LAKE ATITLAN

Group 1: Mathematical and Data Modeling

Pedro Rocha Liedl, Ruoyu Zhi, Zuzka Patacho de Matos Herrero, Xunge Zhang, Qinglu Jiao

March 16, 2024

Mathematical and Data Modelling

School of Engineering Mathematics and Technology

University of Bristol

Declaration of AI use

AI such as ChatGPT has been used to aid the creation of this project. It was exclusively used to accelerate the coding process of creating models. In almost every instance, it was necessary to alter the AI generated response to fit our needs, as such, all AI generated code was thoroughly reviewed.

1 Introduction

Lake Atitlan is one of the deepest and most popular natural lakes in Guatemala, with unique biodiversity and high ecological value. However, in recent years, due to the impact of climate change and human activities, the ecological balance of the lake has faced serious threats. The tropic state of lakes, influenced by inputs from surrounding rivers, changes in climatic conditions, and changes in local ecosystems, has changed significantly. [1] The situation is exacerbated by the fact that Atitlan is an endorheic lake, lacking a visible outlet. Consequently, the lake's accumulated water cannot be discharged and is only lost through evapotranspiration or by sinking ground. [2] For this reason, accurate predictions of future changes in these environmental variables are critical to developing effective lake management and conservation measures.

Many residents in the vicinity of Lake Atitlan rely on its water for daily use. For this reason, "Amigos del Lago Atitlan" are working on preserving the quality of the water of the lake. In recent years, the lake has presented abnormally high levels of chlorophyll, a green pigment prevalent in algae, cyanobacteria, and various aquatic plants, including phytoplankton. Its presence and concentration in water can significantly influence its quality. For this reason, our clients needs a model to forecast the evolution of chlorophyll concentration in the coming years. Such a model is crucial for taking required measures to ensure the potability of the water supply for Atitlan's population.

Our project aims to apply complex Recurrent Neural Network and Time Series Forecasting on combined Lake Atitlan's river, climate and ecosystem data over the past nine years to develop an advanced prediction model. By in-depth analysis of environmental data and development of forecasting models, our research provides new insights about future data that will facilitate lake management and protection.

2 Literature Review

In the study of the water quality of Lake Atitlan, we will focus on the most relevant indicators: chlorophyll concentration, water transparency (sechi), temperature, ph and dissolved oxygen.

In our research we found that previous studies have applied complex methodologies like neural networks and auto regression models to detect and predict the quality of the water.

Simple neural networks face many limitations in time series prediction. Wang et al. introduced a more effective approach using Long Short-Term Memory neural network (LSTM), trained on historical data from Taihu Lake (2000-2006). [3] This LSTM model outperformed traditional neural network methods. This model are more complex and have the ability to analyse non-linear data more accurately. Similarly, Babu et al. focused on water quality prediction using the LSTM algorithm and employed Decision Tree and Naive Bayes classifiers for Water Quality Index (WQI) classification, a methodology similar to our own. [4] Considering the effects of weather and seasonality, R. Xu et al. proposed the SARIMA-LSTM model. [5] Despite its better performance compared to SARIMA and LSTM individually, the combined model still lacks of ideal accuracy. Additionally, Tejoyadav et al. presented the VAR-LSTM hybrid model, which proved more reliable results than traditional forecasting techniques. [6]

In terms of prediction models for water blooms, X. Wang et al. proposed a prediction method based on grey-BP neural network using historical data from Beihai Lake to predict chlorophyll concentration. [7] The model's accuracy of 93.29% is closely associated with its large-scale dataset comprising 2208 instances.

3 The Data

Our data set focus on three key areas related to lake ecosystems:

- Limnology Dataset:

This dataset consists of measurements of the water quality in the lake itself at different depths in two coastal areas, and one in the centre of the lake. The data was collected from February 2014 to December 2023. This is considered to be the most important set of data as these are the values "Amigos del Lago" would like to predict.

- Weather Dataset:

This dataset includes monthly measurements of the weather from January 2014 to October 2023 in different stations around the lake. It includes information about minimum and maximum temperatures, humidity, solar radiation and wind velocity.

- Rivers Dataset:

This dataset represents monthly information related to the quality of the water from incident flowing rivers to the lake. Data goes from February 2015 to December 2023.

3.1 Seasonal Decomposition

One might assume that since the data describes an environmental system over time, there would be a strong patterns of seasonality in the data. In order to investigate this we decomposed the time series of each variable to analyse their trend and seasonality components. The trend represents the long-term linear trajectory of the variable, and the seasonality captures any repeated patterns over approximately a year.

Taking one example shown in 1, as similar patterns are observed with the other features, the monthly average flow of water is seen to have a strong yearly seasonal component. Additionally, we note a rising trend over the years, which is consistent with the endorheic characteristics of the lake.

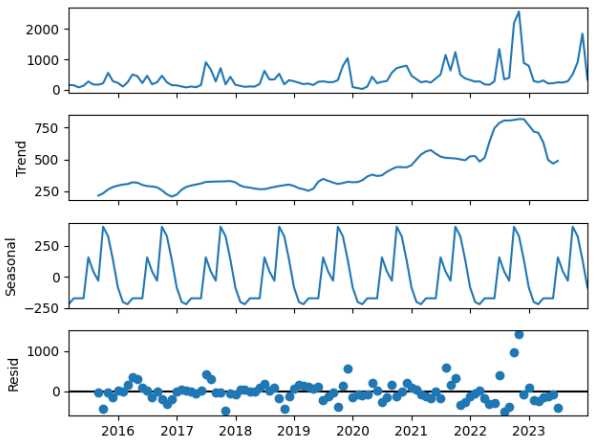

Figure 1: Seasonal decomposition of monthly average flow of water (Caudal)

3.2 Correlation Analysis

Furthermore, in order for us to understand how variables relate, we also analysed the interdependence of variables within each data set. The method of choice was calculating a correlation matrix which measures dependence between variables, defined by Pearson's correlation coefficient. This methodology fits our data as it is numerical, and allows us to find potential dependencies without creating a model. However, we do have to be mindful of what quantities having strong correlations mean in this context where, as always, correlation does not necessarily imply causation. For environmental data like this, correlations can be influenced by a range of factors including seasonal patterns, weather events, and human activities.

The dendrogram (2) allows an easier visualisation correlation matrix by organising the results as a hierarchical cluster. Features which feed into the same parent branch are more related than those which don't. Generally we note that there are no seeming unimportant features or that a given feature depends at least on one other feature. That said, there are clear groups of dependent features, two of which tend to gather around their own subsystems, weather and rivers. When it comes to optimising our model we use this as a starting point to experiment if these clusters of features have any unique predicting importance.

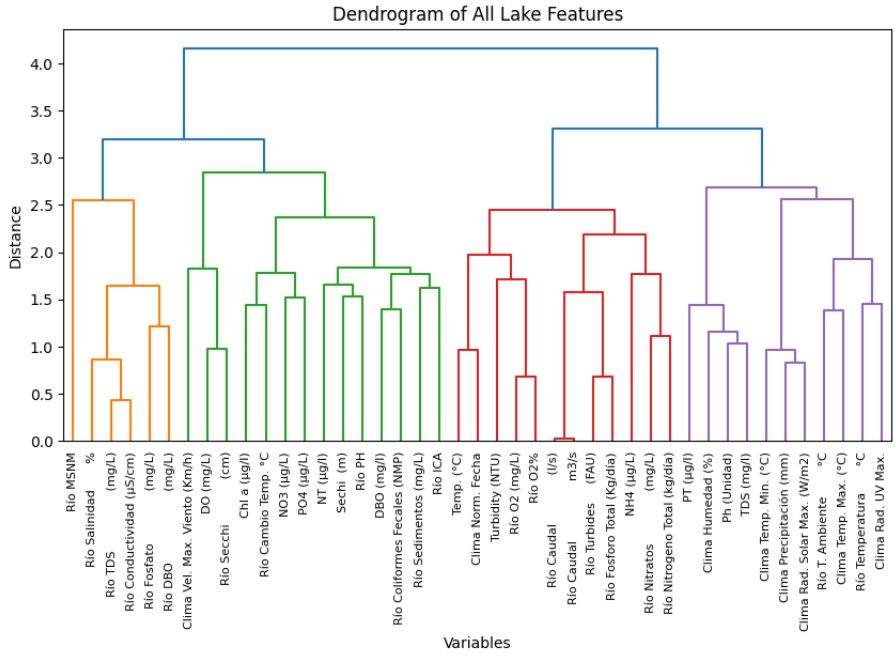

Figure 2: Dendrogram clustering variables of the calculated correlation matrix for combined lake system datasets.

3.3 Data Imputation

When preparing the data for analysis, there were instances where data imputation was necessary for two reasons, missing values and data associated from multiple locations. In an effort to maintain as much information as possible while also introducing the least amount of bias, the following prescription was applied:

1. Average columns of data across the same month and year from all locations where data was available and collect them into a single new row

2. If data was so sparse in the particular data set, simply fill in the missing values with the feature average, then average and collect by month and year as before

This approach was chosen as most of the data was collected on a monthly basis so collecting averages by month maintained the most amount of information. Ideally we would have liked to impute the data in a way which took into account seasonality but most of the data was too sparse in such features. Collecting the data this way also makes it possible to combine the datasets as we have done.

We applied the models in two different datasets. The first combines the three datasets we were given and it is the large one. The other dataset is a reduced version including the most influential variables: chlorophyll, PH, Dissolved Oxygen and temperature. This dataset is organised by depth at which the measurement was taken. We wanted to compare the results from averaging all the values or separating them by different depths. We considered this differentiation could be helpful since lake characteristics differ a lot from the surface to deeper water in the lake. There is a high concentration of chlorophyll in the surface due to the presence of algae which consume oxygen and nutrients like phosphate and nitrate. Deeper in the lake there is more oxygen and the temperature is lower.

4 Models

4.1 Seasonal Auto-Regressive Moving Average (SARIMA)

4.1.1 Auto-regression

In order to predict values that evolve overtime, we use time series forecasting which includes several techniques like Auto Regression (AR). The key aspect of these methodology is that future values are predicted based on analysis of previous trends, assuming the future will be similar to historical data. In this model, the output variable depends linearly on its own past values and a stochastic term:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{1}$$

In this equation, y_{t-i} are the preceding values of the variable of interest, used as predictors. The term c is an autoregressive constant and ϵ_t represents noise. Alterations to ϵ_t will only affect the scale of the series. As you can observe in the equation, the variable of interest is predicted using a linear combination of its previous values.

4.1.2 Moving Average

Time series analysis also includes Moving Average (MA) models, characterized by the use of previous forecast errors in a regression equation to predict future values:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{2}$$

where ϵ_t represents noise and varying parameters θ_i,..., θ_q results in different time series patterns.

4.1.3 ARIMA

Combination of these models results in well-known ARIMA, Auto Regressive Integrated Moving Average:

- Auto regression (AR): variables depends on its own present, or prior, values.
- The Integrated (I) component represents the differentiation of observations so that the time series become stationary, which means that its statistical properties do not depend on the time at which the series is observed. [8] To achieve stationary series, data values are replaced by the difference between the current and the preceding values. Therefore, the integrated element refers to differentiating, allowing the method to support time series data with a trend.
- Moving average (MA): includes the dependency between an observation and a residual error.

This is how an ARIMA model will look like:

$$y'_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{3}$$

where y'_t are the differenced series (which can be differentiated more than once) and ϵ_{t-i} is past noise.

ARIMA(p, d, q) where p is the order of the autoregressive part; d is the degree of first differentiating involved and q is the order of the moving average part. Depending on the values of p, d and q that you set, there will be different variations of ARIMA models.

There are several reasons why ARIMA can be a very useful method for our purpose. Firstly, it only requires historical data, which perfectly aligns with our data set. Another advantage is that it does not need much data, from 50 to 100 observations is enough to build a proper model. Furthermore, it is frequently used for hourly or daily data, matching the nature of our data. [9]

4.1.4 SARIMA

However, there is a feature in our data set that ARIMA cannot handle: seasonality. We can solve this limitation by using Seasonality ARIMA (SARIMA) models. It is an extension of ARIMA used to forecast univariate time series data with seasonality. It is modeled by including additional seasonality terms in the simple ARIMA:

$$X_t = \mu_t + \gamma_t + \phi_t + \sum_{i=1}^{m} \sum_{\tau=0}^{q} \omega_{i,\tau} y_{i,t-\tau} + \varepsilon_t, \tag{4}$$

In this equation, ω_i are the parameters of the model, ϕ_t represents the autoregressive component, γ_t is the seasonal component, μ_t is the trend component and ϵ is noise.

Modeling SARIMA requires selecting parameters for both the seasonal and non-seasonal elements of the series. Choosing proper parameters will make your model successfully forecast the data. This is how you would formulate SARIMA from ARIMA:

$$\text{ARIMA} \quad \underbrace{(p, d, q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \qquad \underbrace{(P, D, Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

where m = number of observations per year. The modeling procedure is almost the same as for ARIMA, except that we need to add seasonal AR and MA terms. The selection of the model parameters is all determined by the minimisation of the AIC (Akaike Information Criterion). [10]

4.2 Perceptron

Frank Rosenblatt introduced the Perceptron model in 1957, setting a foundational step in Artificial Intelligence and Machine Learning.[11]. This model, simulating the behavior of biological neurons, showed that algorithms could perform pattern recognition,laying the basis for future advancements. The Perceptron is a linear classifier that operates by finding a decision boundary to distinguish between different data classes.

The mathematical representation of the Perceptron model is very intuitive:

$$f(x) = \theta(\omega \cdot x + b) \tag{5}$$

where x represents the input feature vector, ω is the weight vector, b denotes the bias term, and the Heaviside step function(θ) maps the weighted sums to category labels. The goal of this model is to accurately predict category labels for all training samples (x_i, y_i) by adjusting weights w and bias b. This objective is achieved by iteratively adapting the parameters over the training data. In each iteration, the model updates the weights and bias for misclassified samples to minimise future classification errors. [11]

The algorithm adjusts the weights and biases when an instance is misclassified, using the update rules: $w = w + \eta y_i x_i$ for weights and $b = b + \eta y_i$ for biases, where η represents the learning rate. This parameter is a small positive value that controls the magnitude of the updates, playing a crucial role in the algorithm's convergence.

4.3 Regression Tree

Regression tree is a machine learning method, which is a variant of decision trees, aimed at fitting data and generating predictions by recursively partitioning input features. In a regression tree, each node represents a feature split, dividing the dataset into different subsets based. This process continues until a stopping criteria is met, such as reaching maximum depth, insufficient samples in a subset, or node impurity falling below a threshold. [12]

When constructing the regression tree model, we utilize optimisation algorithms to determine the optimal division at each node. Commonly used methodologies include mean squared error (MSE) and absolute error, which measure the difference between predicted values and actual values. In this study, we use MSE as the criterion to evaluate the goodness of fit of the model.

Regression tree models present several advantages, such as strong modeling capabilities for nonlinear relationships and robustness to missing values and outliers. However, they also have some limitations, like sensitivity to overfitting or noise and outliers. These limitations can be addressed through adjusting hyperparameters, such as adjusting the maximum depth, minimum samples per leaf, and minimum samples required to split a node. [12]

4.4 Long Short-Term Memory (LSTM)

To deal with the non-linear relationship between the multivariate data and the data itself, and to improve the sensitivity to long-term dependencies, we introduce the LSTM model.

The Long Short-Term Memory (LSTM) model is a special type of recurrent neural networks (RNNs) structure designed to address the shortcomings of traditional RNNs in learning long-term dependencies. [13] LSTM exhibits a deep structure in the time dimension and is equipped with designed gates and memory units as shown in Figure 3. It can determine when to forget state information and how long to preserve it. [13]

The training algorithm of LSTM is similar to traditional neural networks. The principle involves forward propagation to compute the output of each neuron and then backward propagation to calculate the error term of each neuron. Based on these error terms, gradients for each weight are computed.

The key to LSTM lies in the cell state. It acts like a conveyor belt, running directly throughout the entire chain with only a few linear interactions, enabling information to be preserved. The cell state is controlled by three designed gates. At time step t, the input of the memory block (current neuron) includes: the current input variable X_t, the previous hidden state variable h_{t-1}, and the previous memory cell state variable C_{t-1}. Subsequently, the model passes through the forget gate f_c, the input gate i_t, and the output gate o_t in sequence. The output of the memory block includes: the current output variable h_t and the current memory cell state variable C_t.

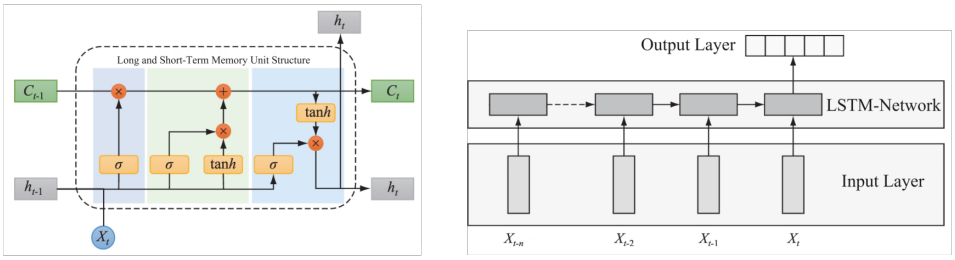

Figure 3: The architecture of LSTM memory cell and Structure of the LSTM model

Deep learning models often involve a large number of hyperparameters, common ones including learning rate, number of training epochs, number of hidden layers, and number of neurons per layer. These hyperparameters significantly impact the performance and effectiveness of the model learning process. Typically, it is necessary to optimize hyperparameters, and we will discuss the setting values of model hyperparameters in the next chapter.

5 Results

As we explained in previous sections, we applied the models in two different datasets. Our results focused on predicting mainly the features 'Sechi' and 'Chlorophyll'. Sechi is a measure of the clarity of the lake. It is of high importance to the client as a lot can be inferred about overall lake health, for example levels of algae, and sediment. Sechi is a disk used to measure the clarity of the water in the lake. It is lowered into the water until it can no longer be seen by the observer. The depth at which it disappears, known as the Sechi depth, is a measure of the water's transparency. Therefore, the deepest it is visible, more clear is the water indicating less contamination. The closer to the surface, less transparency and more contamination.

On the other hand, Chlorophyll is important because it is present in all algae and cyanobacteria, and its concentration is directly related to the amount of these organisms in the water. Thus, measuring chlorophyll can provide an estimate of algal biomass which is important for maintaining recreational use of water bodies, preserving fisheries, and ensuring the safety of drinking water.

5.1 Results of the large dataset

5.1.1 Long Short-Term Model (LSTM)

We will normalise the data before training to standardise features and enhance the model's generalization capability. The hyperparameters of the model are:

Parameter Type	Parameter Name	Tuning Method	Value
Global Hyperparameters	Hidden Layers		2
	Hidden Size		40
	Batch Size	Empirical Tuning	8
	Window Size		16
Training Hyperparameters	Learning Rate		0.01
	Epoch		220

Table 1: LSTM Model Hyperparameters for Large Dataset

To validate the reliability of the model, we will independently run the model 10 times, calculate the average MAE, MSE, RMSE and NRMSE for these four metrics, and show the performance of the prediction results on the test set. Additionally, we will use the entire dataset to make preliminary predictions for the next 5 years.

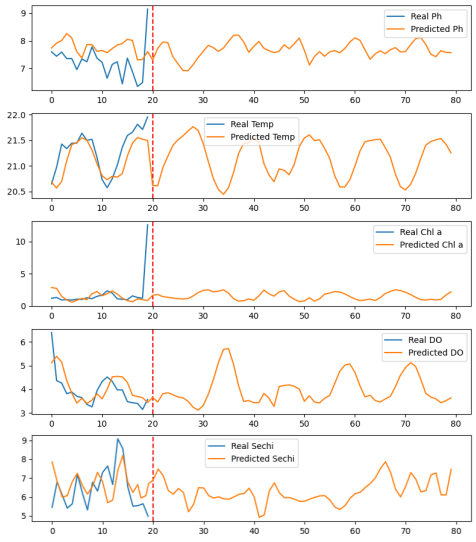

Figure 4: Plots of the original data and 5-year prediction using LSTM model for features: PH, Chlorophyll Concentration, DO, and Sechi.

MAE	MSE	RMSE	NRMSE
0.18	0.061	0.247	0.247

Table 2: Forecasting Performance Evaluation Metrics for the Large Dataset (Normalised)

It can be observed from the figure 4 that, apart from the pH indicator, the LSTM model demonstrates a good fit with the actual data on the test set. However, there are periodic fluctuations in the forecasting results for the next few years. Through multiple experiments, we found that the periodicity is related to the rolling window size. This is because the dataset

has a small amount of data, resulting in the model's decreased generalisation ability for future data, leading to distorted forecasting results.

5.1.2 Perceptron

The dataset contains continuous variables, requiring an investigation into the suitability of the Perceptron model for its application to such continuous datasets. Therefore, as an initial step we need to make these variables discrete. Specifically, we categorise sechi into three distinct classes—low, medium, and high. The categorisation criteria are as follows:

Class	Percentile
Low	0 - 0.25
Medium	0.25 - 0.75
High	0.75 - 1

Table 3: Table for Discretisation of Continuous Variables

This procedure transforms the continuous variables into a finite set of categories, making them compatible with Perceptron model. Furthermore, this approach also preserves the inherent relative information and trends present in the dataset. However, it is important to take into accout that this method may result in a certain level of information loss.

We selected key features for model training and prediction. Using the LabelEncoder, we encoded the sechi category labels into numerical format to follow with the input requirements of the Perceptron.

The model underwent parameter optimisation to minimise prediction errors. This process involved training the model on a given training set and subsequently utilising it to make predictions on a separate test set. The performance of the model was then evaluated by comparing these predictions against the actual data. For the dataset corresponding to Sechi value measurements, the model achieved a prediction accuracy of 0.55, accompanied by a Mean Squared Error (MSE) of 0.9. Figure 5 illustrates the comparison between the predicted values and the actual observations for the test set, showcasing the Perceptron's predictive capability. Low accuracy and high MSE showcase the limitations of the simplicity of Perceptron, as the graph also confirms. For this reason we decided to try more complex models.

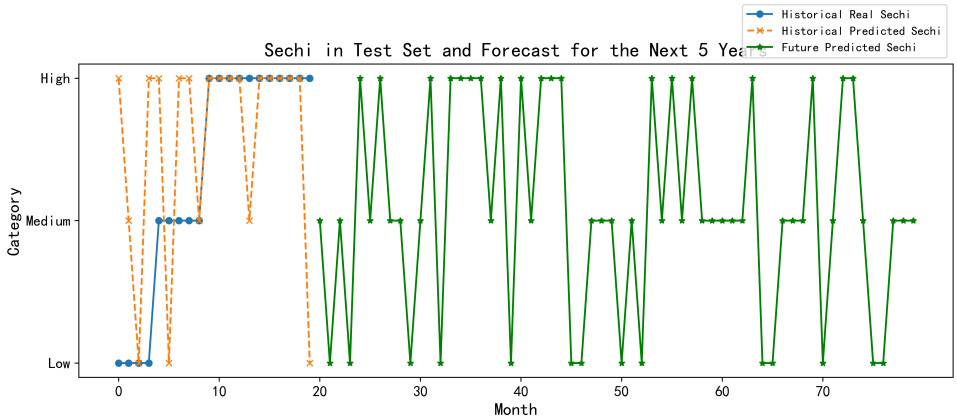

Figure 5: Prediction of the Perceptron for Sechi in the Test Set and the Forecast for the Next 5 Years

5.1.3 Regression Tree

The dataset was partitioned into training and testing sets, with the first 70% designated for training and the remaining 30% for testing. The mean squared error was 1.6136. The prediction results of the test set are illustrated in the following figure. The prediction results in the figure 6 exhibit some discrepancies and lack precision. And for the forecast of Sechi for the next 10 years, refer to the figure 7 :

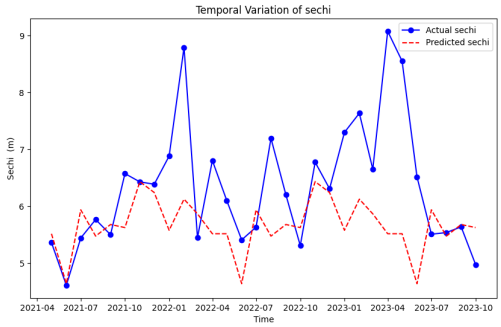

Figure 6: Prediction Results of Sechi by Using Regression Tree in the Test Set

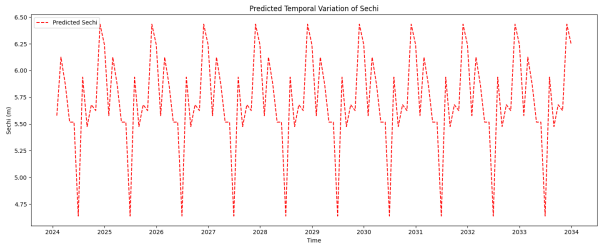

Figure 7: Forecast Results for Sechi by Using Regression Tree Over the Next Decade

5.1.4 SARIMA

We also use SARIMA to forecast Sechi values for the next 10 years: We can observe a seasonality

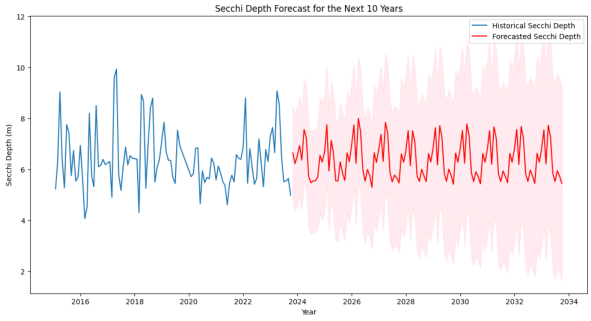

Figure 8: SARIMA prediction of Sechi (m) over the next 10 years

pattern similar to the observed historical data. Sechi values increase at the beginning of the year until a maximum peak around summer, then reduce drastically by the end of the year to then increase back again. The interpretation of these results is that the water is less contaminated the first months of the year until it reaches a maximum and then visibility reduces dramatically.

5.2 Results of the reduced dataset

5.2.1 Long Short-Term Model (LSTM)

To address issues from the fully combined dataset, we will focus on the second dataset. By segmenting various depths into multiple intervals and arranging them sequentially along with the depths of the preceding and succeeding two months, we aim to present a clearer analysis. Furthermore, given the substantial volume of this dataset, deep learning models are expected to perform more effectively.

Furthermore, since we adopt a rolling forecasting approach, where the model's prediction results serve as inputs for subsequent predictions, any errors will gradually accumulate. Additionally, as time progresses, the distribution of time series data may change, potentially leading to a decline in model performance. Therefore, the number of nodes predicted is kept approximately equivalent to the set window size.

The hyperparameters of the second individual model are shown in Table 4 .

Parameter Type	Parameter Name	Tuning Method	Value
Global Hyperparameters	Hidden Layers		2
	Hidden Size		64
	Batch Size	Empirical Tuning	16
	Window Size		36
Training Hyperparameters	Learning Rate		0.01
	Epoch		520

Table 4: LSTM Model Hyperparameters for Reduced Dataset

Similarly, we will independently run the model 10 times and calculate evaluation metrics, as shown in Table 5.

MAE	MSE	RMSE	NRMSE
0.12	0.04	0.191	0.192

Table 5: Forecasting Performance Evaluation Metrics for Reduced Dataset (Normalised)

In the graph, we will analyse other related indicators with depth as a reference. It can be observed that the depth curve exhibits periodicity, with each cycle containing several depth intervals, representing a monthly time step.

Through analysis, it is evident that the model's predictions for the peak values of the other four indicators relative to depth are relatively accurate, but the fit for the valley values of the pH indicator is poor. There is a phenomenon of model drift at the end of the test set, which is due to missing relevant depths in the original dataset, and the prediction model cannot effectively identify such behavioral patterns. However, this has minimal impact on the predictions for the next six months that we need, as we only use the predicted depth values as a reference.

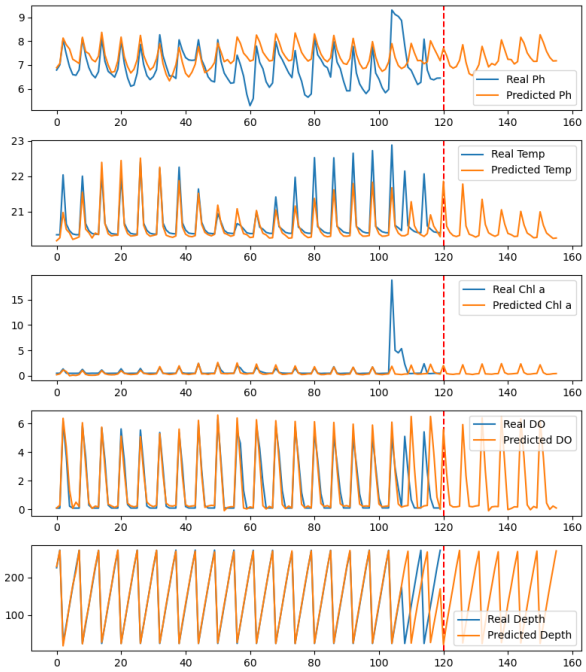

Figure 9: Plots Created by the LSTM Model Fitted and Forecasting pH, Temperature, Chlorophyll Concentration, DO, and Depth.

5.2.2 Perceptron

In this study, Chlorophyll and Dissolved Oxygen (DO) levels are selected as the primary predictors. The dataset is partitioned into an 80% training set and a 20% test set for the model evaluation. The Perceptron is then trained and its performance is evaluated on the test set. Additionally, we forecast the trends of Chlorophyll and Dissolved Oxygen levels in Atitlan Lake's ecosystem over the next decade.

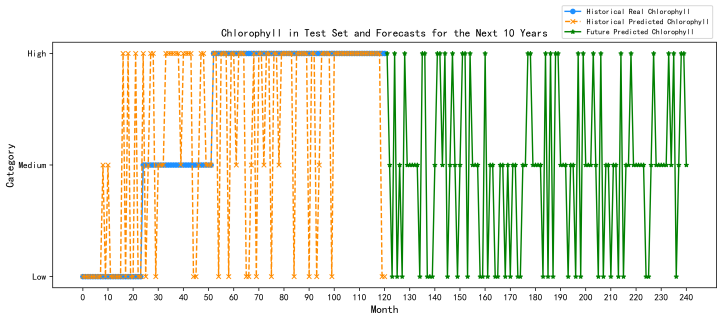

Figure 10: Perceptron prediction of Chlorophyll in the test set and for the next 10 years

The performance evaluation on the test set indicates that the trained Perceptron achieves an accuracy score of 0.6612. Additionally, the Mean Squared Error (MSE) associated with these predictions is 0.7107.

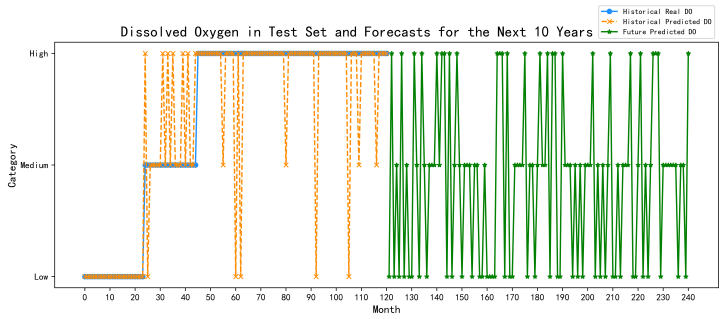

Figure 11: Perceptron prediction of Dissolved Oxygen in the test set and for the next 10 years

Indicator	MSE	Accuracy
Chlorophyll Concentration	0.7107	0.6614
Dissolved Oxygen (DO) Concentration	0.2314	0.8678

Table 6: Mean Square Error and Accuracy of Perceptron on Predicted Values

5.2.3 Regression Tree

This model was also used to forecast changes in chlorophyll concentration, temperature, pH, and dissolved oxygen (DO) concentration.

Indicator	MSE
Chlorophyll Concentration	4.6357
Temperature	0.0180
pH	0.4691
Dissolved Oxygen (DO) Concentration	0.2726

Table 7: Table reporting the mean squared errors for features modelled by the regression tree.

Judging from the mean squared errors 7, except for chlorophyll concentration, the fitting performance is relatively good.

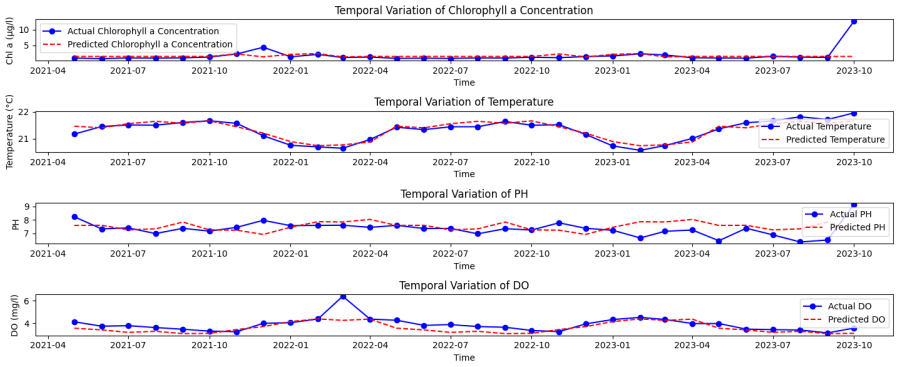

Figure 12: Forecast Results of Regression Tree for Other Indicators in the Test Set.

For the 10 years forecast, refer to the following figure:

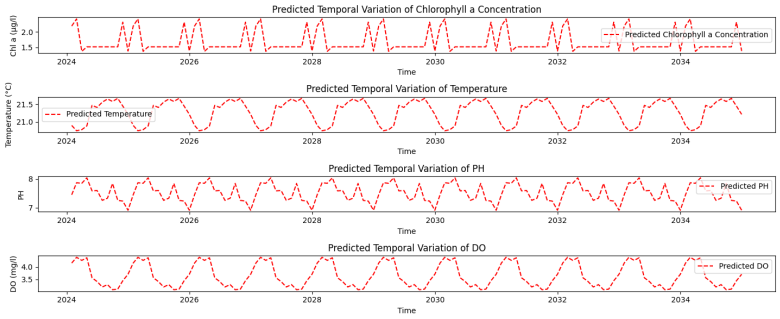

Figure 13: Forecast Results of Regression Tree for Other Indicators Over the Next Decade.

5.2.4 SARIMA

We also performed SARIMA using the reduced dataset obtaining the following predictions:

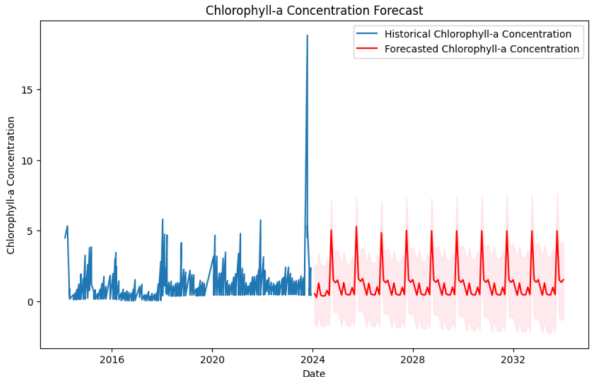

Figure 14: SARIMA prediction of Chlorophyll over the next 10 years using Reduced Dataset.

Once more, it is evident that there is an annual pattern in the chlorophyll concentration, which reaches its peak at the end of each year, coinciding with the rainy season. This observation is logical, considering that organisms tend to prefer colder waters and increased wind conditions. Then, as dry season arrives, chlorophyll concentration reduces again. No increasing trends in chlorophyll concentration are projected for the next decade, a development that is reason to celebrate. Given that SARIMA models rely on historical data, should conditions remain consistent with those of the past nine years, the predicted trends appear to align closely with previous observations.

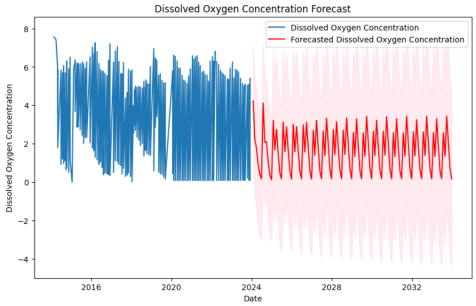

Figure 15: SARIMA prediction of Dissolved Oxygen over the next 10 years using Reduced Dataset

5.3 Comparison of Models

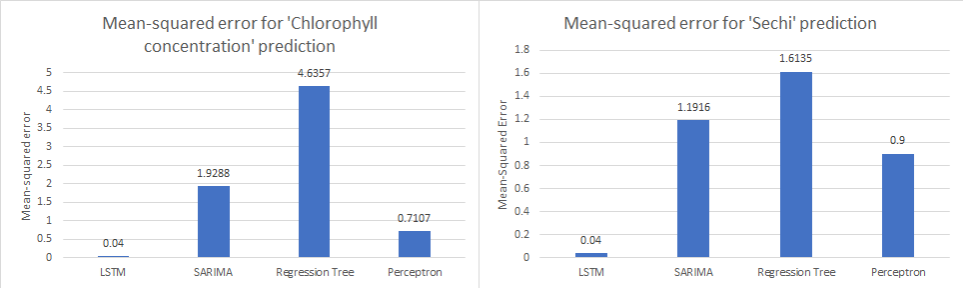

Figure 16: Bar charts comparing the mean-squared errors our LSTM, SARIMA, Regression tree, and Perceptron models using the combined dataset.

In order to compare how our models performs with each other, we use the mean squared error (MSE) as the comparison metric. As seen in figure 16 the LSTM model fits the features Sechi and Chorophyll Concentration the best of all the models with conversely the Regression Tree model performing the worst. We believe this is likely because the Regression Tree is a much simpler model as was previously mentioned, and hence maybe not be able to capture all the hidden data patterns well. We believe the same can be said for the Perceptron as it has used a classified version of the data, despite the second best MSE score. On the other hand, the SARIMA model stands in the third best. Visually inspecting the predicted plot in figure 8 we can see it clearly has found a periodic pattern. As the LSTM also identifies a periodic pattern it may lead us to suggest that even though the LSTM performs better according to the MSE, it holds value as a model for predicting the key features.

6 Conclusion

In this report we attempted to create various models to answer the proposed ask of the client; predict key features of Lake Atitlan for the next ten years. In this process we first started by conducting a review of previous work to find out more about how others have tried to model water systems using time series forecasting. From there we then analysed the data to discover patterns and feature dependencies between the subsystems of the ecosystem. Through seasonal decomposition, we found that most of the data follows a seasonal pattern over a year, which was expected. Additionally, we noted that there were no seemingly isolated features that had no dependency on another one, as identified by the correlation analysis. This information would then aid in the creation of our models. Here we experimented further with creating models which took the combined system data and subsystem only data to see which produced the best results. According to the mean squared error metric and the visual fit of the plots against the original data we conclude the LSTM and SARIMA models perform better at modelling the key features, Sechi and Chlorophyll. In line with all predictive models, we pretext all of our findings with the fact that they are not oracles with absolute accuracy in their predicting power. We do believe however that they may provide useful insights about the trends of the next 10 years (in the absence of any unprecedented events). Hopefully this will be very valuable to the organisation and allows them to take required measures that aid in the preservation of Lake Atitlan.

References

[1] Timothy P. Neher, Michelle L. Soupir, and Rameshwar S. Kanwar. Lake atitlan: A review of the food, energy, and water sustainability of a mountain lake in guatemala. *Sustainability*, 2021.

[2] World Atlas. Fluvial landforms: What is an endorheic basin? 2021.

[3] Yuanyuan Wang, Jian Zhou, Kejia Chen, Yunyun Wang, and Linfeng Liu. Water quality prediction method based on lstm neural network. pages 1–5, 2017. doi: 10.1109/ISKE.2017.8258814.

[4] S. Babu, Banavath Baby Nagaleela, Cheekurimelli Ganesh Karthik, and Lakshmi Narayana Yepuri. Water quality prediction using neural networks. pages 1–6, 2023. doi: 10.1109/ICECONF57129.2023.10084120.

[5] Rui Xu, Qingyu Xiong, Hualing Yi, Chao Wu, and Jianxin Ye. Research on water quality prediction based on sarima-lstm: A case study of beilun estuary. pages 2183–2188, 2019. doi: 10.1109/HPCC/SmartCity/DSS.2019.00302.

[6] Mogarala Tejoyadav, Rashmiranjan Nayak, and Umesh Chandra Pati. Multivariate water quality forecasting of river ganga using var-lstm based hybrid model. pages 1–6, 2022. doi: 10.1109/INDICON56171.2022.10040146.

[7] Xiaoyi Wang, Jun Dai, Zaiwen Liu, Xiaoping Zhao, Suoqi Dong, Zhiyao Zhao, and Miao Zhang. The lake water bloom intelligent prediction method and water quality remote monitoring system. 7:3443–3446, 2010. doi: 10.1109/ICNC.2010.5584552.

[8] G. P. Nason. Stationary and non-stationary time series. In *Statistics in Volcanology*. Geological Society of London, 01 2006. ISBN 9781862392083. doi: 10.1144/IAVCEI001.11. URL https://doi.org/10.1144/IAVCEI001.11.

[9] S. C. Hillmer and G. C. Tiao. An arima-model-based approach to seasonal adjustment. *Journal of the American Statistical Association*, 77(377):63–70, 1982. doi: 10.1080/01621459.1982.10477767.

[10] Asep Rusyana, Nurhasanah, Marzuki, and Mia Flancia. Sarima model for forecasting foreign tourists at the kualanamu international airport. pages 153–158, 2016. doi: 10.1109/ICMSA.2016.7954329.

[11] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi: 10.1037/h0042519.

[12] Bahram Choubin, Gholamreza Zehtabian, Abbas Azareh, et al. Precipitation forecasting using classification and regression trees (cart) model: a comparative study of different approaches. *Environmental Earth Sciences*, 77(9):314, 2018.

[13] R. C. Staudemeyer and E. R. Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks. *arXiv preprint*, Year. doi: 10.48550/arXiv.1909.09586.