Long-Term Predictions of Coastal Dissolved Oxygen Using LSTM

Niamh Murtagh

Applied Research Project submitted in partial fulfilment of the requirements for the degree of

MSc in Data Analytics

at Dublin Business School

Supervisor: Lucas Rizzo

May 2022

Acknowledgements

First and foremost, a massive thank you to my supervisor, Lucas Rizzo, for all of his help, advice and understanding. And special thanks for helping me to choose a more realistic research project when I was over-zealous at the start with my plans.

Thanks also to Joanne O'Donnell for helping me get the academic support I needed. Many thanks to Susan and Peter for taking me in and always having words of encouragement.

The biggest thanks to my family - I couldn't have done any of my studies without the help and support of my parents and so I will always be indebted.

And, of course, massive thanks to Cian for keeping me sane(ish), and for being there with a coffee and a hug at all times.

Contents

1 Introduction

1.1 Background to Research

Climate change is warming our planet and oceans and causing the oceans to choke from warming-induced deoxygenation. Coastal hypoxia (low dissolved oxygen in seawater) is now one of the major threats to marine ecosystems, with hypoxic events continuously increasing in duration and frequency. Worldwide oceanic oxygen content has already decreased by over two percent since 1960, and oceanographers predict a decline in oceanic dissolved oxygen of up to seven per cent by the year 2100 (Schmidtko, Stramma and Visbeck, 2017). Dissolved oxygen (DO) (mg/L), meaning the amount of oxygen dissolved in water and biologically available, is an important measurement in assessing water quality.

Dissolved oxygen concentration is prone to high variation between different locations as it is affected by many factors: water temperature, elevation, industrial runoff, mixing of the water column, to name just a few. For most aquatic animals and plants, access to oxygen is necessary for survival. Generally, most fish need a minimum of 5mg/L of oxygen to survive (Muller-Karger *et al.*, 2018). Water containing between 2 and 6mg/L dissolved oxygen is considered hypoxic (dangerously low), and below 2mg/L is considered anoxic (fully depleted of dissolved oxygen) (*Third Integrated Report on the Eutrophication Status of the OSPAR Maritime Area*, 2017). The Environmental Protection Agency (2020) advise that salmon and trout (which are common off the Irish Atlantic coast) begin to be negatively affected when dissolved oxygen concentration reaches 6mg/L, and death is probable at 1.7mg/L. Where there are low concentrations of dissolved oxygen in water, fish may not meet the energy demands to carry out essential biological and metabolic functions, and so suffocation may occur (O Donncha and Grant, 2020).

1.1.1 Causes and Dangers of Low Oxygen

Oceanic water may be depleted of dissolved oxygen, firstly, for natural reasons, and where this is true, it is unlikely for dissolved oxygen to be so low in concentration that suffocation could occur. However, there are still associated risks. Fish ecology may be affected in the form of reduced energy for motion, reproduction, and growth. Risk of predation may also increase if fish need to rise to surface waters in order to access more oxygen (Kramer, 1987). dissolved oxygen may be low, secondly for artificial reasons. Artificial oxygen-depletion may be more extreme, and so, the risk of fish suffocation increases. Where there is organic waste entering a water system, for example, from nearby industry, sewage, or agricultural activities, dissolved oxygen will be reduced (McGovern, Nash and Hartnet, 2020).

Accurately predicting dissolved oxygen concentration is a necessary activity for the health of marine ecosystems. It enables policy-makers to make effective decisions regarding management and development of coastal areas. dissolved oxygen concentration is non-linear and influenced by many factors - microbial environment, nutrients, climate, presence of aquatic flora and fauna, etc., and therefore can be difficult to predict, (Liu *et al.*, 2021). The ability to accurately forecast harmful levels facilitates the introduction of mitigation measures in advance. It is also imperative for the sustainable development of offshore aquaculture. Aquaculture contributes massively to global food production, providing almost half of the total shellfish and fish humans consume. The UN predicts that by 2050, global demand for dietary protein will have increased by 70%, and increasing aquacultural actvities will help to provide this protein (FAO, 2022). dissolved oxygen concentration affects everything from fish survival, to growth rates, from occurrence of parasites and diseases, to what species are suitable for specific locations (Bailleul, Vacquie-Garcia and Guinet, 2015). The sustainable

development of aquaculture is extremely important, and being able to forecast dissolved oxygen will provide much-needed information for fish farmers.

Hypoxia is a serious and accelerating threat to coastal ecosystems, and it is widely accepted that climate change has been a major player in the decline of oceanic dissolved oxygen. Higher water temperatures mean lower oxygen solubility (Bailleul, Vacquie-Garcia and Guinet, 2015). Warmer water also means less circulation between surface water and deeper water. This leads to less nutrients rising to the surface water, and therefore, less oxygen-producing phytoplankton (Boyce, Lewis and Worm, 2010). The abundance of oceanic 'dead zones' is also increasing, i.e., areas of very low dissolved oxygen concentrations which support little to no life (Schmidtko, Stramma and Visbeck, 2017).

Anthropogenic activities also mean that coastal waters can be prone to heavy metal contamination. The inputs of heavy metals to aquatic systems can inhibit normal aquatic ecosystem activities, and pose a health threat to people who rely on these waters for drinking water and/or recreation (Ouyang *et al.*, 2018). Liu (2019) studied the release of heavy metals from coastal sediment as affected by dissolved oxygen, salinity nitrogen and phosphorus. They found dissolved oxygen concentration to be the strongest influence in the release of lead, cadmium, copper, and chromium. Bioavailability of those, and other heavy metals in the study, was found to be elevated when water was hypoxic (low in DO). It is obvious that dissolved oxygen concentration is an important measurement in the interest of not just fish health, but also the health of the public. It is also well-documented that contamination of seawater by heavy metals often leads to harmful algal blooms, i.e., abnormally fast growth of algae or cyanobacteria that causes harm to people, animals, and local ecosystems (Chen *et al.*, 2018; Ding *et al.*, 2018; Jin *et al.*, 2019).

1.2 Related Work

1.2.1 Machine learning and oceanography

Machine Learning techniques have been increasingly used in the field of oceanography for years and for diverse purposes, for example, modelling of sediment transport, (Goldstein, Coco and Plant, 2019), sea-level fluctuations (Cox, Tissot and Michaud, 2002; Brajard *et al.*, 2006), remote sensing (Krasnopolsky, 2007; Ahmad, 2019b, 2019a) (Ahmad, 2019b), and wave modelling (James, Zhang and O'Donncha, 2018).

The testing of and use of machine learning techniques for prediction of water quality variables in coastal environments is quite limited. There is still a lot of room for research into applying machine learning techniques to predict water quality variables in different environments, using different data, and widely measured input parameters. This is especially true for coastal dissolved oxygen (Yu, Shen and Du, 2020).

Numerous studies have employed satellite data to estimate water quality in coastal regions. Most commonlys they use the satellite data to derive reflectance in water and then a simple linear or non-linear regression model (Ahmad, 2019a). Using this type of data to estimate dissolved oxygen and other water quality parameters is inefficient for coastal areas and would be better suited to open ocean. Coastal waters contain coloured particulate and dissolved organic matter, which affect spectral models in intricate ways. This makes it challenging for a model developed for one coastal waterbody to be effective when applied to data from a different coastal waterbody (Kim *et al.*, 2014).

A study by Bailleul, Vacquie-Garcia and Guinet (2015) provided a creative workaround for an area where in situ DO measurements were not easily measurable; the Southern Ocean. Instead of obtaining data from a mooring with autonomous sensors as is the case with the current study, sensors were deployed on elephant seals. The study found that this type of data collection aided in understanding the biological sources of dissolved oxygen

in the water. It would be interesting to see if combining the two types of data collection when building a dissolved oxygen forecasting model would result in higher accuracy due to the input variables being better understood, but this remains for future work.

1.2.2 Predicting water quality parameters

A parameter other than DO that is sometimes used to measure water quality is phytoplankton biomass. It was successfully predicted (represented by Chlorophyll-a) with Random Forest model (Béjaoui *et al.*, 2018). Similarly, Multilayer Neural Networks (MLNNs) and Support Vector Regressor were used to model Chlorophyll-a (Jimeno-Sáez *et al.*, 2020). The coastal lagoon from which data was used in the study has been experiencing ecological degradation and a water quality crisis for decades. The research in question contributed to the awarding of ~€4 million in funding towards a project that is currently managing the sustainable recovery of the lagoon. This study provides an excellent example of how water quality modelling can ensure productivity of coastal ecosystems and improve the life of citizens (*Europe allocates 4 million to an international research project led by UCAM for the recovery of the Mar Menor*, 2021).

1.2.3 Machine learning vs traditional techniques

The continuous improvement of machine learning techniques means that they have become generally more efficient than traditional numerical models (Krasnopolsky, Chalikov and Tolman, 2002; Fourrier *et al.*, 2020). Numerical modelling is widely used in geology and related fields such as hydrogeology for simulation of physical properties using equations. Where data-based models calculate a logical solution, numerical models essentially guess the

solution using trial and error (*Analytical vs Numerical Solutions in Machine Learning*, 2021). The traditional, numerical models are more suited for use by those with domain knowledge (Valera *et al.*, 2020). Machine Learning models especially outperform numerical ones when there are limited input variables (Valera *et al.*, 2020) which can often be the case with collection of oceanic data. Thus, using machine learning over numeric models for DO predictions, theoretically provides the bonus of being able to limit data collection to variables that are easily measured. Machine learning techniques therefore facilitate almost real-time monitoring, not only of DO but of other biogeochemical parameters. Statistics-based models, as opposed to artificial intelligence- or machine learning-based models, can be more time efficient but are generally less accurate (Liu *et al.*, 2021).

There are two main ways to forecast dissolved oxygen. The first is to approach it as a regression problem, which seems to be generally more comprehensive, and flagging of hypoxic risk can then be automated. The second way is to convert the response variable to categorical by specifying and labelling thresholds of low, medium and high risk of hypoxia. However, setting these thresholds would require knowledge of the geographic area of the study and its inhabitants, and possibly even some guesswork. One study adopted the second approach and created a "Hypoxic Volume Index" (Muller and Muller, 2015). This research provides an excellent example of when using this approach is merited. 26 years of data was used with the aim of extending DO forecasts to the long-term. A wavelet-based neural network model was developed to predict hypoxic volume for years ahead. Monthly averages were analysed and predicted, giving a higher level view of hypoxic risk, meaning the loss in accuracy was not a concern. A second study used this approach but mapped outputs to binary classes indicating presence or absence of hypoxia (Virtanen *et al.*, 2019). They accept that transforming the output restricts the information in predictions but suggest that it is, again, merited for their purposes. As well as allowing forecasts for years ahead, it would allow

visualisation of geographic ranges that are prone to hypoxia, and identification of actionable areas for the needs of management.

1.2.4 Statistic-based models for DO forecasts

With regards to statistical models, Bayesian is represented in recent literature. Huan, Cao and Qin (2018) used a hybrid ensemble empirical mode decomposition (EEMD) with least squares support vector machine (LSSVM). Parameters were optimised using a Bayesian evidence framework. The study found Bayesian evidence framework to reduce prediction error and prediction time when compared with other optimisation techniques (Genetic algorithm and Ant colony algorithm). However the Bayesian regression model had already been improved upon in an earlier study (Khan, 2017). A novel autoregressive fuzzy linear regression method was proposed, and both approaches were capable of grasping the daily DO trend. However, the fuzzy method correctly predicted more low oxygen events than the Bayesian, and this is of greatest importance when modelling DO.

Another study used a statistical model, (grey model) to predict the trend in DO (Liu *et al.*, 2021). The study achieved successful aquacultural DO forecasting using multi-scale methods by decomposing signal features into several multi-scale features. These examples of statistical models have shown capability of accurately predicting DO, however such a recent use of a statistic-based model begs the question, why use them when AI-component models generally perform better?

1.2.5 Predicting DO in Ireland

The closest thing to an Irish example of this study being published is an application of multi-target regression for the prediction of three target variables to represent water quality. Data came from agricultural fields and water quality was represented by biological water quality, nitrogen concentration and phosphorus concentration (Nikoloski *et al.,* 2021). This paper provides useful knowledge regarding clustering and MTR for predicting water but uses three different target variables. The study needed to use multi-target prediction methods as agricultural water quality is more complicated to gauge than sea water. Oxygen is very easy to measure for coastal areas, so there are more options when modelling coastal water quality as it is more feasible to use only one target variable. In any case, the study presents evidence of a regression model that performed faster and more accurately at predicting multiple target variables, than one linear regression model per target variable. It is also useful that the study propses a way to use partially-labelled data that would usually be discarded. All in all, the research is probably more useful to inland water body research than coastal.

1.2.6 ANNs for dissolved oxygen/water quality modelling

Homayoun, Asadollahfardi and Heidarzadeh (2019) used neural networks to predict eutrophication in an Iranian Reservoir. The research presents a time delay neural network (TDNN) as an effective model for water quality variables. An LSTM may have provided a more accurate model as it uses backpropogation where TDNN does not. Also, the use of R-squared as an evaluation metric is questionable. It is not recommended for time series especially non-stationary time series data (Wooldridge, 1991; Davydenko and Fildes, 2016), a bracket which eutrophication data absolutely falls into. The only other metrics used were Root Mean Square Error (RMSE) and Mean Bias Error (MBE). RMSE is a useful metric here

but should be accompanied by other metrics that evaluate the actual magnitude of error. MBE can be useful to indicate model bias, but it does not indicate magnitude of error as positive and negative errors can cancel each other out (Ozyegen, Ilic and Cevik, 2022). It would have been more informative to provide further model evaluations, for example, Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), etc. The present study will make use of all of the mentioned metrics.

Another example combined kernel Principal Component Analysis with RNN for prediction of dissolved oxygen which they present as stable, highly accurate (RMSE of 0.394) and versatile enough to be suitable for predicting water quality indicators other than dissolved oxygen (Zhang, Fitch and Thorburn, 2020). However, there is no evidence in the study that the model was tested on anything other than dissolved oxygen. Another (argubly major) downfall of this study is that the predictive horizons used to test the model only go as far as three hours into the future. The research presented here will use a predictive horizon of 32 hours which would be far better for real life use as a lot more damage control can be carried out in 32 hours than in three. Also the authors suggest that for future work it may be useful to include rainfall data which they did not, but which will be included here.

Zhang *et al.*, (2019) proposed a water quality predictive model based on a multi layer feed-forward neural network, and combined this with mutual information feature selection. Their research presents a very useful comparison of models that do and do not use a dropout layer, which they found to be beneficial when included. Their model performed well, however they again only make very short term predictions of 90 and 120 minutes. The accuracy of their model is excellent, however, it is difficult to imagine that 90 minute forecasts of dissolved oxygen are realistically useful. Also, the ReLU (Rectified Linear Units) activation function was used which is an unusual choice for that kind of an analysis, as it can create dead neurons. Here, hyperbolic tangent (tanh) activation function will be used rather

then ReLU. ReLU is more computationally efficient, but that is an unconvincing reason for its use considering the predictions are for a maximum of 120 minutes, data points are 30 minutes apart and the historical window used was 12 previous data points. Regarding the usefulness of the predictions, it is possible that in contained aquacultural facilities like inland fisheries, ponds, and possibly even in rivers like the one the researchers used data from, there would be a use for such short term predictions as artificial aeration of water could be automated. However, with coastal or oceanic water, this is not the case.

1.3 Rationale & Aims for This Research

This study aims to present an evaluation of multivariate Long Short Term Memory (LSTM) model for the prediction of DO. A simple multiple linear regression (MLR) model will also be presented for comparative purposes. The LSTM model will treat the computational task as one of time series analysis, as opposed to the MLR which will not employ any time series methods nor treat the data as sequential. Both models will use minimal independent variables to reflect the wider availability of ocean chemistry data. The use of LSTM will be presented and examined as a way to model the data sequentially with the help of input variables that influence dissolved oxygen as the target variable. The main focus of the research is the LSTM model but the MLR will provide a comparison by various evaluation measures so as to assess whether treating the data as sequential (LSTM) or non-sequential (MLR) is more effective.

DO models are often specific to an area and do not always translate well from one coastal waterbody to another. This study will optimize DO prediction for a location not yet studied in this way, using data not yet used for related publications. Every study of this type for a yet unused dataset/geographical location adds to the working knowledge of DO

prediction and helps towards a generalised model. Many ecological factors influence DO and each ecosystem is different. Having optimised models for numerous ecosystems and waterbody types contributes towards DO being a universally forecastable water quality measure. Additionally, LSTM models offer the opportunity for a framework that is doubly useful - it may also help to fill in gaps in data. It is inevitable for in situ observations to have gaps for many reasons including biofouling, repairs and servicing of instruments, loss of instruments, changes to funding, etc.(Contractor and Roughan, 2021).

Building on from the limitations of previous research as mentioned previously, this study will also aim to provide a "proof-of-concept" for longer-term predictions of dissolved oxygen. While previous studies provide short-term forecasts, the predictive horizon in this study will be extended to 32 hours. Additional evaluation metrics will be employed here to give a better, more rounded evaluation of model performance.

1.4 Research Methodology

1.4.1 Crisp DM

The main body of this study follows the Cross Industry Standard Proces for Data Mining (Crisp DM) methodolgy. Crisp DM is a very useful six-step process for data mining which is both industry-neutral and technology-neutral. It is very helpful in providing a guideline that can be used in argubly any data science project (Wirth and Hipp, 2000). The six steps which have been laid out as the Crisp DM process and which will be generally followed for this research are as follows:

1. *Business Understanding*

   This first phase involves determining the objectives of the project, taking inventory of data and other resources, formulating a success criteria and a project plan. The project plan for this research will be set out in Figure 1 in this section

2. *Data Understanding*

   The data is collected, described, and exploratory analysis is carried out. This step will be presented in section 2.2

3. *Data Preparation*

   The final dataset is constructed in this step. This involves data cleaning, data selection, and any necesssary restructuring and derivation of attributes. This step will be presented in section 2.3

4. *Modelling*

   The model is compiled in this step and parameters tuned. Preliminary runs of the model are performed so as to determine any aspects of the model that need tweaking. The model building will be presented in section 2.4

5. *Evaluation*

   The model's performance is evaluated through use of various performance metrics. This work will be presented in the section 4

6. *Deployment*

Once a model is running satisfactorily a plan is put in place for deployment. This step is more relevant to a business than to this research as this is more of a proof of concept than a project that will be put into practice by a specific company.

This type of research can be described as quantitative, fixed, and experimental. The research design is illustrated in FIG below.

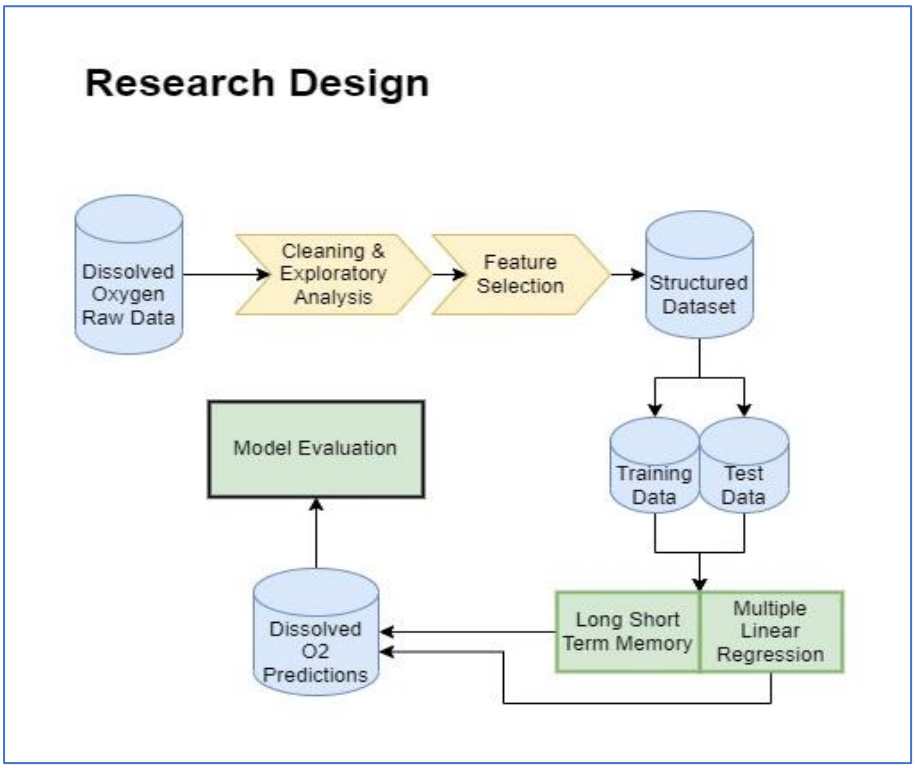

Figure 1 Research Design

Figure 1 above depicts the working order of the present research with steps such as raw data cleaning, feature selection, running of both models, and evaluation of model

predictions. The steps in the diagram correspond loosely with steps of the Crisp-DM methodolgy.

1.4.2 Hypothesis:

Specifically, this research will test the hypothesis that LSTM networks are a viable method of making long-term dissolved oxygen concentration predictions.

Null Hypothesis (H_0) => Long Short Term Memory networks are not useful for predicting coastal dissolved oxygen concentrations. They cannot outperform a much more simple model like multiple linear regression.

Alternative Hypothesis (H_1) => Long Short Term Memory networks are useful for predicting coastal dissolved oxygen concentrations, and moreso than a much simpler multiple linear regression.

1.5 Time Series Data

In order to understand the data that will be used in this research it is important to have an understanding of how time series data works. Time series data is a type of sequential data in which the instances are associated with a time dimension. Each row in timeseries data represents a unique timestamp. A major difference between sequential and non-sequential data is that the instances are not independent of each other. Most supervised learning algorithms assume that input data is Independent and Identically Distributed (IID). This does

not hold true for time series data (Preeti, Bala and Singh, 2019). For example, if a meteorologist were forecasting temperature at a specific hour, it is useful to know the temperature in the timestamps leading up to that hour. With IID data, for example fraudulent and non-fraudulent insurance claims, the previous rows have no bearing on the outcome of the present row.

A popular way of analysing sequence or time series data is with Recurrent Neural Networks (RNN). RNNs have a hidden layer that receives information from the input layer of both the current step and the previous one, as opposed to the standard feed forward neural network which only intakes one input at a time. RNNs are essentially built to learn context and then use this learned context when making predictions (Che *et al.*, 2018).

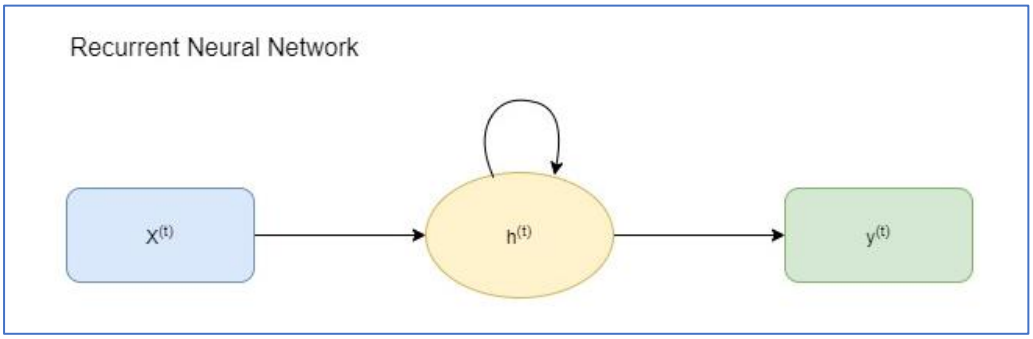

Figure 2 RNN architecture

Figure 2 above displays RNN architecture including three vector nodes. Input (X), hidden (h) and output (y) layers are represented by one node only but would generally contain many units. The hidden layer shows what is known as a Recurrent Edge (represented by an arrow).

There are three requirements for a neural network (Bengio, Simard and Frasconi, 1994):

1 The system has the ability to store information for a non-predetermined duration

2 The system is not overly sensitive to noise, and

3 The parameters can be trained within a reasonable timeframe

1.5.1 Bidirectional RNN (BRNN)

Bidirectional Recurrent Neural Networks (BRNN) solve the problem of modelling an output that depends on the whole input sequence rather than the current and one previous input (Schuster and Paliwal, 1997). BRNNs combine two RNNs: the first moves forward from the beginning of the series; the second moves backward from the end of the series.

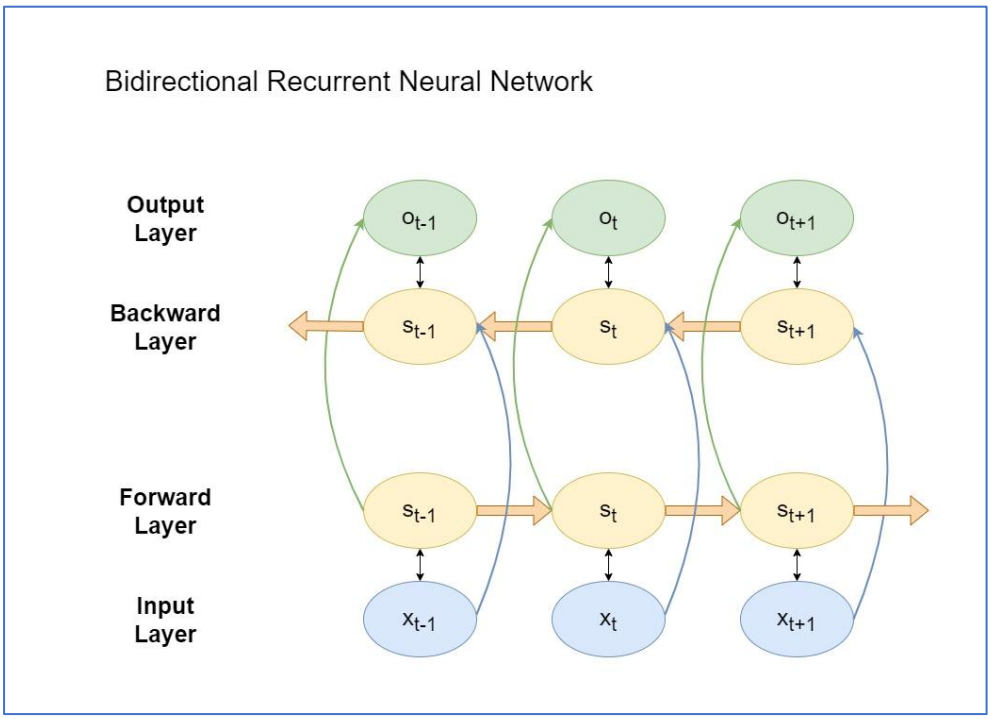

Figure 3 Bidirectional RNN

Figure 3 above displays a simplified example of a BRNN, where:

s = the state variable (a term for the mathematical state of a system which in theory describes the system enough to determine its future state, provided it is without external actors affecting it)

s_t = state variable of the hidden layer with respect to time t

o = state variable of the output layer

o_t = state variable of the output layer with respect to time t

1.5.2 LSTM & The Vanishing Gradient Problem

While adding a bidirectional layer to an RNN solves the issue of lost context from future datapoints, there is still an issue of "vanishing gradient". As information passes from input to output neurons in an artificial neural network (NN), the error is calculated and fed back through the network to inform model weight. The gradient descent algorithm works in neural networks to calculate the minimum cost function (cost function calculates the difference between predicted values and actual values).

With Recurrent NNs, there is added complexity because (1) the information is travelling with respect to time and using data from numerous timestamps; and (2) the cost function is calculated locally at each timestamp rather than globally. With every cost function that is calculated, an update needs to be applied to the weight of every neuron that contributed to the output. (This number of neurons may be very high considering many previous timestamps are used in the neurons). Weights assigned at the inception of the neural network are floating numbers close to 0 and when values are multiplied repeatedly by a wrec (weight recurring) value that is between 0 and 1, the gradient approaches closer and closer 0, hence the term "vanishing gradient".

The further along the neurons the model goes, the lower the gradient and therefore the more difficult training becomes. Weights of preceding steps update much slower than those that follow, and there is a knock-on effect because the ones that follow are relying on information from those that precede. Conversely, there is an opposite problem when weights assigned at the beginning are larger numbers and result in an "exploding gradient". In short, wrec less than 1 results in a vanishing gradient and wrec greater than 1 results in exploding gradient (Bengio, Simard and Frasconi, 1994; Bengio *et al.*, 2003).

Long Short Term Memory networks were invented as a solution to the vanishing gradient problem. Intuitively, it seems that setting wrec as 1 would solve the issue, and although this is a grand over-simplification of LSTMs, it is, in a sense, what they do.

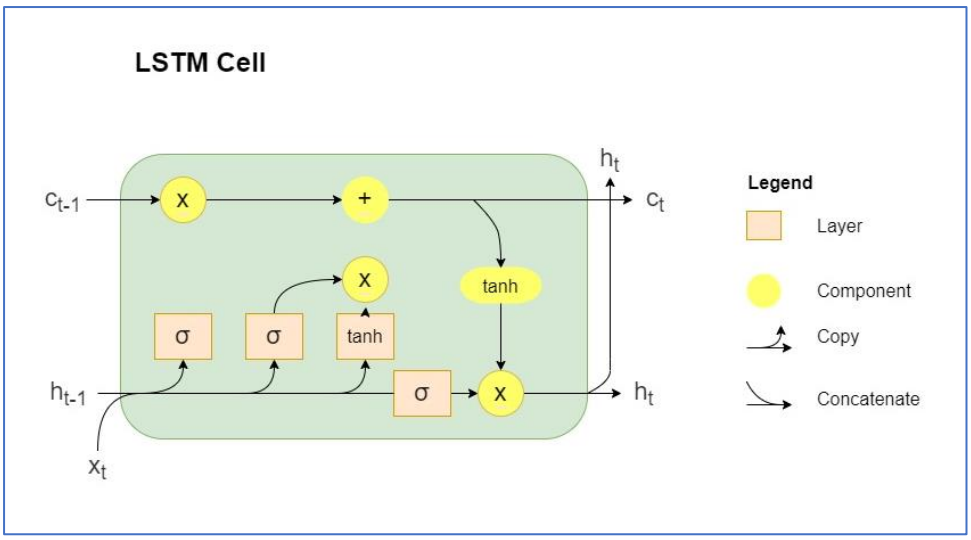

Figure 4 LSTM Cell

Figure 4 above shows the structure of an LSTM cell.

x_t = input at time t

h_t = output at time t

c_t = cell state at time t

σ = sigmoid activation function

tanh = hyperbolic tangent activation function

LSTM controls the cell state using a gate-style structure. Gates include input gates, forget gates, and output gates. Input gates control what information is added to the cell. Forget gates determine what information is unnecessary and left behind. Finally, the output gates are what determines the output value of that cell.

The calculation process of the LSTM memory unit is laid out in the following formulae (Pascanu *et al.*, 2013):

$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$

$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$

$\hat{c}_t = tanh(W_{\hat{c}} h_{t-1} + U_{\hat{c}} x_t + b_{\hat{c}})$

$C_t = C_{t-1} \odot f_t + i_t \odot \hat{c}_t$

$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$

$h_t = o_t \odot tanh\,(C_t)$

$\hat{y}_t = \sigma(V h_t + c)$,

Where;

x_t = input at time *t*,

h_t = output at time *t*,

C_t = cell state at time *t*,

f_t = output of forget gate at time *t*,

i_t = output of input gate at time *t*,

o_t = output of output gate at time *t*,

W = weight of the model,

b = bias of the model,

σ = sigmoid activation function

LSTM networks are very effective for time series analysis (Ozyegen, Ilic and Cevik, 2022). They are capable of learning patterns in non-stationary data (Preeti, Bala and Singh, 2019).

They backpropogate error information through time in an unrolled neural network, meaning the entire input sequence receives this information. This allows temporal dependencies to be better captured. A stacked LSTM network can be built by stacking two or more LSTM layers. The hidden layer in the first LSTM is fed through to the second LSTM layer and a dropout layer may be added for removal of a user-specified amount of data (Che *et al.*, 2018)

2 Materials and Methods

2.1 Materials and Technologies

2.1.1 Study Area: Mace Head Observatory, Co. Galway

In the Crisp-DM methodology, it is important to have an understanding of the business. As this is not strictly a buiness data mining project, presented here is the background to the area of study. This research will use the "COMPASS" dataset collected by the Marine Institute at their Mace Head site. The data is available publically under the CC-BY-4.0 license.

Mace Head Obervatory is an Atmospheric Research station in Carna, about 57 kilometres from Galway city. The coastal habitats around Galway and Mace Head Observatory are high in biodiversity and home to some rare invertebrates. The study location is part of the Cill Chiaráin Bay and Islands SAC (Special Area of Conservation). Covering 213.99km(sq), it has the second greatest diversity of marine species in Ireland, second to Kenmare River (MERC Consultants, 2005).

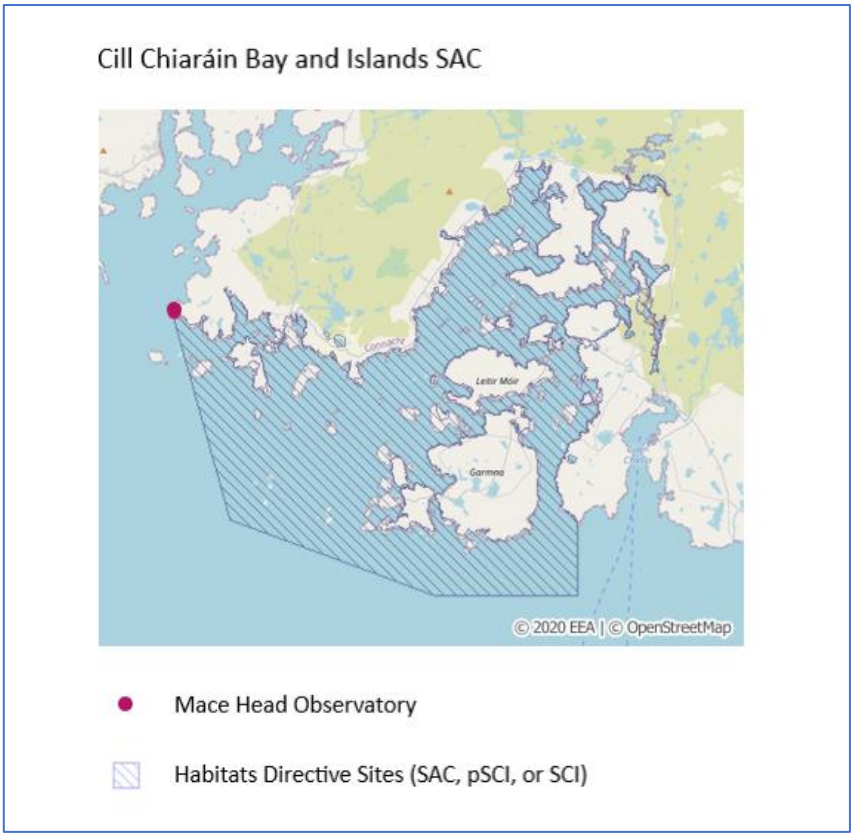

Figure 5 Kilkieran Bay

The area under the SAC is 93% marine. There are 9 habitats types in the SAC that are protected under the Habitats Directive, and 7 species protected under the Nature Directives (Barnacle Goose, Little Tern, Common Tern, Arctic Tern, Naiad, Common Seal, and European Otter). Cill Chiaráin Bay is the only known Irish habitat of Mesacmaea mitchelliiz.

The area is described as showing little effects of human activity. Every ecosystem has been impacted in some ways but the fact that the impact here is minimal may prove useful for getting a more location-versatile model, rather than one built specifically for a habitat with very specific anthropogenic influence.

2.1.2 Technologies: Python & packages

All of the data preparation, modelling, and evaluation activities were carried out in the Google Colaboratory Pro environment, Colab for short. Colab iss an integrated development environment (IDE) and is especially suitable for data analysis activities and machine learning. It was used here as a Python environment as it is user-friendly, cloud-based and can optionally be run on Graphics Processing Unit (GPU). Here, a cloud-based GPU was used as it provided excellent computational speed, which can sometimes be an issue when running deep learning tasks. Colab requires no set-up and easily integrates machine learning packages like TensorFlow, and Keras, which are very important packages in this research. TensorFlow is an open-source machine learning library. Keras runs on top of TensorFlow and is a neural network library. The two can run without each other but Keras can use TensorFlow, among other things, as its backend. All packages and licenses are listed in Appendix A.

2.2 Data Understanding

The data consists of 23 columns and 48494 rows. It has a time component and the range is from 1st June 2019 at 00:00:00 up until 14th April 2022 12:00:00. This is a total of 35 months or 151 weeks of data. Data collection was at 30-minute intervals.

The variables are listed in the table below:

	COLUMN NAME	# NON-NULL ROWS	DATA TYPE	USED IN ANALYSIS: Y/N
0	time	48494	discrete	**Y**
1	latitude	48494	discrete	N
2	longitude	48494	discrete	N
3	air_pressure	48493	continuous	**Y**
4	air_temperature	48493	continuous	N
5	contros_current_avg	48493	continuous	N
6	contros_pco2_avg	48493	continuous	**Y**
7	contros_voltage_avg	48493	continuous	N
8	average_percipitation	43308	continuous	N

9	sbe_conductivity_avg	48493	continuous	N
10	sbe_dissolved_oxygen_avg	48494	continuous	**Y**
11	sbe_salinity_avg	48494	continuous	**Y**
12	sbe_temp_avg	48494	continuous	**Y**
13	scufa_raw_fluorescence_avg	14747	continuous	N
14	seafet_ph_ext_avg	48493	continuous	**Y**
15	seafet_ph_int_avg	48493	continuous	N
16	seafet_temp_avg	48493	continuous	N
17	suna_absorbance_254nm_avg	48493	continuous	N
18	suna_absorbance_350nm_avg	48493	continuous	N
19	suna_bromide_avg	48493	continuous	N
20	suna_dark_value_avg	48493	continuous	N
21	suna_main_current_avg	48493	continuous	N
22	suna_nitrate_conc_avg	48493	continuous	**Y**
23	suna_nitrogen_avg	48493	continuous	N
24	total_precip_avg	43308	continuous	**Y**

Data was collected by multiple pieces of equipment, and there is some overlap in variables, for example the SBE 26 SEAGAUGE records temperature data (*'sbe_temp_avg'*) as does the SeaFET ('*seafet_temp_avg*'). Some variables contained information on technical aspects of the equipment, and were removed, as described in section 2.3 below. The data was quite messy as it contained moderate amounts of missing data and very large amounts of invalid data.

2.2.1 Pearson Correlation

FIG below shows the Pearson corellation values. Most variables have been renamed at this stage for ease of use in Python, and a '*month*' feature was extracted to see if the dependent variable, '*dissolvedO2'* or indeed any of the independent variables had strong correlations with the month and therefore could be estimated to be seasonal. There exists a weak negative correlation between '*month*' and '*dissolvedO2*' (*P*=-0.38). Dissolved oxygen concentrations can definitely be seasonal, especially as there is such a link between water temperature and oxygen solubility in water (Virtanen *et al.*, 2019). However, Pearson

correlation is linear but months are cyclical, which explains why there is not a stronger correlation showing between the two variables.

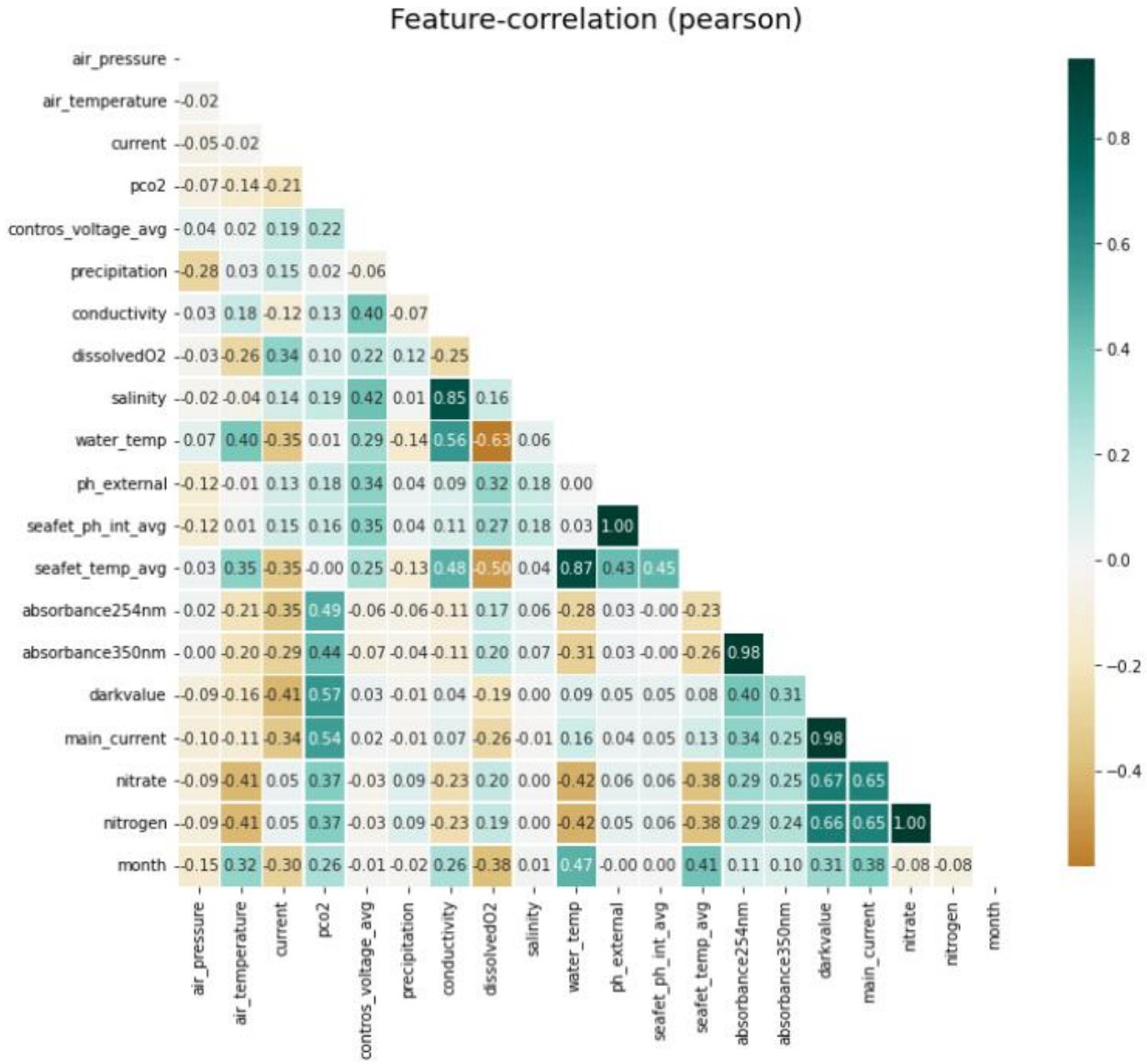

Figure 6 Pearson Correlations

Dissolved oxygen and water temperature show a moderately strong negative correlation (*P*=-0.63). This is to be expected as it is well-documented that oxyegn becomes

less soluble in water as temperatures increase (e.g., Salami and Ehteshami, 2015; Gobler and Baumann, 2016; Virtanen *et al.*, 2019).

2.2.2 Descriptive Statistics

The table below shows descriptive statistics for the dataset. Please note this describes the data after cleansing activities have been carried out. For more information on what changes were made see section 2.3.

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
air_pressure	1010.31	13.68	969.13	1001.02	1011.44	1020.02	1047.86
air_temperature	7.88	13.21	-46.80	8.46	10.45	12.87	23.45
pco2	399.44	44.40	239.77	381.23	404.41	426.98	505.54
rainfall	0.41	1.11	0.00	0.00	0.00	0.13	13.19
conductivity	3.88	0.32	3.08	3.60	3.84	4.17	4.52
dissolvedO2	8.24	1.23	5.01	7.78	8.52	9.20	10.74
salinity	33.92	0.71	28.97	33.57	34.04	34.44	35.13
water_temp	11.79	3.03	6.49	9.06	11.25	14.64	17.88
ph_external	7.92	0.13	7.35	7.85	7.96	8.02	8.25
nitrate	4.87	3.62	-2.30	1.78	4.13	8.13	17.32
nitrogen	0.07	0.05	-0.03	0.02	0.06	0.11	0.24
month	7	4	1	3	8	10	12

2.3 Data Preparation

2.3.1 Missing Values

Missing values of the whole dataset were visualised using matplotlib package, figure 7.

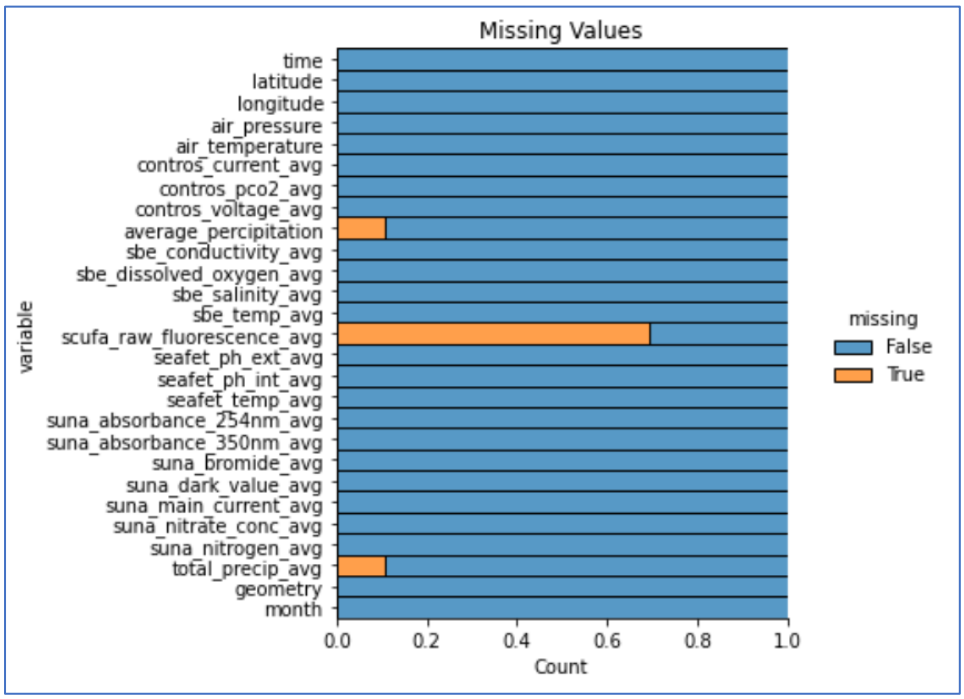

Figure 7 mising values

The *klib* package was used for preliminary data cleaning (e.g., checks for duplicate rows and columns, removes single-valued columns).

A plot was made of distribution of *'average_percipitation'* as visualising would help in deciding method of dealing with missing values, figure 8.

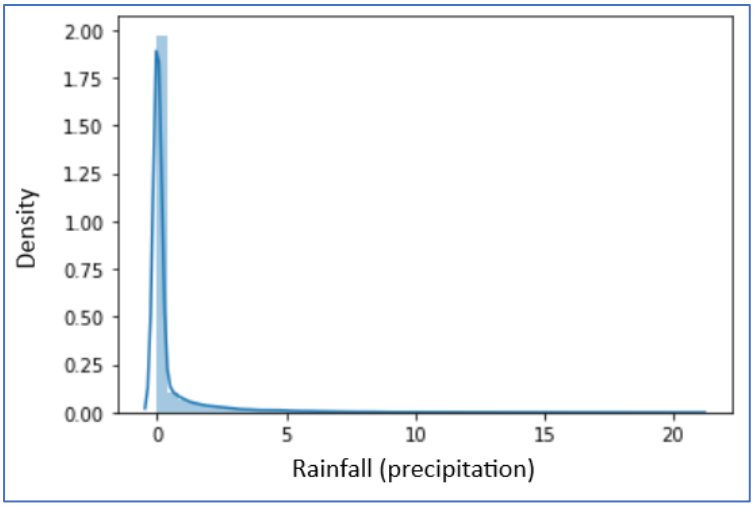

Figure 8 rainfall

Value of 0 was shown to be extremely frequent in the rainfall column compared to all other values (count was 33986). This makes logical sense – all it means is that most of the time there was no rain. Missing values in this column were therefore imputed with the mode. Missing values were plotted again to confirm have they all been dealt with, Figure 9.

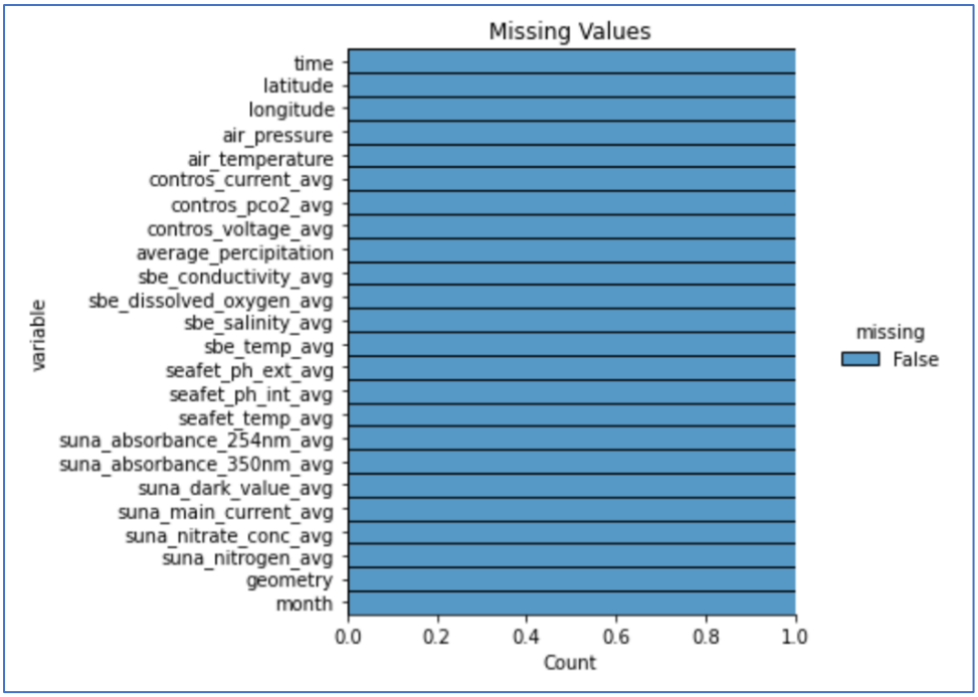

Figure 9 missing values

2.3.2 Invalid Data

Instances of 0 values in columns were checked as some columns should not contain zeros. The dissolved oxygen column contained 0 values, however they are equipment errors. It was visible from the index that DO changed to 0 just for half an hour at a time which is not realistic. If it was actually depleted as low as 0mg/L, it would be 0mg/L for a while either side. The instances where it is 0 for more than 30 minutes (one row) there are also 0s for other values in the row so that means that the equipment lost communication or otherwise temporarily malfunctioned. 137 rows where this was true were dropped for the purpose of the Multiple Linear Regression, as MLR treats data as non-sequential. For LSTM, it is better not to have missing timestamps so deletion of rows was avoided (Schuster and Paliwal, 1997;

Muzaffar and Afshari, 2019). Instead, o2 values between 0 and 2 were forward-filled, FIG.

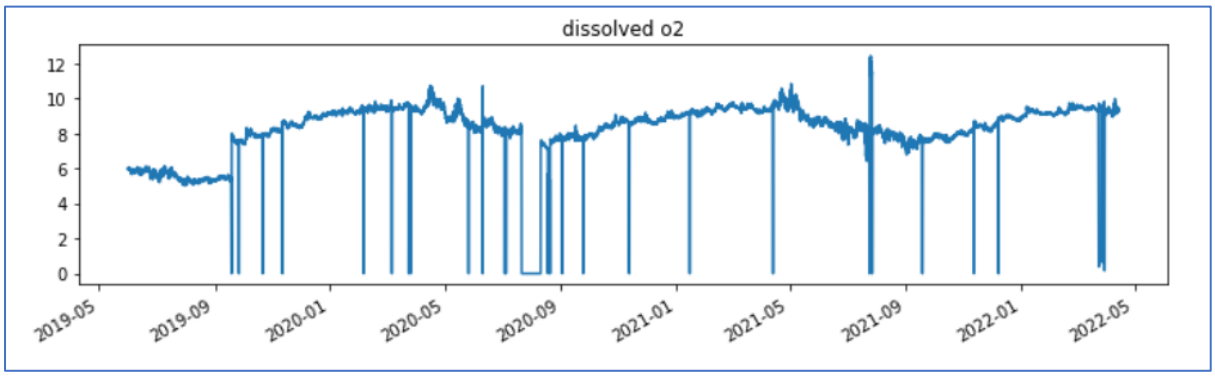

Figure 10 before filling

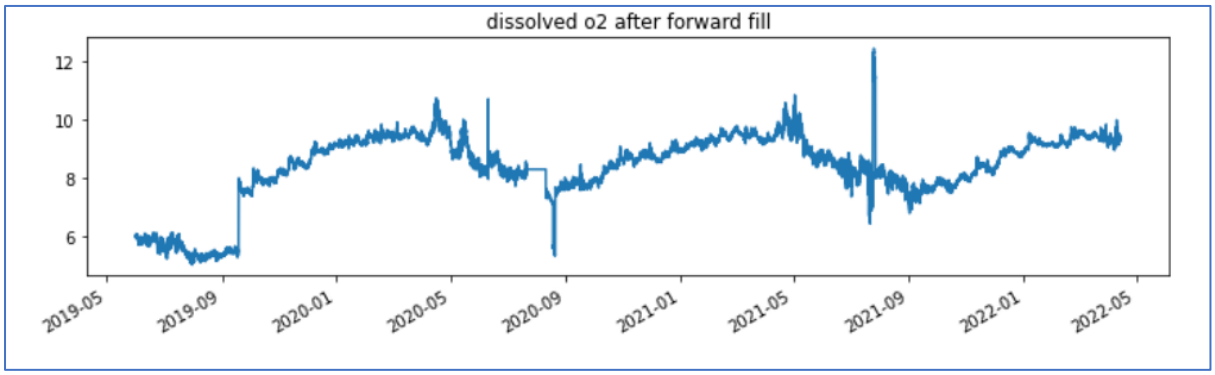

Figure 11 afetr forward filling

The '*pco2'* column also contained many zeros, FIG

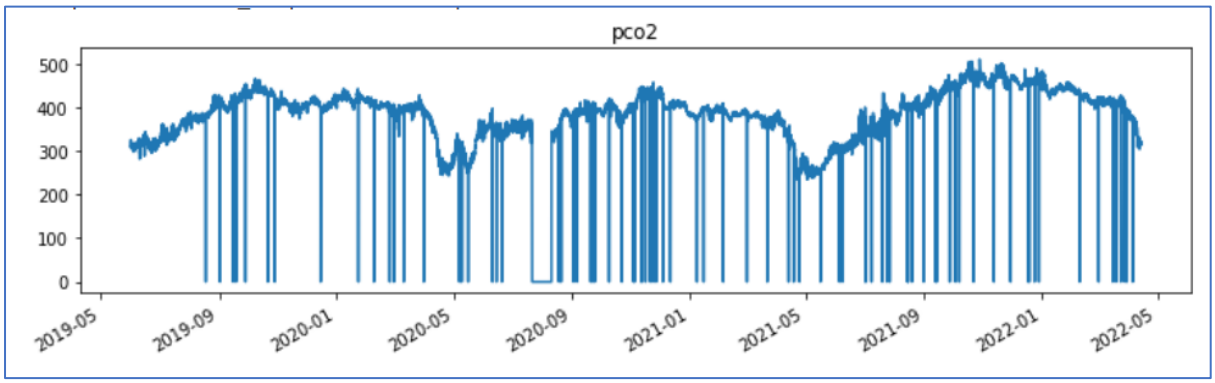

Figure 12

This column was also forward-filled for LSTM as it is relatively robust against outliers (Muzaffar and Afshari, 2019), so a simple forward-fill method was deemed sufficient. FIG shows '*pco2'* after the forward fill.

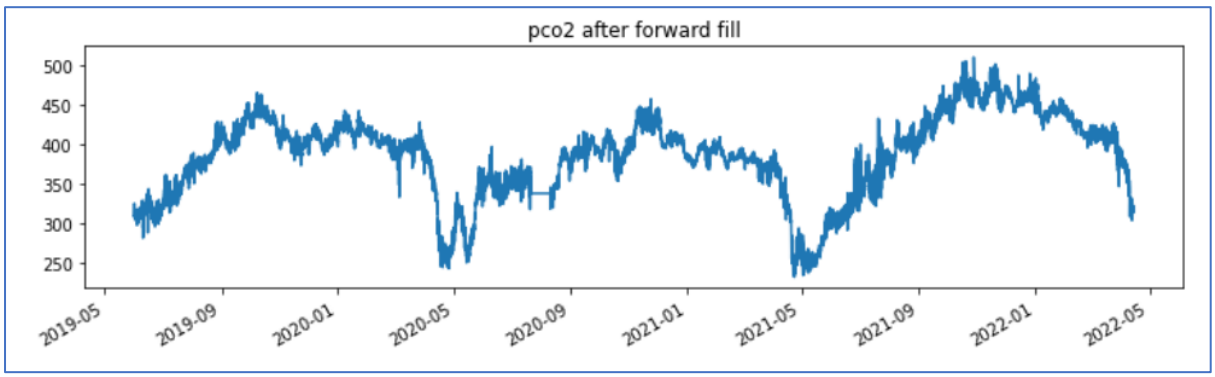

Figure 13

For the MLR, the rows were dropped where '*pco2'* was equal to 0. Outlier detection was also performed for the MLR data using the K Nearest Neighbours and 2660 outliers were removed.

2.3.3 Unsupervised Feature Selection

As part of preliminary (and unsupervised) feature selection, variables that are not useful were removed. Superfluous variables were avoided so as to reflect as closely as possible the Essential Ocean Variables set out by UNESCO in The Global Ocean Observing System. The EOVs provide a framework to adopt common standards for interntional ocean data collection and to maximise its utility and cost effectiveness.

Columns were removed for the following reasons:

'contros_voltage_avg' measurement of voltage of the equipment (Seabird scientific).

'seafet_ph_int_avg' this is the interior pH measurement. It is not needed because the exterior measurement is the main one and the interior is an insurance policy against the exterior one. If they become different from each other that means calibration is needed. Plotting the interior and exterior pH against each other gave a straight line therefore there are no issues and we can use the exterior measurement, FIG.

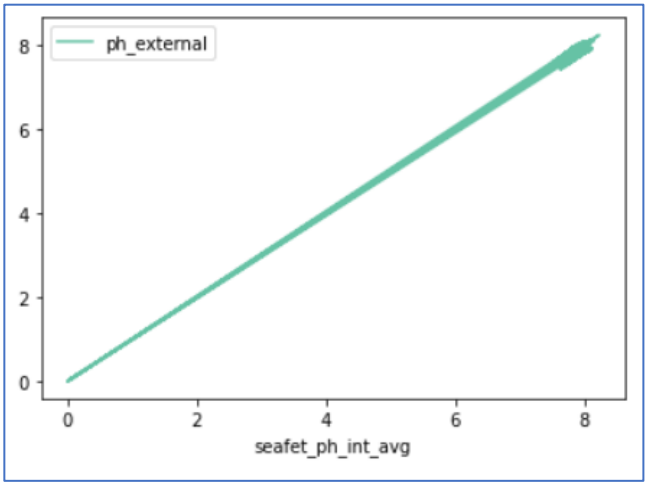

Figure 14 ph internal and external

'suna_main_current_avg' an electrical measurement that is equipment related. The nitrate sensor uses this for nitrate calculations. This is not needed as there is already a column with nitrate info.

'seafet_temp_avg' there is already a water temperature column.

'nitrate_conc' 1.0 correlation with nitrogen; they are two forms of the same thing (nitrogen being an element and nitrate an ion of that element).

'scufa_raw_fluorescence_avg' not widely measured, also contains more than 60% missing values.

'total_precip_avg' same information as '*precipitation*'.

'suna_absorbance_254nm_avg' not widely measured especially for DO forecasting purposes.

'suna_absorbance_350nm_avg' as above.

'suna_bromide_avg' detected by klib as being single-valued.

'suna_dark_value_avg' technical information for nitrate counter.

'suna_main_current_avg' equipment-related technical information.

'sbe_conductivity_avg' this is derived from salinity and salinity is more important to keep in.

This leaves the following columns for analysis:

LSTM: time, dissolvedO2, salt, temp, ph, nitrate, air_pressure, pco2, rainfall

MLR: latitude, longitude, air_pressure, air_temperature, pco2, precipitation (rainfall), conductivity, salinity, water_temp, ph_external, nitrate, nitrogen, month

2.4 Code Development

2.4.1 LSTM

This section will describe the building and stacking of the model and its layers, and a description of some important model arguments. All parameters and arguments are available in the appendices.

Data was resampled from half hourly to hourly as this would cut rows in half and thus save on computation. (When the model was run without hourly resampling, there was no improvement in performance). It was then scaled using a MinMax scaler.

Next, a function was defined to process the time series data in a suitable way for LSTM. (See '*custom_ts_multi_data_prep'* in code.) The purpose of the function was to split the dataset up into 'histories' and 'horizons', to be specified by the user. The function would then return the data as numpy arrays to be fed to the model. 'History' is the number of backward timesteps the cell is shown and bases the prediction off of, and 'horizon' is the number of forward timesteps. Here, history was set to 84 and horizon was set to 32.

Next, the last 32 rows of the dataset were removed as a validation set. It is important in a recurrent neural network to have not only a train set and test set but a validation set also Next, the last 32 rows of the dataset were removed as a validation set. It is important in a recurrent neural network to have not only a train set and test set but a validation set also. The LSTM is training and testing each epoch and using the testing errors to inform the next cells. By holding out data for validation and keeping it seperate from model training, the final evaluation can then be performed on data that has not been used for fine-tuning the model (Davydenko and Fildes, 2016; Contractor and Roughan, 2021).

Batch size can then be specified by the user, a multiple of 32 is best to make use of the GPU. Batch size of 512 was used here. Setting batch size in this way so that it is greater than 1 but less than the total number of rows is called Minibatch Gradient Descent.

The LSTM model was then built by stacking several Keras Sequential layers. Sequential is always used for neural networks in keras. First, a bidirectional layer - this is the forward layer, with 100 LSTM units passed to it and '*return_sequences'* set to 'True' which allows stacking of LSTMs . If set to 'False' then there is no sequence information for a cell to pass onto the next cell. Then a 'dense' layer which is the fully connected layer, set to 20 units. Next, a second bidirectional layer, this time with 80 LSTM units. A second bidirectional layer must be passed for the backpropogation component.

Next, three more dense layers were stacked, each with 20 units. Hyperbolic tangent or '*tanh*' activation function was used in all dense layers as it gave the best performance. Activation function, also known as transfer fucntion, is the method used for transformation from weighted sum of input to output (Rasamoelina, Adjailia and Sincak, 2020). The *tanh* function is effective because it is between -1 and 1 and adds weight to individual values. This avoids the vanishing gradient issue. *tanh*'s second derivative can sustain for a long range before going to 0 (sigmoid's second derivative goes to 0 quite quickly).

Next a dropout layer was added with dropout set to 0.5. Dropout is a regularisation technique which is suggested as a way to reduce overfitting. Random neurons are dropped, and so, complex co-adaptations are avoided (Srivastava *et al.*, 2014). Finally, another dense layer is added to serve as the output layer, and the horizon (32) is passed so that the model will make predictions for 32 hours into the future. However, *sigmoid* activation function is used within the forget gate because it can output 0 to 1 so it can be used to either forget or remember the information (0=forget completely, 1=let everything pass through). A model summary is available in Appendix

The model is compiled with callbacks such as 'early stoppings' which monitors validation loss. This means if validation loss does not improve for (patience=) 10 iterations then it will be stopped. Epochs is set to 100, after which there would likely be some

overfitting, however the callbacks will prevent this. Complete parameters are below in the table.

Parameter	Value passed
train/test split	65/35
history	84
horizon	32
batch size	512
buffer size	150
activation function	tanh
dropout	0.5
optimiser	adam
loss monitor	mean square error
early stop min delta	0 (default)
early stop patience	10
early stop mode	min
epochs	100
steps per epoch	100
validation steps	50

3 Results & Discussion

This section will present results of correlations, model predictions and evaluation metrics.

3.1 Correlations

Note that this correlation analysis was carried out on the cleaned data.

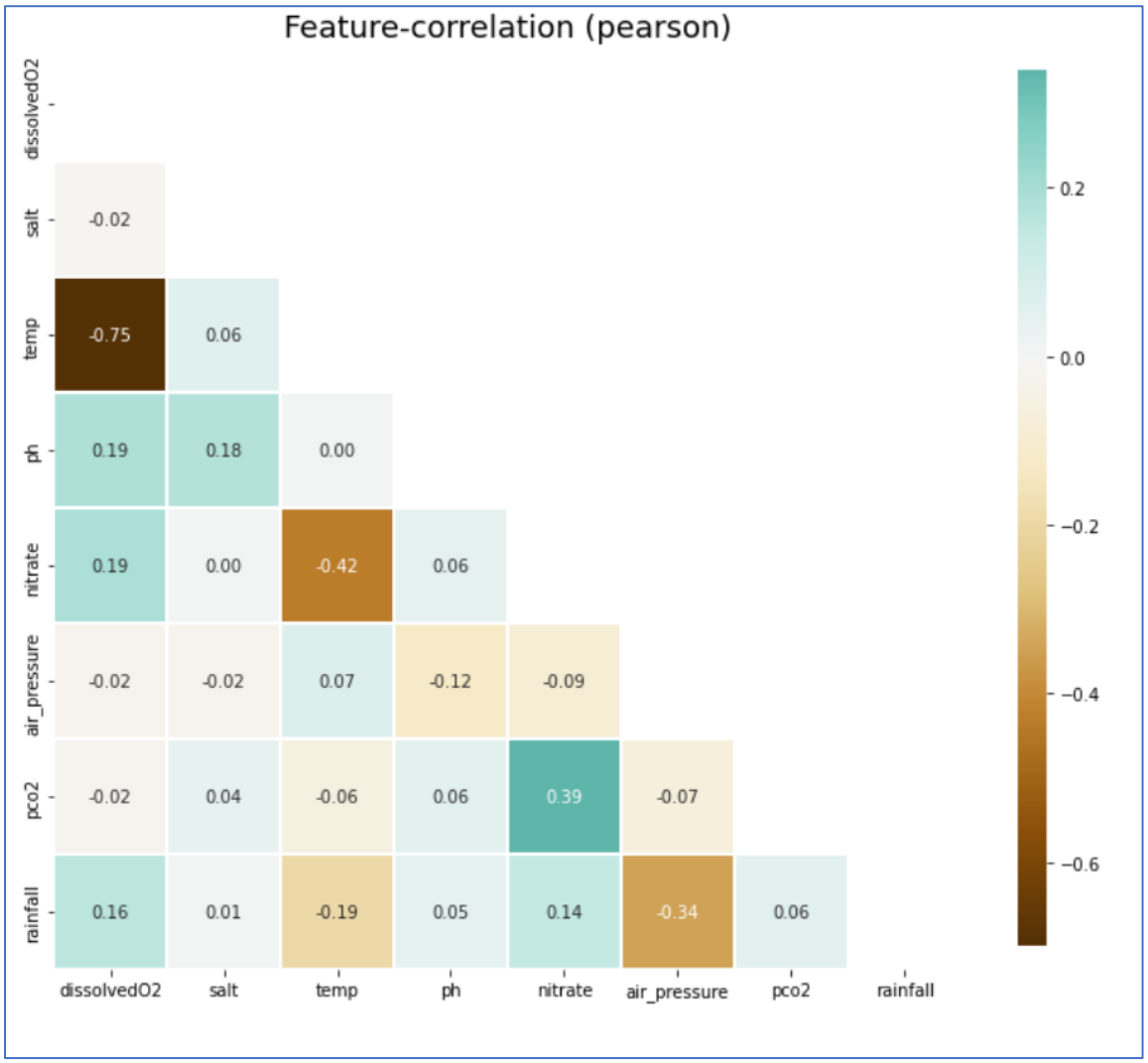

Figure 15 feature correlations

The most important correlation results to note are those between the target variable '*dissolvedO2'* and the others. The target variable shows a strong negative correlation with water temperature which is as expected (*P*=-0.75). That is the only correlation where *P*>0.2. Correlations between the target variable and all other variables were all extremely low (*P*=0.19, *P*=0.16, *P*=0.02). This was also expected as it is well known that dissolved oxygen concentrations have very complex non-linear relationships with other physical and chemical properties of seawater (Muller and Muller, 2015; Gobler and Baumann, 2016).

3.2 Model Predictions & Evaluation

For LSTM, the final 32 hours (rows) of data were used for evaluation as it was necessary to be carried out on unseen data. To see how the model performed on other data, a random 32 rows were chosen from somewhere around the middle of the dataset. Row locations 14000 to 14032 were selected.

For MLR, training and testing was split 70/30 and there is usually no need for a separate validation set, however for continuity's sake the last 32 rows were again removed and predicted. However, it must be noted that these are not the same timestamps between the two models. This is due to the fact that MLR requires a lot more data preparation than LSTM so the row positions were not the same after data cleaning.

3.2.1 Evaluation metrics

The evaluation metrics used are broad because all have their advantages and disadvantages, so the more metrics used, the more rounded a view of model performance. Those included are:

R^2 = coefficient of determination

RMSE = Root Mean Squared Error

MSE = Mean Squared Error

MAE = Mean Absolute Error

MAPE = Mean Absolute Percentage Error

As mentioned previously, R^2 when used with time series data should be taken with a pinch of salt, however it will be included here as it is a popular metric in the literature for dissolved oxygen forecasting. More importance will be placed on RMSE and MAE.

The resulting model performance metrics are depicted below:

	LSTM		MLR
Metric	Final 32 hours	Middle 32 Hours	Final 32 hours
MSE	0.02	0.09	0.42
MAE	0.12	0.29	0.65
RMSE	0.14	0.3	0.5
MAPE	1.33	3.27	0.07
R^2	-2.5	-29.89	0.72

LSTM performed better in terms of MSE, MAE and RMSE. Both sets of predictions achieved better scores in these metrics than MLR. This is what was expected from a much

more complex model. MAPE was actually lower for MLR. As the other metrics were quite unimpressive for MLR, it is likely that a result so low for MAPE is due to overfitting.

MAPE of 1.33 and 3.27 for LSTM is very low and RMSE of 0.14 is also impressive. RMSE is in the units of the original variable, as are MAE and MSE. This translates that LSTM predicted dissolved oxygen at an RMSE of 0.14mg/L and 3.27mg/L for the respective subsets.

Regarding R^2, a score of 0.72 is very mediocre for MLR – this means that 28% of the data (1-0.72=0.28) is unexplained by the model. The R^2 scores for LSTM are very extreme. Often a negative R^2 can result from a model overfitting. A visualisation of the model loss should help with determining if the LSTM was overfitting, Figure 16.

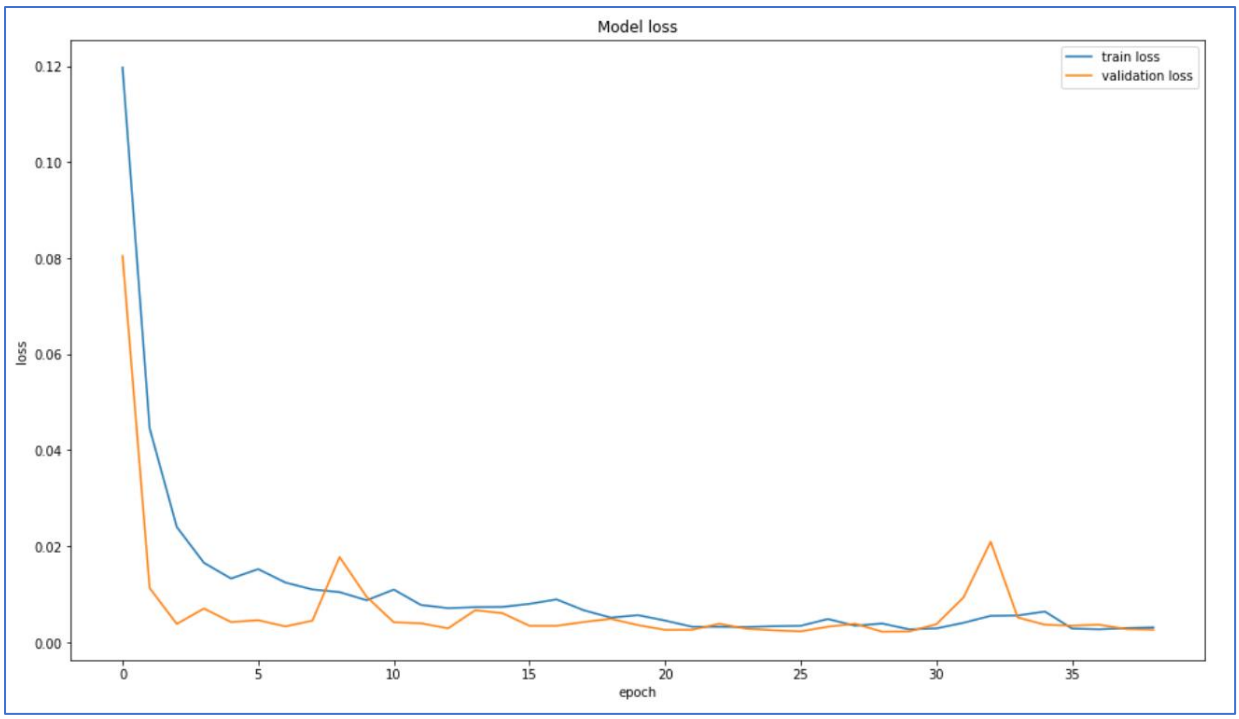

Figure 16 LSTM model loss

If the model was overfitting, the validation loss would decrease, than plateau, as it does in FIG, and then increase again. It makes an increase at the end for perhaps one epoch, so it is possible that overfitting was beginning to occur. Many model parameters were changed in an effort to improve R^2 without success. For a few examples of these changes and their resulting R^2 scores, see Appendix.

Figures 17 and 18 below show the predicted and actual values plotted for LSTM:

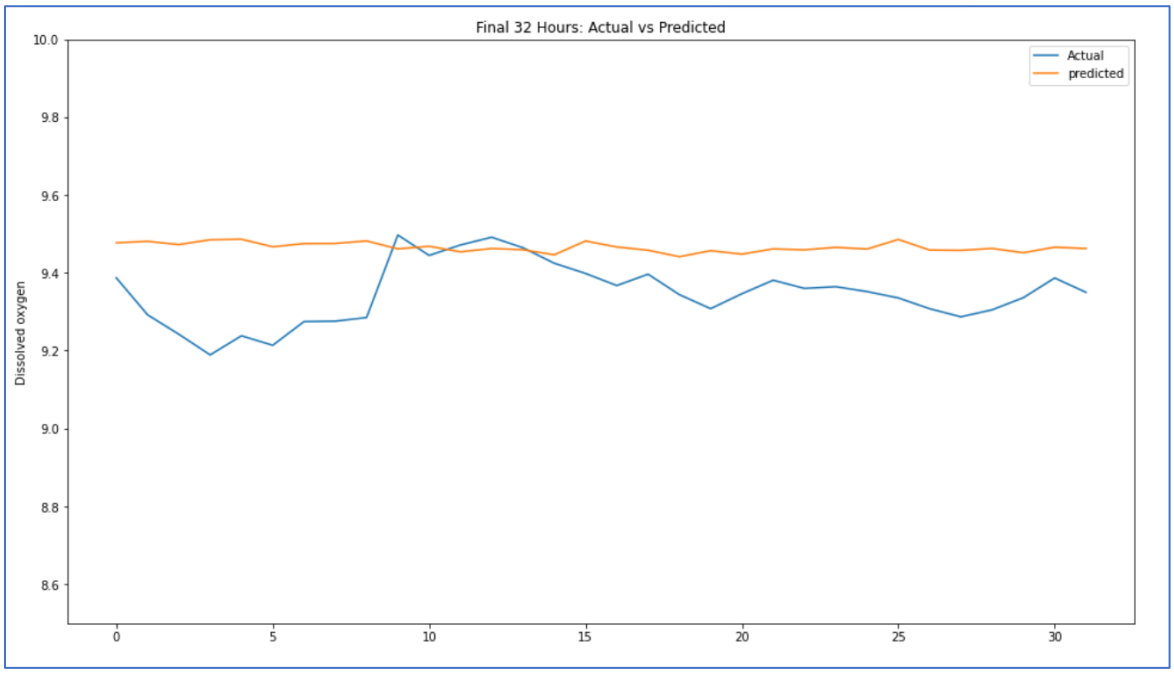

Figure 17 lstm predictions

For the final 32 hours, the predictions are much less fluctuating than the actual values, which contradicts the idea that it may be overfitting. In fact, from Figure 17 it looks closer to underfitting than overfitting. However, the predictions are very close the the actual values so the model should be deemed successful.

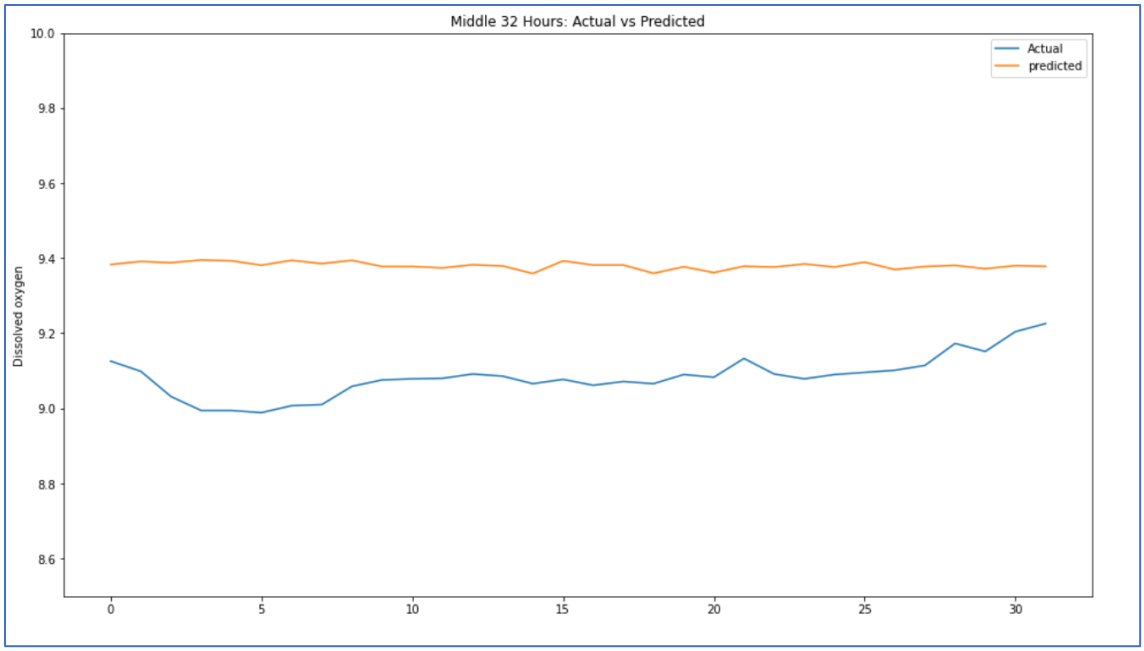

Figure 18 LSTM predictions

It is interesting that the model performed better on the unseen validation set than the random 32 rows from the middle which had been used to train the model (Figure 18).

Figure (19) below shows the predictions made by the MLR:

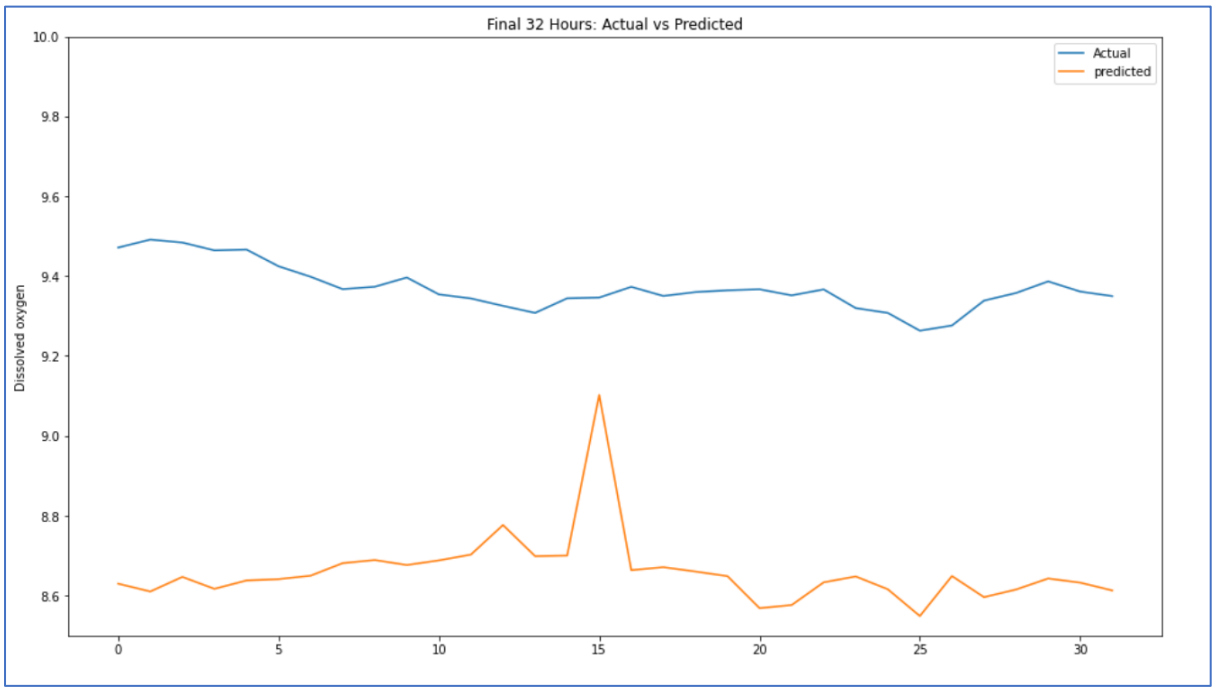

Figure 19 MLR Predictions

Note that the y-axes in Figures (17, 18, 19) do not begin at 0. They are all on the same scale (8.5-10) in order to see the differences more clearly between predicted and actual values.

For LSTM reading off of the plot, the largest disparity between actual and predicted value looks to be approximately 0.4mg/L. For MLR the same thing could be approximated at 0.9mg/L. It would be much more dangerous to overestimate dissolved oxygen content by the latter amount than the former.

All in all, it is still debatable whether the R^2 scores are useful, but the LSTM most definitely outperformed MLR as a predictor of dissolved oxygen concentration.

4 Conclusions

In conclusion, this study aimed to present a proof-of-concept for the use of LSTM as a suitable model for the prediction of dissolved oxygen concentrations in coastal waterbodies, using variables that are commonly measured and made available. We argue that this has been achieved, although there are limitations. LSTM can take a moderately long time, however if used with a GPU this can be greatly reduced. The trade-off between computation and the predictive accuracy demonstrated should be considered. Additionally, there is minimal pre-processing needed. The MLR in this study received data that could have benefited from a lot more preprocessing and still there was much more done for it than LSTM. Further works should compare evaluation metrics specifically for LSTM. This research improved upon the predictive horizons prevalent in the literature, and further works can improve upon this even further.

5 References

Ahmad, H. (2019a) ‘Application of Machine learning in Oceanography View project International Journal of Oceanography & Aquaculture Applications of Remote Sensing in Oceanographic Research’, *International Journal of Oceanography & Aquaculture*. doi: 10.23880/ijoac-16000159.

Ahmad, H. (2019b) ‘Machine learning applications in oceanography’, *Review Article Aquatic Research*, 2(3), pp. 161–169. doi: 10.3153/AR19014.

Analytical vs Numerical Solutions in Machine Learning (no date). Available at: https://machinelearningmastery.com/analytical-vs-numerical-solutions-in-machine-learning/ (Accessed: 20 April 2022).

Bailleul, F., Vacquie-Garcia, J. and Guinet, C. (2015) ‘Dissolved Oxygen Sensor in Animal-Borne Instruments: An Innovation for Monitoring the Health of Oceans and Investigating the Functioning of Marine Ecosystems’, *PLOS ONE*, 10(7), p. e0132681. doi: 10.1371/JOURNAL.PONE.0132681.

Béjaoui, B. *et al.* (2018) ‘Machine learning predictions of trophic status indicators and plankton dynamic in coastal lagoons’, *Ecological Indicators*, 95, pp. 765–774. doi: 10.1016/J.ECOLIND.2018.08.041.

Bengio, Y., Simard, P. and Frasconi, P. (1994) ‘Learning Long-Term Dependencies with Gradient Descent is Difficult’, *IEEE Transactions on Neural Networks*, 5(2), pp. 157–166. doi: 10.1109/72.279181.

Bengio, Y *et al.* (2003) ‘Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies Unsupervised Learning of Speech Representations View project Gradient Flow in Recurrent Nets: the Diiculty of Learning Long-Term Dependencies’.

Boyce, D. G., Lewis, M. R. and Worm, B. (2010) ‘Global phytoplankton decline over the past century’, *Nature 2010 466:7306*, 466(7306), pp. 591–596. doi: 10.1038/nature09268.

Brajard, J. *et al.* (2006) ‘Use of a neuro-variational inversion for retrieving oceanic and atmospheric constituents from satellite ocean colour sensor: Application to absorbing aerosols’, *Neural Networks*, 19(2), pp. 178–185. doi: 10.1016/J.NEUNET.2006.01.015.

Che, Z. *et al.* (2018) ‘Recurrent Neural Networks for Multivariate Time Series with Missing Values’, *Scientific Reports*, 8(1), pp. 1–12. doi: 10.1038/s41598-018-24271-9.

Chen, M. *et al.* (2018) ‘Mechanisms driving phosphorus release during algal blooms based on hourly changes in iron and phosphorus concentrations in sediments’, *Water Research*, 133, pp. 153–164. doi: 10.1016/j.watres.2018.01.040.

Contractor, S. and Roughan, M. (2021) ‘Efficacy of Feedforward and LSTM Neural Networks at Predicting and Gap Filling Coastal Ocean Timeseries: Oxygen, Nutrients, and Temperature’, *Frontiers in Marine Science*, 8, p. 368. doi: 10.3389/FMARS.2021.637759/BIBTEX.

Cox, D. T., Tissot, P. and Michaud, P. (2002) ‘Water Level Observations and Short-Term Predictions Including Meteorological Events for Entrance of Galveston Bay, Texas’, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 128(1), pp. 21–29. doi: 10.1061/(ASCE)0733-950X(2002)128:1(21).

Davydenko, A. and Fildes, R. (2016) ‘Forecast error measures : Critical review and practical recommendations’, *Business Forecasting: Practical Problems and Solutions*, (January), pp. 1–12. doi: 10.13140/RG.2.1.4539.5281.

Ding, S. *et al.* (2018) ‘Internal phosphorus loading from sediments causes seasonal nitrogen limitation for harmful algal blooms’, *Science of the Total Environment*, 625, pp. 872–884. doi: 10.1016/j.scitotenv.2017.12.348.

Europe allocates 4 million to an international research project led by UCAM for the recovery of the Mar Menor (2021) *Universidad Católica de Murcia*. Available at: https://international.ucam.edu/university-news/europe-allocates-4-million-international-

research-project-led-ucam-recovery-mar (Accessed: 14 April 2022).

FAO (2022) ‘Framework for Action on Biodiversity for Food and Agriculture’, *FAO Commission on Genetic Resources for Food and Agriculture*. doi: 10.4060/CB8338EN.

Fourrier, M. *et al.* (2020) ‘A Regional Neural Network Approach to Estimate Water-Column Nutrient Concentrations and Carbonate System Variables in the Mediterranean Sea: CANYON-MED’, *Frontiers in Marine Science*, 7, p. 620. doi: 10.3389/FMARS.2020.00620/BIBTEX.

Gobler, C. J. and Baumann, H. (2016) ‘Hypoxia and acidification in ocean ecosystems: coupled dynamics and effects on marine life’, *Biology Letters*, 12(5). doi: 10.1098/RSBL.2015.0976.

Goldstein, E. B., Coco, G. and Plant, N. G. (2019) ‘A review of machine learning applications to coastal sediment transport and morphodynamics’, *Earth-Science Reviews*, 194, pp. 97–108. doi: 10.1016/J.EARSCIREV.2019.04.022.

Homayoun Aria, S., Asadollahfardi, G. and Heidarzadeh, N. (2019) ‘Eutrophication modelling of Amirkabir Reservoir (Iran) using an artificial neural network approach’, *Lakes and Reservoirs: Research and Management*, 24(1), pp. 48–58. doi: 10.1111/lre.12254.

Huan, J., Cao, W. and Qin, Y. (2018) ‘Prediction of dissolved oxygen in aquaculture based on EEMD and LSSVM optimized by the Bayesian evidence framework’, *Computers and Electronics in Agriculture*, 150, pp. 257–265. doi: 10.1016/J.COMPAG.2018.04.022.

James, S. C., Zhang, Y. and O’Donncha, F. (2018) ‘A machine learning framework to forecast wave conditions’, *Coastal Engineering*, 137, pp. 1–10. doi: 10.1016/J.COASTALENG.2018.03.004.

Jimeno-Sáez, P. *et al.* (2020) ‘Using Machine-Learning Algorithms for Eutrophication Modeling: Case Study of Mar Menor Lagoon (Spain)’, *International Journal of Environmental Research and Public Health 2020, Vol. 17, Page 1189*, 17(4), p. 1189. doi:

10.3390/IJERPH17041189.

Jin, Z. *et al.* (2019) ‘High resolution spatiotemporal sampling as a tool for comprehensive assessment of zinc mobility and pollution in sediments of a eutrophic lake’, *Journal of Hazardous Materials*, 364, pp. 182–191. doi: 10.1016/j.jhazmat.2018.09.067.

Khan, U. (2017) ‘Comparing A Bayesian and Fuzzy Number Approach to Uncertainty Quantification in Short-Term Dissolved Oxygen Prediction.’, *Journal of Environmental Informatics*, 30(1), pp. 1–16. doi: 10.3808/jei.

Kim, Y. H. *et al.* (2014) ‘Machine learning approaches to coastal water quality monitoring using GOCI satellite data’, *GIScience and Remote Sensing*, 51(2), pp. 158–174. doi: 10.1080/15481603.2014.900983.

Kramer, D. L. (1987) ‘Dissolved oxygen and fish behavior’, *Environmental Biology of Fishes*, 18(2), pp. 81–92. doi: 10.1007/BF00002597.

Krasnopolsky, V. M. (2007) ‘Neural network emulations for complex multidimensional geophysical mappings: Applications of neural network techniques to atmospheric and oceanic satellite retrievals and numerical modeling’, *Reviews of Geophysics*, 45(3), p. 3009. doi: 10.1029/2006RG000200.

Krasnopolsky, V. M., Chalikov, D. V. and Tolman, H. L. (2002) ‘A neural network technique to improve computational efficiency of numerical oceanic models’, *Ocean Modelling*, 4(3–4), pp. 363–383. doi: 10.1016/S1463-5003(02)00010-0.

Liu, H. *et al.* (2021) ‘A Hybrid Neural Network Model for Marine Dissolved Oxygen Concentrations Time-Series Forecasting Based on Multi-Factor Analysis and a Multi-Model Ensemble’, *Engineering*, 7(12), pp. 1751–1765. doi: 10.1016/j.eng.2020.10.023.

Liu, Y. *et al.* (2019) ‘Attention-based recurrent neural networks for accurate short-term and long-term dissolved oxygen prediction’, *Computers and Electronics in Agriculture*, 165, p. 104964. doi: 10.1016/J.COMPAG.2019.104964.

McGovern, J. V., Nash, S. and Hartnet, M. (2020) *Modelling Irish Transitional and Coastal Systems to Determine Nutrient Reduction Measures to Achieve Good Status*. Wexford.

MERC Consultants (2005) 'Biodiversity of Cill Chiaráin Bay, Co. Galway', (14087), p. 115.

Muller-Karger, F. E. *et al.* (2018) 'Advancing marine biological observations and data requirements of the complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) frameworks', *Frontiers in Marine Science*, 5(JUN). doi: 10.3389/FMARS.2018.00211.

Muller, A. C. and Muller, D. L. (2015) 'Forecasting future estuarine hypoxia using a wavelet based neural network model', *Ocean Modelling*, 96, pp. 314–323. doi: 10.1016/J.OCEMOD.2015.11.003.

Muzaffar, S. and Afshari, A. (2019) 'Short-term load forecasts using LSTM networks', *Energy Procedia*, 158, pp. 2922–2927. doi: 10.1016/J.EGYPRO.2019.01.952.

Nikoloski, S. *et al.* (2021) 'Exploiting partially-labeled data in learning predictive clustering trees for multi-target regression: A case study of water quality assessment in Ireland', *Ecological Informatics*, 61. doi: 10.1016/J.ECOINF.2020.101161.

O Donncha, F. and Grant, J. (2020) 'Precision Aquaculture', *IEEE Internet of Things Magazine*, (December 2019), pp. 26–30.

Ouyang, W. *et al.* (2018) 'Heavy metal loss from agricultural watershed to aquatic system: A scientometrics review', *Science of the Total Environment*, 637–638, pp. 208–220. doi: 10.1016/j.scitotenv.2018.04.434.

Ozyegen, O., Ilic, I. and Cevik, M. (2022) 'Evaluation of interpretability methods for multivariate time series forecasting', *Applied Intelligence*, 52(5), pp. 4727–4743. doi: 10.1007/S10489-021-02662-2/FIGURES/6.

Pascanu, R. *et al.* (2013) ‘How to Construct Deep Recurrent Neural Networks’.

Preeti, Bala, R. and Singh, R. P. (2019) ‘Financial and Non-Stationary Time Series Forecasting using LSTM Recurrent Neural Network for Short and Long Horizon’, *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*. doi: 10.1109/ICCCNT45670.2019.8944624.

Rasamoelina, A. D., Adjailia, F. and Sincak, P. (2020) ‘A Review of Activation Function for Artificial Neural Network’, *SAMI 2020 - IEEE 18th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, pp. 281–286. doi: 10.1109/SAMI48414.2020.9108717.

Salami, E. S. and Ehteshami, M. (2015) ‘Simulation, evaluation and prediction modeling of river water quality properties (case study: Ireland Rivers)’, *International Journal of Environmental Science and Technology*, 12(10), pp. 3235–3242. doi: 10.1007/s13762-015-0800-7.

Schmidtko, S., Stramma, L. and Visbeck, M. (2017) ‘Decline in global oceanic oxygen content during the past five decades’, *Nature*, 542(7641), pp. 335–339. doi: 10.1038/nature21399.

Schuster, M. and Paliwal, K. K. (1997) ‘Bidirectional recurrent neural networks’, *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681. doi: 10.1109/78.650093.

Srivastava, N. *et al.* (2014) ‘Dropout: A Simple Way to Prevent Neural Networks from Overfitting’, *Journal of Machine Learning Research*, 15.

Third Integrated Report on the Eutrophication Status of the OSPAR Maritime Area (2017).

Valera, M. *et al.* (2020) ‘Machine learning based predictions of dissolved oxygen in a small coastal embayment’, *Journal of Marine Science and Engineering*, 8(12), pp. 1–16. doi: 10.3390/jmse8121007.

Virtanen, E. A. *et al.* (2019) ‘Identifying areas prone to coastal hypoxia - The role of topography’, *Biogeosciences*, 16(16), pp. 3183–3195. doi: 10.5194/BG-16-3183-2019.

Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’.

Wooldridge, J. M. (1991) ‘A note on computing r-squared and adjusted r-squared for trending and seasonal data’, *Economics Letters*, 36(1), pp. 49–54. doi: 10.1016/0165-1765(91)90054-O.

Yu, X., Shen, J. and Du, J. (2020) *A Machine-Learning-Based Model for Water Quality in Coastal Waters, Taking Dissolved Oxygen and Hypoxia in Chesapeake Bay as an Example*, *Water Resources Research*. doi: 10.1029/2020WR027227.

Zhang, Y. *et al.* (2019) ‘Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved Oxygen’, *Frontiers in Environmental Science*, 7(MAR), pp. 1–11. doi: 10.3389/fenvs.2019.00046.

Zhang, Y. F., Fitch, P. and Thorburn, P. J. (2020) ‘Predicting the trend of dissolved oxygen based on the kPCA-RNN model’, *Water (Switzerland)*, 12(2), pp. 1–15. doi: 10.3390/w12020585.

6 Appendices

Appendix A

Package	License	Use
tensorflow	Apache 2.0	LSTM
numpy	Apache 2.0	mathematical functions for
pandas	BSD 3-Clause	data manipulation
os	Apache 2.0	operating system interaction
matplotlib, pyplot	Matplotlib	data visualisation
statsmodels	BSD 3-Clause	STL
plotly express	MIT + file	data visualisation
sci kit learn	BSD 3-Clause	machine learning library
keras	Apache 2.0	LSTM layers
pyod	BSD-2-Clause	outlier detection
klib	MIT/X11	data cleaning
seaborn	BSD 3-Clause	data visualisation
missingno	Apache 2.0	visualises missing data
geopandas	BSD-3-Clause	geospatial data functions

Appendix B: Model Summary

Layer (type)	**Output Shape**	**Param #**
bidirectional_6 (Bidirectional)	(None, 84, 200)	86400
dense_15 (Dense)	(None, 84, 20)	4020
bidirectional_7 (Bidirectional)	(None, 160)	64640
dense_16 (Dense)	(None, 20)	3220
dense_17 (Dense)	(None, 20)	420
dense_18 (Dense)	(None, 20)	420
dropout_3 (Dropout)	(None, 20)	0
dense_19 (Dense)	(None, 32)	672
Total params: 159,792		
Trainable params: 159,792		
Non-trainable params: 0		

Appendix C R-squared

Change Made	**Resulting R^2**
Return sequences=false	Not compiling
Batch size=128 buffer=75	-634
Remove dropout &decrease units	-273
LSTM Units=120,80, dense units = 15	-289
Add extra dense layer	-409
Change history to 48	-289
Change dropout to 0.15	-1189
Change dropout to 0.5	-11