Multimodal Approach to Early Detection of Harmful Algal Blooms

Onatkut Dagtekin[1,*], Nina Dethlefs[1]

[1]Department of Computer Science and Technology, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom

Abstract

A rise in ecological anomalous events will be observed due to climate change. One such event is the harmful algal bloom which occurs due to an increase in nutrients from anthropogenic activities and has economic and ecological effects. Algae thrive in warmer temperatures which will lead to an increase in the frequency of harmful algal blooms. To overcome this increasing frequency, early detection tools are essential. Deep learning and frequent monitoring have been used to detect this phenomenon with a focus on unimodal approaches. In this work, we propose using multiple sources of satellite and in-situ data for detecting algal blooms with a joint multimodal learning approach, focusing on the North Sea and the Irish Sea. This work will aid domain experts to monitor potential changes to the ecosystem done by human interference and to take action when necessary.

Keywords

harmful algal blooms, multimodal fusion, transformer networks, convolutional neural networks

1. Introduction

Harmful algal blooms (HABs) occur due to nutrient overloading causing the phytoplankton population to rise rapidly, affecting the environment, and resulting in issues such as oxygen depletion and sunlight blocking [1]. The occurrence of algal blooms will increase in currently observed locations and will start to occur in new locations due to the rising temperatures [2]. HABs have severe negative impacts as they reduce the income from fisheries and touristic activities and increase the cost of presentation of biodiversity [3]. These blooms lower the income from tourism and fisheries [4, 5]. Machine learning could be utilized to detect and prevent the occurrence of HABs using the data gathered from various collection programmes.

Different forms of data could be used for detection; In-situ data such as buoys or water samples analysed in labs or satellite data such as Moderate Resolution Imaging Spectroradiometer (MODIS) for detecting colour or nutrient changes from various bands. The in-situ data is collected frequently but only covers a small area whereas the satellite data covers a larger area with infrequent data. Both data sources might have lower data quality due to Q&A processes or external factors such as cloud cover or biofouling. The data sources can be combined to partially

MACLEAN: MAChine Learning for EArth ObservatioN Workshop 2022, in conjunction with ECML/PKDD (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) 2022, Grenoble, September 19, 2022

*Corresponding author.

o.dagtekin-2019@hull.ac.uk (O. Dagtekin); n.dethlefs@hull.ac.uk (N. Dethlefs)

alleviate these problems. In the current approaches, each modality is analysed separately and independently from each other to detect blooms. To improve the detection process, the information from different data modalities could be analysed together. In this work, we propose a multimodal fusion approach where each type of data is analysed simultaneously to detect HABs, one to seven days before they occur.

2. Current Approaches

There are two approaches for detecting algal blooms: thresholding [6, 7], which divides data into predetermined categories and predicts future HAB occurrences or regression, which directly predicts the continuous value. The variable to be predicted is either dissolved oxygen or chlorophyll-a (chl-a), both of which rise during higher photosynthetic activity from algae, as chl-a captures sunlight to perform photosynthesis and produce glucose and oxygen. These changes could also be observed using satellite imagery such as color changes in RGB format or an increase in the intensity of chl-a measurements during a bloom.

The current approaches use each data modality separately. [8] use MODIS data to predict three variables related to HABs: chl-a, sea surface temperature and fluorescence line-height with SARIMA, regression and ANNs. [9] classify HAB events using twelve different MODIS channels with CNN, LSTM and ML methods. [10] combine Medium Resolution Imaging Spectrometer (MERIS), MODIS and in-situ data before training a genetic programming model to measure phycocyanin. [11] use XGBoost to predict chl-a levels in several lakes in China and uses in-situ data for validation. [12] use sensory data to forecast the chl-a concentration one day and four days ahead in Geum River, South Korea with stacked LSTMs. [13] use attention LSTMs to forecast the chl-a values half a day ahead in Fujian, China. [14] use in-situ data to predict the chl-a values 1 to 3 days ahead, with an ensemble of ANNs and Discrete Wavelet Transform. [7] utilize AdaBoost for SVM and RF to predict HABs using sensory data in Yuyuantan Lake, China. [15] forecast chl-a values one week or two weeks ahead in Tolo Harbour, Hong Kong using ANN, generalized regression network and SVMs.

It can be noticed that for this task either satellite data or sensory data is used for analysis but not both. In some cases where satellite data is used, in-situ data is used for verification purposes. Using only satellite data reduces the temporal prediction capabilities of models as the data is infrequent. Using only in-situ data reduces the spatial extendibility of the predictions as the observations are location specific. The span of data used for early HAB detection is usually shorter than a year, reducing the generalisability of the model [16]. The study locations are mostly focused on China, the Great Lakes, the Gulf of Florida, the Baltic Sea and the Mediterranean [16].

In this study, we focus on the North Sea and the Irish Sea. The dataset we used encompasses 10 years; between 2009-2019. The modelling is done using different modes of data; with two different satellite imagery sources and in-situ data, aiming to predict 1 to 7 days ahead, only using data from a single day of observation. Detecting only a single variable such as chl-a or dissolved oxygen has no applicability for the end-user and other contextual information is needed. The proposed model predicts additional variables, temperature and salinity which affect the maximum amount of oxygen the water can contain. Using the predicted variables, the oxygen saturation at time t is calculated, providing more information to the end-users.

3. Methodology

3.1. Data

In-situ Data: The data for this work was collected by ESM2 and ESMx data loggers at four different moorings depicted in Figure 1. The data was collected as a part of The National Marine Monitoring Programme (NMMP) to monitor eutrophication regarding The Convention for the Protection of the Marine Environment of the North-East Atlantic (OSPAR) and Marine Strategy Framework Directive (MSFD) assessments. The whole dataset was partitioned into four fractions based on location. Each of the locations has different characteristics such that the Liverpool buoy is near a maritime route, the WestGab buoy is near a wind farm, the TH1 buoy is near the delta of the River Thames and the Dowsing buoy is in the open sea. It is known that the chl-a concentration has been decreasing in certain hot spots in the Southern North Sea [17].

The periodicity and the relationship between the variables were analysed by [18, 19, 20] with varying date ranges and locations by performing wavelet analysis. The periodicities of variables depend on the season and range between 6 hours to 24 hours. The data consists of eight features; chl fluorescence (*fluors*), turbidity (*ftu*), dissolved oxygen concentration (*o2conc*), salinity (*sal*), temperature (*temp*) and photosynthetically active radiation (PAR) at depths 0, 1 and 2 meters (*depth_0*, *depth_1*, *depth_2*). The data was collected at 20-30 minute intervals at each station. The data used spans the range between Jan 2009 and Dec 2019. Before being given as input, the data was normalized with z-score normalization.

Figure 1: Locations of moorings

MODIS: MODIS satellites, Terra and Aqua, gather data using 36 different bands. Various bands can be used for detecting changes in water/soil. RGB bands 1, 4 and 3 could be used for these purposes. During algal blooms, the water colour changes to a different colour based on the species. Therefore color changes could be observed and analysed using these groups of bands. The data collected by these sensors have to go through extensive preprocessing since there will be many issues regarding cloud cover, aerosol loading etc. resulting in low quality and less frequent data.

Copernicus Marine Service (CMS): Additional satellite data was obtained via the CMS. The information includes variables such as chlorophyll, nitrate and phosphate in seawater around the observation sites. The source of the data will depend on the dataset used. Data sources include a mixture of MODIS, Visible Infrared Imaging Radiometer Suite (VIIRS), Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) etc. Different levels of processed data are supplied by CMS ranging from raw data (L1) to cloud-free (L4) data.

Both types of satellite data were collected as a daily mean with a resolution of 1 km. A region of 6x6 km is gathered around each monitoring site, upsampled to 256x256 with bicubic interpolation. MODIS data contains the best information from a 16-day period depending on several factors such as observation coverage, cloud coverage, view angles etc for each pixel. The CMS data used is *OCEANCOLOUR_ATL_CHL_L4_REP_OBSERVATIONS_009_098*[1] which is gathered by several satellites. The data gathered is Level 4 which went through the process of interpolation. Each pixel contains the daily mean value for chl-a. Both modalities' data is upsampled due to the used CNN models, as they are designed to receive 256x256 images as input. A single image was used for each observation day.

3.2. Proposed Approach

he first proposed approach uses CNNs to analyse the RGB bands from MODIS and chlorophyll data from Copernicus Marine Service (CMS) separately. Transformer networks were used to analyse the time series data from the in-situ buoys. Figure 2 visualizes the proposed architecture for algal bloom detection. The proposed approach makes use of joint representations combined with data fusion using each modality's representation to alleviate each modality's disadvantages such as differences in data collection frequencies and data sampling disruptions. Unlike [10], our approach does not concatenate multiple data sources before learning rather it will use middle or late fusion of modality representations. With the use of middle fusion, the model will be able to learn the relationship between modalities, resulting in a better model. Due to the different characteristics of data sources, early fusion cannot be applied and alternate approaches, i.e, intermediate or late fusion can be used. After fusing the intermediate hidden states, a linear layer is used to obtain the predictions. Each component of the model is trained simultaneously. The pseudocode is presented in Algorithm 1.

The linear layer is also replaced with an XGBoost model. The input to the XGBoost is supplied by concatenating the output of each modality, transferring the learned individual representations to a different model. Each output variable will have a separate XGBoost model since XGBoost models cannot be used for multi-value regression.

The variables are used to calculate the predicted oxygen saturation and the real oxygen saturation using Equation 1, where $A_0, ..., A_5$, $B_0, ..., B_3$ and C are coefficients of the equation

[1] now renamed *OCEANCOLOUR_ATL_BGC_L4_MY_009_118*

Algorithm 1 Multimodal approach forward propagation (single batch)

Ensure: X_{src} = tensor of (seq_len,batch_size,num_features)
Ensure: X_{tgt} = tensor of (seq_len,batch_size,num_features - 1)
Ensure: X_{modis} = tensor of (batch_size,in_channels = 3,height=256,width=256)
Ensure: X_{cms} = tensor of (batch_size,in_channels = 2,height=256,width=256)

```
X_src ← time2vec(X_src)
X_tgt ← time2vec(X_tgt)
X_src ← transformer_encode(X_src)
X_src ← transformer_decode(X_src, X_tgt, masks)
X_src ← avg_pool(GeLU(conv_1d(X_src)))
X_modis ← modis_cnn(X_modis)
X_cms ← cms_cnn(X_cms)
X ← torch.concat(X_src, X_modis, X_cms)
if last_layer == linear then
    Y ← linear(X)
else
    for n in output_variables do
        Y_n ← XGBoost_n(X)
    end for
end if
```

given in Table 1, S is the salinity and T is $In[(298.15 - T_O)(273.15 + T_O)^{-1}]$ where T_O is the observed temperature value at time t [21]. Mean Absolute Error (MAE) is used to compare the saturation percentages. This method is followed to give insight into the results of the predictions for domain experts.

$$
\begin{aligned}
D_O = In(A_0 + A_1 T_{+} A_2 T^2 + A_3 T^2 + \\
A_3 T^3 + A_4 T^4 + A_5 T^5 + \\
S(B_0 + B_1 T + B_2 T^2 + B_3 T^3) + CS^2)
\end{aligned} \tag{1}
$$

Coefficient	Value
A_0	2.00907
A_1	3.22014
A_2	4.05010
A_3	4.944457
A_4	$-2.56847 * 10^{-1}$
A_5	3.887674
B_0	$-6.24523 * 10^{-3}$
B_1	$-7.37614 * 10^{-3}$
B_2	$-1.03410 * 10^{-2}$
B_3	$-8.17083 * 10^{-3}$
C	$-4.88682 * 10^{-7}$

Table 1
Coefficients for Equation 1

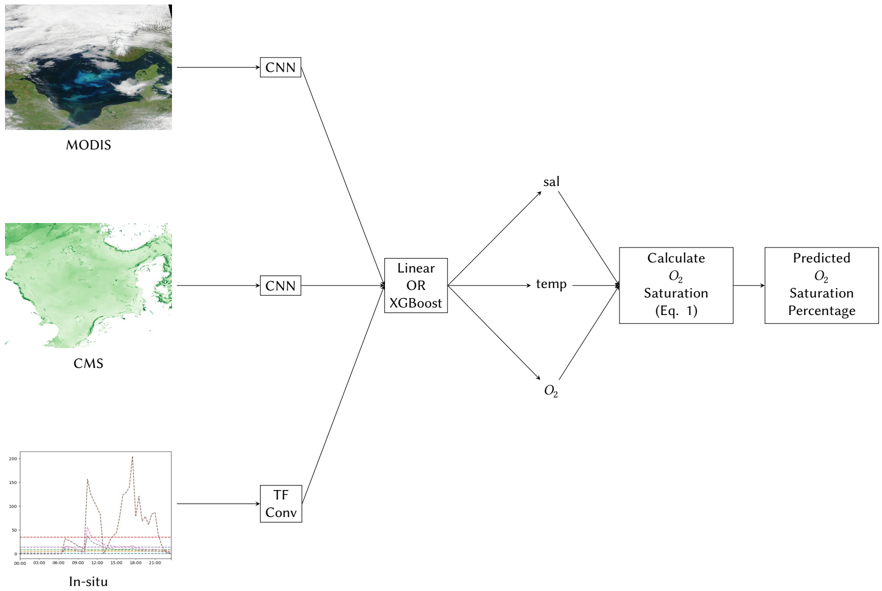

Figure 2: Proposed Multimodal Fusion Approach

4. Results & Discussion

The predictions are done between one day to one week into the future given the observation at day *n*. 70% of TH1 buoy data was used for training, and 30% for validation. This location was chosen due to nutrient flow as it is located near the delta of the River Thames. The reason behind the location choice is to create a more generalized model using the nutrient flows. We used a single location to observe if the model would be able to perform satisfactorily for other locations with different properties, therefore testing the generalisability of the model. The other three sites are used for testing. The baseline models are: Support Vector Regression (SVR), K-Neighbours Regression (KNR), Multilayer Perceptron (MLP) and Luong attention model [22, 23, 24]. The baseline models were chosen due to their frequent use for this task. The input into the baselines models is the in-situ data. The models are trained using Mean Square Error (MSE) as the loss function.

Figure 3 illustrates the MSE values for each model based on the number of days into the future and Figure 4 illustrates the MSE values for each site based on the number of days into the future. For all models, a hyperparameter search was done based on the prediction day. For the SVR model, a model was created for each predicted variable. For deep learning models, an Adam optimizer was used for this task with 200 epochs and earlystopping with a patience of 15 epochs [25]. The model in [26] is used for the in-situ data. The CNN models tested for MODIS and CMS data are: ResNet18, ResNet152, MobileNet_v2, VGG19, VGG19_bn, and AlexNet. These architectures were chosen due to their frequent and off-the-shelf use. Two comparisons are made one with MSE for the three predicted variables and one with MAE to compare oxygen saturation percentages.

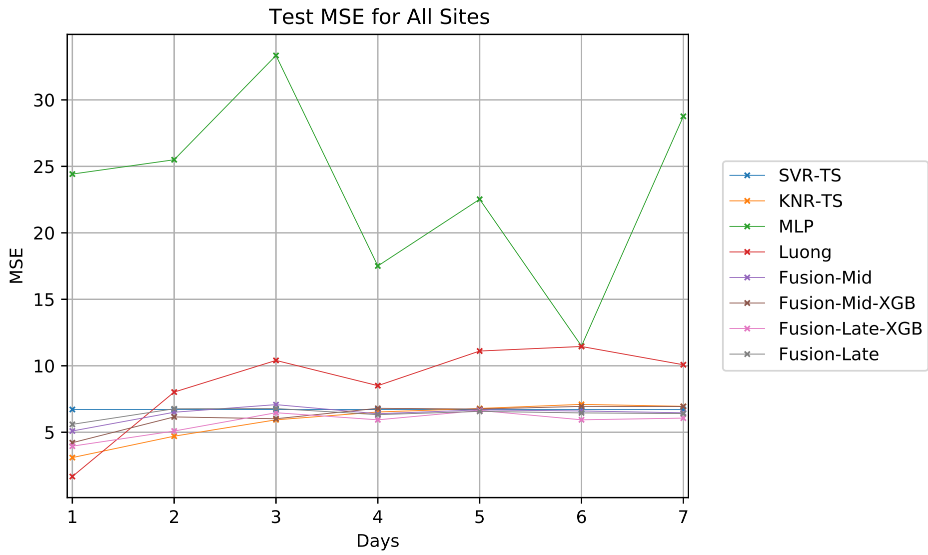

Figure 3: Mean MSE for test locations

Day	Model Type	MODIS CNN Model	CMS CNN Model	MAE
1	Luong	-	-	5.182
2	KNR	-	-	7.954
3	KNR	-	-	7.952
4	XGB-Late	ResNet152	MobileNet_v2	8.41
5	Fusion-Late	VGG19_bn	MobileNet_v2	7.855
6	XGB-Late	VGG19	VGG19	8.300
7	XGB-Late	VGG19	VGG19	8.115

Table 2
MAE results for each day with the best performing model

From Figure 3, it can be deduced that the Luong attention model is suitable for predicting the next day and k-nn is suitable for predicting two and three days ahead. For the rest of the days, the most suitable model is one of the multimodal approaches we have proposed, either using the late fusion approach or transferring the learned representations from the late fusion approach and using XGBoost as the final classifier. Table 2 indicates an error rate between 7.855- 8.3% for multimodal approaches for prediction days 4-7 which is on par with unimodal approaches for days 2 & 3. The models for calculating MAE were chosen based on the best performing model for each day (i.e. with lowest MSE). It can be deduced that using only in-situ data for predicting the near future is suitable whereas additional modalities are needed to predict further future days. From Figure 4, it can be deduced that the baseline algorithms only perform adequately in the training site. For applicability, generalisability is required. The proposed approaches are able to generalize using multiple modalities, resulting in better performing models.

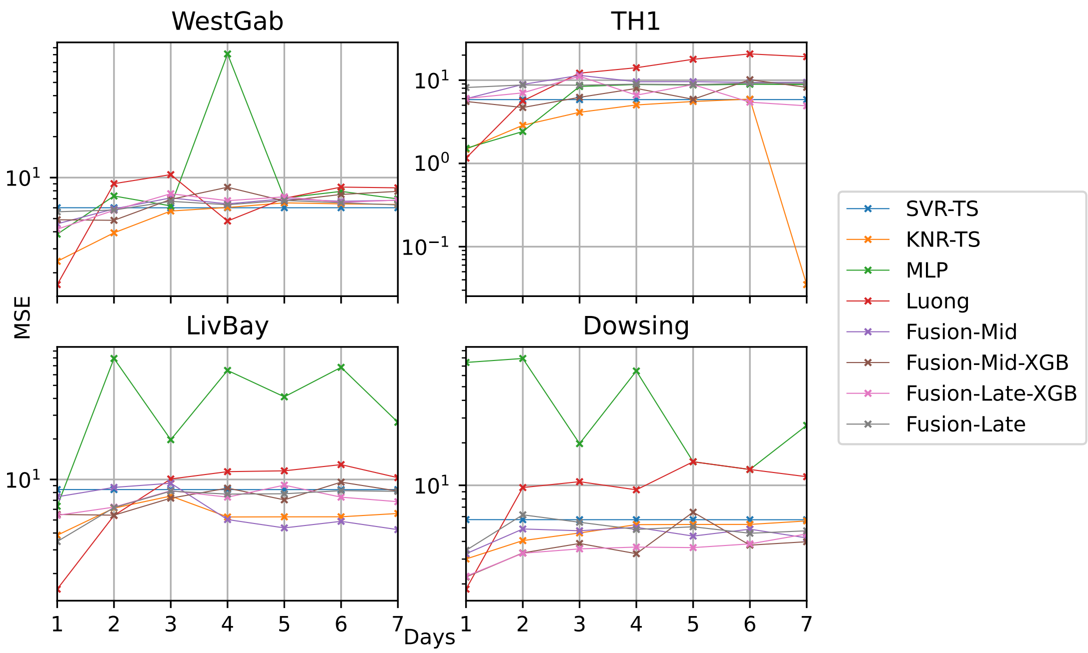

Figure 4: Mean MSE for each location

5. Future Work

Due to increasing temperatures, algae will have favourable conditions to reproduce resulting in a higher frequency of algal blooms for at-risk areas and will occur in new locations. These blooms will have negative ecological and economic impacts. To combat the effect of this phenomenon, early detection tools are essential. In this work, we propose a new approach to early algal bloom detection that uses various modalities of data. The proposed model uses in-situ data and various satellite data to predict algal blooms before they occur for a predefined time range. The aim is to use each modality to extract essential information to make a more reliable model. The area of focus was around the North Sea and the Irish Sea which are less studied areas. Different types of multimodal learning techniques could be experimented in further iterations. Other than prediction, model explainability and interpretability could be explored in the future.

References

[1] M. Kahru, B. G. Mitchell, Ocean color reveals increased blooms in various parts of the world, Eos, Transactions American Geophysical Union 89 (2008) 170–170.

[2] M. L. Wells, V. L. Trainer, T. J. Smayda, B. S. Karlson, C. G. Trick, R. M. Kudela, A. Ishikawa, S. Bernard, A. Wulff, D. M. Anderson, et al., Harmful algal blooms and climate change: Learning from the past and present to forecast the future, Harmful algae 49 (2015) 68–93.

[3] W. K. Dodds, W. W. Bouska, J. L. Eitzmann, T. J. Pilger, K. L. Pitts, A. J. Riley, J. T. Schloesser,

D. J. Thornbrugh, Eutrophication of us freshwaters: analysis of potential economic damages, 2009.

[4] A. Bechard, The economic impacts of harmful algal blooms on tourism: an examination of southwest florida using a spline regression approach, Natural Hazards 104 (2020) 593–609.

[5] B. Karlson, P. Andersen, L. Arneborg, A. Cembella, W. Eikrem, U. John, J. J. West, K. Klemm, J. Kobos, S. Lehtinen, et al., Harmful algal blooms and their effects in coastal seas of northern europe, Harmful Algae (2021) 101989.

[6] N. Mellios, S. J. Moe, C. Laspidou, Machine learning approaches for predicting health risk of cyanobacterial blooms in northern european lakes, Water 12 (2020) 1191.

[7] Y. Yang, Y. Bai, X. Wang, L. Wang, X. Jin, Q. Sun, Group decision-making support for sustainable governance of algal bloom in urban lakes, Sustainability 12 (2020) 1494.

[8] M. R. A. Shehhi, A. Kaya, Time series and machine learning to forecast the water quality from satellite data, arXiv preprint arXiv:2003.11923 (2020).

[9] P. R. Hill, A. Kumar, M. Temimi, D. R. Bull, Habnet: machine learning, remote sensing-based detection of harmful algal blooms, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 (2020) 3229–3239.

[10] B. Vannah, N.-B. Chang, Fusion of hyperspectral remote sensing data for near real-time monitoring of microcystin distribution in lake erie, in: Satellite Data Compression, Communications, and Processing IX, volume 8871, International Society for Optics and Photonics, 2013, p. 88710M.

[11] Z. Cao, R. Ma, H. Duan, N. Pahlevan, J. Melack, M. Shen, K. Xue, A machine learning approach to estimate chlorophyll-a from landsat-8 measurements in inland lakes, Remote Sensing of Environment 248 (2020) 111974.

[12] H. Cho, U. Choi, H. Park, Deep learning application to time-series prediction of daily chlorophyll-a concentration, WIT Trans. Ecol. Environ. 215 (2018) 157–63.

[13] X. Wang, L. Xu, Unsteady multi-element time series analysis and prediction based on spatial-temporal attention and error forecast fusion, Future Internet 12 (2020) 34.

[14] S. Shamshirband, E. Jafari Nodoushan, J. E. Adolf, A. Abdul Manaf, A. Mosavi, K.-w. Chau, Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters, Engineering Applications of Computational Fluid Mechanics 13 (2019) 91–101.

[15] X. Li, J. Yu, Z. Jia, J. Song, Harmful algal blooms prediction with machine learning models in tolo harbour, in: 2014 International Conference on Smart Computing, IEEE, 2014, pp. 245–250.

[16] M.-T. Sebastiá-Frasquet, J.-A. Aguilar-Maldonado, I. Herrero-Durá, E. Santamaría-del Ángel, S. Morell-Monzó, J. Estornell, Advances in the monitoring of algal blooms by remote sensing: A bibliometric analysis, Applied Sciences 10 (2020) 7877.

[17] K. Von Schuckmann, P.-Y. Le Traon, N. Smith, A. Pascual, S. Djavidnia, J.-P. Gattuso, M. Grégoire, G. Nolan, S. Aaboe, E. Á. Fanjul, et al., Copernicus marine service ocean state report, issue 4, Journal of Operational Oceanography 13 (2020) S1–S172.

[18] A. N. Blauw, E. Benincà, R. W. Laane, N. Greenwood, J. Huisman, Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the north sea, Progress in Oceanography 161 (2018) 1–18.

[19] A. N. Blauw, E. Beninca, R. W. Laane, N. Greenwood, J. Huisman, Dancing with the tides:

fluctuations of coastal phytoplankton orchestrated by different oscillatory modes of the tidal cycle, PLoS One 7 (2012).

[20] J. Heffernan, J. Barry, M. Devlin, R. Fryer, A simulation tool for designing nutrient monitoring programmes for eutrophication assessments, Environmetrics: The official journal of the International Environmetrics Society 21 (2010) 3–20.

[21] H. E. Garcia, L. I. Gordon, Oxygen solubility in seawater: Better fitting equations, Limnology and oceanography 37 (1992) 1307–1312.

[22] S. Lauguico, R. Concepcion, R. R. Tobias, J. Alejandrino, D. Macasaet, E. Dadios, Indirect measurement of dissolved oxygen based on algae growth factors using machine learning models, in: 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), IEEE, 2020, pp. 1–6.

[23] Q. V. Ly, X. C. Nguyen, N. C. Lê, T.-D. Truong, T.-H. T. Hoang, T. J. Park, T. Maqbool, J. Pyo, K. H. Cho, K.-S. Lee, et al., Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the han river, south korea, Science of The Total Environment 797 (2021) 149040.

[24] H. Juan, M. Li, X. Xu, Z. Hao, B. Yang, J. Jianming, S. Bing, Multi-step prediction of dissolved oxygen in river based on random forest missing value imputation and attention mechanism coupled with recurrent neural network, Water Supply (2022).

[25] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[26] O. Dagtekin, N. Dethlefs, Modelling phytoplankton behaviour in the north and irish sea with transformer networks, in: Proceedings of the Northern Lights Deep Learning Workshop, volume 3, 2022.