Forecasting of Water Quality Parameters of Lake Atitlán

Group 1: Lake Atitlán

Evan Steer, Nicholas Saremi, Sidney O'neill, Talia Saltmarsh

March 27, 2025

Mathematical and Data Modelling 3

School of Engineering Mathematics and Technology

University of Bristol

Declaration of AI use

In this report, AI was used for preliminary research to suggest and explain concepts and sources for fact-checking and further research, which were then verified. ChatGPT was also used to develop code to reduce time when programming, all code was then extensively tested to ensure correct implementation.

1 Introduction

Lake Atitlán, the deepest lake in Central America, is located in the Sierra Madre mountains in the central highlands of Guatemala. Its shores are home to more than 7000 people, many of whom live in extreme poverty, and depend on the lake for drinking water, washing, food, and the cultivation of crops. The lake is also the major transportation network in the area, as roads and footpaths are limited, and hold symbolic importance to various Maya indigenous communities [1]. However, the environmental quality of the lake is declining due to human activity, as the watershed of the lake is used mainly for agriculture. Due to technological limitations, bureaucratic barriers, and steep topography water treatment is minimal, so most of the nutrient-rich waste containing nitrogen and phosphorus is washed into the lake untreated [2]. Since 2008, the uncontrolled input of nutrients into the lake has resulted in yearly occurrences of cyanobacterial blooms, which can produce toxins harmful to humans. In 2009, a bloom in October covered almost half the lake's surface for more than 2 months [3]. These blooms can be detrimental to aquatic wildlife as they deplete the dissolved oxygen levels in the lake water, in turn threatening both the food and water supply of the inhabitants of the lake's shores [4, 5].

Predicting negatively changing lake conditions can enable the early implementation of preventative policies (such as banning environmentally harmful pesticides) and motivate long-term water quality management strategies and infrastructure. These steps must be taken to ensure the prosperity of the lake, its surrounding areas and the local people and wildlife that depend on it. In this report, the cleaning and imputation of existing limnological data from Lake Atitlán is explored and the use of time series and regression models to predict the future values of these data is evaluated. It is found that a feature-specific processing method most effectively interpolates missing data, and a variety of limnological features can then be predicted using time-series forecasting. The LSTM model generally performs better than the SARIMA model in predicting trends in temperature and turbidity, however, a combination of both is the most effective method for prediction. We then demonstrate the use of machine learning regression models to use this data to further predict other parameters, such as dissolved oxygen and chlorophyll amounts.

2 Review of Existing Literature

This section focuses on past research on forecasting limnological characteristics. Gupta et al. forecasted various parameters of water pollution at nine locations along a river in Dehli (pH, COD, BOD and DO) by carrying out time series analysis using the ARIMA (Auto Regressive Integrated Moving Average) model. They noted a significant amount of missing data from their data set, emphasising the importance of properly processing environmental data [6]. Jinxin et al. also notes the crucial steps that data cleaning plays in climate forecasting when describing their successful application of Long Short-Term Memory (LSTM) models in climate prediction, particularly in predicting temperature [7]. As our data has significant gaps due to constraints during data collection, processing the dataset will be crucial to achieving accurate results from any modelling.

Past research has used a variety of methods to predict climate data. Liu et al. find that the LSTM model is more effective than other prediction models when applied to marine temperature modelling in multiple seas [8]. Islam et. al use both the LSTM and the seasonality autoregressive integrated moving average (SARIMA) model when predicting the quality of water in two rivers in China, as these methods are effective for data with strong seasonality [9]. As recent previous research has found that either LSTM or SARIMA models give the best results, these methods

are used for predicting the future water quality of Lake Atitlán in this paper.

Ahmed et al. tested a variety of supervised machine-learning models to estimate the water quality index (WQI). They found that gradient boosting and polynomial regressing predict the WQI most effectively while employing a minimal number of input parameters of temperature, humidity, pH, and total dissolved solids (TDS) [10]. Similarly, Trolle et al. use empirical regressions between lake water quality attributes (e.g. biological indicators, water temperature, and nutrient load) to enable extrapolation to future scenarios [11].

This research identifies three key areas in which our paper can be focused to effectively predict the future water quality of Lake Atitlán. These are in the pre-processing of the data to make it useable with the selected prediction methods, implementing the LSTM and SARIMA to predict future values of variables, and utilising regression methods to infer values of target variables from other parameters.

3 Data Pre-processing and Imputation

Models in this report are trained on a large limnological dataset provided by collaborators from the Association of Friends of Lake Atitlán. This data details several factors that affect the condition of the lake across depth and time, such as Temperature, Turbidity and nutrient levels. Readings were taken from 3 sampling stations labelled SA, WG and WP, with WG being in the deepest part of the lake towards the centre and SA and WP close to the shore. Data collection was prioritised at the WG location, therefore this report focuses on the data collected at this location.

Depth data was grouped into three categories: 0-10m 10-30m and 30m+, as these align with natural stratification and allowed for a more ecologically meaningful interpretation of the data. The 0-10m range represents the epilimnion layer where surface mixing, light penetration and biological activity like algal blooms are the highest. This layer is critical for nutrient cycling as it interacts directly with atmospheric inputs. The 10-30m range captures the metalimnion, where temperature gradients influence dissolved oxygen and nutrient availability and the 30m+ range corresponds to the hypolimnion a more stable layer with limited mixing, lower temperatures and reduced light penetration [12, 13].

Figure 1 shows the seasonal decomposition of the Dissolved Oxygen (DO) in the surface depth group (0-10m) at the WG location from January 2014 to December 2023. DO is a key variable in quantifying water quality and must be between 6.5 & 8 mg/L to support aquatic wildlife [14]. There is a high negative correlation between DO and nutrients present in pesticides such as Phosphate (PO4), at -0.79 and Nitrogen (NT) at -0.73, corroborating the current understanding that pesticides are harming the lake's water quality (see figure 6 in appendix A). The seasonal decomposition indicates a modest downward trend in DO at the 0-10m depth, although there are two sharp dips towards the lower limit for healthy water (6.5 mg/L). The issues inferred in this paragraph from the seasonal decomposition of DO alone exemplify the need for policies and infrastructure to prevent the worsening of the lake's water quality. Our modelling hopes to motivate the implementation of such policies and infrastructure.

Due to logistical constraints, ecological relevance and resource limitations, there are an extensive number of missing entries in the data. Most notably, there were large gaps (at least 12 consecutive months) between 2015 and 2018 that can be seen for many key variables such as Turbidity and nutrient concentrations. These extensive gaps pose significant challenges to effective data imputation (see figure 8 in Appendix A). This in turn affects our ability to make accurate future

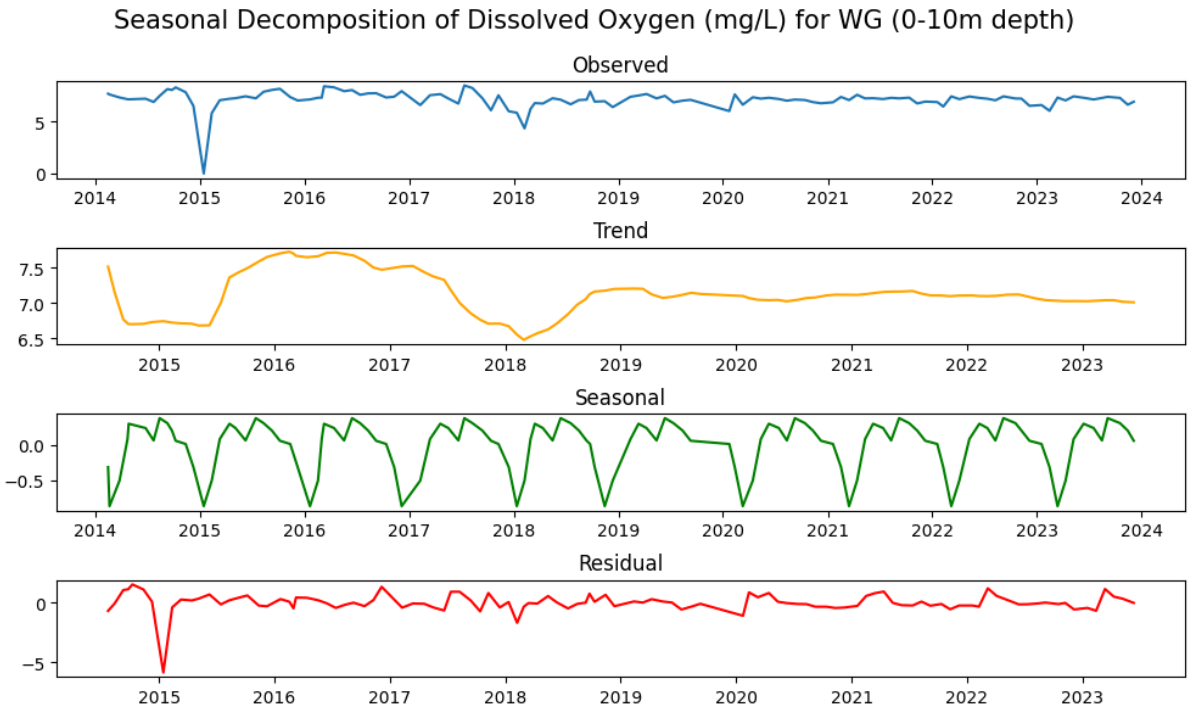

Figure 1: Seasonal Decomposition of DO at the 0-10m depth group. There is an anomaly in January 2015 showing an unrealistic dip to 0 mg/L, which is removed as an outlier in our final dataset. The variable shows clear seasonality and a slightly negative trend that dips towards the lower limit (6.5 mg/L) for water to harbour aquatic life.

predictions. For example, the LSTM model is incompatible with missing data points as they invalidate the model's calculations, memory states and back-propagation. Therefore, imputing the missing data whilst maintaining the overall trend and seasonality of the variables is necessary, and with the nonlinearity present and the size of the gaps in data, linear interpolation will not suffice. Instead, two different interpolation techniques are implemented and compared. These are using singular value decomposition and a feature-specific processing (FSP) method.

3.1 Technique 1 - Singular Value Decomposition

Singular Value Decomposition (SVD) can be used to impute the data due to its ability to capture underlying patterns, its computational speed and its accuracy when compared with other data imputation models [15]. Furthermore, with the presence of many large gaps in data, the ability of SVD to impute data using information from the dataset as a whole, instead of isolating each variable, was seen as advantageous. SVD was implemented by replacing missing entries with the feature-wise mean and applying a low-rank approximation matrix X_k to fill the missing entries [16], where $k = 10$ in our model, reapplying and iterating this process until convergence. $k = 10$ was used as it was found that of the 13 features needing imputation, 10 principle components captured 90% of the variation in the data, as found by performing Principle Component Analysis (PCA) [17] on the raw data (see figure 7 in appendix A). 90% was used as the percentage of retained variation as this value captured the relevant characteristics of the data without the imputed values regressing to the mean.

The aforementioned large gaps present in the data before 2018 for 7 out of 13 features, combined with the need for SVD to treat the data as one matrix initially resulted in unrealistic imputation

(see figure 8 in the appendix). Imputing the raw data while it still contained large gaps would be detrimental to our modelling outcomes if our data did not accurately represent reality. Therefore, SVD processes were reapplied to the data available between 2018 to the end of 2023, as all features had good coverage across this period. SVD imputation of Turbidity can be seen in more detail in figure 9 of the appendix and can be seen for Chlorophyll-a (Chl-a) in figure 2.

3.2 Technique 2 - Feature-Specific Processing (FSP)

An alternative approach to SVD is to adapt the interpolation and imputation technique used based on the feature being processed. This allows the implementation of domain knowledge in choosing the technique used as well as the ability to address unique data characteristics while ensuring spatial and temporal consistency across features. With limnological data, this is particularly important as features vary massively in how they are measured and collected, as well as in how they vary with depth. FSP creates a complete, standardised dataset with the same temporal and spatial coverage for all features.

Limnological data for a specific feature at a specific location is extracted from the raw data. A measurement matrix is formed representing all measurements, with dates as columns and depths as rows. A standardised depth grid is formulated, where the spacing between depths is determined by the data collection methodology. Distance between measurements increased at deeper levels, where a space of 1m was used between 0-50m and and unit spacing of 10m was used for 50m+. Then, missing values between the maximum and minimum depths (to prevent extrapolation) were interpolated, using a feature-specific interpolation method. The unique set of dates associated with a feature was then cross-checked with the unique set of dates of every other feature, and 'missing' dates were added as a column of empty values. This ensured uniform temporal coverage across each feature at a defined location. Finally, k-nearest-neighbours (knn) is used to perform horizontal imputation across time to fill in any missing dates, where the standard value of $k = 5$ was chosen [18].

Temperature, Chl-a, pH and DO followed the method mentioned above as the data was collected consistently across different depths and regularly across time, and linear interpolation was chosen to interpolate across depth [19]. Turbidity and nutrient concentration lacked any viable data before 11/01/2018 so the processing only was applied to data after this date, with any data before this date discarded. Similarly, processing for Total Dissolved Solids (TDS) was limited to post-12/03/2019 due to extreme data sparsity and linear interpolation was used to impute missing values at different depths. One feature, Nitrogen, had to be excluded completely due to insufficient data. Finally, Secchi and Biochemical Oxygen Demand (BDO) were treated as surface values and then measurements were replicated across all depth groups. KNN imputation was implemented on all features in the same way.

The final dataset was formed by combining all the processed data for each feature at each location into on large dataset with the the columns being the features and the rows representing all feature values at a specific location, within a specific depth group on a specific date.

3.3 Comparison of Imputation Techniques

Figure 2 compares the raw and processed data of Chl-a at WG in the 0-10m depth group. The processed datasets are compared to the raw data which was grouped into depth groups by averaging values that fall with each group. Both methods proved effective in imputing missing values, with effective imputation by the FSP method seen in late 2015. Both SVD and FSP lead to similar but slightly differing processed datasets. The first discrepancy can be seen in that

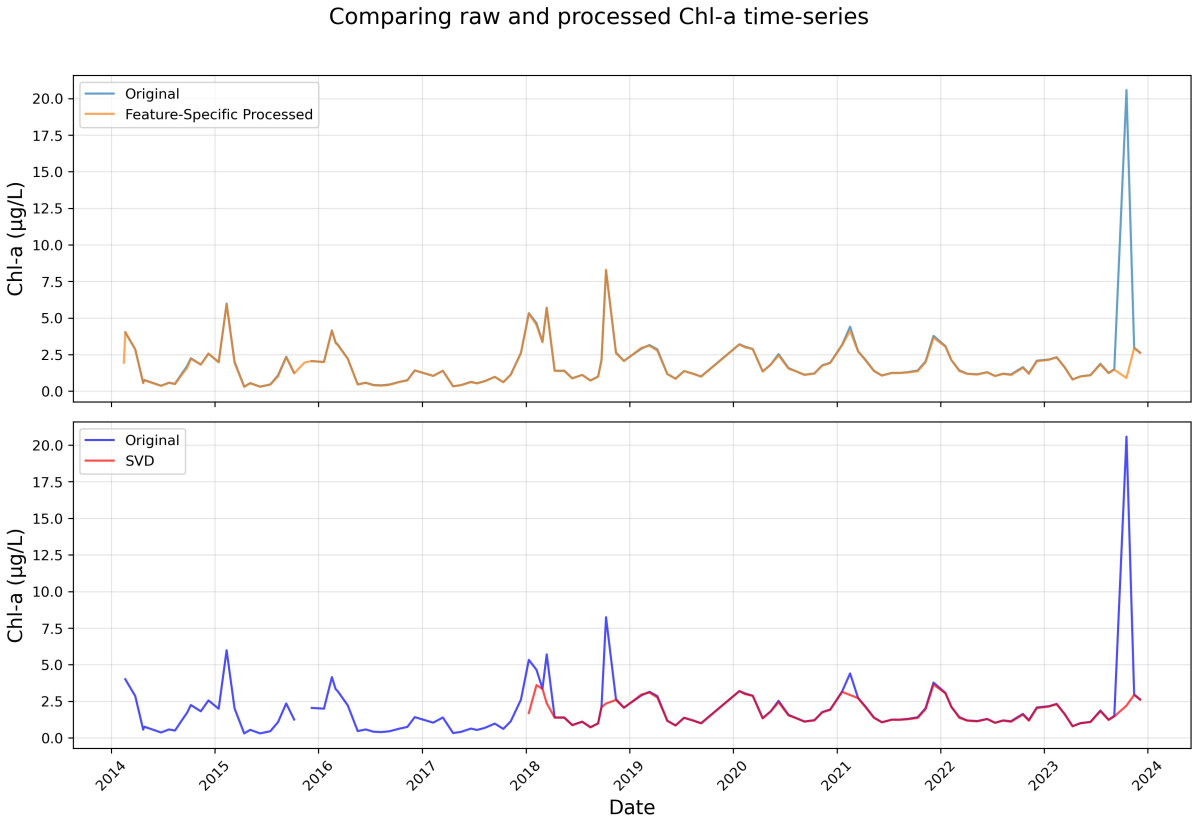

Figure 2: Comparison of Chl-a concentrations (0-10m) at the WG location between the original dataset and two imputation methods. The top panel shows the FSP dataset, which handles each variable independently, while the bottom panel shows the SVD (Singular Value Decomposition) method, which considers correlations between variables. Both methods preserve the general temporal patterns while filling gaps in the original measurements.

SVD removes all data before 2018, whereas the FSP maintains this data for certain features like Chl-a.

Another discrepancy can be seen in the differing outlier removal methods as SVD removed outliers in late 2018 and early 2021 that FSP did not. Potential outliers are present within the data and have to be treated carefully. The exact reasons for these outliers are unknown, such as an improbably large spike in Chl-a in late 2023. These events could be errors during data collection or entry, however, they also could be 'black swan' events of extreme environmental deviation and would have had severe impacts on the environment of the lake. In the SVD method, outliers were removed if their z-score was greater than 2. A more effective method was utilised in the FSP method, where time-series plots of each feature were manually examined, and clear outliers were removed individually. This massively reduced the misclassification of outliers seen in the SVD method.

To compare the accuracy of the two processing techniques, 5% of the data points were randomly removed. After processing, the original removed values were compared to the new imputed values, and the root mean squared error for each feature was found. Each feature's RMSE was normalised by dividing by the feature's mean to maintain consistency and allow for meaningful comparison. This was repeated 10 times with the mean RMSE and variance calculated. The results from this can be observed in table 1. The FSP method had a lower mean RMSE than

Imputation Method	Mean RMSE	RMSE Variance
SVD	0.6365	0.0022
Feature Specific	0.4413	0.0087

Table 1: Comparison of accuracy of data imputation methods using imputation validation. The RMSE has been normalised by column and one RMSE value was calculated over all features, as with the variance. The validation was run 10 times with the mean RMSE and the variance over the 10 runs making up the content of the table.

SVD but a higher variation. The lower mean RMSE is likely due to the adaptive nature of the feature-specific imputation method, in that it chooses the most appropriate method to impute each feature. Conversely, the variance of the SVD imputation may have been lower as it incorporates information from all variables to be able to fill in missing data with consistent accuracy.

The combination of a more effective outlier removal methodology, the inclusion of pre-2018 data for certain variables, and the lower mean RMSE associated with the FSP method when compared to the SVD method, was significant enough to conclude that the FSP dataset should used for all modelling throughout the rest of the report.

4 Time-Series Forecasting of Features

4.1 SARIMA

SARIMA is well-suited for modelling water quality variables such as temperature, which are governed by climatic seasonality. Previous studies such as those by Ghaemi et al. and Gupta et al. have demonstrated SARIMA's efficacy in capturing nonlinear dynamics in similar ecological datasets [20, 21]. For Turbidity, which shows no clear seasonality, SARIMA was used to assess whether any underlying seasonality could be detected despite its inherently irregular patterns. SARIMA's parameters take the form of $(P, D, Q)(p, d, q)(m)$, as explained by Liu et al [22]. The modelling used temperature and turbidity values from the final dataset described in Section 3.

A grid search approach was employed to identify optimal choices for the remaining model parameters. Other methods such as auto-ARIMA, or Bayesian optimisation could have been used for parameter selection, however, a grid search method was preferred due to its systematic approach and exhaustive evaluation [23]. The range of hyperparameter values for the grid search was selected to balance accuracy and computational efficiency. The chosen ranges for parameters are as follows: For Non-Seasonal Orders (p, d, q): $p, q \in [0, 2]$, $d \in [0, 1]$. For Seasonal Orders (P, D, Q): $P, Q \in [0, 2]$, $D \in [0, 1]$.For the Seasonal Period (m): $m = 12$ to reflect monthly seasonality common in ecological datasets (see Appendix B).

Limiting the SARIMA model parameters to small values helps maintain computational feasibility by reducing the number of parameter combinations. This approach is often sufficient for modelling real-world time series data, as higher-order parameters can lead to over-fitting without significant improvements in model performance [24].

A 5-fold time series cross-validation was first performed on 80% of the data, partitioned chronologically. The model was trained on earlier subsets and validated on subsequent ones in each fold, and the root mean squared error (RMSE), averaged across folds, was chosen as the primary metric for model selection. This procedure provided a robust, initial measure of forecasting performance on data not used for fitting. After choosing the best hyperparameters, a 20% holdout

test set that had never been used in either training or cross-validation was tested. The model (with its chosen hyperparameters) was retrained on the entire training pool and evaluated exactly once on this holdout test set. This additional step ensured an unbiased assessment of how well the final model generalises to completely unseen data. RMSE was favoured over metrics like the Akaike Information Criterion (AIC) because it penalizes large forecast errors more heavily, which is crucial for capturing time-series dynamics [25]. Once the final model's performance was verified on the holdout set, it was retrained on the full dataset.

4.2 LSTM

Variables that do not display clear seasonality can be difficult to model for SARIMA because they lack the strong predetermined periodicity that the model uses to form its predictions. Additional modelling complexity may be required to capture underlying patterns and project into the future for more nonlinear variables. The Long Short Term Memory (LSTM) Recurrent Neural Network model emerges as a well-suited candidate for this task as, having originally been presented with applications in sentiment analysis and music composition [26], the model's architecture allows it to store and analyse both long and short term patterns in data and use this stored information to form future predictions [27]. The ability to memorise long and short-term patterns simultaneously is particularly important to our goal of modelling future lake water quality due to the interdependence of many variables over time in the data, as shown in the correlation matrix in figure 6 in the appendix. For example, nutrient concentrations often have a delayed effect on future values of Chl-a and DO due to the biological processes they can trigger [28]. Furthermore, it has been found that time series forecasting with LSTM tends to outperform traditional forecasting models such as the ARIMA model [29], making LSTM a sensible choice for forecasting lake water quality variables into the future and trying to improve on the forecasting accuracy of the SARIMA model.

The LSTM model was applied at the WG location to model the values of Temperature and Turbidity. These variables were chosen because of their inherent differences in seasonality. Also, Temperature was taken over the whole timescale from the original data, as it had no missing data and turbidity was taken only from 2018, presenting potential caveats to the forecasting power of the LSTM model. The activation function used in the output layer of the model was the linear activation function, as this allows the unscaled numerical outputs needed for forecasting real-world data. A grid search was performed to optimise over the number of epochs (range of 100 to 400 moving in intervals of 100), learning rate (log range of 0.01 to 0.0001) and number of neurons (range of 32 to 128 in powers of 2), using ranges that aligned with typical empirical ranges for hyperparameter tuning of LSTM models [30]. The inclusion of dropout and batch size into the grid search was also considered, but this was not pursued due to computational and time constraints. Instead, the batch size was assumed to be 24 as this represented 24 months or two years, allowing the model to see two full seasonal cycles at any one time. A dropout value of 0 was used as, due to our data only being 69 data points long (for the WG location), it could not be justified to limit the learning power of our model with such a short dataset. The data was split into train, validation and test subsets in a 70:10:20 percentage ratio, training the data on the train set, validating the optimal choice of hyperparameter from the grid search using the validation set, and evaluating our predictions on the test set. In contrast to the SARIMA model, this method is an alternative to the 5-fold time series cross-validation to reduce processing time.

Model	Variable	Depth Group	Optimal Hyperparameters	MAE	MSE
SARIMA	Turbidity	0-10m	[1, 1, 1] [0, 0, 0, 12]	0.1087	0.0161
SARIMA	Turbidity	10-30m	[0, 1, 1] [0, 0, 0, 12]	0.2419	0.0680
SARIMA	Turbidity	30m+	[1, 1, 1] [0, 0, 0, 12]	0.0355	0.0019
LSTM	Turbidity	0-10m	[128, 0.001, 100]	0.1651	0.0350
LSTM	Turbidity	10-30m	[64, 0.01, 300]	0.1889	0.0357
LSTM	Turbidity	30+m	[32, 0.001, 200]	0.1737	0.0427
SARIMA	Temp.	0-10m	[0, 0, 0] [0, 1, 1, 12]	0.3508	0.1931
SARIMA	Temp.	10-30m	[0, 0, 1] [0, 1, 1, 12]	0.3150	0.1693
SARIMA	Temp.	30m+	[1, 0, 0] [0, 1, 0, 12]	0.0533	0.0039
LSTM	Temp.	0-10m	[64, 0.01, 300]	0.1272	0.0239
LSTM	Temp.	10-30m	[64, 0.01, 300]	0.1001	0.0160
LSTM	Temp.	30+m	[32, 0.001, 200]	0.0908	0.0111

Table 2: Results of the SARIMA and LSTM models Mean Absolute Error (MAE) and Mean Squared Error (MSE) after grid search hyperparameter optimisation and test set forecast. Each model is evaluated on the test and train subsets of the Turbidity and Temperature variables from the FSP dataset. The Optimal Hyperparameters column for the SARIMA model shows non-seasonal and seasonal values respectively as [p,d,q] [P,D,Q,m]. For the LSTM model, this column shows the grid search-optimised values for [Neurons, Learning Rate, Epochs] from the validation set, as well as the MAE and MSE of test set predictions.

4.3 Results and Comparison

Table 2 shows the results of predicting the unseen 20% test subset of the data for the temperature and turbidity variables. The test set has been used to evaluate the future predicting power of both models as it was completely unseen in training. MAE and MSE were used in conjunction to evaluate model performance, as MAE is an indication of absolute error, whereas MSE indicates outliers in the models' predictions, allowing us to detect cases of over-fitting. SARIMA and LSTM exhibit differing performance profiles across the Turbidity and Temperature variables at varying depths. For Turbidity, SARIMA often achieves lower MAE and MSE—particularly at 0–10m depth, yet the time-series plots (Figure 3) reveal a tendency to revert to near-mean forecasts. This is due to the lack of seasonal parameters ($[P, D, Q, m] = [0, 0, 0, 12]$ for all depths), meaning that SARIMA was not able to capture the seasonality of the variable because the variable did not exhibit it, compromising the model's ability to perform future predictions while seemingly giving favourable MAE and MSE results. By contrast, LSTM captures these abrupt variations more responsively, though it may overshoot on occasion, a result of over-fitting to previous patterns, leading to slightly higher average errors. Furthermore, the short timescale of the Turbidity variable, which is the case due to missing data, is likely to have limited the model's forecasting power. There were only 69 data points for the WG location, 30% of which were reserved for validation and test sets. It could be argued that this is an insufficient amount of data to train an LSTM model to accurately make future predictions, a hypothesis that is backed up by a study on how training data size affects LSTM model performance for rainfall-runoff modelling by Boulmaiz et al., who found that LSTM's predictive accuracy improves with longer datasets to train on [31]. In real-world applications, such as detecting sudden increases in turbidity, an approach that captures the qualitative trend of the data will likely be preferable over a model with lower overall MAE or MSE, suggesting that the LSTM model would be chosen over the SARIMA model for forecasting variables that do not exhibit seasonality.

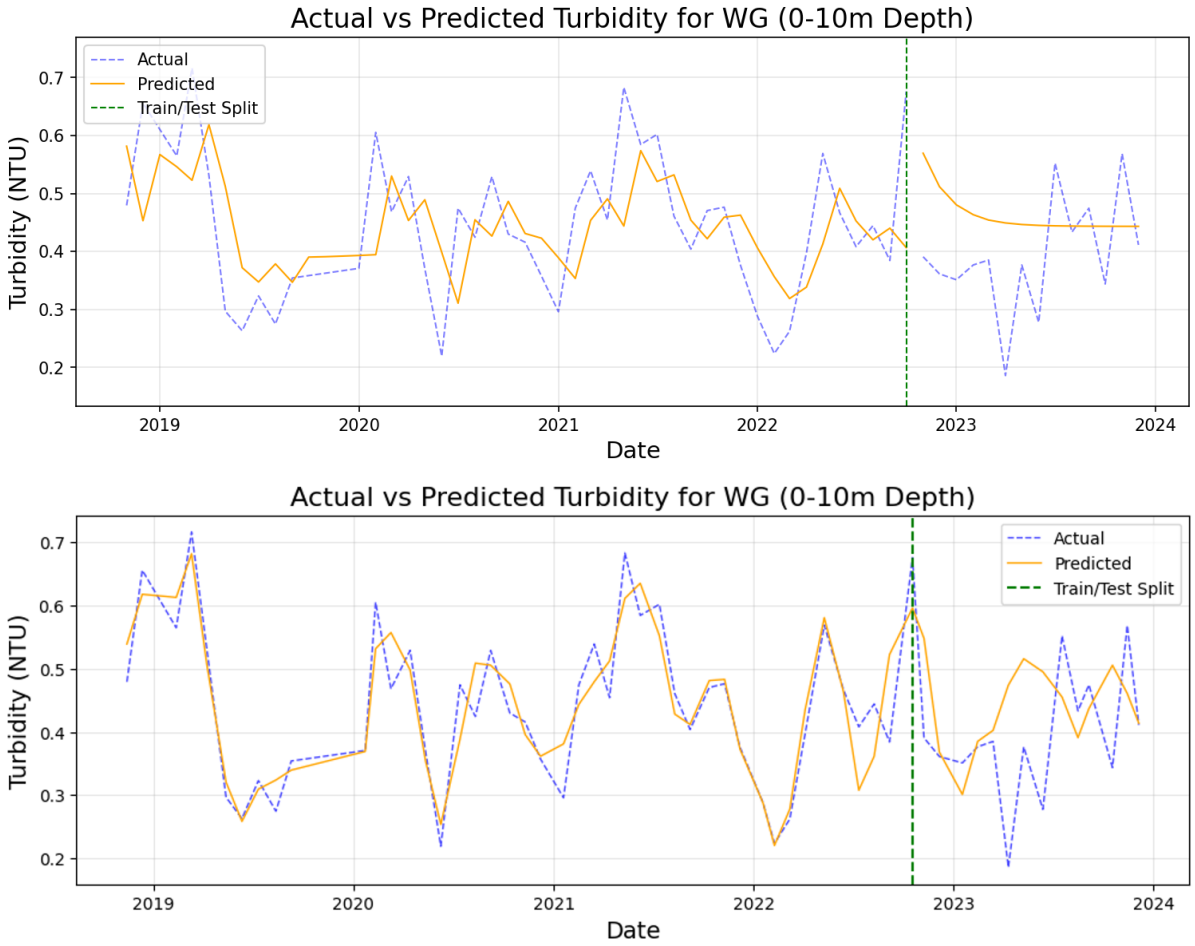

Figure 3: Comparison of SARIMA (top) versus LSTM (bottom) predictions of monthly turbidity at the WG location (0–10m depth). In both plots, the dashed blue line indicates actual turbidity observations, while the solid line shows the model's predicted turbidity and the green dashed vertical line marks the train–test split date. The top chart illustrates how the SARIMA model primarily tracks long-term patterns, whereas the bottom chart shows LSTM predictions that more closely follow short-term fluctuations.

In contrast, for Temperature, both SARIMA and LSTM produce visually similar forecasts (see Appendix B) because temperature typically displays a strong annual cycle. Here, each model's seasonal structure or memory mechanisms can capture the main peaks and troughs effectively. Since the forecasts for Temperature are largely comparable in their shape and timing of major fluctuations, comparing MSE and MAE becomes a sufficient way to determine which model is better for prediction accuracy. Notably, at shallow depths (0–10m and 10–30m), the LSTM model achieves lower MAE and MSE compared to SARIMA, indicating that it handles variations to seasonality more accurately. However, for 30m+ depth, where temperature variation is naturally more limited, SARIMA outperforms LSTM, posting lower errors in both metrics (MAE and MSE). This likely reflects the reduced dynamic range of temperature at deeper levels (i.e. smaller shifts from minimum to maximum), which SARIMA's simpler structure can model effectively. As a result, while LSTM appears superior for shallow depths with greater temperature swings, SARIMA remains sufficient or even preferable where the water column experiences only mild temperature fluctuations.

5 Using Machine Learning Models to capture relationships between variables

Dissolved oxygen and chlorophyll are very important to the quality of Lake Atitlán due to the effect of cyanobacterial blooms on these features. In this section, machine learning approaches are utilised to predict the values of these features. This allows DO and Chl-a to be predicted into the future, using the predictions of other features as found by time-series forecasting. The two models used to do this are Random Forest and Gradient boosting. Both these methods are powerful tools suitable to the domain of limnological data as they handle non-linear relationships and interactions between environmental variables well.

Both approaches involve taking the pre-processed and depth-aggregated data. Then, the data is temporally sorted which prevents data leakage. A 70-30 train-test split was used for both models. Specific variables are selected as input features based on their ecological relevance as well as their completeness in the data set. In the case of the processed limnology data set, six variables were selected as input features for both models. These were: Temperature, Chl-a, pH, DO, Secchi and BDO as these had no data missing, allowing for the maximum time frame of data to be used. For both models, a time series split was implemented to maintain temporal dependences during cross-validation, which works by training the model on earlier periods and validating it on subsequent periods.

5.1 Random Forests

The Random Forest algorithm is a powerful modelling tool for the prediction of lake water quality [32]. In the Random Forest algorithm decision trees split data into decision regions based on feature thresholds. These trees capture complex and non-linear interactions between variables. Each tree in the ensemble attempts to minimise a chosen cost function. At each decision node, the feature and threshold are selected in such a way that the data is split into groups where this cost is minimised,

A grid search was used to tune the hyper-parameters to prevent under-fitting or over-fitting. The hyper-parameters tuned were the $n\_estimators$ (number of trees in the forest) and max_depth (maximum depth of each tree). The grid searched was nine combinations, with $n\_estimators$ of 100, 200, 300 and max_depth of 5, 10 and 15. MSE was selected as the cost function for optimisation. An optimised model is formed from the best hyper-parameters and applied to the test set to evaluate the model.

5.2 Gradient Boosting

A second method to predict the features of the lake is by using the method of gradient boosting. The Gradient Boosting model uses an ensemble of weak decision tree prediction models, each of which learns from the errors of the previous model as described by Otchere et al. [33].

The gradient boosting regressor in this paper is optimised by minimising the absolute error loss, as it is the most robust method. As gradient boosting is robust to over-fitting from the number of learners, and so a larger number of learners usually results in better performance, the number of estimators was set as constant at 100, which is high enough to maximise accuracy without overly impacting processing speed. The maximum depth of the regressor, limiting the number of nodes in the tree, is tuned to minimise the MSE using time series splits for validation, tested for values between 1 and 100 to cover the range of estimators.

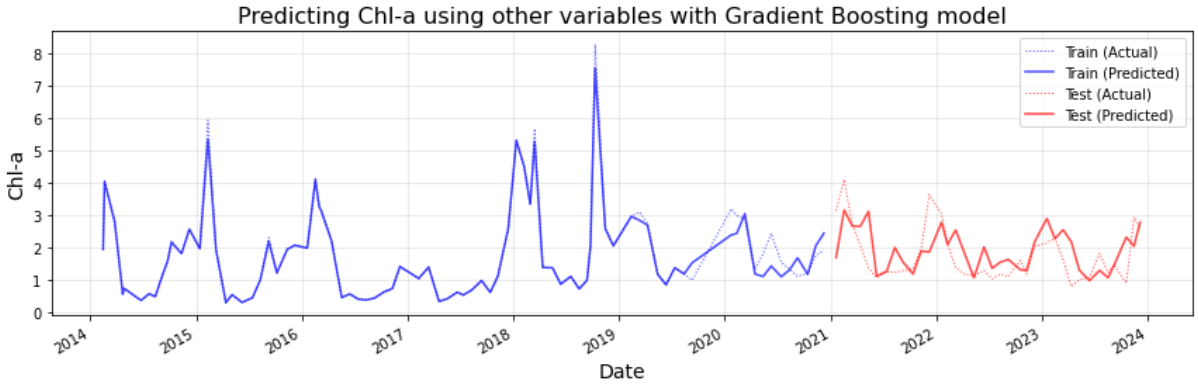

Figure 4: Gradient Boosting model predictions of Chl-a (μg/L) over time, using environmental parameters as predictors. The model was trained with a 70-30 time series split, shown in blue (training) and red (testing) respectively, with dotted lines representing actual values and solid lines showing predictions. Hyperparameter optimisation was performed using TimeSeriesSplit.

5.3 Results and Comparison

To test the two regression models' viability three metrics have been used: MSE, MAE, and the Nash-Sutcliffe Efficiency (NSE). For each location-depth group with sufficient data (at least 50 samples), the data was split into 70% training 30% testing whilst preserving temporal order. This ensures that the models are trained on historical data and then evaluated on unseen future data, which mimics real-world forecasting. Individual regression models were implemented for each depth group and feature. This supports scalability into alternative depth groups while preserving the utility of the model. Appendix C summarises the accuracy of the two models in estimating Chl-a and DO for different depth groups.

By comparing the MSE, it can be said that the RF model is more accurate for predicting DO between 0-10m, with a lower MSE of 0.2991 compared to gradient boosting which has 0.3893. MAE is also lower, while NSE is higher as expected for greater accuracy. Figure 5 shows the predicted values for DO, compared to the actual values. When predicting Chl-a at a depth of 0-10m, the RMSE for the gradient boosting model is slightly lower, at 0.6735, compared to 0.7197 for RF. Figure 4 shows the predicted values for Chl-a. The low MSE scores demonstrate the effectiveness of using regression to predict DO and Chl-a values, as they can be found more accurately using RF and gradient boosting models respectively than with time-series forecasting using predictions from other features.

6 Discussion and Conclusion

When testing imputation techniques, a feature-specific process works better than SVD for a variety of reasons. As well as producing a lower RMSE value when tested against removed data, the requirement of SVD to treat data as a single matrix limited the amount of data usable across all variables. Due to complications in data collection, several variables were almost entirely missing before 2018, and the more individualised pre-processing techniques allowed us to keep earlier data for certain features.

The SARIMA model is effective for predicting seasonal data, such as temperature, but has limitations when predicting data that shows a less seasonal trend and is outperformed by the LSTM model for temperature at shallower depths. The LSTM model proved more accurate

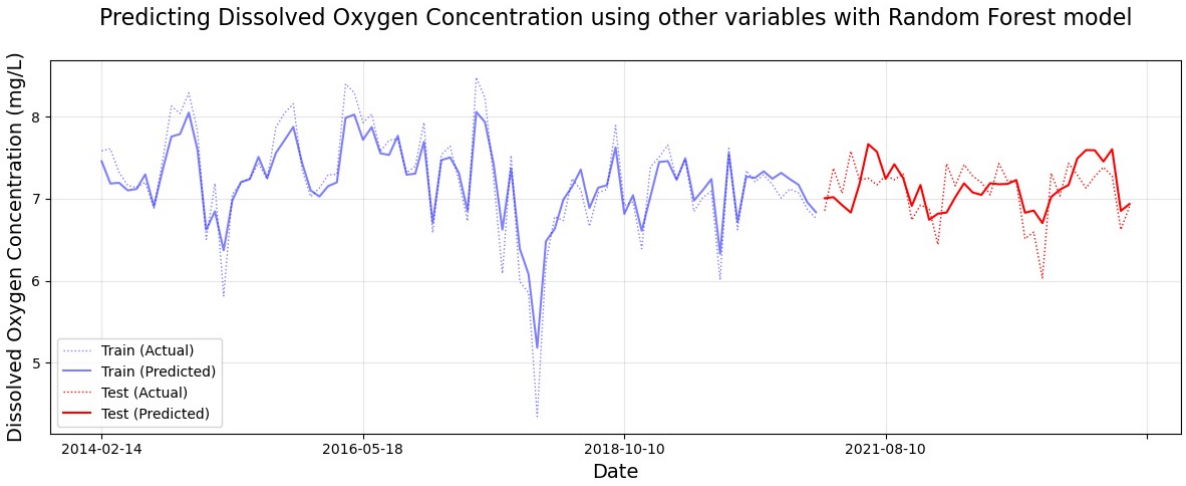

Figure 5: Random model predictions of DO (μg/L) over time, using environmental parameters as predictors. The model was trained with a 70-30 time series split, shown in blue (training) and red (testing) respectively, with dotted lines representing actual values and solid lines showing predictions. Hyperparameter optimisation was performed using TimeSeriesSplit cross-validation with GridSearchCV.

in predicting the non-seasonal turbidity, however, encountered some difficulties in over-fitting, resulting in worse MSE and MAE scores despite visually following the correct trend more closely than SARIMA. Further research could reduce the inaccuracies in these models. Implementing the multivariate SARIMAX model, which incorporates both seasonal patterns and exogenous variables, could improve the accuracy of the SARIMA model. Further research could be done in implementing a better grid search method for LSTM, free of time constraints, that includes batch size and drop-out. Further, a more robust test-train split using a method similar to the one used for the SARIMA model could reduce the errors present in the LSTM model.

One limitation applicable to all models in the report is that due to the short size of the data set, predictions can only be shown to be accurate in the short term. There is insufficient data to properly extrapolate long-term effects, such as climate change, and our test sets only represent up to three years to allow sufficient data for training (and less for some features, such as turbidity, with significant amounts of missing data). Because of this, long-term predictions will have a high level of uncertainty. This is something that is backed up by existing research, as discussed in the comparison of LSTM and SARIMA, and further work involving a larger dataset for training would improve the future forecasting ability of the models, particularly LSTM.

Gradient boosting and random forest models were both effective in predicting DO and Chl-a, displaying a high accuracy with low RMSE and MAE values, using just five other variables. Given this high accuracy, one potential avenue for further research could be the use of these models in removing outliers, as peaks in the data due to an environmental event would likely be mirrored in other variables, however, errors in data collection would not.

In conclusion, this report shows predictions can be made for the future climate of Lake Atitlán by implementing a feature-specific processing method for data imputation and using this data to train SARIMA and LSTM models. A regression algorithm, such as the random forest model, can then be used to extrapolate these predictions to other variables, such as dissolved oxygen concentration.

A Data Preprocessing and Imputation

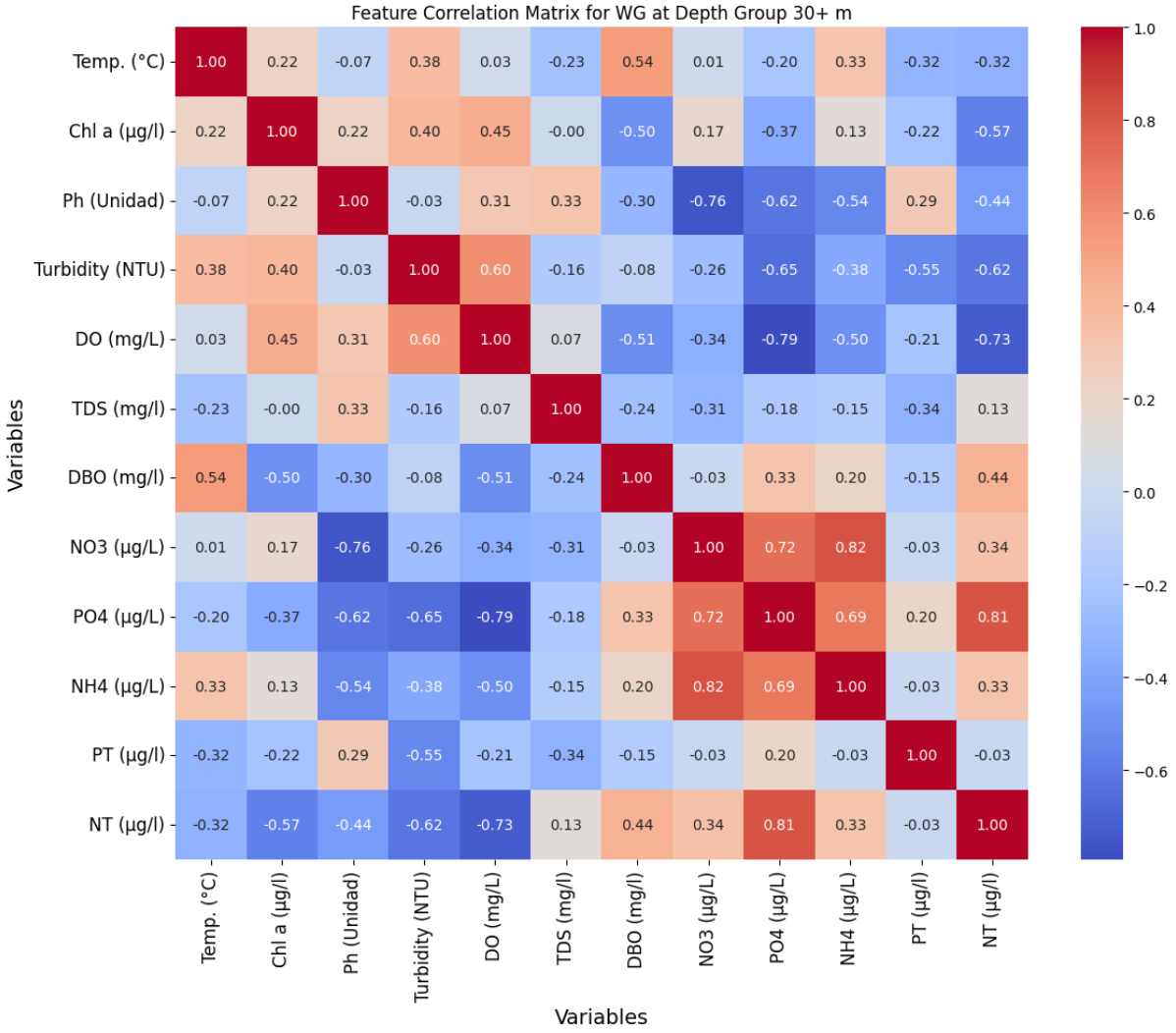

Figure 6: Correlation Matrix of all variables at the 30m+ depth group for the WG location. There is a high positive correlation between most of the nutrients (bottom right corner) and a high negative correlation between Turbidity & Dissolved Oxygen with nutrients such as Phosphate (PO4) and Nitrogen, both of which are key ingredients in the pesticides being washed into the lake from surrounding areas.

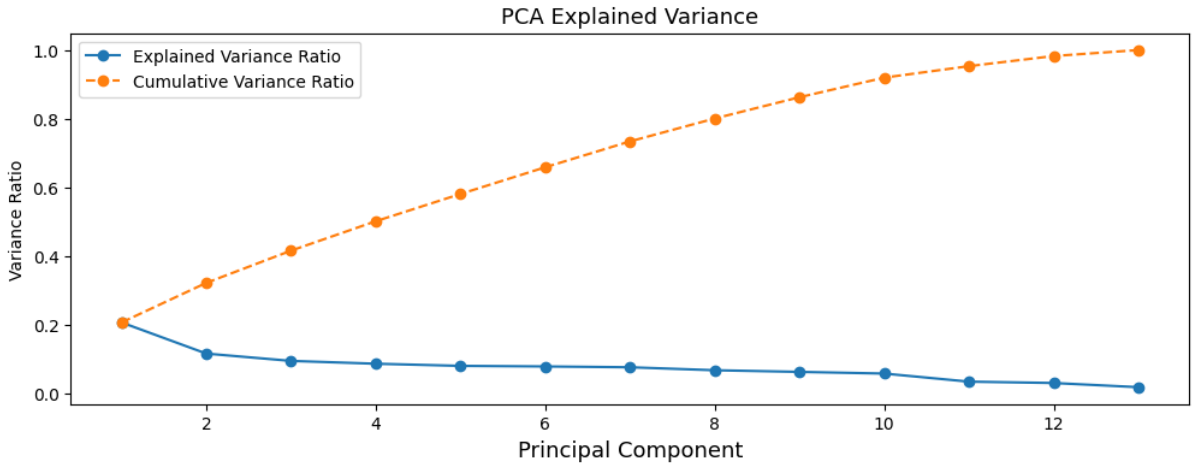

Figure 7: Principle Component Analysis (PCA) on the lake dataset. The cumulative variation shows that 90% of the variation in the data is expressed by the first 10 principal components.

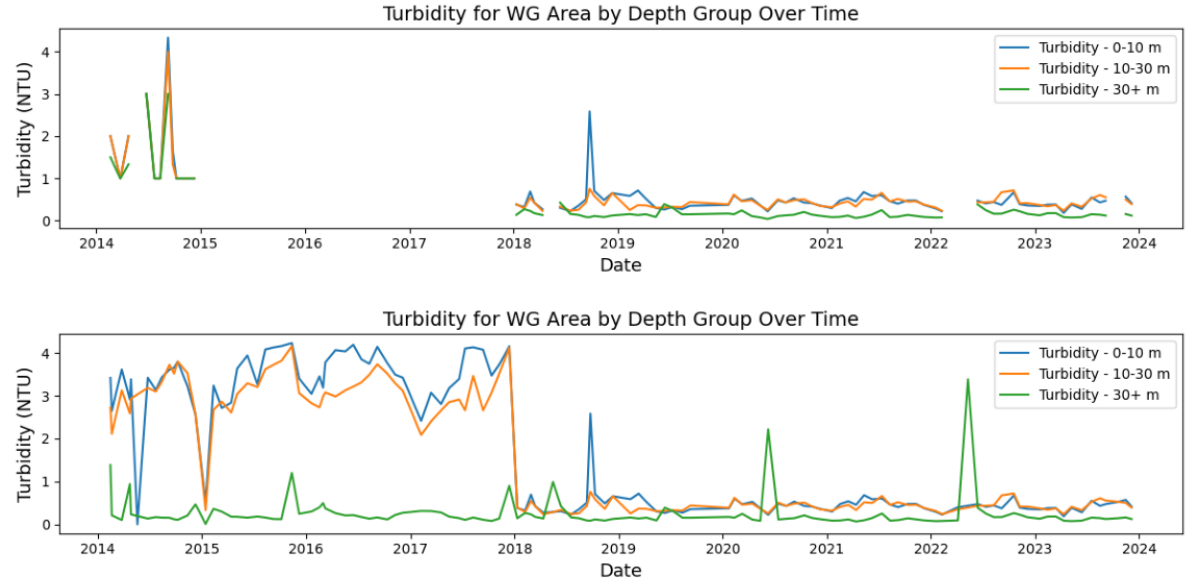

Figure 8: Values for Turbidity (a measure of water clarity and indicator of the quality of water) at the WG location for the three depth groups from January 2014 to December 2023 before and after (top and bottom respectively) the dataset had been processed using the techniques detailed above. There are many gaps in the data including a large gap between January 2015 and January 2018. Similar gaps in data were present for 7 out of 13 features at all three sampling stations, which has resulted in the imputed data for large gaps being inconsistent with the rest of the raw data and in turn inhibiting the removal of outliers, which is exemplified in this figure but also present in other variables with large gaps in raw data. This figure shows the shortfalls of SVD for imputing large gaps in data.

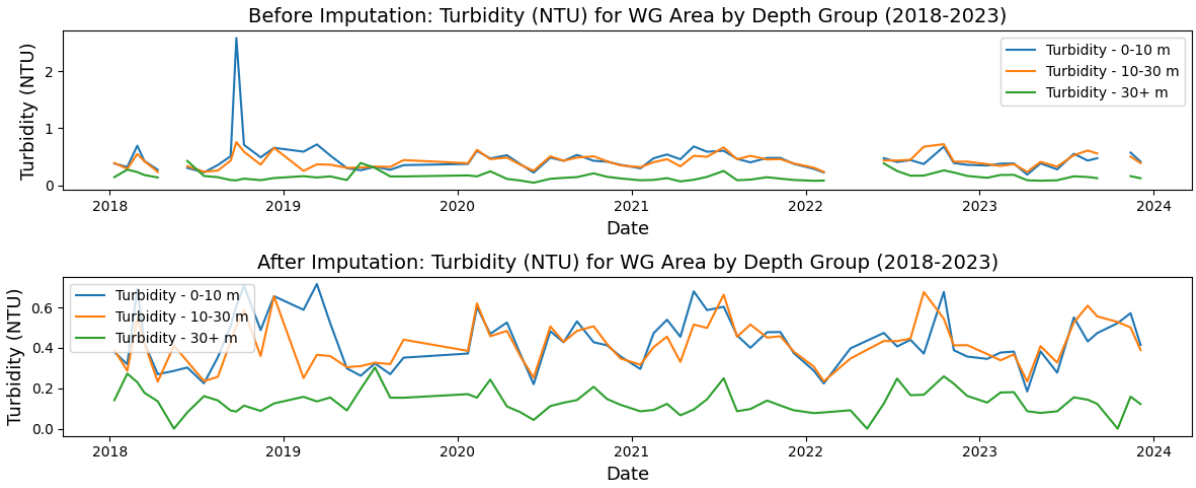

Figure 9: Two graphs for Turbidity from January 2018 to December 2023, showing before and after the implementation of the SVD imputation and outlier removal processes on the raw grouped data. The imputed data is more fitting with the trend of the surrounding data and less likely to spike uncharacteristically than in figure 8 due to the omission from the SVD process of periods prone to large data gaps. Further, removing outliers removed spikes present both in the raw data and the imputed data, a process that is essential before applying models such as LSTM to the data. The smaller scale of the imputed graph shows the trend of the Turbidity variable in more detail, hence the more varying lines than before imputation when the outliers obscured the trend of the data.

B Time Series Modelling

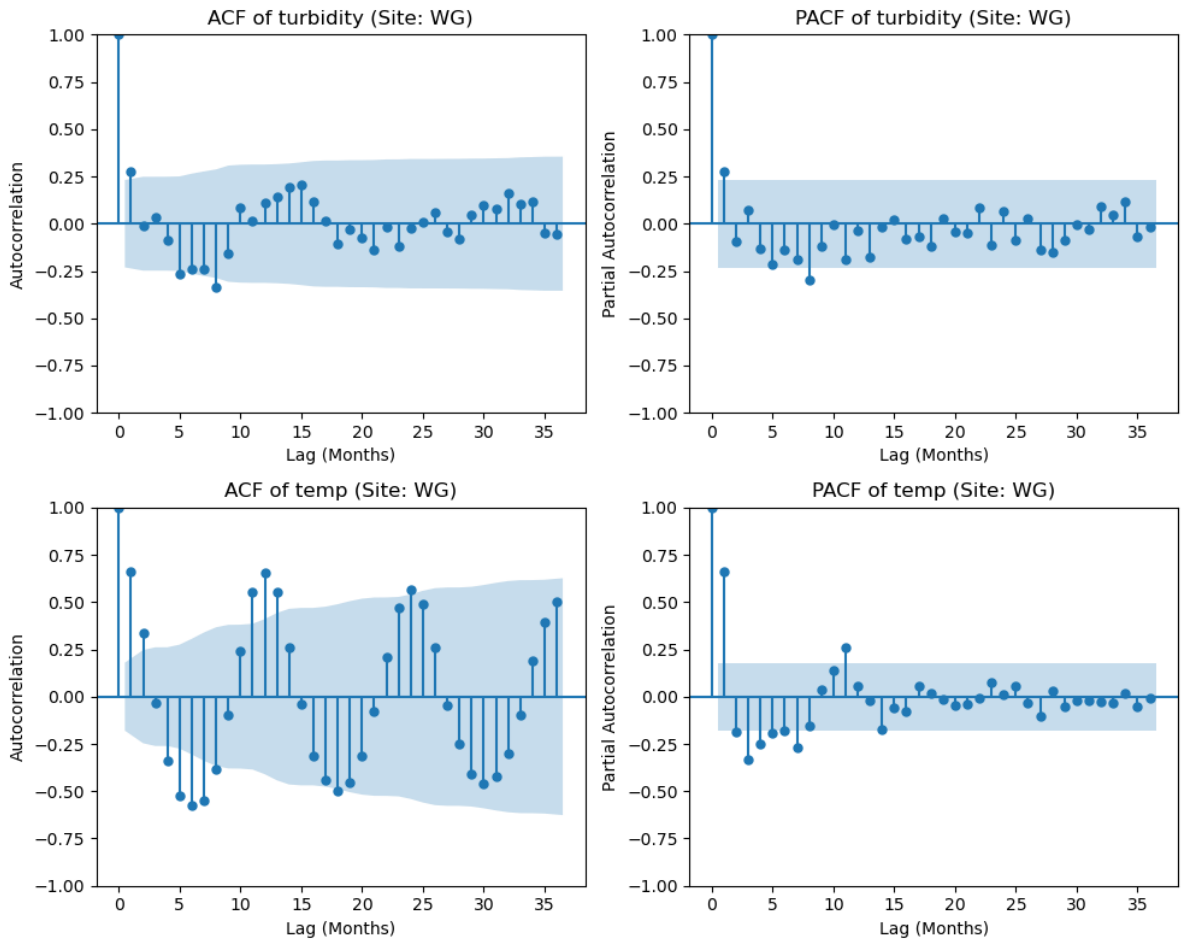

Figure 10: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for turbidity (top row) and temperature (bottom row) at the WG location. The ACF and PACF for turbidity do not exhibit strong seasonal peaks, indicating weak dependency on past values and no significant annual cycles. In contrast, the ACF and PACF for temperature show pronounced peaks at lags of 12, 24, and 36 months, confirming strong annual seasonality consistent with climatic cycles.

To choose the seasonal period m, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were generated. For temperature at the WG location, the ACF shows clear peaks at lags 12, 24, and 36 months—indicating strong annual seasonality ($m = 12$) - and the PACF has significant spikes at lags 1 and 12. In contrast, the ACF for turbidity lacks notable seasonal peaks and the PACF only shows weak short-lag dependency, suggesting it is more influenced by stochastic events. Nonetheless, $m = 12$ was chosen for turbidity for consistency and to test for any latent annual effects. This led to the hypothesis that SARIMA models would perform better for temperature, while turbidity might yield less accurate forecasts or revert to mean-level predictions.

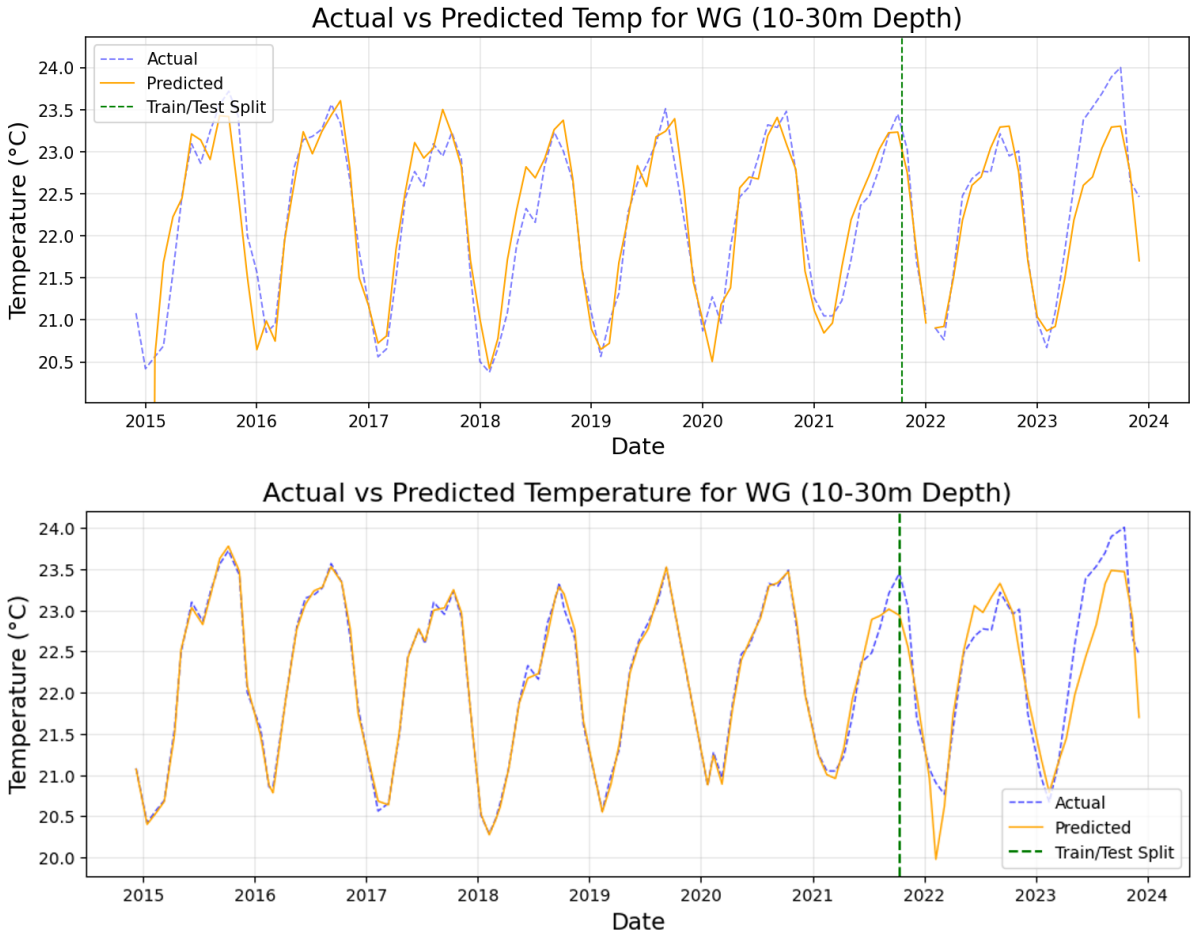

Figure 11: Comparison of SARIMA (top) versus LSTM (bottom) predictions of monthly temperature at WG location (10-30m depth). In both plots, the dashed blue line indicates actual temperature observations, while the solid line shows the model's predicted turbidity and the green dashed vertical line marks the train–test split date.

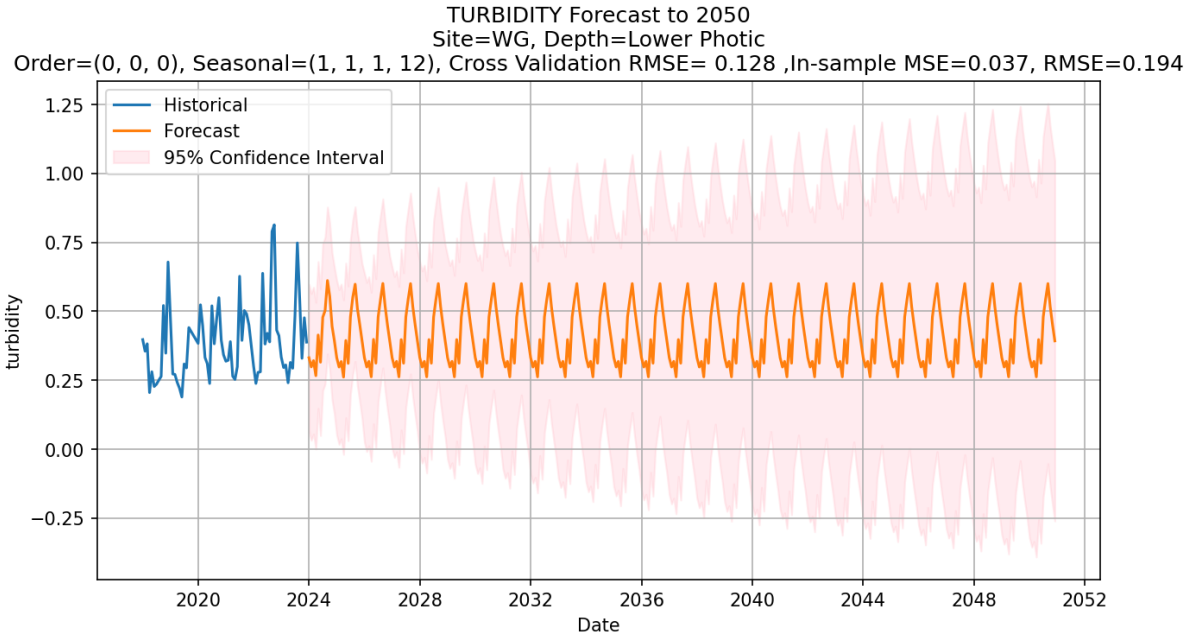

Figure 12: This figure presents the extended SARIMA forecast for turbidity at WG location, Lower-Photic depth (using the feature-specific dataset). Historical data (blue) extend until late 2023, after which the orange line denotes the predicted mean turbidity up to 2050, and the surrounding pink region shows the 95% confidence intervals. Notably, this model's best-fit parameters included a seasonal component (1,1,1,12), causing the forecast to oscillate with an annual cycle, an outcome that differs substantially from other WG location turbidity depths, whose solutions tended to flatten out at a near-constant value.

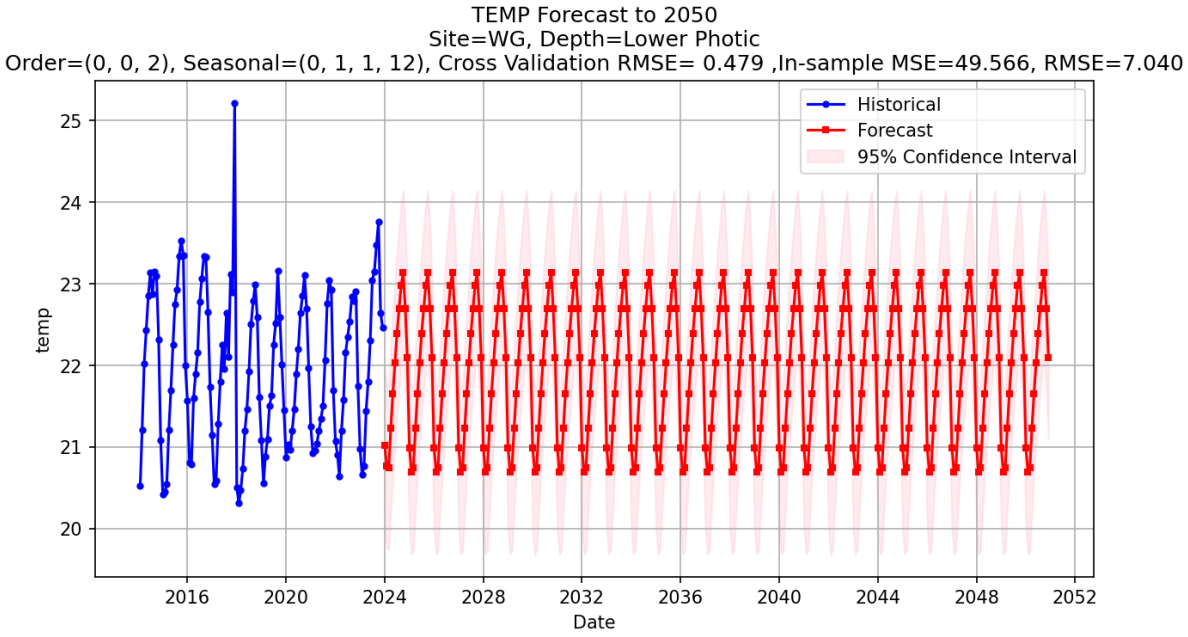

Figure 13: This figure illustrates the SARIMA forecast for temperature at the WG location, Lower-Photic depth (again from the feature-specific dataset). The model parameters are (0,0,2) for the non-seasonal part and (0,1,1,12) for the seasonal part, which leads to a pronounced sinusoidal pattern over future years. The solid red line denotes the monthly mean prediction out to 2050, while the pink ribbon indicates the 95% confidence region. Apart from an outlier spike in 2018–2019, the training data show moderate seasonal variation, which the forecast extends regularly into the future with repeating annual cycles.

C Regression Models

Model	Depth Group	RMSE	MAE	NSE	Best Hyper-Parameters
RF	0-10m	0.7197	0.55	0.22	{'max_depth': 10, 'n_estimators': 100}
GB	0-10m	0.6735	0.51	0.32	max_depth{14}
RF	10-30m	1.4296	0.94	-0.05	{'max_depth': 5, 'n_estimators': 100}
GB	10-30m	1.4129	0.86	-0.03	max_depth{32}
RF	30m+	0.4363	0.24	-0.27	{'max_depth': 5, 'n_estimators': 300}
GB	30m+	0.4668	0.29	-0.46	max_depth{40}

Table 3: A comparison of the two regression models, random forest (RF) and gradient boosting (GB) for predicting Chl-a, using the data from the WG location. The GB model has a higher accuracy, by its lower RMSE and MSE and an NSE value closer to 1 than the random forest model for both 0-10m and 10-30m.

Model	Depth	RMSE	MAE	NSE	Best hyperparameters
RF	0-10m	0.2991	0.2428	0.1670	{'max_depth': 15, 'n_estimators': 100}
DO	0-10m	0.3893	0.3000	-0.41	{'max_depth': 24}
RF	10-30m	0.3474	0.2664	-0.8460	{'max_depth': 5, 'n_estimators': 300}
DO	10-30m	0.3533	0.3000	-0.91	{'max_depth': 14}
RF	30m+	0.5414	0.4142	0.0683	{'max_depth': 15, 'n_estimators': 300}
DO	30m+	0.6679	0.5100	-0.42	{'max_depth': 84}

Table 4: A comparison of the two regression models, random forest (RF) and gradient boosting (GB) for predicting DO, using the data from the WG location. The RF model has a higher accuracy, by its lower RMSE and MSE and an NSE value closer to 1 than the random forest model for 0-10m, 10-30m and 30m+.

References

[1] Mengqi Liu Hannah Nichols Brien K. Ashdown, Meghan E. Brown and Isabel Urquiza. Indigenous community members' views about water quality in lake atitlán, guatemala. *Local Environment*, 27(1):32–45, 2022. doi: 10.1080/13549839.2021.2001447. URL https://doi.org/10.1080/13549839.2021.2001447.

[2] Margaret Dix Jaroslava Komárková Nancy Girón Eliška Rejmánková, Jiří Komárek. Cyanobacterial blooms in lake atitlan, guatemala. *Limnologica*, 41(4):296–302, 2011. ISSN 0075-9511. doi: https://doi.org/10.1016/j.limno.2010.12.003. URL https://www.sciencedirect.com/science/article/pii/S0075951110000800.

[3] Margaret Dix Nancy Girón Amber Roegner Jana Veselá Sudeep Chandra James J. Elser Jessica R. Corman, Emily Carlson and Eliška Rejmánková. Nutrient dynamics and phytoplankton resource limitation in a deep tropical mountain lake. *Inland Waters*, 5(4): 371–386, 2015. doi: 10.5268/IW-5.4.843. URL https://www.tandfonline.com/doi/abs/10.5268/IW-5.4.843.

[4] David Harper. *What is eutrophication?*, pages 1–28. Springer Netherlands, Dordrecht, 1992. ISBN 978-94-011-3082-0. doi: 10.1007/978-94-011-3082-0_1. URL https://doi.org/10.1007/978-94-011-3082-0_1.

[5] Mohamed M. Dorgham. *Effects of Eutrophication*, pages 29–44. Springer Netherlands, Dordrecht, 2014. ISBN 978-94-007-7814-6. doi: 10.1007/978-94-007-7814-6_3. URL https://doi.org/10.1007/978-94-007-7814-6_3.

[6] Neetu Gupta, Surendra Yadav, and Neha Chaudhary. Time series analysis and forecasting of water quality parameters along yamuna river in delhi. *Procedia Comput. Sci.*, 235(C): 3191–3206, July 2024. ISSN 1877-0509. doi: 10.1016/j.procs.2024.04.302. URL https://doi.org/10.1016/j.procs.2024.04.302.

[7] Jinxin Xu; Zhuoyue Wang; Xinjin Li; Zichao Li; Zhenglin Li. Prediction of daily climate using long short-term memory (lstm) model. *International Journal of Innovative Science and Research Technology (IJISRT)*, 2024. doi: https://doi.org/10.38124/ijisrt/IJISRT24JUL073. URL https://api.semanticscholar.org/CorpusID:208624849.

[8] Jun Liu, Tong Zhang, Guangjie Han, and Yu Gou. Td-lstm: Temporal dependence-based lstm networks for marine temperature prediction. *Sensors (Basel, Switzerland)*, 18(11):E3797, November 2018. ISSN 1424-8220. doi: 10.3390/s18113797. URL https://europepmc.org/articles/PMC6263690.

[9] Md Sahidul Islam, Hailong Yin, and Mustafizur Rahman. Long-term trend prediction of surface water quality of two main river basins of china using machine learning method. *Procedia Computer Science*, 236:257–264, 2024. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2024.05.029. URL https://www.sciencedirect.com/science/article/pii/S1877050924010457. International Symposium on Green Technologies and Applications (ISGTA'2023).

[10] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad Ali Shah, Rabia Irfan, and José García-Nieto. Efficient water quality prediction using supervised machine learning. *Water*, 2019. URL https://api.semanticscholar.org/CorpusID:208624849.

[11] Dennis Trolle, Anders Nielsen, Jonas Rolighed, Hans Thodsen, Hans E. Andersen, Ida Bjørnholt Karlsson, Jens Christian Refsgaard, Jørgen Eivind Olesen, Karsten Bolding, Brian Kronvang, Martin Søndergaard, and Erik Jeppesen. Projecting the future ecological state of lakes in denmark in a 6 degree warming scenario. *Climate Research*, 64:55–72, 2015. URL https://api.semanticscholar.org/CorpusID:55098801.

[12] A. Gavin, S. Nelson, J. Saros, M. SanClements, and I. Fernandez. Depth moderates doc impact on cold-water refugia in small, northern temperate lakes. *Water Resources Research*, 59, 2023. doi: 10.1029/2022wr033430.

[13] Y. Zhang, Z. Wu, M. Liu, J. He, K. Shi, M. Wang, and Z. Yu. Thermal structure and response to long-term climatic changes in lake qiandaohu, a deep subtropical reservoir in china. *Limnology and Oceanography*, 59:1193–1202, 2014. doi: 10.4319/lo.2014.59.4.1193.

[14] Datastream.org. Dissolved oxygen, 2025. URL https://datastream.org/en-ca/guidebook/dissolved-oxygen-do#:~:text=Healthy%20water%20should%20generally%20have%20dissolved%20oxygen%20concentrations,aquatic%20animals%20depend%20on%20this%20oxygen%20to%20breathe. Accessed: 2025-02.

[15] P. et al. Nguyen. Faster imputation using singular value decomposition for sparse data. In *Proceedings of the International Conference on Artificial Intelligence and Data Science*, pages 135–147. Springer, 2023. URL https://link.springer.com/chapter/10.1007/978-981-99-5834-4_11.

[16] Carl Eckart Gale Young. The approximation of one matrix by another of lower rank. *Springer*, 1:211–218, 1936. URL https://link.springer.com/article/10.1007/BF02288367#citeas.

[17] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. ISSN 0169-7439. doi: https://doi.org/10.1016/0169-7439(87)80084-9. URL https://www.sciencedirect.com/science/article/pii/0169743987800849.

[18] Ramu Gautam and Shahram Latifi. Comparison of simple missing data imputation techniques for numerical and categorical datasets. *Journal of Research in Engineering and Applied Sciences*, 8(1):468–475, 2023. doi: 10.46565/jreas.202381468-475. URL https://qtanalytics.in/journals/index.php/JREAS/article/view/1821.

[19] Ryan P. North and David M. Livingstone. Comparison of linear and cubic spline methods of interpolating lake water column profiles. *Limnology and Oceanography: Methods*, 11: 213–224, 2013. doi: 10.4319/lom.2013.11.213.

[20] A. Ghaemi et al. Effects of climate change on water quality parameters in lakes. *Journal of Hydrology*, 372(1-4):1–12, 2009. doi: 10.1007/s00484-008-0167-2. URL https://link.springer.com/article/10.1007/s00484-008-0167-2.

[21] D. Gupta et al. Environmental monitoring and assessment of water quality parameters. *Environmental Monitoring and Assessment*, 193(6):1–14, 2021. doi: 10.1080/11104929.2020.1839345. URL https://link.springer.com/article/10.1007/s00484-008-0167-2.

[22] D. Liu. The prediction and analysis of global climate change based on sarima. *Applied and Computational Engineering*, 40(1):268–273, 2024. doi: 10.54254/2755-2721/40/20230665. URL https://www.researchgate.net/publication/378359394_The_prediction_and_analysis_of_global_climate_change_based_on_SARIMA.

[23] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012. URL http://jmlr.org/papers/v13/bergstra12a.html.

[24] Jason Brownlee. How to grid search sarima model hyperparameters for time series forecasting in python, 2020. URL https://machinelearningmastery.com/how-to-grid-search-sarima-model-hyperparameters-for-time-series-forecasting-in-python/. Accessed: 2025-01-28.

[25] ML Pills. How to train a sarima model step by step, 2023. URL https://mlpills.dev/time-series/how-to-train-a-sarima-model-step-by-step/. Accessed: 2025-01-28.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

[27] geeksforgeeks.org. Long short-term memory (lstm) rnn in tensorflow, 2023. URL https://www.geeksforgeeks.org/long-short-term-memory-lstm-rnn-in-tensorflow/. Accessed: 2025-02.

[28] Walter K. Dodds Wayne A. Wurtsbaugh, Hans W. Paerl. Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. *WIREs*, 6, 2019. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wat2.1373?casa_token=L0EDFti8fMMAAAAA%3Ac44JCtiaY5E1Gr0zb8Af6DQCUamSjoxZCyeUQpXUQb4YNtYD215_aQlcofxMcI97mRGRBY4zklB4Zw.

[29] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1394–1401, 2018. doi: 10.1109/ICMLA.2018.00227.

[30] GeeksforGeeks. How to choose batch size and number of epochs when fitting a model, 2024. URL https://www.geeksforgeeks.org/how-to-choose-batch-size-and-number-of-epochs-when-fitting-a-model/. Accessed: [02/2025].

[31] Tayeb Boulmaiz, Mawloud Guermoui, and Boutaghane Hamouda. Impact of training data size on the lstm performances for rainfall–runoff modeling. *Modeling Earth Systems and Environment*, 6, 12 2020. doi: 10.1007/s40808-020-00830-w.

[32] E. K. Read, V. P. Patil, S. K. Oliver, A. L. Hetherington, J. A. Brentrup, J. A. Zwart, K. M. Winters, J. R. Corman, E. R. Nodine, R. I. Woolway, H. A. Dugan, A. Jaimes, A. B. Santoso, G. S. Hong, L. Winslow, P. C. Hanson, and K. C. Weathers. The importance of lake-specific characteristics for water quality across the continental united states. *Ecological Applications*, 25:943–955, 2015. doi: 10.1890/14-0935.1.

[33] Daniel Asante Otchere, Tarek Omar Arbi Ganat, Jude Oghenerurie Ojero, Bennet Nii Tackie-Otoo, and Mohamed Yassir Taki. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, 208:109244, 2022. ISSN 0920-4105. doi: https://doi.org/10.1016/j.petrol.2021.109244. URL https://www.sciencedirect.com/science/article/pii/S0920410521008998.