Estimation of chlorophyll concentration for environment monitoring in Scottish marine water.

YAN, Y., ZHANG, Y., REN, J., HADJAL, M., MCKEE, D., KAO, F.-J., and DURRANI, T.

2022

Estimation of chlorophyll concentration for environment monitoring in Scottish marine waters

Yijun Yan[1], Yixin Zhang[2], Jinchang Ren[1*], Madjid Hadjal[3], David McKee[3],
Fu-jen Kao[4], Tariq Durrani[2]

[1] National Subsea Centre, Robert Gordon University, Aberdeen, U.K.
[2] Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK
[3] Dept. of Physics, University of Strathclyde, Glasgow, UK
[4] National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC

Abstract. The Scottish Government is tasked with reporting on the environmental status of Scottish marine waters, an enormous area of water extending from the shoreline to deep oceanic waters. As one of the most important variables, chlorophyll concentration (Chl) plays an important role in seawater quality monitoring as an indicator of eutrophication. Currently, Chl observations are mostly done by expensive ship-based surveys that have very limited spatio-temporal coverage. Satellite based ocean colour remote sensing has the potential to significantly enhance monitoring capabilities but this opportunity has not been widely adopted by statutory reporting bodies across Europe due to concerns over satellite data quality. To break through this bottleneck, in this paper we explore implementation of advanced machine learning techniques to automatically estimate Chl via the historic time series of ocean colour remote sensing data from July 2002 to September 2019.

Keywords: Environment monitoring, Scottish marine waters, Chlorophyll, Multispectral remote sensing.

1 Introduction

Scotland is a maritime nation with territorial waters that have a combined areal extent six times greater than the land area. This massive area ranges from the seashore to deep abyssal ocean depths, and supports a diverse range of industries including energy production, oil and gas, fisheries and aquaculture and tourism, bringing contributions of tens of billions of pounds to the Scottish economy annually. These waters are protected under various pieces of legislation including the EU Marine Strategy Framework directive [1], the Water Framework Directive [2] and other international obligations such as the OSPAR convention. Marine Scotland Science (MSS) is the Scottish Government Directorate with responsibility to report the water quality regularly. However, such a vast area of ocean can only be sampled at relatively sparse spatial and temporal resolution using traditional *in situ* sampling approaches. For example, the Scottish

adfa, p. 1, 2011. Corresponding author, Prof. J. Ren (J.Ren@rgu.ac.uk).

Coastal Observatory consists of only 11 monitoring sites, from which conditions around Scotland are extrapolated [3]. Chlorophyll concentration (Chl) is commonly used as an indicator for observing primary productivity (PP), formation of harmful algal blooms (HABs) and monitoring the occurrence of eutrophication events (EE) [4]. It can be retrieved from satellite-based ocean colour remote sensing data with daily, global resolution (subject to cloud cover and daylight restrictions) [5]. Unfortunately, uncertainty and variability of data quality has proved to be a major hurdle hindering widespread uptake of ocean colour data by MSS and other public environmental monitoring bodies. In this paper, we will use existing machine learning techniques to estimate Chl from satellite ocean colour reflectance signals. This will inform evaluation of the historic ocean colour time series, with Marine Scotland Science's existing database of *in situ* Chl data for testing Chl algorithms' performance.

The outline of this paper is as follows: Section 2 introduces the data and methods used in this work. Section 3 presents some preliminary experimental results. Finally, some concluding remarks and future work are summarized in Section 4.

2 Methods

2.1 Data source

The North Sea area near Scottish coastal is used as a case study in this paper. The *in situ* Chl data was sourced from Marine Scotland and then matched up against remote sensing optical data (15 spectral bands of remote sensing reflectance – CMES_PML processing) obtained from CMEMS (Copernicus Marine Environmental Monitoring Service) [6]. There are 16609 cloud-free matchup samples between *in situ* and remotely sensed data in total covering the period from July 2002 to September 2019. *In situ* Chl data are measured at different depth (0-30m). An illustration of the dataset is shown in

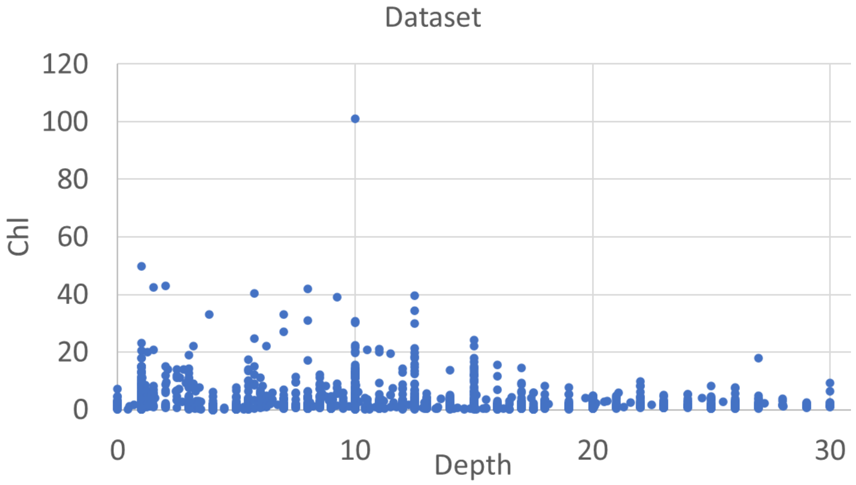


Fig. 1. Illustration of Chl (mg m^{-3}) v.s. Depth

Fig. 1. In this work, all Chl data are categorized into three classes, i.e. Chl-30, Chl-20 and Chl-10, which represents averaged Chl data that are available between 0 to 30, 20 and 10 meters, respectively. For example, when the *in situ* measurement depth is less than 10 meter, Chl-10, Chl-20 and Chl-30 have the same value. While the *in situ* measurement depth is higher than 10 meters, the Chl-10 will be zero and the Chl-20 will equal to Chl-30. To this end, each spectral sample has three associated Chl values.

2.2 Statistical analysis

In this work, four machine learning (ML) models i.e. Lasso regression, Ridge regression [7], Support Vector Machine (SVM) [8, 9] and Random Forest (RF) [10, 11] are used to evaluate the prediction performance of chlorophyll concentration. In the process of model training, the data set is randomly divided into a training set and testing set at a ratio of 8:2, resulting 13287 training samples and 3322 testing samples. The training set is used for model training and fitting, and the testing set is used to evaluate the accuracy of the model.

To simplify the computation and facilitate the convergence speed and accuracy of the training models, all data X in this work is normalized through the zero-mean normalization processing (Eq.(1)).

$$X_{nor} = \frac{X - \text{mean } (X)}{\text{std } (X)} \tag{1}$$

During the test, Mean absolute error (MAE), and R-square (R^2) score are used as the evaluation criterion [12] (Eq.(2-3)). The definition of R^2 score is shown below. It reflects the proportion of the total variation of the dependent variable that can be explained by the independent variable through the regression relationship. The numerator is the mean square error, and the denominator is the variance.

$$MAE = 10\text{^}\left(\frac{\sum_{i=1}^{n}(|log_{10}(y_i) - log_{10}(y_i')|)}{n}\right) \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(log_{10}(y_i) - log_{10}(y_i')\right)^2}{\sum_{i=1}^{N}(log_{10}(y_i) - y_i'')^2} \tag{3}$$

where y_i and y_i' represent the actual value and predicted value of the i^{th} Sample, respectively; y_i'' is the mean of all $log_{10}(y_i)$ values.

3 Results and discussion

In this section, we present statistical results for inter-comparison of machine learning methods in Table 1 and log-transformed Chl estimation results for specific matchups in Fig. 2, respectively. It can be seen that two linear regression methods (i.e. Lasso regression and Ridge regression) have very fast computation speed but yield relatively poor estimation accuracy. As seen their R^2 value are all negative, which means the regression model cannot be fitted on the data very well. SVM has better prediction accuracy than linear regression methods, but it has the highest computation cost. Among these ML models, random forest yields the highest R^2 score and lowest MAE, and has much

Table 1. Performance evaluation of chlorophyll estimation on 3322 testing matchups

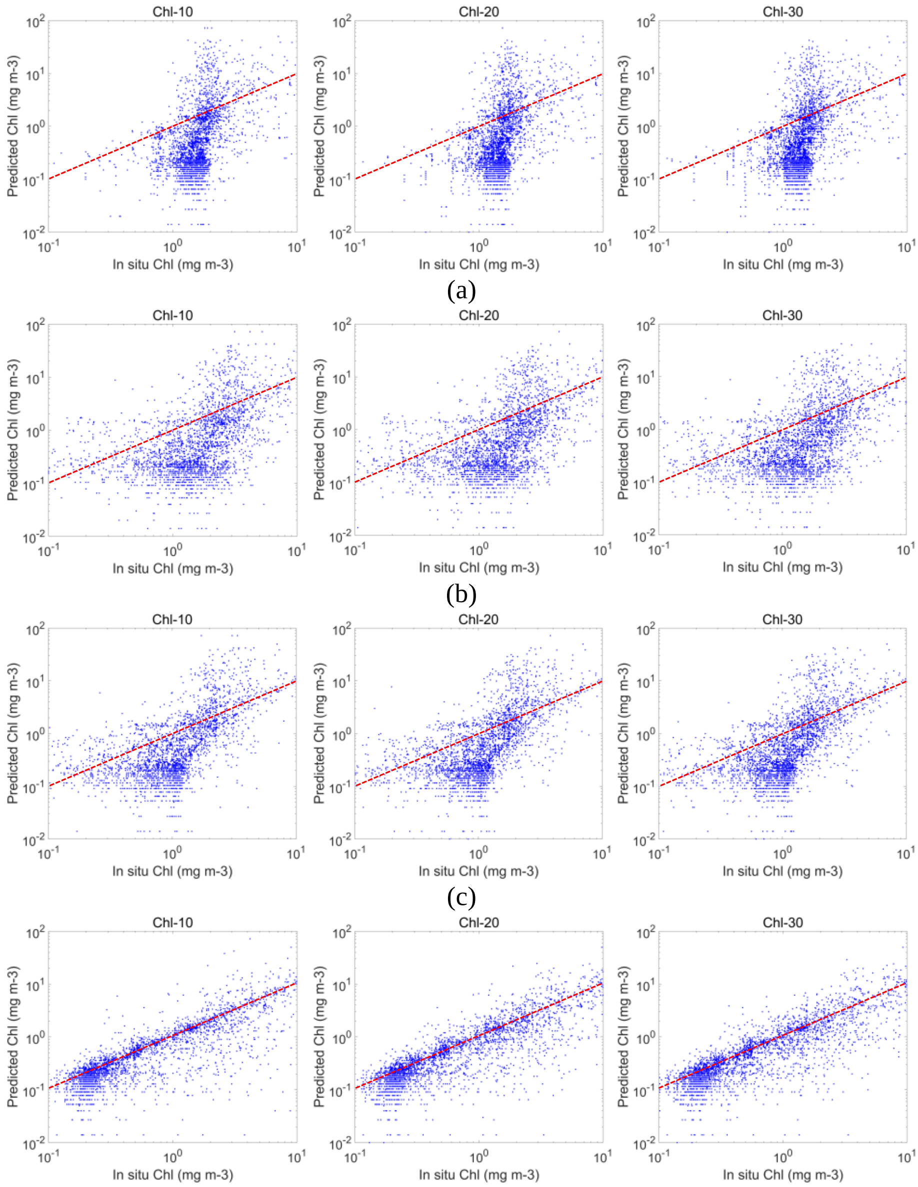
Regressors		**Lasso**	**Ridge**	**SVM**	**Random forest**
Parameters		Alpha=0.1	Alpha=0.1	Kernel='rbf', cost=8, gamma=0.5	No. tree = 40 Max depth=20
Time(s)		0.043	0.012	9.74	3.64
MAE	Chl-30	3.7795	3.4262	2.8781	1.6587
	Chl-20	3.9619	3.5744	2.8960	1.6408
	Chl-10	4.0457	3.7228	2.8661	1.5813
R^2	Chl-30	-0.3451	-0.1617	0.1051	0.6832
	Chl-20	-0.3635	-0.1716	0.1224	0.6923
	Chl-10	-0.3478	-0.1864	0.1878	0.7178

reduced computation cost than SVM. Fig. 2 shows the predicted Chl results in the range from 0.01 to 10 mg m^{-3} as most *in-situ* Chl values were lower than 10 mg m^{-3} (Fig. 1). As seen in Fig. 2, linear regression cannot predict the Chl very well as the blue dots are far from red 1:1 line which means most predicted results different against actual value. SVM has some good prediction results when the Chl is between 1 mg m^{-3} to 2 mg m^{-3}. However, when the Chl is larger (i.e. >2 mg m^{-3}), the blue dots and the red line have a great distance, which indicates poor prediction performance. Random forest has the lowest prediction bias when the Chl is lower than 4 mg m^{-3}, but uncertainties increase for higher values of Chl. The main reason is that there is an imbalance problem in the data, as most matchup samples have low Chl values (<5 mg m^{-3}). Therefore, Chl in the range from 0 to 5 mg m^{-3} will have better prediction accuracy. There are very few data available for Chl values greater than 5 mg m^{-3} which means that the machine learning models cannot be well trained, resulting in poor prediction accuracy and low R^2 values. Overall, Random forest produces the worst prediction performance on Chl-30 data in terms of higher MAE and lower R^2. In this case, Chl-20 data and Chl-10 data are recommended to use. Chl-10 data has the best prediction results in our study because it usually works well in open ocean waters. On the different note, using Chl-20 data can avoid freshwater inputs from land to make the satellite fail.

4 Conclusions

This research examines the performance of four common learning machines in the task of estimating chlorophyll levels from ocean colour remote sensing. Some outcomes are summarized as follows from the preliminary experimental results:

1) Random forest yields the best prediction accuracy on the existing data



(a)

(b)

(c)

(d)

Fig. 2. Visualized testing results of Chl estimation using (a) Lasso regression, (b) Ridge regression, (c) SVM, (d) Random Forest. Data were log10 transformed for display.

2) Better prediction accuracy can be achieved by ML methods when Chl is in the range from 0 to 4 mg m^{-3}.

3) Current performance levels are constrained by imbalanced data in training and limited amount of training data for Chl concentrations > 5 mg m^{-3}.

4) Given current data limitations, better feature extraction techniques may improve the prediction accuracy.

For the future work, we will work to collect more high-quality spectral data and *in-situ* data. Once an expanded training data set is available, we will test novel band selection methods [13] and feature extraction methods [9] to extract the most useful information and help to get more accurate prediction results. Some chlorophyll algorithms such as [14] and [15] will be also useful for the data from specific ocean color sensors such as sentinel-3 and NASA, etc. Super resolution model [16] will be also beneficial to the spatial-level Chl estimation. Additional options to improve prediction performance include use of novel deep learning frameworks such as multi-scale feature framework [17] and optimization framework [18].

Acknowledge

This work is partially supported by the exchange project on ocean color remote sensing funded by Royal Society of Edinburgh and Ministry of Science and Technology of Taiwan, ROC.

References

1. Directive SF. Directive 2008/56/EC of the European Parliament and of the Council. Journal) Council Decision of. 2008.
2. Directive WF. Water Framework Directive. Journal reference OJL. 2000;327:1-73.
3. Marine Scotland. Available from: https://marine.gov.scot/data/scottish-coastal-observatory-data.
4. Harvey ET, Kratzer S, Philipson P. Satellite-based water quality monitoring for improved spatial and temporal retrieval of chlorophyll-a in coastal waters. Remote Sensing of Environment. 2015;158:417-30.
5. Ruddick K, Lacroix G, Lancelot C, Nechad B, Park Y, Peters S, et al. Optical remote sensing of the North Sea. Remote Sensing of the European Seas: Springer; 2008. p. 79-90.
6. Copernicus Marine Environmental Monitoring Service. Available from: https://www.pml.ac.uk/.
7. Polat E, Gunay S. The comparison of partial least squares regression, principal component regression and ridge regression with multiple linear regression for predicting pm10 concentration level based on meteorological parameters. Journal of Data Science. 2015;13(4):663-92.
8. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27.
9. Yan Y, Ren J, Tschannerl J, Zhao H, Harrison B, Jack F. Nondestructive Phenolic Compounds Measurement and Origin Discrimination of Peated Barley Malt using Near-infrared Hyperspectral Imagery and Machine Learning. IEEE Transactions on Instrumentation and Measurement. 2021.

10. Shah SH, Angel Y, Houborg R, Ali S, McCabe MF. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. Remote Sensing. 2019;11(8):920.
11. Hu C, Feng L, Guan Q. A Machine Learning Approach to Estimate Surface Chlorophyll a Concentrations in Global Oceans From Satellite Measurements. IEEE Transactions on Geoscience and Remote Sensing. 2020;59(6):4590-607.
12. Seegers BN, Stumpf RP, Schaeffer BA, Loftin KA, Werdell PJ. Performance metrics for the assessment of satellite data products: an ocean color case study. Optics Express. 2018;26(6):7404-22.
13. Sun H, Ren J, Zhao H, Yuen P, Tschannerl J. Novel Gumbel-Softmax Trick Enabled Concrete Autoencoder With Entropy Constraints for Unsupervised Hyperspectral Band Selection. IEEE Transactions on Geoscience and Remote Sensing. 2021.
14. O'Reilly JE, Werdell PJ. Chlorophyll algorithms for ocean color sensors-OC4, OC5 & OC6. Remote Sensing of Environment. 2019;229:32-47.
15. Pahlevan N, Smith B, Schalles J, Binding C, Cao Z, Ma R, et al. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. Remote Sensing of Environment. 2020;240:111604.
16. Ha VK, Ren J-C, Xu X-Y, Zhao S, Xie G, Masero V, et al. Deep Learning Based Single Image Super-resolution: A Survey. International Journal of Automation and Computing. 2019;16(4):413-26.
17. Sun G, Zhang X, Jia X, Ren J, Zhang A, Yao Y, et al. Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images. International Journal of Applied Earth Observation and Geoinformation. 2020;91:102157.
18. Fang Z, Ren J, Marshall S, Zhao H, Wang S, Li X. Topological optimization of the densenet with pretrained-weights inheritance and genetic channel selection. Pattern Recognition. 2021;109:107608.