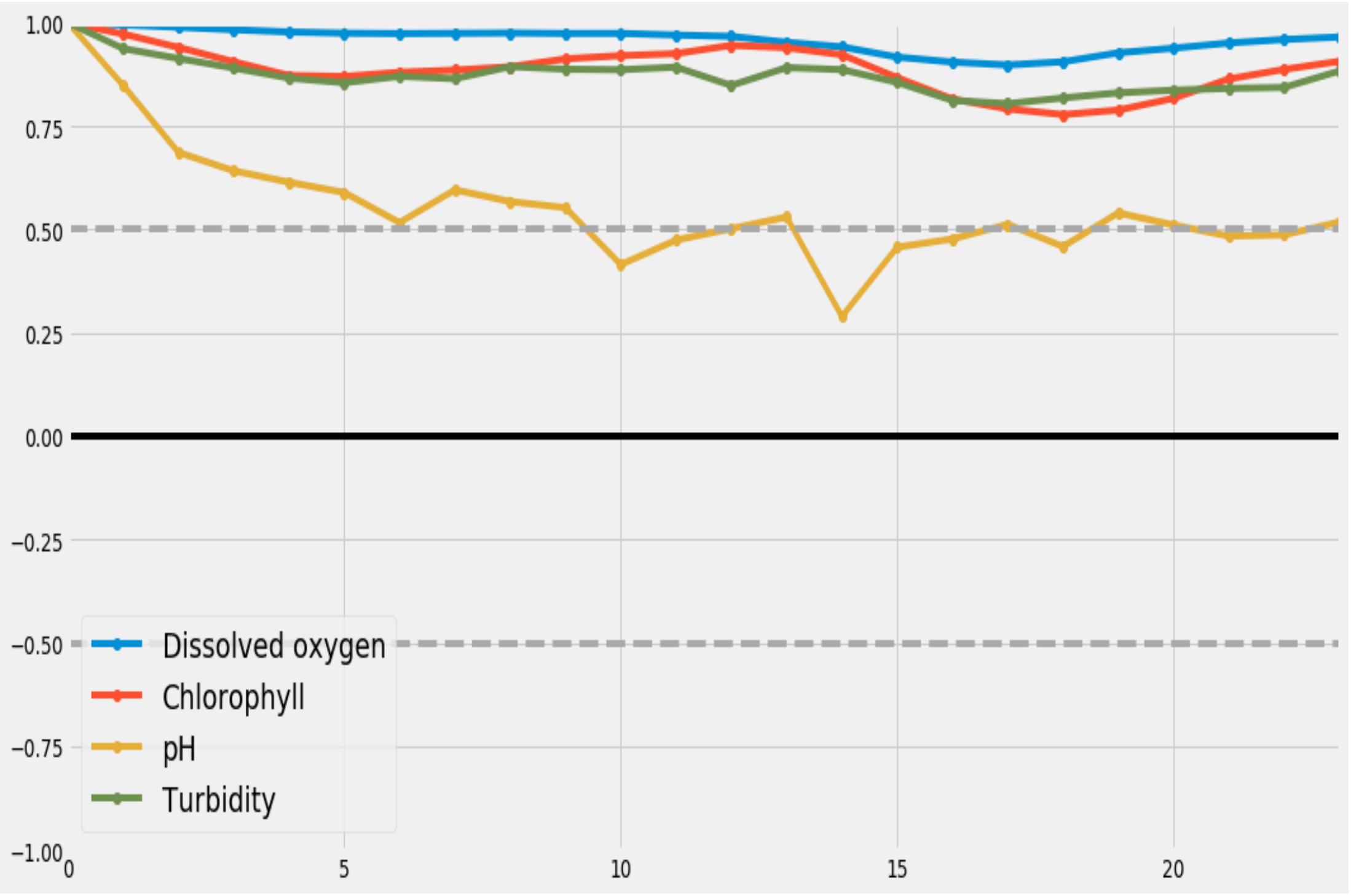


Figure 2. Auto-correlation of four sample variables, based on daily observations.

Twelve variables were used to train and test the models: dissolved oxygen, chlorophyll, specific conductance, turbidity, pH, temperature, nitrate, sampling depth, elevation, tidal prediction, salinity, chlorophylls. Auto-correlation was used to determine the relationship between measurements at different time intervals.

PCA was applied to evaluate the relative importance of variables over each other.

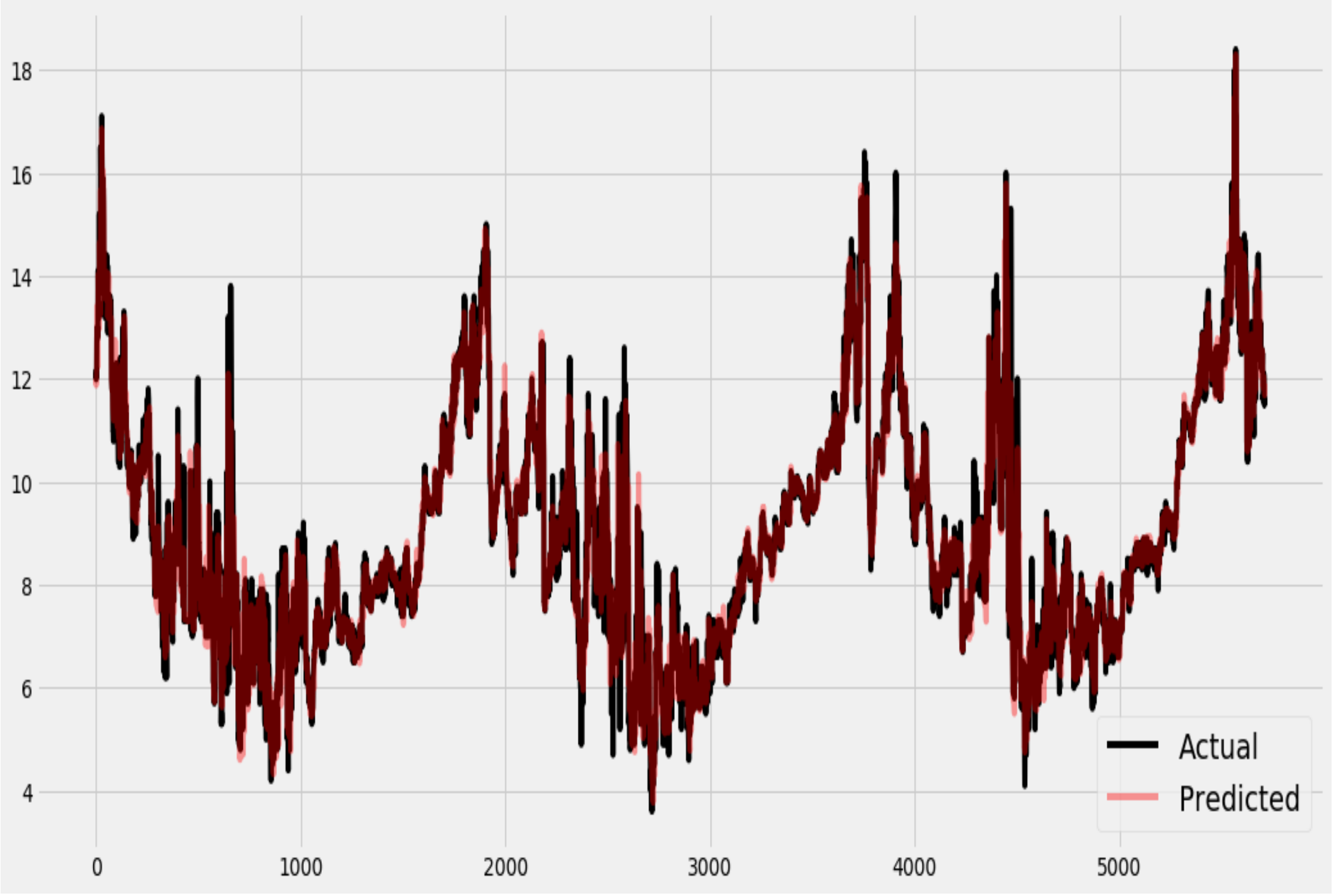


Figure 3. MLP Reg of actual and predicted DO.

Prediction

Based on the correlation results, 18 hours of data is used to predict the following 6 hours on daily data.

Figures 2 and 3 show the actual and predicted DO values using MLP Reg and LSTM respectively.

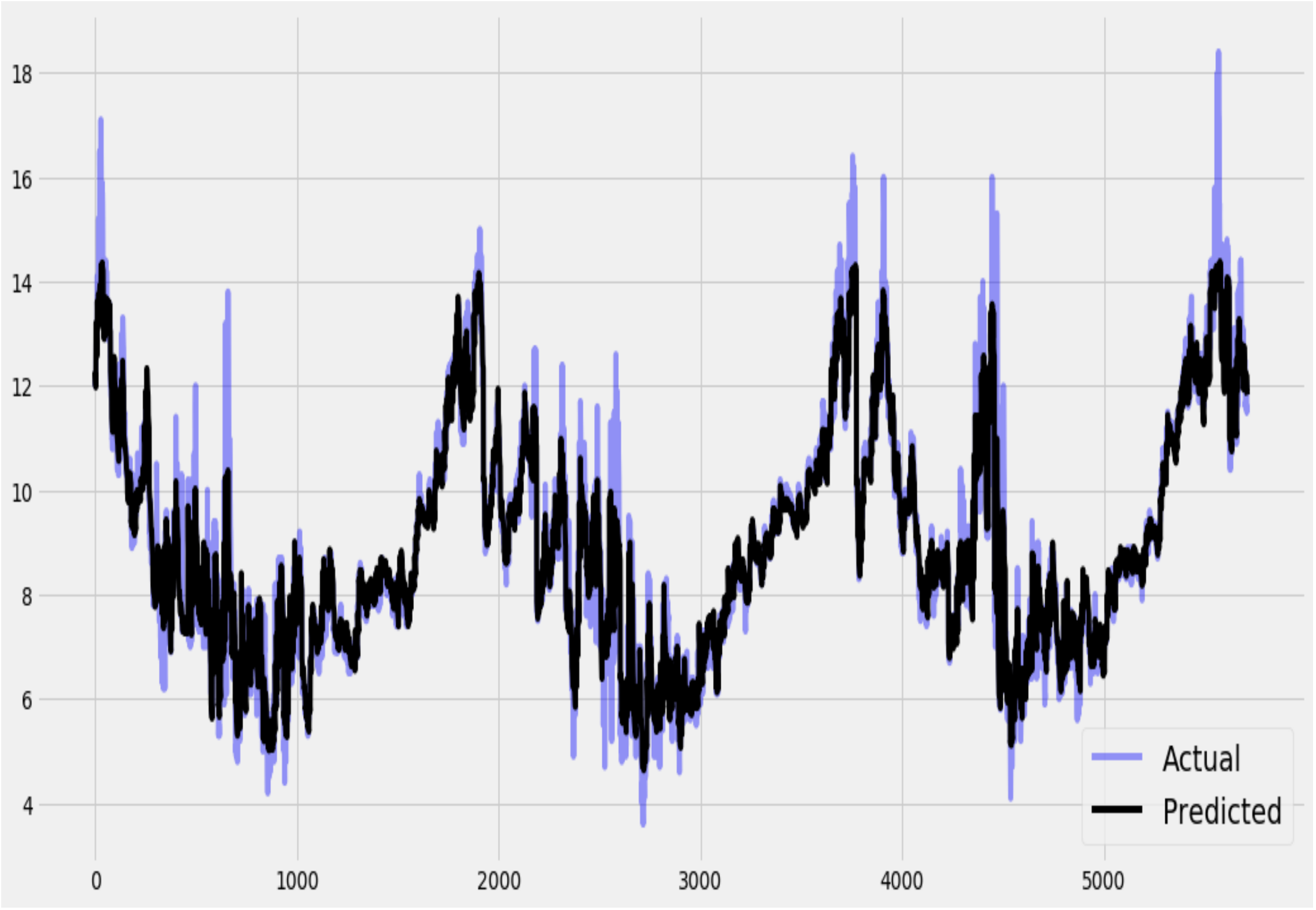


Figure 4. LSTM of actual and predicted DO.

Comparison

Numerous regression metrics were used to measure performance of the models **[3]**. These specify the losses or scores for each measurement. The best possible EVS score is 1.0, and other metrics have different ranges.

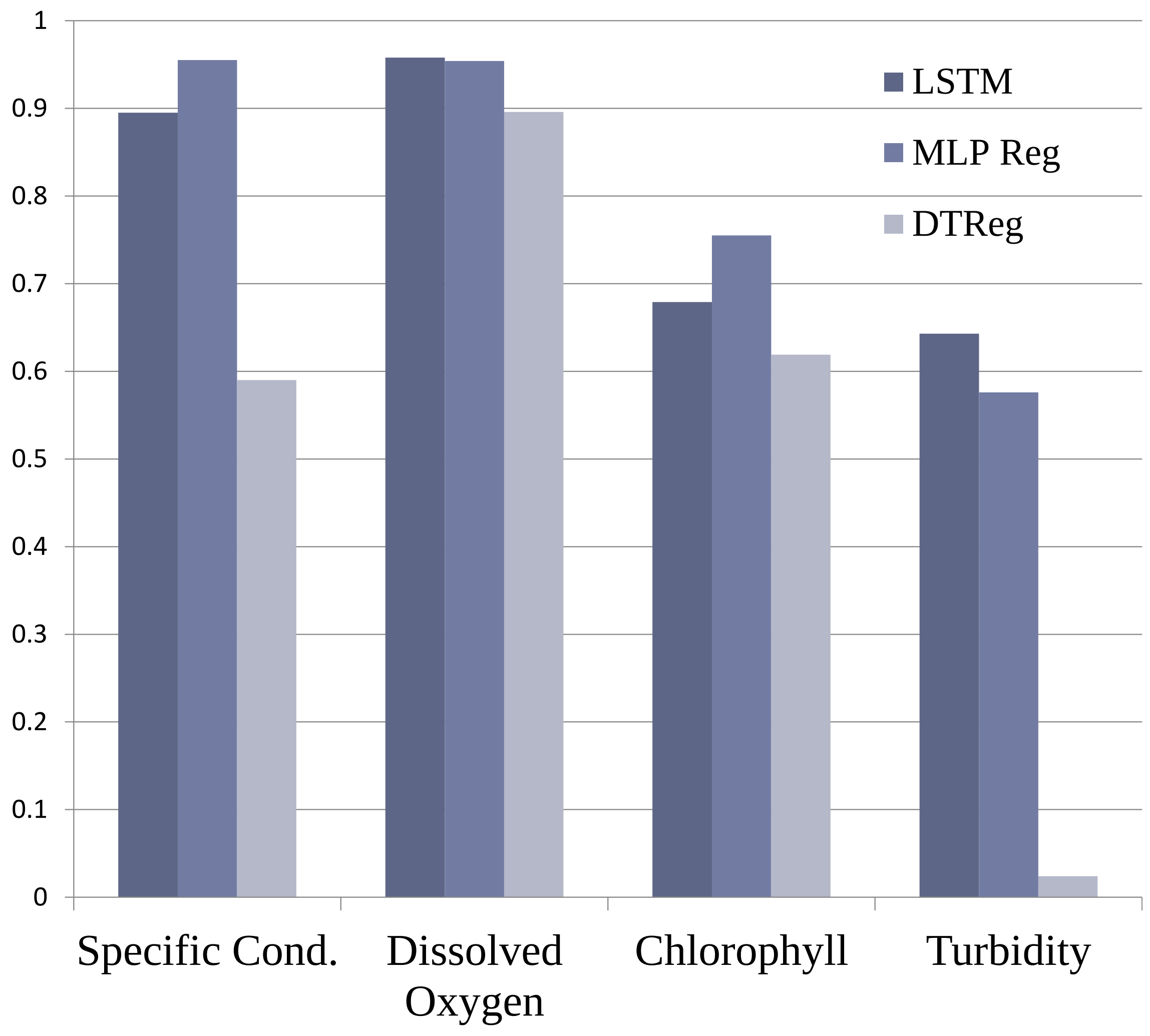


Figure 5. EVS accuracy of four sample variables.

Conclusion

This work presented various models applied to water quality data. LSTM and MLP Reg resulted in a highest EVS accuracy, while DTReg showed the lowest accuracy.

Future Work

The findings lay the groundwork for a number of natural extensions: application of the trained models to nearby water sources as well as model construction with different time frames.

References

[1] http://www.who.int/water_sanitation_health/resourcesquality/wqachapter3.pdf Accessed: 25/04/2018

[2] Predicting and analyzing water quality using Machine Learning: A comprehensive model, authors: Yafra Khan and Chai Soo See, IEEE Journal 2016

[3] http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics Accessed: 10/07/2018

Contact

Reem Elmahdi: reemomer@aims.ac.za
Willie Brink: wbrink@sun.ac.za
Josefine Wilms: j.wilms@csir.co.za

Scan to Get More Results!