Algal Bloom Forecasting using Remote Sensing

Discovering the most predictive data modalities for Algal Bloom Forecasting

Kadir Tolga Gökçe[1]

Supervisor(s): Dr. Jan van Gemert[1]**, Attila Lengyel**[1]**, Robert-Jan Bruintjes**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2023

Name of the student: Kadir Tolga Gökçe
Final project course: CSE3000 Research Project
Thesis committee: Dr. Jan van Gemert, Attila Lengyel, Robert-Jan Bruintjes, Prof. Dr. K.G. Langendoen

Abstract

An algal bloom is defined as a rapid increase in common algae (phytoplankton) abundance in water bodies and it can occur when a group of certain environmental factors is combined. If the algae populations grow out of control, such algal blooms become problematic and cause damage to the ecosystem, such phenomena are called harmful algal blooms. For this reason, it is important to detect and forecast these phenomena to be able to take action beforehand. Remote sensing is measuring and monitoring the characteristics of an area at a distance and it is typically done by satellites. Remote sensed data containing various environmental measurements can be used as the input for a machine learning system to estimate chlorophyll-a concentrations which is the main indicator used for detecting algal blooms. The main question this research aims to answer is: *Which input modality is the most predictive for estimating chlorophyll-a concentrations for water bodies in Uruguay?* This research presents the step-by-step construction of a system to pre-process the environmental data collected through remote sensing and use this data to train and test a machine learning system to assess and compare 11 different environmental factors or so-called data modalities individually against each other to find out the most predictive one. Carrying out the machine learning experiments brings the results into the open that radiation mean and turbidity of water are the two most predictive data modalities for algal bloom forecasting with accuracy scores of approximately 34%, while radiation mean is performing slightly better.

1 Introduction

An algal bloom is defined as a rapid increase in common algae (phytoplankton) abundance in water bodies and it can occur when a group of certain environmental factors is combined, such as but not limited to increased nutrients, warmer temperature, abundant light, and stable wind conditions [1]. Algal blooms are identified as problematic when the algae populations grow out of control. As a result of this growth, water bodies become contaminated and intoxicated, thus, damaging the whole ecosystem, together with all the creatures which depend on it to exist, including humans [2]. To this end, the reliability of detecting and forecasting these algal blooms has been crucial for researchers, graduate students, and professional engineers who are engaged in monitoring and modeling water quality [3].

Even though the most straightforward method of detecting algal blooms is direct water sampling followed by a biological analysis of the sample, it can be quite cumbersome and limited by spatial and temporal factors [4]. Remote sensing on the other hand, which is measuring and monitoring the characteristics of an area at a distance and is typically done by satellites [5], can hold an advantage over direct water sampling as vast areas can be covered with this method, although results could not be as precise as laboratory measurements.

Algal bloom monitoring and forecasting using remotely sensed data was initially proposed by Steidiner and Haddad in 1981 [6]. Since then, there have been lots of optimizations and various methods to forecast algal blooms. Furthermore, one of the most important data sources to detect algal blooms has been chlorophyll-a concentration for many investigations, as it directly refers to the algae abundance [7], which is also used and referred to as the "ground truth" in this research.

"Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities." [8]. As the nature of algal bloom forecasting is a multimodal problem since multiple environmental factors need to be considered while attempting to detect such events [9], this research aims to analyze each one of these modalities and make comparisons between them to answer the question of *which input modality is the most predictive for estimating chlorophyll-a concentrations* in water bodies located in Uruguay by:

1. Constructing a simple machine learning model,
2. Processing the sampled biological, water temperature, and meteorological data to enable a computer system to derive meaningful information from it,
3. Training and testing the machine learning model to analyze and compare the information derived from the data.

Comparing the results shows that radiation mean and turbidity of water are the most predictive data modalities among all the 11 that are assessed, however, per-class accuracy scores show that individual data modalities fail to predict certain intervals alone and make more meaningful predictions when they are combined.

This research paper is organized as follows. Section 2 gives example works conducted in this area which could ideally help answer the research question. Section 3 describes the methods used to answer the research question. Section 4 gives details of how the experimental setup has been constructed and presents the results that are obtained from the experiments. Section 5 dives into the ethical aspects of the research and is followed by Section 6 which makes a discussion and conclusion over the presented results. Finally, the limitations and recommendations for future work are presented in Section 7.

2 Related Work

This section focuses on example works that cover similar topics that are tackled in this research as well. The decisions taken for the methods are also compared to the existing strategies in algal bloom forecasting and they are presented together with the motivation choices.

2.1 Multi-dimensional Data Structure

Using multi-dimensional data sources for harmful algal bloom (HAB) detection is one of the most effective ways of characterizing HAB events in terms of how they are defined within the time and the space variables [10]. The two-dimensional data structure is commonly referred to as a table, when these two-dimensional tables are extended with more dimensions and they are stacked on top of each other, the data structure becomes a data cube [11]. Data cubes are especially important in Earth sciences as they contribute greatly to the challenge of working with big data since organizing all data sources into a single data cube makes it possible to access both multi-temporal and multi-spacial information in a simple and logical manner [12, 13]. Similar to the way how data cubes have been constructed in these papers, methods presented in this paper combine various different data sources and organize this data according to its temporal information to create multiple batches which form the data cubes to be used as the main source of data for the presented system.

2.2 Machine Learning Techniques

Due to the challenges of applying machine learning techniques on satellite data images, numerous different machine learning techniques have been utilized to overcome the issue of having to deal with big data and make accurate predictions [14]. CNN (Convolutional Neural Network) [15] is particularly important for image recognition [16] and can resolve the problem of the non-linear relationship between the spectrum of algal blooms [17], thus, the technique has been widely used by the experts working in this area. Moreover, there are more complicated machine learning techniques that build upon CNNs, to better process the high dimensionality of spatiotemporal data cubes for the outcome predictions, in that regard, a proposed solution makes use of ConvLSTM which is, in essence, "the deep learning model long short-term memory powered by convolutional structures for the transitions to tackle the now-casting problem." [18].

Validating the results obtained by machine learning algorithms requires analytically logical and indicative methods. ROC (receiver operating characteristic) curve is one of the widely known and used methods for assessing the performance of predictions obtained through classification, also for HAB forecasting [19, 20], which exploits the chosen system's strengths and weaknesses. In addition to that, accuracy score, Kappa coefficient and F-Score are also used as indicative evaluation metrics [9].

Much like the given research papers, this paper too utilizes the power of machine learning algorithms to train and test the model with high dimensional data, however, unlike the examples, the machine learning model has been constructed in a simple manner, taking into consideration that this research has been conducted in a limited amount of time and the chosen model is somewhat irrelevant for comparing the different environmental factors but is important for accurate HAB predictions. Furthermore, for evaluating the results, the total accuracy score and accuracy per class have been calculated and presented together with loss graphs of the training phases to judge the performance of the machine learning model.

2.3 Classification vs. Regression

The common practice for the presented previous works is making use of classification algorithms for the machine learning model to decide whether the situations indicate the presence of HABs or not, even though using classification introduces bias in the predictions because of the dependence on manually set values [21]. Using regression algorithms, on the other hand, allows for precise predictions on chlorophyll-a levels, and can even be further improved with using multiple linear regression (MLR) when "multicollinearity of the independent variables is removed" [22]. Nonetheless, using both methods is applicable for the detection of HAB events and seems to achieve satisfactory results with the help of accurate pre-processing of the data and properly done hyperparameter tuning [23]. This paper carries out the experiment with a classification-based machine learning model using the pre-determined class separator values defined by the domain experts.

3 Discovering the Most Predictive Data Modalities

This section describes the main concepts and presents the methods to answer the research question.

To assess different environmental factors of algal blooms, there needs to be a machine learning system constructed for predicting chlorophyll-a levels using a single data modality. Furthermore, appropriate data needs to be processed and used as this system's input, which consists of different types of meteorological, water temperature, and biological measurements that are sampled from one of the water reservoirs of the river Rio Negro in Uruguay. From the water reservoirs Palmar, Baygorria, and Bonete, Palmar reservoir has been selected as the water body under analysis and its data is used for the machine learning system because there are more samples belonging to the reservoir and using more data to train the model would ideally benefit input diversity, thus, the prediction accuracy [24].

Identifying the data modalities. In the case of algal bloom forecasting, each environmental measurement refers to a different data modality, and the final prediction is obtained by using a multimodal machine learning system that makes use of each data modality to learn and predict. The relevant data to forecast algal blooms are mainly retrieved from various satellites including but not limited to Sentinel 2, NOAA-GFS, and MODIS. The raw data of the Palmar reservoir which was collected by the satellites have been converted into operable formats and further organized concerning their timestamps by the domain experts. The modalities that are individually assessed and relevant to the research question are demonstrated in Table 1.

Biological Measurements	**Water Temperature Measurements**	**Meteorological Measurements**
Chlorophyll-a	Water temperature	Mean air temperature
Turbidity of the water		Mean cloud coverage
CDOM (Colored dissolved organic matter)		Precipitation sum
		Radiation mean
		Relative humidity mean
		U wind mean (Eastward wind vector)
		V wind mean (Northward wind vector)

Table 1: Data modalities categorized by their measurement types.

Using data cube as the data structure. The training data is organized as data cubes where each data cube contains the measurement values for each one of the data modalities corresponding to pixel locations in a 224×224 cropped square images that together compose the complete water reservoir image. To produce the data cubes, image samples for data modalities are stacked on top of each other to create multi-layered data structures. Depending on their batch size, which is the number of image samples taken from the complete water reservoir image, and the window size, which is the number of days of measurements to be used for the prediction, multi-layered structures are once again stacked to form the final data cubes. A visual representation of how the data is structured is presented in Figure 1. This structure of the input data allows for easy access to each layer simply by using indices.

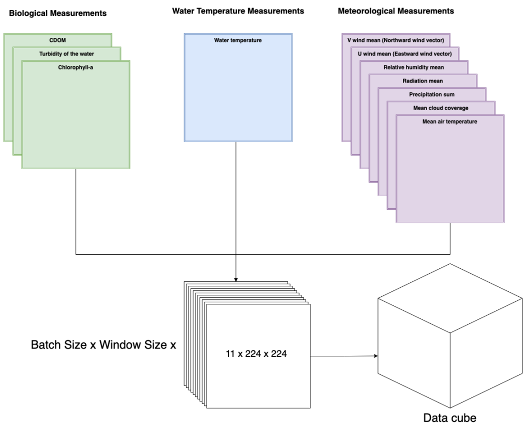

Figure 1: Upmost, each different data modality image is categorized under its measurement types, and below that the construction of the data cube with the multi-layered data modality measurements, batch size, and window size is demonstrated.

Test data is also organized similarly, however, as for the ground truth, there is merely one data modality which is the chlorophyll-a measurements of the day to predict. Furthermore, there is no window size since in the training data, multiple samples taken from different days could be used to predict the outcome but not vice versa.

Pre-processing of the data. Pre-processing is performed on both train and test data. For the train data, data modality images are first clipped between 10^{-6} and 150 to avoid extreme chlorophyll-a measurements while also allowing minor negative values since some of the data modalities such as water temperature, negative measurements are possible and valid. Clipping is followed by replacing the NaN (not a number) values with pre-calculated mean values for each modality to handle missing values without causing a skew in the data set. NaN values occur because of impossibility in sampling for some data modalities such as water temperature not being able to be sampled on land or CDOM (colored dissolved organic matter) not being able to be detected when there is a cloud band blocking the sensors of the satellite. After the NaN values are handled, the Yeo-Johnson transform is applied to reduce the skew in the raw variables [25]. Following that, the data is normalized with pre-calculated mean and standard deviation values using the Z-score normalization technique [26] to transform the features to be on the same scale. The last step applied to the training data is that a data modality is chosen from the data cube to be assessed, and as a result of this, one dimension in the data cube is collapsed. Finally, the dimensions of the pre-processed data cube are arranged correctly to distinguish the features of the machine learning model, which are the window size and measurement values of the chosen data modality.

For test data, values are first clipped between 0 and 150 to eliminate extreme or negative samples in the chlorophyll-a measurements. This time, NaN values are replaced with -1 as they are, later in the process, ignored and not used for training the machine learning model. In the last step of pre-processing, all ground truth data is binned to pre-defined intervals of 0-10 ug/L, 10-30 ug/L, 30-75 ug/L, and 75+ ug/L chlorophyll-a concentrations which are the ideal threshold values for chlorophyll-a classification and they have been again provided by the domain experts.

Linear classifier as the machine learning model. To be able to compare the effectiveness of various data modalities relative to each other, the chosen machine learning model is constructed in a simple manner using a linear classifier. The model has an input dimension of 1 for each measurement in the input image of each data modality and an output dimension of 5 referring to each class of intervals in binned ground truth measurements indicating the class label. The visual representation of chosen classes is demonstrated in Figure 2. NaN values that are converted to -1 in pre-processing step are assigned to the class representing the values lower than 0, which is also the class to be ignored while training the model. Moreover, for the sake of simplicity, the linear classifier is trained with samples of the data modalities taken one day before the day to predict, thus, the model always predicts one day ahead.

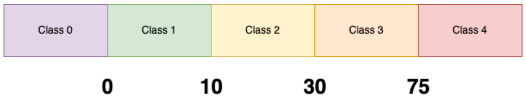

Figure 2: Pre-determined intervals of the ground truth data for the chlorophyll-a classification task, interval values are represented in ug/L units.

Evaluation of the machine learning model. In order to analyze the results, two common evaluation metrics for classification tasks are used. Cross-entropy values are calculated for the model's predictions with $-\sum_{c=1}^{M} y_c \log(p_c)$ where M denotes the number of classes, p_c denotes the prediction for the class c and y_c denotes the actual ground truth value of the class c [27]. Cross-entropy loss indicates the performance of the linear classifier and the loss values are plotted for each data modality over the cycles to observe an expected decrease in the gradual loss for training predictions. Furthermore, accuracy scores are calculated with $\text{Accuracy} = \frac{\text{Number of correct predictions of the class}}{\text{Number of occurrences of the class}}$ for all classes and total accuracy score of each data modality is calculated with $\text{Total Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of non-ignored data}}$. The results are later compared against each other to demonstrate how accurate each data modality with the outcome predictions is.

	Accuracy score (%) of class 1 (0-10 ug/L)	Accuracy score (%) of class 2 (10-30 ug/L)	Accuracy score (%) of class 3 (30-75 ug/L)	Accuracy score (%) of class 4 (75+ ug/L)	Total Accuracy (%)
Chlorophyll-a	25.19	53.18	0	0	27.47
Turbidity of the water	59.38	42.14	0	0	34.33
CDOM (Colored dissolved organic matter)	38.68	0	82	0	25.87
Water temperature	43.93	0	0	56.77	22.56
Mean air temperature	0	31.81	27.85	9.62	17.67
Mean cloud coverage	64.07	11.12	0	21.01	27.66
Precipitation sum	96.33	0	0	0	30.71
Radiation mean	17.87	70.78	0	0	34.86
Relative humidity mean	0	46.6	0	42.12	23.38
U wind mean (Eastward wind vector)	0	77.97	20.89	0	31.95
V wind mean (Northward wind vector)	95.92	0	3.38	0	31.14

Table 2: Results of the experiment that are carried out with single data modalities, presenting the accuracy scores of each data modality for chlorophyll-a concentration prediction of the next day in Palmar reservoir. Calculated accuracy scores are rounded to the nearest 2 decimal places for the sake of legibility.

4 Experimental Setup and Results

This section is dedicated to explanations of how the experimental setup has been constructed with the presented methods by diving into some implementation details and it further presents the results and their explanations.

Machine learning experiment is independently performed on each data modality as well as when all of them are combined. The presented linear model is trained using the Adam optimizer [28] to calculate the gradients over 25 epochs with a learning rate of 0.0001 and a batch size of 16. The criterion `torch.nn.CrossEntropyLoss` for calculating loss [29] is initialized with 2 optional arguments. The first optional argument is `ignore_index=0` to exclude class 0 (see Figure 2) from contributing to the learning process. The second optional argument is `weight` to give a manual rescaling to each class due to the ground truth chlorophyll-a measurements are being skewed towards low values which are discovered by Bayraktar by performing data analysis on the Palmar reservoir data [30]. The distribution of the classes is given in Table 3. Following the training process, the model is tested using unseen data with a batch size of 1, and the per-class accuracy scores together with total accuracy scores are calculated to be used as the evaluation metric.

Interval values for chlorophyll-a measurements (ug/L)	Occurrence in the dataset (%)
0-10	63.57
10-30	23.19
30-75	7.96
75+	5.27

Table 3: Distribution of chlorophyll-a measurements in the dataset of Palmar reservoir (NaN values excluded from the calculation) [30].

4.1 Predictions based on Single Data Modalities

This experiment aims to exploit the correlation between each data modality measurement and their respective chlorophyll-a concentration measurements belonging to the next day which are algal bloom forecasting predictions. The results in Table 2 show the accuracy scores retrieved after the testing process for each data modality. Overall, the total accuracy scores for all data modalities are between 15-35%. Radiation mean and turbidity of water have been the most predictive methods when the total accuracy scores are taken into account. For some modalities, predictions are heavily skewed on one class such as precipitation sum where the accuracy score of class 1 is 96.33% and the accuracy scores of the other classes are all 0%, which means that the model fails to predict any values outside the class 1. Mean air temperature and mean cloud coverage are the only two data modalities to predict values in 3 distinct classes, while the total accuracy score of mean air temperature is the lowest among all data modalities.

The plot showing the decrease in the loss for each epoch for the data modality chlorophyll-a is presented in Figure 3 and the loss plots for other data modalities can be found in Appendix A through Appendix J.

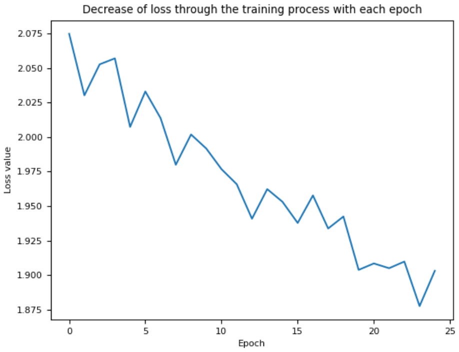

Figure 3: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with chlorophyll-a measurements.

4.2 Predictions based on Multiple Modalities

This experiment aims to combine all discussed data modalities to ideally obtain results with higher accuracy scores and more even distribution over the classes for the outcome predictions than the results of the experiments presented in Table 2. The results in Table 4 show the accuracy scores of the outcome predictions for each one of the classes as well as the total accuracy. The first notable matter is that there are predictions for each class which was not the case in the first experiment, although the total accuracy score is not any higher than the total accuracy scores of individual data modalities. Furthermore, the predictions are skewed towards class 1, where there is a significant decrease in accuracy scores from class 1 through class 4.

The plot showing the decrease in the loss for each epoch for all data modalities combined is presented in Figure 4.

	Accuracy score (%)
Class 1 (0-10 ug/L)	54.19
Class 2 (10-30 ug/L)	18.45
Class 3 (30-75 ug/L)	18.58
Class 4 (75+ ug/L)	3.61
Total Accuracy	27.63

Table 4: Results of the experiment that is carried out when all data modalities are used together as the input for the linear classifier for predicting the chlorophyll-a measurements which belong to the next day. Calculated accuracy scores are rounded to the nearest 2 decimal places for the sake of legibility.

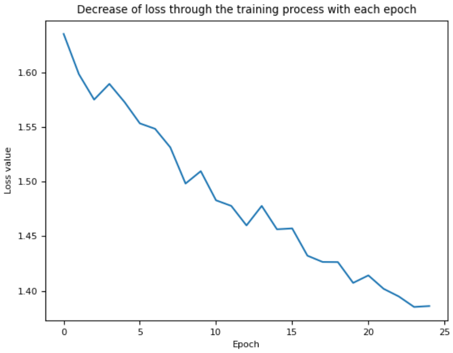

Figure 4: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with all data modalities.

5 Responsible Research

One of the most important aspects of conducting responsible research is that throughout the experiments, using data that is not subject to abuse or violence against privacy. In that regard, the data used to perform the experiments presented in the paper are collected by the domain experts and have been transmitted to the research group, thus there has been no data collection throughout the research. The raw data has been provided to TU Delft by various parties. Biological data has been provided by the Ministry of Environment of Uruguay. Data regarding meteorology has been collected by NOAA Global Forecast System and is publicly available online [31]. Moreover, all data that is relevant for the surface water temperature has been collected by the satellites "AQUA" and "TERRA" which are owned and operated by NASA [32].

Another important aspect of presenting responsible research is the reproducibility of the introduced experimental setup. What it means to the other researchers or interested parties who would like to carry out the same experiments and verify the results of this research is, they are able to do it with the knowledge and methods conveyed in the paper. Throughout the research, it has been made sure that all introduced methods are reproducible. Especially, extra attention has been paid to the random sampling when loading the dataset and with the help of `torch.manual_seed` method, a manual seed has been set for generating random numbers. All the code that has been used for carrying out the experiments is also publicly accessible and is kept in TU Delft's servers. In that regard, the repository is also well documented and each important operation has a comment describing its purpose.

6 Discussion and Conclusion

The main aim of this research is to compare different environmental factors as data modalities and discover how predictive each one of them is for estimating chlorophyll-a concentrations. The results of this research have provided different correlations between the assessed data modalities and chlorophyll-a concentrations. When only total accuracy scores are taken into account, it can be concluded that turbidity of water and radiation mean are the most predictive data modalities among all the 11 with accuracy scores both higher than 34%. However, it is important to highlight that these modalities failed to produce any instances of two classes, Moreover, none of the data modalities managed to predict chlorophyll-a concentrations with high and reliable accuracy scores. On the other hand, the results which are presented in Table 4 show that data modalities are more meaningful in terms of predicting different outcomes when they are used together to train a machine learning system, although the accuracy scores are still not indicative precise predictions.

The results are not quite comparable to similar studies since a group of different data modalities and domain values are used, however, data modalities such as chlorophyll-a and photosynthetically available radiation also play an important role in much more advanced HAB detection systems [10]. Another important aspect to highlight is the performance constructed system, as it is extremely limited in accurately predicting HABs even when the presented data modalities are combined. Nonetheless, this research provides methods and results to satisfy the contributions described in Section 1.

1	Radiation mean
2	Turbidity of the water
3	U wind mean (Eastward wind vector)
4	V wind mean (Northward wind vector)
5	Precipitation sum
6	Mean cloud coverage
7	Chlorophyll-a
8	CDOM (Colored dissolved organic matter)
9	Relative humidity mean
10	Water temperature
11	Mean air temperature

Table 5: The rank list of data modalities concerning their total accuracy scores.

To compare and assess different data modalities used for algal bloom forecasting, a piece of necessary knowledge is provided throughout this research together with a reproducible methodology. The research explains how a multi-dimensional data structure named a data cube is utilized to organize the data provided by various sources and how this data is further modified to be meaningful for a computer system. Finally, it builds a system that is able to produce comparable and interpretable results which use the data cubes as its input to do so.

This research is concluded by presenting a rank list of all 11 different modalities to predict chlorophyll-a concentrations for algal bloom forecasting in Table 5, created with regard to their total accuracy scores, thus, revealing how predictive each one of them is in comparison to each other.

7 Limitations and Future Work

The machine learning system presented and constructed in this research is limited by what a linear classifier is capable of, which is not extremely efficient in terms of both operating speed and prediction accuracy. A very useful way to overcome this limitation is to utilize a more powerful machine learning model for the scenario presented in this paper, as discussed in Section 2 with the related works, such as a CNN or ConvLSTM [18] which would perform better with an image dataset [17].

The dataset of Palmar water reservoir which is used to create data cubes in this research is not evenly distributed over the given classes as also mentioned in Section 4 [30]. A solution is already applied in the current system to prevent bias for the prediction outcome, more accurate predictions can be obtained by further extending the dataset or using a different domain with the presented system as a method of sanity check.

The selected evaluation metrics already form a good basis for comparing the results, however, class accuracy is not indicative enough in terms of what is classified incorrectly and can sometimes be misleading. For more feasible comparisons, different evaluation metrics such as confusion matrices can be calculated for each data modality and incorrect classifications can be exploited [33].

References

[1] M. of Environment and C. C. Strategy, *What causes an algae bloom? - Province of British Columbia*, Jun. 2022. [Online]. Available: https://www2.gov.bc.ca/gov/content/environment/air-land-water/water/water-quality/algae-watch/what-are-algae/causes-of-an-algae-bloom.

[2] R. . Santoleri, "Year-to-year variability of the phytoplankton bloom in the southern Adriatic Sea (1998–2000): Sea-viewing Wide Field-of-view Sensor observations and modeling study," *Journal of Geophysical Research*, vol. 108, no. C9, 2003. DOI: 10.1029/2002jc001636. [Online]. Available: http://dx.doi.org/10.1029/2002jc001636.

[3] W. Zhang and I. Lou, "Monitoring and modeling algal blooms," in *Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs: General Principles and a Case study of Macau*, I. Lou, B. Han, and W. Zhang, Eds. Dordrecht: Springer Netherlands, 2017, pp. 1–14, ISBN: 978-94-024-0933-8. DOI: 10.1007/978-94-024-0933-8_1. [Online]. Available: https://doi.org/10.1007/978-94-024-0933-8_1.

[4] S. E. Craig *et al.*, "Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, Karenia brevis," *Applied Optics*, vol. 45, no. 21, p. 5414, Jul. 2006. DOI: 10.1364/ao.45.005414. [Online]. Available: http://dx.doi.org/10.1364/ao.45.005414.

[5] *What is remote sensing and what is it used for? — U.S. Geological Survey*, Jan. 2022. [Online]. Available: https://www.usgs.gov/faqs/what-remote-sensing-and-what-it-used.

[6] K. A. Steidinger and K. . Haddad, "Biologic and Hydrographic Aspects of Red Tides," *BioScience*, vol. 31, no. 11, pp. 814–819, Dec. 1981. DOI: 10.2307/1308678. [Online]. Available: http://dx.doi.org/10.2307/1308678.

[7] D. . Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress in Oceanography*, vol. 123, pp. 123–144, Apr. 2014. DOI: 10.1016/j.pocean.2013.12.008. [Online]. Available: http://dx.doi.org/10.1016/j.pocean.2013.12.008.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019. DOI: 10.1109/TPAMI.2018.2798607.

[9] M. Izadi, M. Sultan, R. E. Kadiri, A. Ghannadi, and K. Abdelmohsen, "A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom," *Remote Sensing*, vol. 13, no. 19, 2021, ISSN: 2072-4292. DOI: 10.3390/rs13193863. [Online]. Available: https://www.mdpi.com/2072-4292/13/19/3863.

[10] P. R. Hill, A. Kumar, M. Temimi, and 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*, vol. 13, pp. 3229–3239, 2020. DOI: 10.1109/JSTARS.2020.3001445.

[11] H. Hristova, "What Is a Data Cube?," Jul. 2022. [Online]. Available: https://365datascience.com/trending/data-cube/.

[12] P. Baumann, "Datacube Standards and their Contribution to Analysis-Ready Data," *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2018. DOI: 10.1109/igarss.2018.8518994. [Online]. Available: http://dx.doi.org/10.1109/igarss.2018.8518994.

[13] P. Baumann, "Big Data Standards and Analysis-Readiness: Status and Evolution," *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Sep. 2020. DOI: 10.1109/igarss39084.2020.9323424. [Online]. Available: http://dx.doi.org/10.1109/igarss39084.2020.9323424.

[14] X. Li, J. Yu, Z. Jia, and J. Song, "Harmful algal blooms prediction with machine learning models in tolo harbour," in *2014 International Conference on Smart Computing*, 2014, pp. 245–250. DOI: 10.1109/SMARTCOMP.2014.7043865.

[15] I. C. Education, *Convolutional Neural Networks*, Jan. 2021. [Online]. Available: https://www.ibm.com/cloud/learn/convolutional-neural-networks.

[16] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Jun. 2018. DOI: 10.1007/s13244-018-0639-9. [Online]. Available: http://dx.doi.org/10.1007/s13244-018-0639-9.

[17] J. Shin, B.-K. Khim, L.-H. Jang, J. Lim, and Y.-H. Jo, "Convolutional neural network model for discrimination of harmful algal bloom (HAB) from non-HABs using Sentinel-3 OLCI imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 250–262, Sep. 2022. DOI: 10.1016/j.isprsjprs.2022.07.012. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2022.07.012.

[18] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, *Convolutional lstm network: A machine learning approach for precipitation nowcasting*, 2015. DOI: 10.48550/ARXIV.1506.04214. [Online]. Available: https://arxiv.org/abs/1506.04214.

[19] S. Malek, S. M. Syed Ahmad, S. K. K. Singh, P. Milow, and A. Salleh, "Assessment of predictive models for chlorophyll-a concentration of a tropical lake," *BMC Bioinformatics*, vol. 12, no. S13, Nov. 2011. DOI: 10.1186/1471-2105-12-s13-s12. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-12-s13-s12.

[20] "Bayesian model averaging for harmful algal bloom prediction," vol. 19, no. 7, pp. 1805–1814, 2009, ISSN:

10510761. [Online]. Available: http://www.jstor.org/stable/40346289 (visited on 12/22/2022).

[21] R. Elkadiri *et al.*, "Development of a Coupled Spatiotemporal Algal Bloom Model for Coastal Areas: A Remote Sensing and Data Mining-Based Approach," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 11, pp. 5159–5171, 2016. DOI: 10.1109/JSTARS.2016.2555898.

[22] I. In Ieong, I. Lou, W. K. Ung, and K. M. Mok, "Using Principle Component Regression, Artificial Neural Network, and Hybrid Models for Predicting Phytoplankton Abundance in Macau Storage Reservoir," *Environmental Modeling & Assessment*, vol. 20, no. 4, pp. 355–365, Oct. 2014. DOI: 10.1007/s10666-014-9433-3. [Online]. Available: http://dx.doi.org/10.1007/s10666-014-9433-3.

[23] S. Es, "Hyperparameter Tuning in Python: a Complete Guide," Dec. 2022. [Online]. Available: https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide.

[24] *How Much Data Is Required for Machine Learning? – PostIndustria*. [Online]. Available: https://postindustria.com/how-much-data-is-required-for-machine-learning/.

[25] I.-K. Yeo, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, Dec. 2000. DOI: 10.1093/biomet/87.4.954. [Online]. Available: http://dx.doi.org/10.1093/biomet/87.4.954.

[26] S. Gupta, "Z-Score Normalization - Machine Learning Concepts," Dec. 2022. [Online]. Available: https://ml-concepts.com/2021/10/08/z-score-normalization/.

[27] S. Maheshkar, "What Is Cross Entropy Loss? A Tutorial With Code," Mar. 2022. [Online]. Available: https://wandb.ai/sauravmaheshkar/cross-entropy/reports/What-Is-Cross-Entropy-Loss-A-Tutorial-With-Code--VmlldzoxMDA5NTMx.

[28] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Cornell University - arXiv*, Jan. 2015. DOI: 10.48550/arxiv.1412.6980. [Online]. Available: http://arxiv.org/pdf/1412.6980.

[29] V. Yathish, "Loss Functions and Their Use In Neural Networks - Towards Data Science," Aug. 2022. [Online]. Available: https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9.

[30] K. C. Bayraktar, "Spatio-Temporal Embedding in Deep Learning for Algal Bloom Forecasting," *CSE3000 Research Project*, in press.

[31] *Global forecast system (gfs)*, May 2022. [Online]. Available: https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast.

[32] *MODIS Web*. [Online]. Available: https://modis.gsfc.nasa.gov/about/.

[33] J. Brownlee, *What is a confusion matrix in machine learning*, Aug. 2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/.

A Loss graph of Turbidity of Water

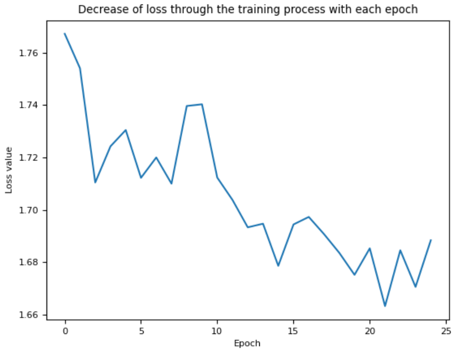

Figure 5: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with turbidity of water measurements.

B Loss graph of CDOM

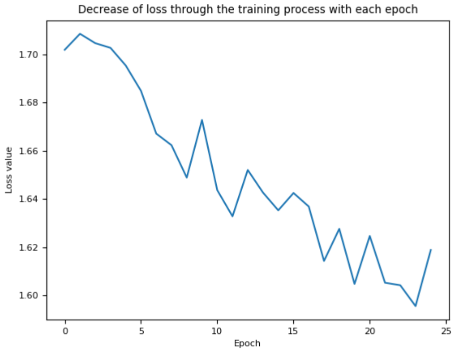

Figure 6: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with CDOM measurements.

C Loss graph of Water Temperature

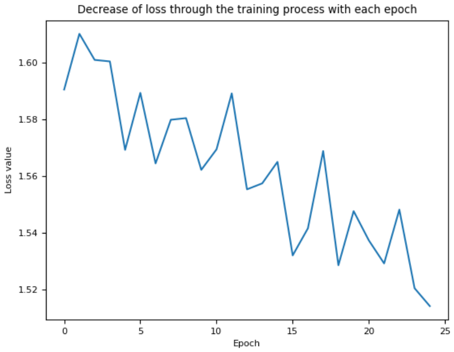

Figure 7: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with water temperature measurements.

D Loss graph of Mean Air Temperature

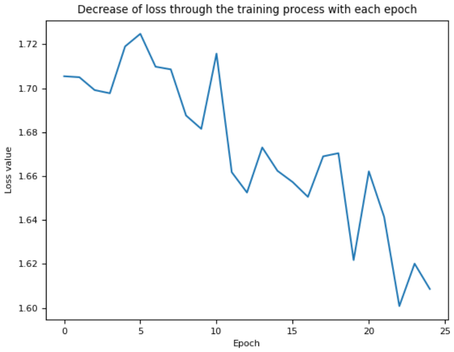

Figure 8: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with mean air temperature measurements.

E Loss graph of Mean Cloud Coverage

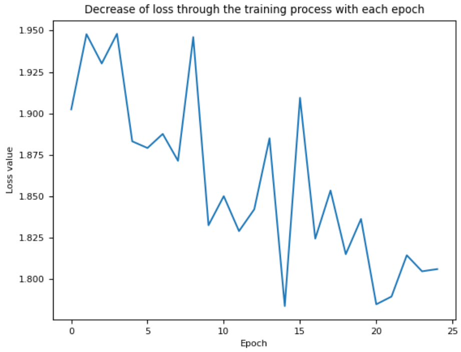

Figure 9: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with mean cloud coverage measurements.

F Loss graph of Precipitation Sum

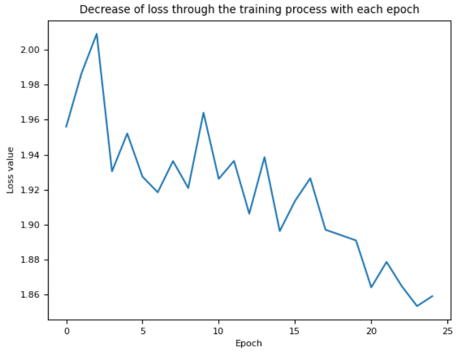

Figure 10: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with precipitation sum measurements.

G Loss graph of Radiation Mean

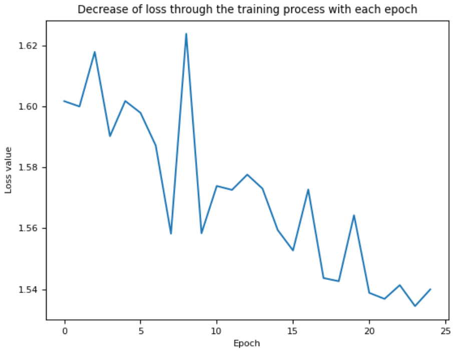

Figure 11: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with radiation mean measurements.

H Loss graph of Relative Humidity Mean

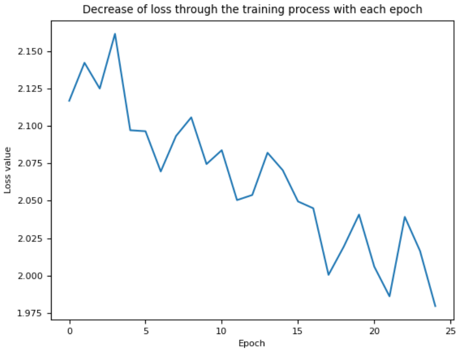

Figure 12: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with relative humidity mean measurements.

I Loss graph of U Wind Mean

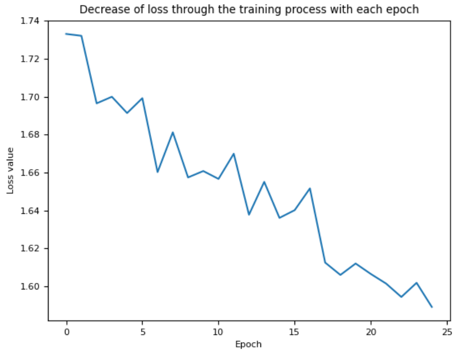

Figure 13: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with U wind mean measurements.

J Loss graph of V Wind Mean

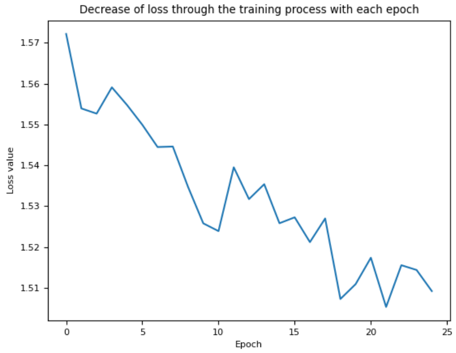

Figure 14: Plot showing the decrease in loss, thus difference from the desired target with each epoch when training the system with V wind mean measurements.