# PREDICTION OF WATER HYACINTH COVERAGE ON THE HARTBEESPOORT DAM

**C.D. Camacho de Gouveia[1] & J.H. Bührmann[2*]**

## ARTICLE INFO

***Article details***
Presented at the 34th annual conference of the Southern African Institute for Industrial Engineering, held from 14 to 16 October 2024 in Vanderbijlpark, South Africa

Available online 29 Nov 2024

***Contact details***
* Corresponding author
Joke.Buhrmann@nwu.ac.za

***Author affiliations***
1 School of Mechanical, Industrial and Aeronautical Engineering, University of Witwatersrand, South Africa

2 School of Industrial Engineering, North-West University, South Africa

***ORCID® identifiers***
C.D. Camacho de Gouveia
https://orcid.org/0000-0001-9176-7677

J.H. Bührmann
https://orcid.org/0000-0003-0657-9933

***DOI***
http://dx.doi.org//10.7166/35-3-3090

## ABSTRACT

Water hyacinth is an invasive weed that contributes to the Hartbeespoort Dam's poor water quality. Although biological control is the most effective and sustainable method of controlling water hyacinth, a prediction model to plan the biological controls is essential for successful intervention. The literature shows that mathematical models and remote sensing have been used successfully in the past to estimate plant growth rates in similar applications. This study presents various machine-learning models that were investigated to predict water hyacinth coverage.

The complex relationships of water hyacinth growth were simplified to focus on the most influential factors: temperature and nutrients. Missing data were imputed using the multiple k-nearest neighbours imputation. The nutrient datasets were extrapolated to the correct timeline using Monte Carlo simulations and seasonal patterns. Ensemble learning, decision trees, artificial neural networks, and support vector machine models were developed, with ensemble learning (bag algorithm) resulting in the best predictions.

## OPSOMMING

Waterhiasint is 'n indringer wat bydra tot die Hartbeespoortdam se swak watergehalte. Alhoewel biologiese beheer die mees doeltreffende en volhoubare metode is om waterhiasint te beheer, is 'n voorspellingsmodel om die biologiese beheermaatreëls te beplan noodsaaklik vir suksesvolle ingryping. Die literatuur toon dat wiskundige modelle en afstandswaarneming in die verlede suksesvol gebruik is om plantgroeitempo's in soortgelyke toepassings te skat. Hierdie studie bied verskeie masjienleermodelle aan wat ondersoek is om waterhiasintbedekking te voorspel.

Die komplekse verhoudings van waterhiasintgroei is vereenvoudig om op die mees invloedryke faktore te fokus: temperatuur en voedingstowwe. Vermiste data is toegereken met behulp van die veelvuldige k-naaste bure-toerekening. Die voedingstofdatastelle is na die korrekte tydlyn geëkstrapoleer deur Monte Carlo-simulasies en seisoenale patrone te gebruik. Ensembleleer, besluitnemingsbome, kunsmatige neurale netwerke en ondersteuningsvektormasjienmodelle is ontwikkel, met ensembleleer (sakalgoritme) wat tot die beste voorspellings gelei het.

## 1. INTRODUCTION

The Hartbeespoort dam in South Africa has been eutrophic since the 1960s [1]. When water is eutrophic, it has a high concentration of nutrients, particularly total phosphorous (TP) and total nitrogen (TN), that are above the standards set out by the Department of Water Affairs and Forestry [2]. The dam's eutrophic state is caused by wastewater from surrounding areas and traces of mining effluent [3]. The presence of nutrients forms favourable conditions for water hyacinth growth [1],[2].

Water hyacinth can absorb excess nutrients and toxins in water and so can be used to improve water quality [4],[5]. However, it is also highly invasive, and its rapid growth rate results in thick mats that suffocate waters, harm ecosystems, and can become a breeding place for disease-carrying organisms [6]-[8]. Thus research focuses on controlling water hyacinth growth using remote sensing, spatial mapping, and biological control methods [9]-[11].

The Centre for Biological Control (CBC) has found that introducing a biological agent such as the plant hopper is the cheapest and most sustainable control method [12]. However, conditions at the Hartbeespoort dam are unfavourable for stabilising these biological control populations [13]. This means that the correct number of bugs needs to be introduced with intermittent frequency to control the water hyacinth regrowth effectively [14],[15]. Placing the biological control agent too frequently or in excess is unsustainable. Not only is it an unnecessary cost, but it could also reduce the water hyacinth too much, which would mean the return of excess nutrients and toxins in water [13]. Gutiérrez *et al.* [16] suggest that biological control is more effective when using a planned, proactive approach. Therefore, a prediction model for water hyacinth coverage (WHC) is essential to plan the successful introduction of biological agents.

Supervised machine learning (ML) models have been used successfully in the past for predictions of regression models, and can provide more accurate results than traditional forecasting models [17]. 'Supervised ML' refers to ML prediction models that are trained on datasets in which the outcome is known and then applied to unseen data. 'Regression' refers to models with continuous outcome variables, unlike classification models in which the outcomes need to fall into a specific category. This study aimed to illuminate how well various ML models could predict WHC with the limited data that was available.

The remainder of this article is structured as follows. A brief overview of case studies on predicting plant coverage or biomass using machine learning is provided in Section 2. This is followed by a description of the identified datasets that were used to predict WHC in this study. Next, an explanation of the method, including the data preparation and ML model building phases, is presented. Finally, the results are presented in Section 6, followed by a discussion and conclusion.

## 2. LITERATURE REVIEW

Artificial neural networks (ANNs), support vector machines (SVMs), k-nearest neighbours (KNN), and ensemble algorithms were repeatedly used in multiple case studies to predict plant coverage or biomass [18]-[20]. According to [19], ML is a dominant tool for predicting crop yields. Those researchers identified ANN as the most popular algorithm in this field, with temperature and nutrients being the most common features for predicting crop yields.

Bayable *et al.* [20] used ML algorithms to detect water hyacinth in a body of water and to estimate its seasonal spatial coverage. The regression ML algorithms tested in this study were SVMs, random forest, and decision trees. [21] proved that ensemble algorithms, specifically random forest, were the best-performing algorithms for predicting vegetation compared with SVMs and other regression algorithms.

[22] used images from multiple satellites, remote sensing, and image classification ML algorithms to map WHC on Lake Tana in Ethiopia. The researchers analysed the relationships between water hyacinth and climate variability, such as the evaporation of water in the lake. This was achieved by analysing trends using Pearson's correlation. However, this study did not predict WHC in the future; instead, it used algorithms to enhance poor-quality images and to handle missing images.

## 3. DATA DESCRIPTION

To predict WHC accurately, an investigation was launched to identify the necessary data and availability. According to [23] and [24], water hyacinth growth depends heavily on extreme temperatures as well as the eutrophic levels in the dam, and there is a hyperbolic relationship between a dam's eutrophic levels and the phosphate and nitrogen levels. The factors that influence water hyacinth growth the most, therefore, are the minimum and maximum daily ambient temperature, phosphate levels, and the nitrogen levels, composed of nitrite ($NO_2$), nitrate ($NO_3$) and ammonium ($NH_4$), in the dam. Based on this, the datasets described below were identified for inclusion in the study.

### 3.1. Water hyacinth coverage

WHC was the predicted output variable. Access to recorded data of WHC on the Hartbeespoort dam was obtained from the CBC at Rhodes University [12]. The full dataset is not publicly available, but summaries have been published by the CBC [15],[25].

The dataset contained 587 data points with the date and percentage coverage between 1 August 2015 and 27 March 2021. During 2020 and 2021, the WHC was influenced by restrictions on human movement during the COVID-19 pandemic, and did not reflect normal WHC levels. It was, therefore, excluded from the study. The remaining 461 data points were not sampled during regular intervals. Figure 1 shows an example of the count of data points per month for a section of the dataset from January 2017 to January 2019.

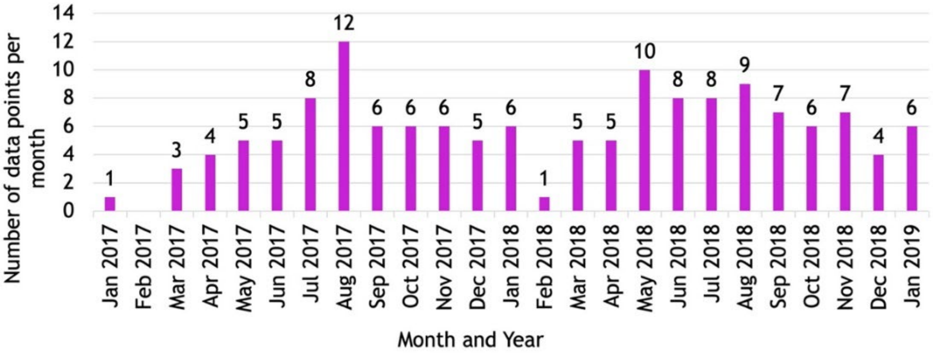

**Figure 1: Number of WHC data points recorded per month**

The figure shows that some months, such as February 2017, had no recorded data points, while others, such as August 2017, had 12. Therefore, a fixed time interval was needed to make all datasets consistent and comparable. As the WHC dataset represents the target variable, it was used to determine the best interval, based on the average time interval between recorded WHC dates. The average was 5.35. All datasets were therefore imputed to have data every five days from 1 August 2015 until 28 December 2019.

The percentage of missing data in the WHC dataset was 37.7%. This was not seen as a problem, because frequent data on other intermittent dates could be used during the imputation. This value was also lower than the recommended 40% threshold used to determine whether a dataset would remain reliable when missing data was imputed [26]. A time series plot of the WHC dataset after the imputation of the missing data can be viewed in Figure 4 in Section 6.2.

### 3.2. Ambient temperature

The daily maximum (MaxTemp) and minimum (MinTemp) temperatures were obtained from the South African Weather Service [27] for the recommended weather station (05125544) closest to the dam [23]. The MaxTemp and MinTemp datasets were kept separate because WHC is sensitive to both high and low extreme temperatures [23],[28]. Both datasets had 1.8% missing data over the five-day intervals from 1 August 2015 to 28 December 2019. A representation of the final temperature datasets, after imputations, can be seen in Figures 5 and 6 in Section 6.2.

### 3.3. Total phosphorous

Two datasets were used to determine the total phosphorous (TP) levels on the Hartbeespoort dam. The first was a water quality dataset downloaded from the National Eutrophication Monitoring Programme website [29]. It contained 2 668 relevant phosphate data points measured in mg/L at stations on the dam between 11 March 1980 and 18 July 2018.

A second dataset was obtained from the Centre for Water Sciences and Management at North-West University [30],[31]. It contained the chemical and physical properties of many South African dams, including recordings of phosphate ($PO_4$) levels from stations on the Hartbeespoort dam. The average was taken for instances with multiple recordings for any given day. This resulted in 315 data points representing the $PO_4$ levels between 1 December 1999 and 14 December 2011 on the dam.

The two datasets were compared, and after verifying that the data were indeed similar, the datasets were combined to create a single TP dataset with 2 706 averaged daily data points between 11 March 1980 and 18 July 2018. Unfortunately, no data could be found for the period after 19 July 2018. The data points in the TP dataset, on average, were recorded every 5.18 days. This was very similar to the sample frequency of the WHC dataset. The percentage of missing data for the TP dataset that ended on 18 July 2018 was 8.45%. Section 4.3 discusses how the data was extrapolated for the additional period from 19 July 2018 to 28 December 2019. The final dataset, after imputations of the current missing data and extrapolation to extend the data to the required timeline, can be viewed in Figure 7 in Section 6.2.

### 3.4. Total nitrogen

The second dataset downloaded from the Centre for Water Sciences website for the TP data also contained data on nitrogen compounds [30],[31]. The dataset included recordings of the sum of nitrite and nitrate ($NO_2$+$NO_3$) and of ammonium ($NH_4$) levels, which, when added together, represent the total nitrogen (TN) levels of the dam.

However, before the two nitrogen compounds could be combined to form one TN dataset, the missing data of each needed to be imputed separately. The average number of days between data points was 15. This meant that, on average, there were two recordings per month, with some months completely missing, such as October and November 2011. The percentage of missing data for the two compounds was 7.6% and 0.3% for $NO_2$+$NO_3$ and $NH_4$ respectively.

The datasets represented data from 1 December 1999 to 14 December 2011. The data had to be extrapolated from 1 August 2015 to 28 December 2019. This is discussed in Section 4.4. A time series graph of the final TN dataset can be seen in Figure 8 in Section 6.2.

## 4. DATA PREPARATION

After the available datasets had been identified, the data had to be prepared in a data preparation phase before the machine learning model could be built. Figure 2 illustrates the method followed in this research. The method was based on various techniques (which are discussed below) in order to impute the missing data and to extrapolate the TP and TN datasets needed to prepare the data, combined with the standard regression ML model building steps [17]. Coding in both phases was done in MATLAB [32].

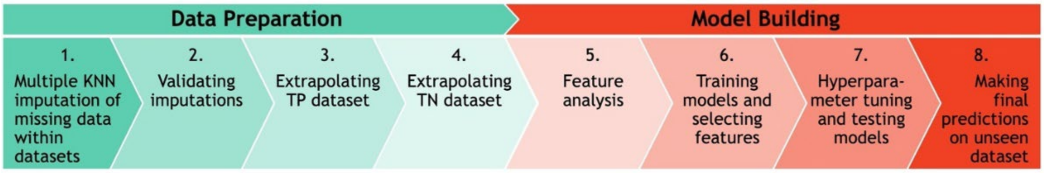
Data Preparation
Model Building
1. Multiple KNN imputation of missing data within datasets
2. Validating imputations
3. Extrapolating TP dataset
4. Extrapolating TN dataset
5. Feature analysis
6. Training models and selecting features
7. Hyperpara-meter tuning and testing models
8. Making final predictions on unseen dataset

**Figure 2: Process flow diagram used to make final WHC predictions**

During the data preparation phase, missing data in the identified datasets was first imputed using the multiple KNN imputation method, and then validated. Next, the TN and TP datasets were extrapolated to the same timeline as the WHC dataset, using Monte Carlo simulations and identified seasonal patterns. After the data preparation, the finalised MaxTemp, MinTemp, and simulated TN and TP datasets were used as the input features in the model building phase to create ML models, with WHC as the target variable to be predicted. The model building phase is discussed further in Section 5.

### 4.1. Multiple KNN imputation of missing data within datasets

Imputation was used to deal with the missing data in the current timeline [33]-[36]. Multiple KNN imputation was used, as it is an efficient imputation method that imputes the missing data with its KNN multiple times [37]-[40].

The multiple KNN imputation was coded in MATLAB [41]. The dataset was split into groups, and the missing point was imputed with the KNN value in the group, using the built-in *knnimpute* function. This was done iteratively with variations in the groupings so that the missing point was imputed with a range of values. The average of all the imputed values per point was then used as the final imputed value. The number of iterations was initially set to 100 and increased in increments of 100. More than 500 iterations did not result in large differences between imputations, as the differences were less than 0.001; thus 500 was deemed the best number of iterations to use.

### 4.2. Validating imputations

The imputations were visually inspected to determine whether they were valid and represented the observed dataset well. The descriptive statistics of the observed datasets, both with and without imputations, were also calculated and compared. Finally, we also checked that the imputed data points fell within the 95% confidence intervals of the original dataset. This was done for all the imputed feature datasets and for the WHC target variable dataset.

### 4.3. Extrapolating TP dataset

The TP dataset did not coincide with the timeline of the WHC dataset, and had to be extrapolated to the required dates between 1 August 2015 and 28 December 2019. Five scenarios were developed in which the TP dataset was generated differently. Table 1 describes the methods used to generate the TP dataset in each scenario.

Usually, the Kolmogorov–Smirnov test (KS-test) is used to compare similarities between datasets [42]. In the case of the TP scenarios, the KS-test values were too close to one another, and could not be used to differentiate between the scenarios. Therefore, the generated datasets were compared using the two-sided T-test [43], which tested whether the generated dataset came from a distribution with a similar variance and mean as the observed TP dataset. The closer the p-value was to 1, the better the fit between the generated and the observed TP datasets, with a similar mean and variance. Scenario 5 had the highest T-test p-value of 0.9, and was thus used to create the extrapolated TP dataset.

**Table 1: Scenarios used to generate the TP dataset**

| Scenario | Description of method |
|---|---|
| 1 | Monte Carlo simulation was used together with the seasonal curve of the observed TP dataset. The curve was fitted to the daily average TP. Coefficients for the curve with a 95% confidence interval were generated using MATLAB. |
| 2 | Simulated deltas were added to the TP dataset generated using Scenario 1 to obtain a dataset with a wider variance. Repetitive patterns were analysed, and a dataset lagging by 1 506 points from the observed TP dataset was created. This number was identified by the partial autocorrelation function (PACF) in MATLAB as a significant lag for the dataset. The difference between the observed TP dataset and the lagged dataset, referred to as ‘the deltas’, was calculated. A probability distribution was fitted to the deltas and simulated using Monte Carlo simulation. |

| | |
|---|---|
| 3 | Simulated deltas were added to the TP dataset generated in Scenario 2, but lagged by 2 220 points. This lag was identified during time series analysis as the point where the seasonal cycle started to repeat. |
| 4 | TP values were extrapolated from the curve fitted to the daily average TP, similar to Scenario 1. The simulated deltas generated in Scenario 2 were then added to the extrapolated TP data. |
| 5 | TP values were again extrapolated from the curve fitted to the daily average TP, similar to Scenario 1. Simulated deltas from Scenario 3 were added to the extrapolated TP data. |

### 4.4. Extrapolating TN dataset

Similar to the TP dataset, five scenarios were developed to extrapolate the TN dataset. A time series analysis was conducted on the observed TN dataset to determine the significant number of lags to be used and to identify repetitive patterns. This time series analysis revealed a repetitive annual seasonal pattern that could be represented by a polynomial curve. The plotted PACF also identified lags 1 and 2 as significant and, thus, were used to generate the TN dataset in Scenarios 2 and 3, respectively.

The KS-test was used to compare the generated datasets, and similarly to the T-test values, a value close to 1 meant the two datasets had similar probability distribution functions. The scenario with the highest p-value was scenario 3, and so was used to extrapolate the final TN dataset.

## 5. MODEL BUILDING

### 5.1. Feature analysis

The features identified as important for WHC predictions were Date, MaxTemp, MinTemp, and the generated TP and TN datasets. Based on a suggestion from [44], the features were analysed to ensure that they resulted in good final predictions. Feature analysis was conducted to determine the features that would positively impact the model's performance.

The first analysis was to calculate Pearson's correlation between all features. Correlations of 70% or higher indicate a very strong correlation [45], and could imply that one feature is redundant in the model. Features with a high correlation can also be combined into one feature to represent both correlating features.

The second analysis was to calculate the significance of each feature. MATLAB was used to perform an F-test and to calculate the corresponding p-value for each feature. Features with a p-value below 0.05 are considered significant [46]. The F-test statistic's p-value reveals how well the feature describes the target variable, with high values indicating high significance.

The information obtained from the Pearson's correlation analysis and F-test statistic was used to identify various combinations of features for the model building phase, to develop the best possible ML model.

### 5.2. Training models and selecting features

The datasets were split into train, test, and unseen datasets during model building, as described in Section 6.2. The unseen data used to make final predictions were all the data from 1 June 2018 to 28 December 2019. Data from 1 August 2015 to 31 May 2018 were used to train and test after being split randomly using an 80:20 ratio.

Five types of regression ML model were chosen, based on similar case studies [18]-[20]. These models were the ensemble (boost), ensemble (bag), decision tree, ANN, and SVM. 'Ensemble (boost) model' refers to an ensemble model that uses the boosting algorithm, while the 'ensemble (bag)' uses the bagging algorithm and builds on multiple decision trees [45]-[47]. The latter was expected to perform better than the decision tree model, but was included here for completeness. Each model was trained on the four identified feature combinations, and the RMSE for each model-feature combination was determined. An RMSE value close to zero indicates a model with good performance and that the model is a good fit [48],[49]. The results of the four feature combinations were compared to determine the best feature combination to use for prediction.

### 5.3. Hyperparameter tuning and testing models

The hyperparameters of the models were optimised using the Bayesian optimisation function in MATLAB. This function used a grid search functionality to test different combinations of hyperparameter values, and calculated the RMSE for 30 iterations on the test dataset. The hyperparameters with the lowest test RMSE were chosen as the best ML model and used to make the final predictions.

### 5.4. Making final predictions on unseen dataset

The final predictions were made on the unseen dataset, which contained data from 1 June 2018 to 28 December 2019. The best ML model was used to make the final predictions and to calculate the RMSE. The final predictions were also depicted on a graph together with the observed target values for visual inspection. The final results of the data preparation and model building phases are presented in the next section.

## 6. RESULTS

In this section we first present the results of the feature analysis and identified feature combinations. Next, the finalised datasets, after the data preparation phase had been completed, are presented. Finally, the results of the various ML models using the identified feature combinations, as well as the final predictions of unseen data of the chosen ML model and feature combination, are provided.

### 6.1. Feature analysis

Pearson's correlation was calculated to compare the input features with one another. The correlations were calculated in MATLAB, with the results shown in Table 2. The light pink shaded cells in Table 2 indicate that the MaxTemp and MinTemp variables had a very strong correlation of 73%. An alternative to having two correlated features in the model was to replace the variables with an average value called AvgTemp. Figure 9 shows the AvgTemp dataset and its breakdown into the training, test, and unseen datasets used in the ML model building.

**Table 2: Pearson's correlation matrix for features**

| | Date | MaxTemp | MinTemp | TP | TN |
|---|---|---|---|---|---|
| Date | 1 | 0.02 | -0.14 | 0.04 | 0.01 |
| MaxTemp | 0.02 | 1 | 0.73 | 0.27 | -0.05 |
| MinTemp | -0.14 | 0.73 | 1 | 0.21 | -0.20 |
| TP | 0.04 | 0.27 | 0.21 | 1 | 0.03 |
| TN | 0.01 | -0.05 | -0.20 | 0.03 | 1 |

Next, the significance of the features was tested by calculating the corresponding p-values of the F-test statistic. Figure 3 shows the p-values for all features. Values below 0.05 represent features of significance. The figure shows that the Date, MaxTemp, MinTemp, and AvgTemp features all had p-values of 0 and thus had high significance.

The TN and TP features had p-values of 0.67 and 0.98 respectively. These values were greater than 0.05 and could, therefore, be considered insignificant. However, the literature has shown that nutrients play a vital role in WHC and growth [7],[50]. Thus an additional feature named TNTP was introduced. This additional feature is the ratio of TN to TP, calculated by dividing TN by TP. It was added, as it is often used in the literature to determine the limiting factor in eutrophication [51],[52]. In Figure 3, the TNTP feature has a p-value of 0, which means that it is significant. Figure 10 in Section 6.2 shows the TNTP dataset.

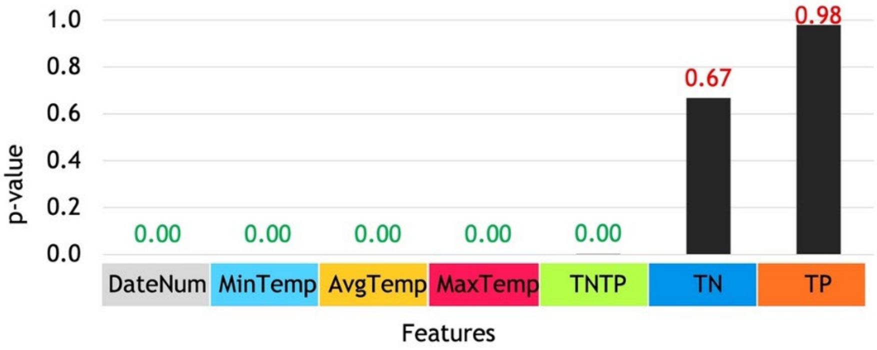

**Figure 3: F-test statistic's corresponding p-value of features**

Using this information and the correlation results, four combinations of features were identified, as shown in Table 3. The four feature combinations were used to train each of the five models to determine the best combination of features. Various representations of the Date feature were tested (for example, having month, day, and year separate), but representing this feature as a serial number resulted in the best RMSE values, and so the feature was left as-is.

**Table 3: Combinations of features used to train models**

| | | Features | | | | | | | Reason for choice |
|---|---|---|---|---|---|---|---|---|---|
| | | Date | Max-Temp | Avg-Temp | Min-Temp | TP | TN | TNTP | |
| Combination | A | X | X | | X | X | X | | Original features |
| | B | X | | X | | X | X | | Replacing correlating features with one that combines both |
| | C | X | X | | X | | | X | Replacing insignificant features with one that combines both |
| | D | X | | X | | | | X | Replacing all correlating and insignificant features |

## 6.2. Finalised datasets

The imputed WHC target dataset was split into training, testing, and unseen datasets in preparation for the model building phase. The unseen dataset was WHC data from 1 June 2018 to 28 December 2019, as shown in Figure 4. The remaining data were split randomly at a ratio of 80:20 to form the training and testing datasets respectively, as done in [53]. To make this split, the *cvpartition* function in MATLAB was used. Figure 4 shows how the dataset was split.

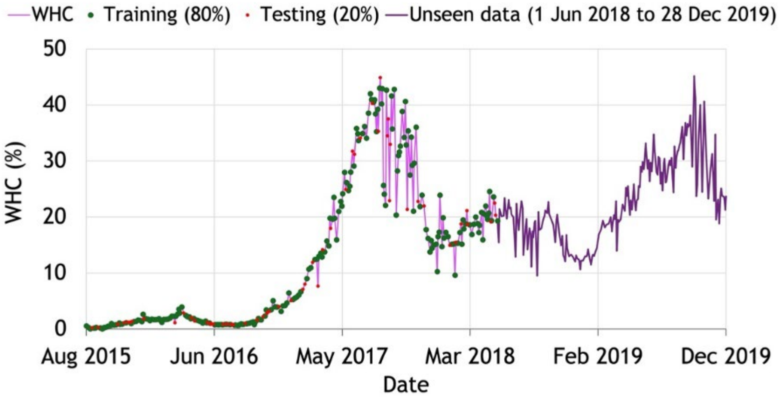

**Figure 4: Time series plot of WHC dataset used as the target variable**

The datasets used as input features in the model building phase are shown in Figures 5 to 10. These figures show how the datasets were split into training, testing, and unseen datasets in the same way that was done for the target variable.

### 6.3. Model building

The validation RMSE of each model that was trained using the different feature combinations is shown in Figure 11. All the SVM models had a much higher validation RMSE than the other models, with an average RMSE value of 8.94. It was clear that the SVM model did not perform well compared with the other models, and it was not developed further. The ensemble (bag) and decision tree models performed similarly, with an average RMSE of 3.20 and 3.70 respectively. The ensemble (bag) consists of multiple decision tree models, and thus it is logical that the ensemble (bag) would outperform the decision tree models. Therefore, the decision tree model was also excluded from further development.

The average RMSE on the remaining three ML models was calculated for each feature combination to determine which resulted in the lowest RMSE across the three remaining models. The results are shown in Figure 12.

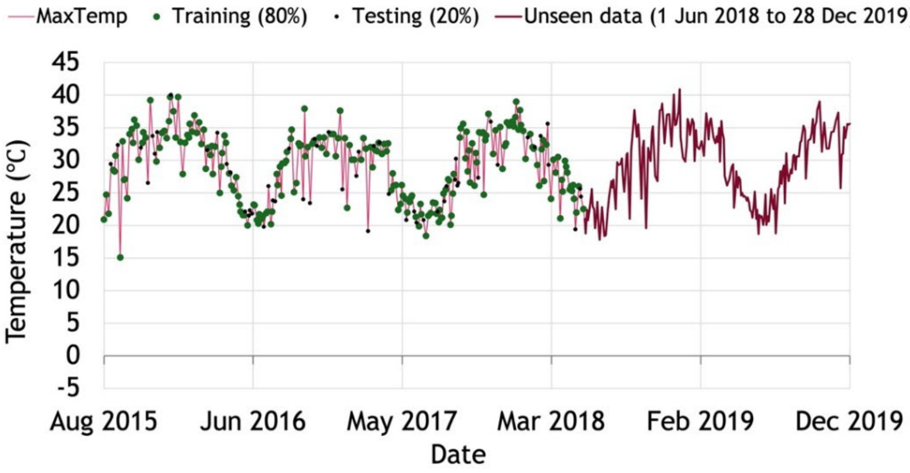

**Figure 5: Time series plot of MaxTemp dataset used as a feature**

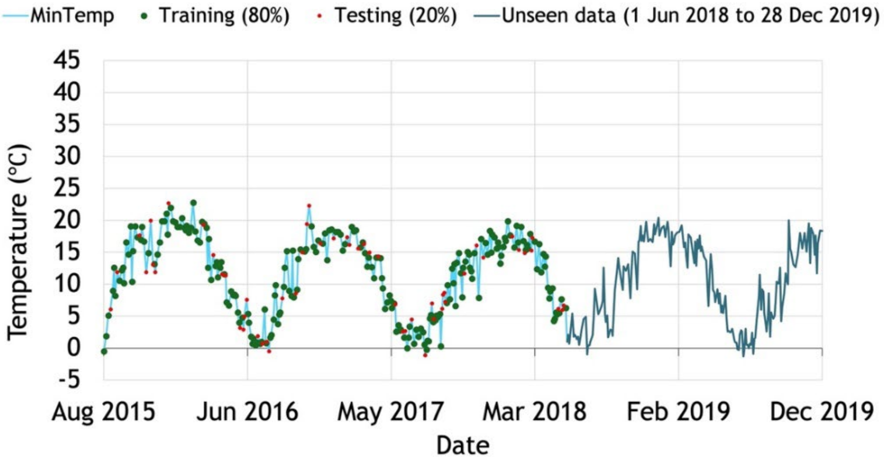

**Figure 6: Time series plot of MinTemp dataset used as a feature**

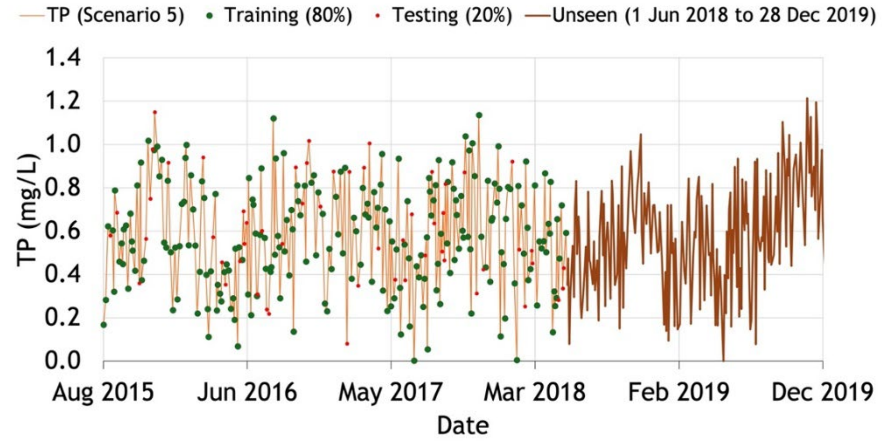

**Figure 7: Time series plot of TP dataset used as a feature**

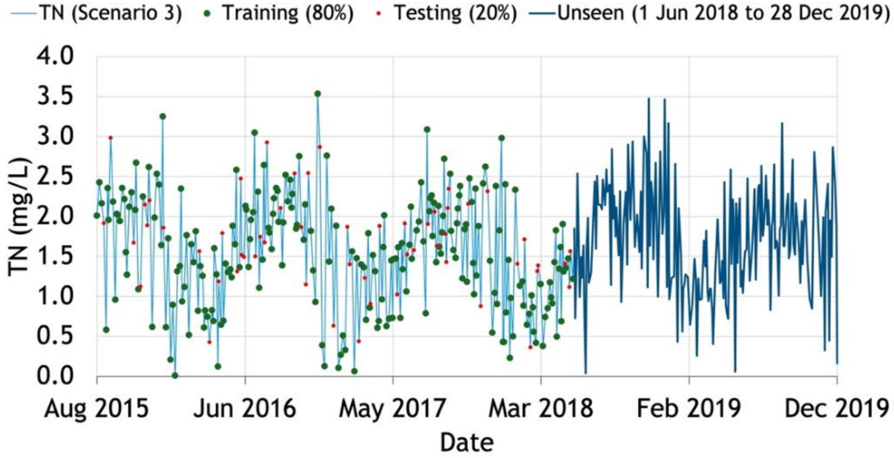

**Figure 8: Time series plot of TN dataset used as a feature**

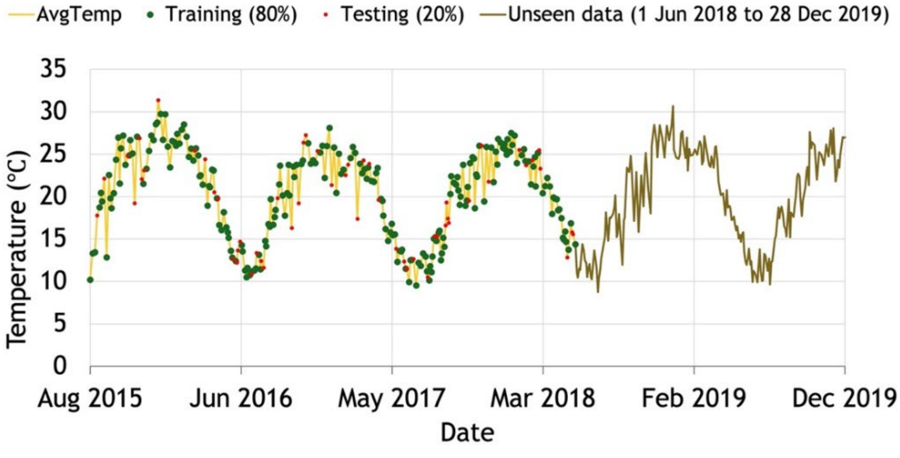

**Figure 9: Time series plot of AvgTemp dataset used as an additional feature**

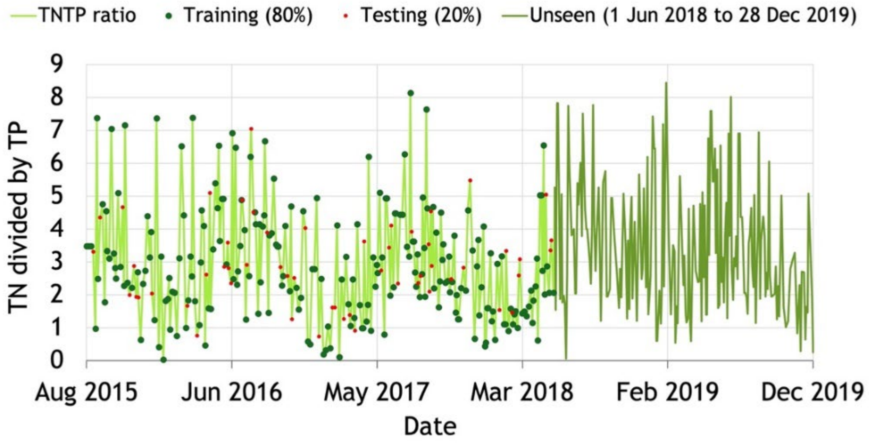

**Figure 10: Time series plot of TNTP dataset used as an additional feature**

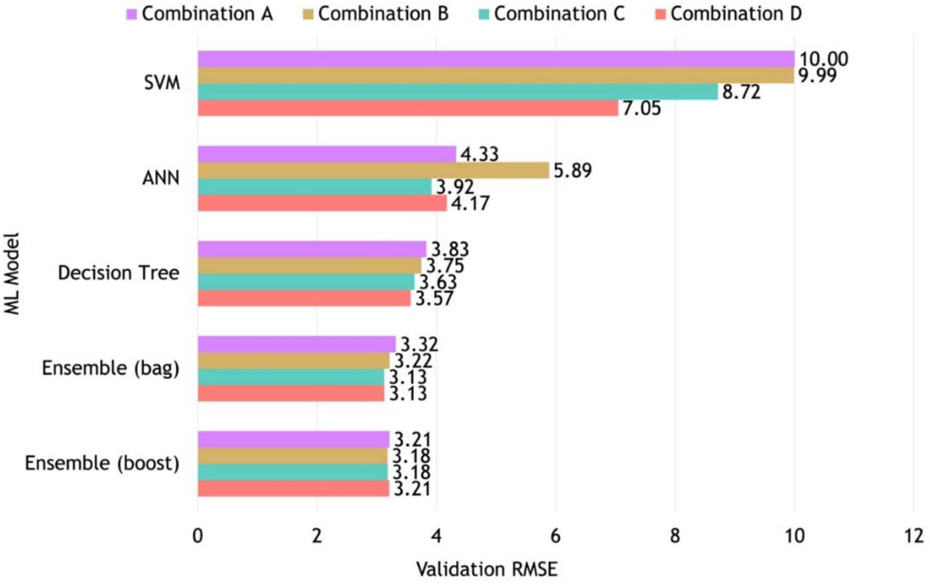

**Figure 11: Validation of RMSE values for each model trained using different features**

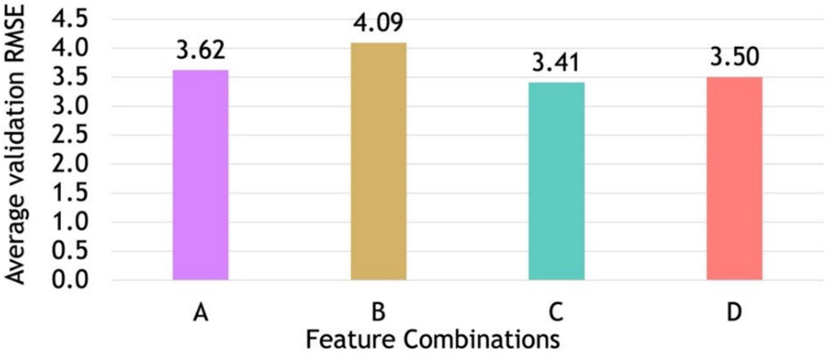

**Figure 12: Average validation RMSE of each model using different features**

Combination C resulted in the lowest average validation RMSE of 3.41. Therefore, the Date, MaxTemp, MinTemp, and TNTP features were used to train the final model. The ensemble (boost), ensemble (bag), and ANN models were optimised using hyperparameter tuning on the training dataset with the Bayesian optimiser function in MATLAB, and then tested using the test dataset. The ensemble (bag) algorithm with the hyperparameters of minimum leaf size = 2 and number of learners = 13 resulted in the lowest test RMSE of 2.73, and it was therefore selected as the best ML model.

### 6.4. Final predictions

The final predictions made on unseen data are shown in Figure 13 together with the observed unseen WHC dataset. These predictions resulted in an RMSE of 7.31. This was much higher than the results of the test dataset with an RMSE of 2.73. However, it is clear in Figure 13 that the model became inaccurate after May 2019. When excluding predictions after May 2019, the RMSE was 4.01.

Further analysis was performed by calculating the 95% confidence intervals of the observed data. Figure 14 depicts these intervals and the observed and predicted values from June 2018 to December 2019. Even though the values became inaccurate after May 2019, they remained within the 95% confidence interval for all except two data points near October 2019. The observed WHC values were more volatile from September 2019, causing the confidence interval to fluctuate significantly after this date.

A final test was conducted to evaluate the model's performance if the training, testing, and unseen datasets were randomly selected points over the complete WHC dataset. The data were again split into three datasets: training, testing, and unseen. The model returned an RMSE of 3.63, confirming that day-to-day predictions instead of months in advance result in more accurate results. The plot in Figure 15 shows the predictions on the randomly selected unseen dataset.

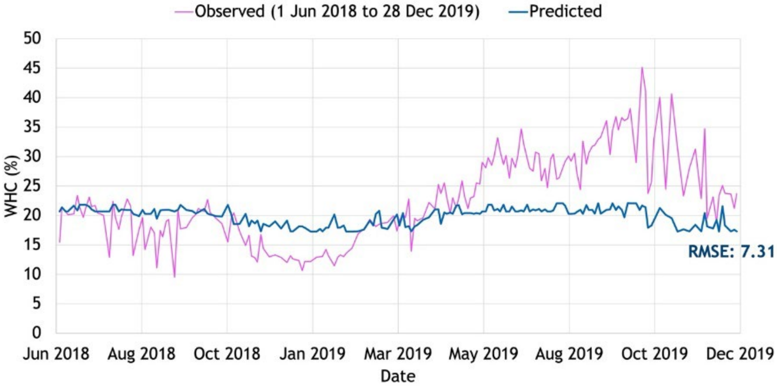

**Figure 13: Time series plot of final WHC predictions**

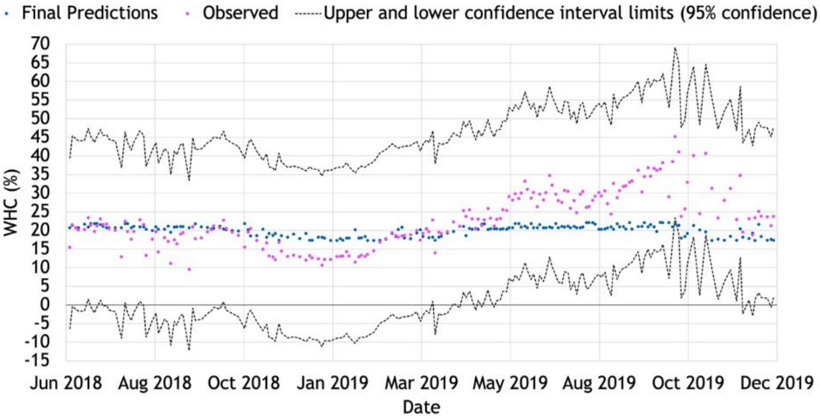

**Figure 14: Confidence interval limits of unseen WHC data**

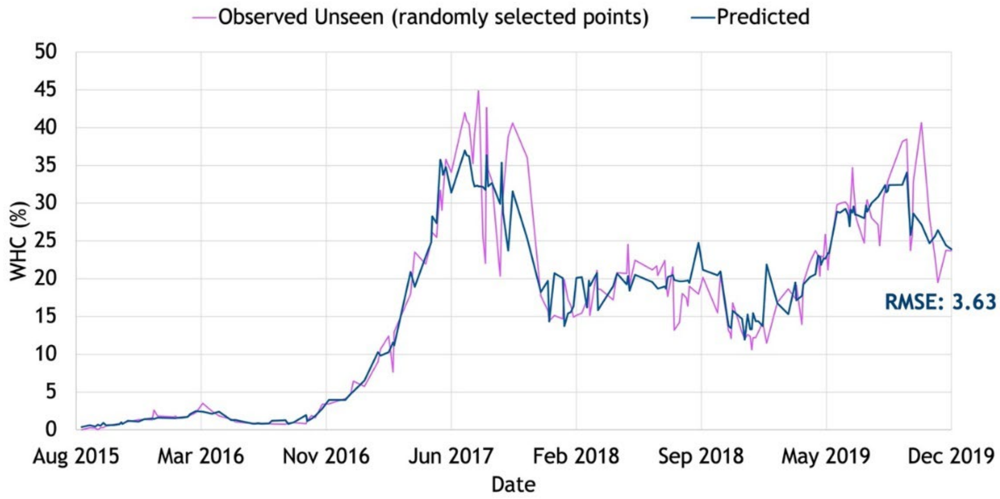

**Figure 15: Plot of WHC predictions on the randomly selected unseen dataset**

## 7. DISCUSSION

WHC predictions for 1 June 2018 to 28 December 2019 were made using the ensemble (bag) model combining Date, MaxTemp, MinTemp, and TNTP features. The feature analysis indicated that the ratio of the features TN to TP had a higher significance than keeping the features separate, and allowed the model to learn patterns from both nutrients that were not evident when kept separately.

The ensemble (bag) model was the chosen final model, as it had the lowest test RMSE compared with the ensemble (boost) and ANN models. This was expected from the literature, as the ensemble (bag) model outperformed other models in most cases [54]. The RMSE of the predictions made by the final ensemble (bag) model was 7.31. This was much higher than the test RMSE of 2.73.

The final predictions were significantly centred around the average observed WHC, but the model was able to mimic the trends seen in the observed. It was evident that, from May 2019, the model did not make predictions that were close to the observed WHC. There was also more volatility in the last section of the observed WHC, which could explain why the predictions were inaccurate. The RMSE calculated for the predictions up to May 2019 was 4.01, which was significantly lower. This meant that the model's performance was close to its performance during testing until May 2019. Therefore, the model made good predictions up to 11 months after the known data – from 1 June 2018 to 1 May 2019.

## 8. CONCLUSION

The research aimed to illuminate whether WHC could be accurately predicted for 1 June 2018 to 28 December 2019 using ML models based on the limited available data. Data from 2020 to 2022 were excluded from this research owing to the impact of the COVID-19 pandemic and the consequent limitations on the human movements that indirectly influence WHC on the dam. The most influential factors for WHC were identified. Those factors included date, minimum and maximum temperature per day, and estimations of the nutrients found in the dam in the form of TN and TP.

Missing data in the datasets was imputed using multiple KNN imputations. The quality of the imputations was visually validated, and fell well within the 95% confidence interval limits of the observed data. Reliable and consistent data for the nutrients in the dam were limited; so those datasets had to be extrapolated to the correct timeline. To extrapolate, various simulation methods were developed and tested. The chosen scenarios were the ones that resulted in the highest statistical test p-values.

Five different ML models were built with different combinations of features that were based on their significance and correlation with one another. The features chosen were Date, MaxTemp, MinTemp, and TNTP. The ensemble model using the bag algorithm was the chosen final model with the lowest test RMSE of 2.73. Visual inspection also validated that the model made accurate predictions within the 95%

confidence interval of the observed WHC data from 1 June 2018 to 1 May 2019, with an RMSE of 4.01. This implied that the best ML model could provide good estimate predictions up to 11 months in advance. Therefore, we believe that using the ensemble bag algorithm would illuminate the planning of WHC biological control in future decision-making.

## ACKNOWLEDGEMENTS

We would like to thank David Kinsler and Gillian McGregor from Rhodes University, and Prof. Julie Coetzee from the Centre for Biological Control, for supplying the WHC data and all their expertise and advice.

## REFERENCES

[1] Water and Sanitation South Africa (WASA), "Bugs beat back water hyacinth," *Water & Sanitation Africa March/April 2020*, vol. 15, no.1, pp. 32–33, Mar. 2020.

[2] J. N. Rossouw, W. R. Harding, and O. S. Fatoki, "A guide to catchment-scale eutrophication assessments for rivers, reservoirs and lacustrine wetlands," Pretoria, Water Research Commission report, TT352/08, Apr. 2008.

[3] S. A. Mitchell, J. G. Crafford, M. van Veelen, J. Maree, and E. Reyneke, "Review of the Hartbeespoort Dam integrated biological remediation programme (Harties Metsi a Me)," Water Research Committee report, KV 357/16, July 2016. Accessed: June 5, 2024. Available: https://www.wrc.org.za/wp-content/uploads/mdocs/KV%20357.pdf

[4] J. Auchterlonie, C. Eden, and C. Sheridan, "The phytoremediation potential of water hyacinth: A case study from Hartbeespoort Dam, South Africa," *South African Journal of Chemical Engineering*, vol. 37, no. 1, pp. 31–36, Jul. 2021.

[5] M. M. Petrucio and F. A. Esteves, "Uptake rates of nitrogen and phosphorus in the water by *Eichhornia crassipes* and *Salvinia auriculata*," *Revista Brasileira de Biologia*, vol. 60, no. 2, pp. 229–236, May 2000.

[6] K. R. Reddy and J. C. Tucker, "Productivity and nutrient uptake of water hyacinth, *Eichhornia crassipes* I. Effect of nitrogen source," *Economic Botany*, vol. 37, no. 2, pp. 237–247, Apr. 1983.

[7] J. Rojas-Sandoval and P. Acevedo-Rodríguez, "*Eichhornia crassipes* (water hyacinth)," CABI Compendium, Jan. 2022. Accessed: June 5, 2024. Available: https://www.cabidigitallibrary.org /doi/10.1079/cabicompendium.20544

[8] M. K. Mzuza, L. Chapola, F. Kapute, I. Chikopa, and J. Gondwe, "Analysis of the impact of aquatic weeds in the Shire River on generation of electricity in Malawi: A case of Nkula Falls hydro-electric power station in Mwanza District, Southern Malawi," *International Journal of Geosciences*, vol. 6, no. 6, pp. 636–643, 2015.

[9] G. Singh, C. Reynolds, M. Byrne, and B. Rosman, "A remote sensing method to monitor water, aquatic vegetation, and invasive water hyacinth at national extents," *Remote Sensing*, vol. 12, no. 24, pp. 1–24, Dec. 2020.

[10] G. Keebine, "Mapping and monitoring the spatial distribution of *Eichhornia crassipes* (water hyacinth) in the Hartbeespoort Dam, South Africa, using remote sensing data," Master's thesis, University of the Witwatersrand, 2019. Accessed: February 28, 2024. Available: https://wiredspace.wits.ac.za/server/api/core/bitstreams/34e318e2-438c-4145-a385-008da30cd99f/content.

[11] R. Moffat *et al.*, "Bridging boundaries: Six years of community engagement with biological control implementation and monitoring of water hyacinth on Hartbeespoort Dam, South Africa," *Biological Control*, vol. 194, 105544, 2024.

[12] Rhodes University Centre for Biological Control (CBC), "CBC releases status and future expectations for its Hartbeespoort Dam water hyacinth project," *Rhodes Latest News: November 2020*, 2020. Accessed: June 5, 2024. Available: https://www.ru.ac.za/latestnews/archives/2020/ cbcreleasesstatusandfutureexpectationsforitshartbeespoortdamwaterhyac.html

[13] J. A. Coetzee and M. P. Hill, "The role of eutrophication in the biological control of water hyacinth, *Eichhornia crassipes*, in South Africa," *BioControl*, vol. 57, no. 2, pp. 247–261, Nov. 2011.

[14] B. E. Miller, "Post-release evaluation of *Megamelus scutellaris* Berg. (Hemiptera: Delphacidae): A biological control agent of water hyacinth *Eichhornia crassipes* (Mart.) Solms-Laub (Pontederiaceae) in South Africa," Master's thesis, Rhodes University, Grahamstown, South Africa, 2020.

[15] J. A. Coetzee, B. E. Miller, D. Kinsler, K. Sebola, and M. P. Hill, "It's a numbers game: Inundative biological control of water hyacinth (*Pontederia crassipes*), using *Megamelus scutellaris* (Hemiptera: Delphacidae) yields success at a high elevation, hypertrophic reservoir in South Africa," *Biocontrol Science and Technology*, vol. 32, no. 11, pp. 1302–1311, Aug. 2022.

[16] E. L. Gutiérrez, E. F. Ruiz, E. G. Uribe, and J. M. Martínez, “Biomass and productivity of water hyacinth and their application in control programs,” in *ACIAR Proceedings 102*, 2001, pp. 109–119.

[17] V. E. Maluleke and J. H. Bührmann, “Corporate failure prediction of JSE listed South African firms using machine learning (A credit risk management approach),” in *SAIIE32 Proceedings*, 4-6 October 2021, Glenburn Lodge, Gauteng, South Africa, pp. 545-555.

[18] J. Liu, C. Yue, C. Pei, X. Li, and Q. Zhang, “Prediction of regional forest biomass using machine learning: A case study of Beijing, China,” *Forests*, vol. 14, no. 5, 1008, May 2023.

[19] T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” *Computers and Electronics in Agriculture*, vol. 177, no. 1, 105709, Oct. 2020.

[20] G. Bayable *et al.*, “Detection of water hyacinth (*Eichhornia crassipes*) in Lake Tana, Ethiopia, using machine learning algorithms,” *Water*, vol. 15, no. 5, 880, Feb. 2023.

[21] B. Roy, “Optimum machine learning algorithm selection for forecasting vegetation indices: MODIS NDVI & EVI,” *Remote Sensing Applications: Society and Environment*, vol. 23, no.1, 100582, Aug. 2021.

[22] T. Abebe, B. G. Awoke, and W. Nega, “Spatiotemporal patterns of water hyacinth dynamics as a response to seasonal climate variability in Lake Tana, Ethiopia,” *Applied Water Science*, vol. 13, 170, Aug. 2023.

[23] J. A. Coetzee, “Meteorological weather station data can be used in climate matching studies of biological control agents,” *Biocontrol Science and Technology*, vol. 22, no. 4, pp. 419–427, Apr. 2012.

[24] I. Aoyama and H. Nishizaki, “Uptake of nitrogen and phosphate, and water purification by water hyacinth *Eichhornia crassipes* (Mart.) Solms,” *Water Science and Technology*, vol. 28 no. 7, pp. 47–53. Oct. 1993.

[25] D. Kinsler, “Water hyacinth status,” *Rhodes University, 2021*. Accessed: February 19, 2024. Available: https://www.ru.ac.za/centreforbiologicalcontrol/resources/waterhyacinthstatus/

[26] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, “When and how should multiple imputation be used for handling missing data in randomised clinical trials: A practical guide with flowcharts,” *BMC Medical Research Methodology*, vol. 17, no. 1, 162, Dec. 2017.

[27] South African Weather Service (SAWS), “Hartebeespoort,” *SAWS Home — WeatherSA Portal*, 2022. Accessed: February 29, 2024. Available: https://www.weathersa.co.za/

[28] K. R. Reddy and W. F. Debusk, “Growth characteristics of aquatic macrophytes cultured in nutrient-enriched water: I. Water hyacinth, water lettuce, and pennywort,” *Economic Botany*, vol. 38, no. 2, pp. 229–239, Apr. 1984.

[29] Department of Water and Sanitation (DWS), “National eutrophication monitoring programme phosphorus and chlorophyll summaries,” *National Eutrophication Monitoring Programme*, 2024. Accessed: February 28, 2024. Available: https://www.dws.gov.za/iwqs/eutrophication/NEMP/report/NEMPyears.aspx

[30] J. M. Huizenga, “An inorganic water chemistry dataset of rivers, dams and lakes in South Africa,” *Centre for Water Sciences and Management*, Jan. 2013. Accessed: May 24, 2024. Available: https://www.waterscience.co.za/waterchemistry/data.html

[31] J. M. Huizenga, M. Silberbauer, R. Dennis, and I. Dennis, “Technical note: An inorganic water chemistry dataset (1972–2011) of rivers, dams and lakes in South Africa,” *Water SA*, vol. 39, no. 2, Apr. 2013.

[32] MATLAB, “evalclusters: Evaluate clustering solutions,” *MathWorks*, 2013. Accessed: February 24, 2024. Available: https://www.mathworks.com/help/stats/evalclusters.html

[33] S. Diouf, E. H. Dème, and A. Dème, “Imputation methods for missing values: The case of Senegalese meteorological data,” *African Journal of Applied Statistics*, vol. 9, no. 1, pp. 1245–1278, Jan. 2022.

[34] B. N. Eskelson, H. Temesgen, V. Lemay, T. M. Barrett, N. L. Crookston, and A. T. Hudak, “The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases,” *Scandinavian Journal of Forest Research*, vol. 24, no. 3, pp. 235–246, Jun. 2009.

[35] G. T. Ferrari and V. Ozaki, “Missing data imputation of climate datasets: Implications to modeling extreme drought events,” *Revista Brasileira de Meteorologia*, vol. 29, no. 1, pp. 21–28, Mar. 2014.

[36] R. Yang, “Analyses of approaches to deal with missing data in water quality data set,” in *Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)*, vol. 215, no. 1, pp. 1102-1108, Apr. 2022.

[37] G. E. Batista and M. C. Monard, “A study of k-nearest neighbour as an imputation method,” in *Proceedings of Soft Computing Systems: Design, Management and Applications*, 2002, vol. 87, pp. 251-260.

[38] Y. Dong and C.-Y. J. Peng, “Principled missing data methods for researchers,” *SpringerPlus*, vol. 2, no. 1, 222, May 2013.

[39] B. Zhang, D. Zhi, K. Zhang, G. Gao, N. N. Limdi, and N. Liu, “Practical consideration of genotype imputation: Sample size, window size, reference choice, and untyped rate,” *Statistics and its Interface*, vol. 4, no. 3, pp. 339–351, 2011.

[40] M. B. Mohammed, H. S. Zulkafli, M. B. Adam, N. Ali, and I. A. Baba, “Comparison of five imputation methods in handling missing data in a continuous frequency table,” in *Proceedings of Sciemathic 2020*, vol. 2355, no. 1, May 2021.

[41] MATLAB, “knnimpute: Impute missing data using nearest-neighbor method,” *MathWorks,* 2006. Accessed: March 23, 2024. Available: https://www.mathworks.com/help/bioinfo/ref/knnimpute.html

[42] C. D. Nguyen, J. B. Carlin, and K. J. Lee, “Model checking in multiple imputation: An overview and case study,” *Emerging Themes in Epidemiology*, vol. 14, no. 8, Aug. 2017.

[43] J. M. Curran, “The frequentist approach to forensic evidence interpretation,” *Encyclopedia of Forensic Sciences*, vol.1, no. 2, pp. 286–291, Dec. 2013.

[44] F. Marini, R. Bucci, A. L. Magrì, and A. D. Magrì, “Artificial neural networks in chemometrics: History, examples and perspectives,” *Microchemical Journal*, vol. 88, no. 2, pp. 178–185, Apr. 2008.

[45] S. Raschka, *Python machine learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Birmingham, United Kingdom: Packt Publishing, 2015.

[46] D. P. Kroese, Z. I. Botev, T. Taimre, and R. Vaisman, *Data science and machine learning: Mathematical and statistical methods*. Boca Raton, FL: Chapman & Hall/CRC, 2020.

[47] I. D. Dinov, *Data science and predictive analytics: Biomedical and health applications using R*. Cham, Switzerland: Springer International Publishing, 2018.

[48] S. K. Singh, S. K. Jha, and R. Gupta, “Enhancing the accuracy of wind speed estimation model using an efficient hybrid deep learning algorithm,” *Sustainable Energy Technologies and Assessments*, vol. 61, no. 1, 103603, Jan. 2024.

[49] M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, and A. Mukhopadhyay, “RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 218, no. 1, 105624, Jul. 2021.

[50] P. Fernando, “Modelling of growth cycle of water hyacinth: An application to Bolgoda Lake,” Master’s thesis, University of Moratuwa, Sri Lanka, 2003. Accessed: February 28, 2024. Available: http://dl.lib.uom.lk/handle/123/1317

[51] A. K. Bergström, “The use of TN:TP and DIN:TP ratios as indicators for phytoplankton nutrient limitation in oligotrophic lakes affected by N deposition,” *Aquatic Sciences*, vol. 72, no. 3, pp. 277–281, Mar. 2010.

[52] R. Ptacnik, T. Andersen, and T. Tamminen, “Performance of the redfield ratio and a family of nutrient limitation indicators as thresholds for phytoplankton N vs. P limitation,” *Ecosystems*, vol. 13, no. 8, pp. 1201–1214, Oct. 2010.

[53] B. Oancea and Ş. C. Ciucu, “Time series forecasting using neural networks,” *Challenges of the Knowledge Society*, pp. 1402–1408, 2014. Accessed: February 28, 2024. Available: https://arxiv.org/pdf/1401.1333.

[54] N. H. Agjee, O. Mutanga, M. Gebreselasie, and R. Ismail, “A comparison of regression tree approaches to modelling the efficacy of water hyacinth biocontrol using multitemporal spectral datasets,” *Journal of Spectroscopy*, vol. 2018, pp. 1–11, May 2018.