CAPSTONE PROJECT BY TEAM 12

A DATA SCIENCE APPROACH TO MODEL PHYTOPLANKTON CONCENTRATION

Xinyu Xu (z5175081), Yuewen Mao (z5210649), Zilin Li (z5158442), Cheng Qian (z5158272), Henry Jiang (z5205963).

School of Mathematics and Statistics
UNSW Sydney

November 2020

Submitted in partial fulfilment of the requirements of the capstone course DATA3001

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: Clare Xinyu Xu Date: 21/11/2020

Signed: Date: 21/11/20

Signed: Date: 21/11/20

Signed: Date: 21/11/20

Signed: Date: 21/11/2020

Acknowledgements

By far the greatest thanks must go to my supervisor for the guidance, care and support they provided.

Thanks must also go to Pierre, Amendia, Michele, Tom, Batista and David who helped by proof-reading the document in the final stages of preparation.

Although we have not lived with them for a number of years, our families also deserve many thanks for their encouragement. Thanks go to Robert Taggart for allowing his thesis style to be shamelessly copied.

20/11/2020.

Abstract

Australia's coast is home to over 80% of its population, so for a large majority of Australian's the ocean is an important and ever-present factor in their lives. Phytoplankton are microscopic marine algae pivotal to the health of the marine biosphere, it is a core building block of the entire marine food web, as a result fluctuations in their concentration are of chief concern to researchers and governmental bodies who study, monitor and manage oceanic biospheres, as well as fisheries all along the coast of Australia. Tasked by the Fluid Dynamics, Oceanic and Atmospheric Sciences research group at UNSW, we aimed to model the relationship between phytoplankton growth and changes in environmental factors such as, depth, temperature and light intensity, and in doing so we attempted to gain an understanding of the important factors affecting phytoplankton growth.In order to understand the given dataset and possible angles of approaching our modelling process we first conducted a review of the relevant literature and preliminary data exploration. The combination of the two both exposed the flaws in the raw dataset, as well as provided starting points for further modelling construction and refinement. A preliminary data exploration facilitated through the use of Python exposed a significant percentage of missing or bad data points, however through a case by case process we were able to determine appropriate solutions to each instance, whether that be partial or complete deletion or imputation. This pre-processing phase was informed by a review of relevant literature on best practices and similar methodology. The basis of the modelling process was informed by what was learnt during the literature review from those starting points a more in-depth review of our own time and complexity limitations informed our final decision on our modelling process, ultimately selecting a generalized linear modelling approach. In order to simplify the modelling process, the original dataset was split up by their respective glider missions into two sets,TwoRocks and Leeuwin, from there our models were slowly built from the ground up in R, being cautious of the fit of the model and issues such as multicollinearity, ultimately resulting in two models for each mission,(TWOROCKS MODEL),(LEEUWIN MODEL). These models shed light on the covariation between phytoplankton concentration and environmental factors. The presence of light penetration (VBSC) and light intensity (IRRAD) reflect the importance of light as an energy source for phytoplankton. Furthermore, the presence of salinity and conductivity, reflects the existence of a relationship, however, the current understanding of its role in the growth of phytoplankton is not as robust, presenting a possibility of further research.

Contents

Chapter 1

Introduction

The project proposed by the Fluid Dynamics, Oceanic and Atmospheric Sciences research group at UNSW involves exploring and attempting to model the growth of phytoplankton off the east coast of Australia with the use of a wide breadth of related environmental variables. The focal point of the project revolves around the variation of phytoplankton growth, phytoplankton are the core building blocks of the entire marine ecosystem, acting as the base of the food web suppling energy and nutrients to all other marine life. As such phytoplankton growth in the ocean is of great concern to the health of marine ecosystems, marine biologists, government bodies who study and monitor such ecosystems, as well as fisheries all along the coast of Australia.

The ultimate goal for our project is to produce a model that best explains the growth of phytoplankton as a function of more readily available and easily obtainable data and environmental measures, due to the difficulty of autonomously directly measuring phytoplankton biomass, a major component of phytoplankton, chlorophyll-a, is measured instead as a proxy. Increases and decreases in chlorophyll-a concentration indicate growth, or lack thereof, of phytoplankton and will be used interchangeably throughout this paper.

In order to accomplish this goal, we have been provided with millions of data points taken from research vessels called 'gliders', these gliders are autonomous diving vessels that traverse the coasts of Australia while periodically changing depth. During this process gliders record a multitude of environmental variables ranging from their position, depth and water temperature to light penetration, intensity, water conductivity, oxygen concentration and chlorophyll concentration. In addition to the variables, their associated quality control classifiers are included, indicating whether the recorded data is good, bad, correctable, missing or interpolated. These data points are what we aim to use in order to model and gain insight into the relationship between phytoplankton growth and the surrounding environment.

Chapter 2

Literature Review

The focus of this project is the modelling of various environmental variables and their interaction with our variable of interest; concentration of chlorophyll-a. The identification of chlorophyll-a concentration as our variable of interest is a result of its use as a proxy for the presence of algal biomass, the real-world factor of interest. Our understanding of said interaction and the subsequent modelling process is informed by the surrounding literature, of which, we will discuss in the current chapter.

Our understanding of the relationship between chlorophyll-a concentrations and other environmental variables such as temperature, light intensity and depth hinges on the assumption that chlorophyll-a concentration is a useful indicator of phytoplankton growth/blooms. Examining the literature concerning this topic we can see that many studies have used chlorophyll-a concentrations as a proxy for phytoplankton growth. Desortova, 1981[1], studied the relationship between chlorophyll-a concentration and phytoplankton biomass, water samples were taken from five water reservoirs at weekly or three week intervals over two year periods, measures of chlorophyll-a concentration and phytoplankton biomass were obtained and results were used in linear regression resulting in strong positive correlation coefficients. Furthermore, in Huot et al., 2007[2], chlorophyll-a concentration was compared to five other measures as an index for primary production ultimately finding that chlorophyll-a concentration provided an equal or more accurate estimate than all others, on the basis of correlation coefficient, root mean square error and mean absolute percent error. Many studies on the relationship between primary production/phytoplankton growth and an assortment of environmental variables have used chlorophyll-a concentration as a proxy, Moore and Abbott, 2002[3], Gong et al., 2003[4], Gameiro et al., 2004[5], all of these case studies examined different bodies of water, phytoplankton growth in said body of water (in terms of chlorophyll a concentration) and the influences upon its variation.

Due to the number of explanatory variables in our data set (18) it was imperative to have a base understanding of the most relevant factors that we could examine. The above-mentioned Moore and Abbott, 2002[3] and Gong et al., 2003[4] both provide insight into some of the most relevant environmental variables that other researchers have observed in the study of variations in chlorophyll-a concentration. Moore and Abbott, 2002[3], examine the seasonal and spatial patterns of chlorophyll-a concentrations in relation to the Polar Front, whereas Gong et al., 2003[4], examines the seasonal variation of chlorophyll-a in relation to the subtropical East China Sea. While both look at vastly different bodies of water, both use satellite imaging of sea surface temperatures and ocean colour data to examine fluctuation in chlorophyll-a

concentration. Both find that light limitations during some seasons were a likely cause of low primary production/phytoplankton growth. In a similar case study by Ji et al., 2018[6], sea surface temperature and ocean colour imaging we used again to explore the relationship between temperature and chlorophyll-a concentration, finding significant positive correlation between surface temperatures and CPHL on the north side of the East China Sea, but found negative correlations in the south, citing low nutrient density in the area as a possible cause.

In terms of possible modelling approaches there exist a vast number of different ideas ranging from multiple linear regression to primitive equation models beyond our purview, both data-driven and model-driven approaches. Due to our restrictions in terms of time and ability require us to seriously consider approaches only accessible and reasonably understandable given the time frame, the relevant literature concerning these are thus explored. In Wu et al., 2014[7], researchers develop both an artificial neural network and multiple linear regression modelling approach in order to predict daily dynamics of chlorophyll-a concentrations, finding good prediction results for both. In Phillips et al., 2008[8], univariate generalised linear modelling was used to analyse the relationship between certain categorical variables, log transformed nutrient levels and log transformed chlorophyll-a concentrations, however they found high degrees of scatter in these relationships, indicating many confounding factors. Additionally, time series modelling as came up as a possible avenue of interest in Vantrepotte and Melin, 2009[9], Jeong et al., 2008[10], where both use more complex time series models, Temporal Autoregressive Recurrent Neural Network and Census I/II in conjunction with SARIMA to model the seasonality of variations in chlorophyll-a concentration. These provide good starting points for further investigation into modelling approaches which can be further expounded upon in later chapters.

Chapter 3

Material and Methods

3.1 Software

We used Python and R as our core software languages.
The data pre-processing was completed with Python through Jupyter notebook due to the convenience of Python modules. It can easily display data visualizations and impute missing values.Our use of R focuses primarily on modelling and diagnosis as R can select and assess models through the use of different libraries.

3.2 Description of the Data

The raw dataset is one 1.16 GB CSV file. There are 3,123,117 observations with 57 variables, including the dependent variable CPHL.The dataset records values on seven distinct gliders distributed across Australia from 2013 to 2015. However, a large number of data, around 19%, is missing, spread unevenly across the different variables. Take one feature, VCUR, which is the value of seawater velocity northward as an example, a significant amount (97% overall) is missing.

3.3 Pre-processing Steps

3.3.1 Import libraries and explore the raw data

First of all, we visualize the raw data by plotting different diagrams; missing value tables, dendrogram, correlation diagram, histogram, distribution mapping and time-series. These methods give insights into the distribution on each glider and feature. We then filter data within the valid range to generate the legal dataset according to the standard variable instance table[11]. For individual values, whose missing percentage is significantly high at more than 90% (e.g. UCUR), ignoring the entire feature is an option after comparing the variable's representation and diagnosing the missing data mechanism. Nevertheless, not all features can be deleted straightforward. Since the raw dataset is enormous and contains many missing values that are distributed unequally over different glider missions, splitting the dataset by glider mission allows us to have a better understanding of the missing values for the later imputation process.

3.3.2 Analyze features by features on each glider

After the spliting of raw data into seven subsets based upon seven different glider missions, we start to analyze each feature one by one. Depending on the corresponding quality control types defined by the raw dataset which are eight categories in

total, we first keep all values classified as good data and then calculate the percentage of the remaining categories, such as bad or missing data. These data point would either be dropped or imputed according to the importance and the missing percentage.

3.3.3 Missing value imputation

For the missing data that is expected to be imputed, we plot the time-series for the variable against the remaining features, finding the possible correlation and diagnosing the missing mechanisms. After comparing different imputations methods, including KNN, mean and median, KNN was selected to fill in the table as it gives the highest accuracy on the training model.

3.3.4 Explore data after imputation and compared the differences between each glider

Previous diagrams in step 1 have been plotted again to make sure there is no missing values for each variable. Further, the filled data on separate gliders was analyzed to compare the difference. Finally, we generate CSV files for further modelling steps.

3.4 Data Cleaning

For missing data, we either delete them or impute them based on the percentage and the meaning of the missing data.
It is always better to keep data than to discard it, but if the missing data is limited to a small number of observations (less than 5%)[12], we may delete those cases from the analysis.
If the percentage of missing data is more significant than 5%, but less than 40% and the data is neither missing entirely at random nor missing not at random, we should consider filling the missing data by imputation.
If the data is missing for more than 40% of observations, we should consider deleting the missing values.
If the data is missing for more than 90% of observations, this entire feature could be considered for deletion
For the method of missing value imputation, after comparing the strength and shortcomings of mean imputation, most frequent value imputation, and KNN imputation, we decided to implement KNN imputation. In KNN imputation, the k nearest neighbors algorithm can be used for imputing missing data by finding the k closest neighbors to the observation with missing data and then imputing them based on the non-missing values in the neighbors. The reasoning behind the decision is because we have high dimensional data, and KNN imputation is specifically designed for this situation, and it could also interpret this situation the most. Besides that, KNN is also the most widely used imputation method.

3.5 Assumptions

Since we have over 3,123,117 observations, we assume the independence between each observation, which could support the modeling step.
Since the observations were collected from different gliders by different devices, and those data were trained together during the modelling step, we assume the change of device has no impact on the measurement.
Since there are seven different gliders from different regions, the data collected from different gliders were regarded to be under the same measurement standards. Thus, we assume there is no measurement bias between seven gliders.
Since the observations are from 2013 to 2015 with a 2-year duration, and the time change was regarded to have no impact on the data measurement. Thus, we assume there is no measurement bias due to the time change for each glider.

3.6 Modelling Methods

3.6.1 Generalized Linear Model method (GLM)

A generalized Linear Model is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

And two functions.
A link function that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor$g(\mu_i) = \eta_\phi$
A variance function that describes how the variance, $var(Y_i)$, depends on the mean $var(Y_i) = \phi V(\mu)$, Where the dispersion parameter ϕ is a constant.
Generalized Linear Model is very useful by building a relationship between the predictors and the response via some link functions. However, the Generalized Linear Model can only be applied when the response belongs to the exponential family distribution and the link function describes the relationship between the response and other linear combinations of predictors according to Julian.J[13]. Hence, for trying this method, we must plot the distribution of the response CPHL first to check whether this dataset could be applied to the Generalized Linear Model method. If possible, we will compare, fit, improve this model by analyzing residuals vs fitted diagrams, Q-Q plot, R^2, p-values for each predictor, and ANOVA table.

3.6.2 Variance Inflation Factor method (VIF)

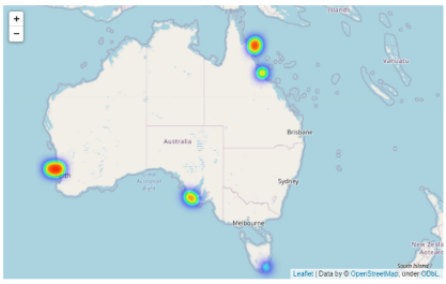
Catalina B states that the Variance Inflation Factor (VIF) is a useful tool to detect whether there is multicollinearity in our models[14]. Multicollinearity exists when there are at least two predictors that have a strong correlation with each other which will reduce the precision of the estimate coefficients and weaken the model's accuracy. The existence of multicollinearity is troubling when VIF is greater than 10. Hence, we need to plot the correlation plot to have a brief understanding of correlation among variables and check the VIF when fitting our model, ensuring VIF values of all predictors in our final model are less than 10.

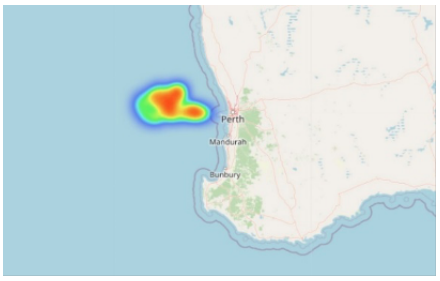
Chapter 4

Exploratory Data Analysis

4.1 Using Python

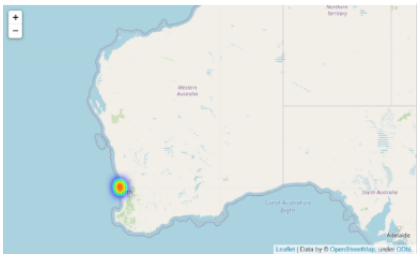
To begin with, we explore the number of unique gliders, totaling seven, and plot the location maps.The unique gliders named are shown in Appendix such as TwoRocks20130215, Leeuwin20131017, AIMS20151127 and TwoRocks20140808.
We firstly visualized the distribution of gliders, acknowledging that they spread across all of Australia (figure 4.1.1). Further, three of them named TwoRocks20130215, TwoRocks20140808, and Leeuwin20131017 are all located around Perth, from figure 4.1.2 to figure 4.1.4.

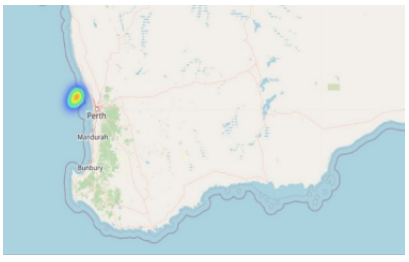




Left:Figure 4.1.1 The distribution of gliders in Australia

Right:Figure 4.1.2 The distribution of gliders of Leeuwin20131017





Left:Figure 4.1.3 The distribution of gliders of TwoRocks20140808

Right:Figure 4.1.4 The distribution of gliders of TwoRocks20130215

Then we started to explore the raw dataset, plotting a heatmap of the correlation of different numerical features as described in figure 4.1.5 and dendrogram 4.1.6 points out the similarity between each variable.

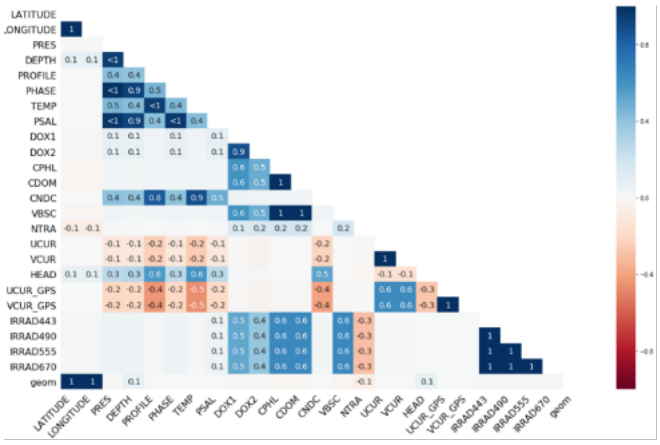


Figure 4.1.5 Correlation between Numerical variables

The above heatmap shows the quantitative correlation between the numerical features. As the coefficient goes higher, block color changes from red to blue. For example, DOX1 has 0.1 relationships with DEPTH, PROFILE, TEMP and VBSC, 0.9 with DOX2, 0.6 with CPHL, CDOM, VBSC and 0.5 with all the IRRAD related variables. It can be concluded that DOX1 has a high correlation with DOX2 and CPHL however, features like DEPTH has little impact on it.

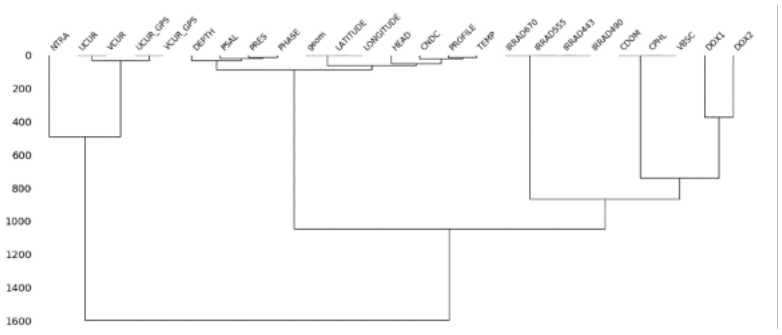
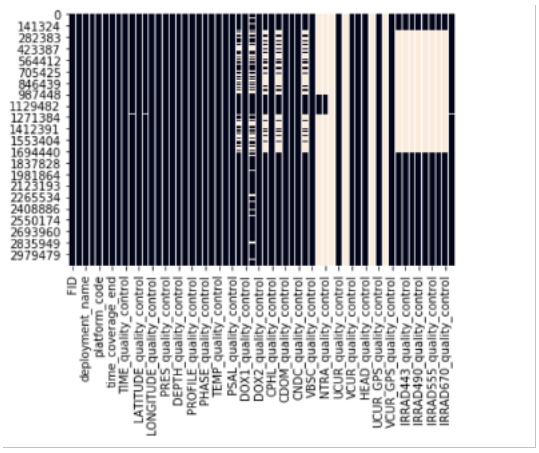
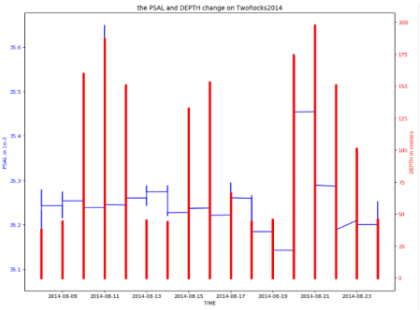


Figure 4.1.6 Dendrogram of Numerical Variables

The dendrogram 4.1.6 diagram indicates the attribute distance between each feature in terms of similarity. The key to interpret this is to focus on the height at which any two objects are joined. In the example above, we can see that UCUR and UCUR_GPS are similar. UCUR records the eastward seawater velocity, and the latter is velocity on the eastward sea surface. However, the raw data includes a significant amount of missing values. According to figure 4.1.7, white vertical stripes represent the missing data. In that case, variables including NTRA, UCUR, VCUR and their related quality controls have a significant number of missing values while the missing data on IRRAD features, like IRRAD443, are concentrated on the first half of raw data. According to variable instance data from AODN[11], we first filter data with the valid range to generate the legal dataset. After selection, the total number of rows has decreased from 3,123,117 to 3,101,188.





Left: Figure 4.1.7 Concentration of Missing Values in Raw Data

Right: Figure 4.1.8 Timeseries on the change of PSAL and DEPTH

For variables consisting of more than 90% missing values, including UCUR, VCVR, UCUR_GPS, VCVR_GPS, and NTRA, we calculate the percentage of missing values against the type of glider and time slot. The results show that the missing values are distributed equally, spread over almost every glider and time. It can be concluded that it is missing at random for the above features.

As shown in the IMOS project by the Australian National Facility for Ocean Gliders [15], UCUR and VCUR are the zonal and meridional components of the depth-integrated current velocity. They can only be calculated when the glider is close to the surface, which explains the frequency of missing data. NTRA, after communicating with the project supervisor, was determined to be a less relevant variable; therefore, we decided to delete the features and their quality control counterparts.

Due to the enormous size of the raw data and the different percentages of missing variables from each glider, it is better to analyze data from gliders located in different areas separately. Thus, we will explore the data sourced from Perth named TwoRocks20140808.

Assessing each feature, we acknowledge that PSAL, which represents the seawater salinity, has five quality control types as defined in the raw dataset, including 0 (No QC performed), 1(Good data), 4 (Bad data), 3(Bad data that are potentially corrected) and 9 (Missing Data).

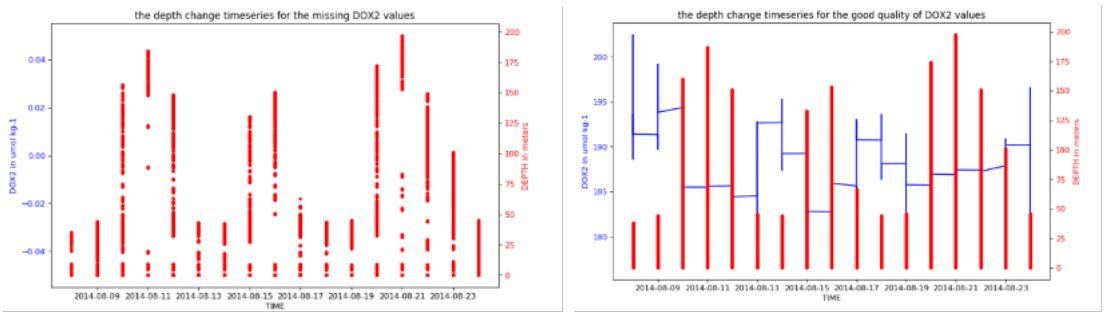
As there also exists independent variable TIME, it is necessary to investigate the relationship between the variables and time. In such a case, we plot the diagram of PSAL against time and depth movement.

It is clear from figure 4.1.8 that the glider TwoRocks2014 dove to it's deepest point at almost 200 meters on 21st Aug 2014, indicated by the right vertical line in red. Seawater salinity, however, distributed evenly between 35.15 to 35.45 in August.

We can observe that the PSAL variables is only missing two data points, a relatively minuscule amount. Therefore, it is a good idea to ignore these invalid rows. Similarly, we can also delete the bad quality corresponding values since the relative percentage of bad data is 0.7%.

However, not all features can be straightforwardly deleted on the missing rows. Take the following variable DOX2, which is the value of moles of oxygen per unit

mass in seawater, for example in result states in Appendix, the number of missing values is 55,056, a significant percentage of the data set with 8.08%. This moderate amount of missing data cannot easily be ignored; therefore, we continue to further imputation methods.
Likewise, we can plot the time-series diagram to describe the variation of DOX2 against the basic properties, including depth of gliders and seawater temperature. After comparing the figures from 4.1.9 to 4.1.10 and time-series of DOX2 against temperature in Appendix, it can be seen that the data is missing during the gliders' diving process. The gap in data is caused by the gliders' movement as depth and temperature changes instead of specific or accidental reasons.



Left: Result 4.1.9 timeseries on the missing DOX2 values against depth

Right: Result 4.1.10 timeseries on the good type DOX2 values against depth

Concerning a conference paper about time series imputation methods [16], k-nearest-neighbors (KNN) imputation, which is to match data points with its closest k neighbours in a multi-dimensional space and replace the non-value obtained from related cases in the whole set of records is an effective and accurate way of dealing with all variety of missing data.

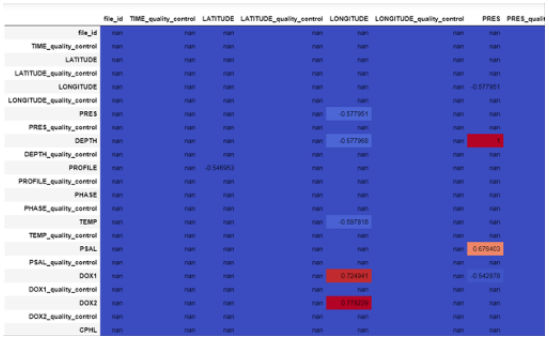
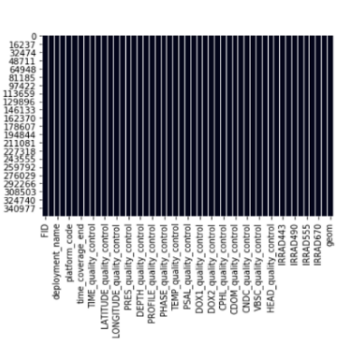
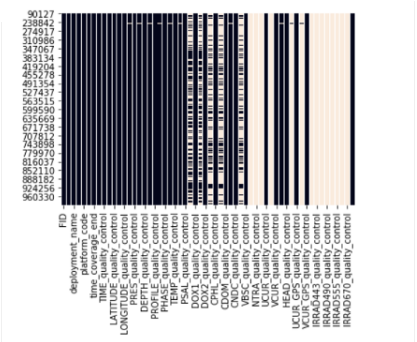


Figure 4.1.11 Partial correlation plot between each variable

The above figure 4.1.11, indicates the correlation values between every feature. High correlation with a coefficient between 0.5 and 1 can be considered further, the dark blue entries represents non or less correlated features. As the coefficient goes higher, the color of the entry turn light blue reflecting 0.5 correlated variables, orange correlation entries reflect between 0.6 and 0.7 coefficients and dark red entries reflect coefficients of 0.7 and above. Variables with coefficients of 0.5+ in addition

to non-missing entries like PRES and DEPTH can be defined as predictable sets and can be used to evaluate the DOX2 missing values as a target set. First splitting the non-missing data into training and testing groups of 80% and 20%, we fit the model with the training set, predict the value of DOX2 and compare with their real values, which gives a significantly high accuracy fitted model at 89.2%. After acknowledging the good-fit model, it can be applied to the set of missing DOX2 values using the same predictable set, giving us the imputation result.
By similar methods, either imputation or deletion, we analyzed all the variables that contained the missing values. We then re-run the heatmap concentration of missing value, figure 4.1.12, finding are no white gaps (which represents missing) anymore.





Left: Figure 4.1.12 Concentration of missing value after data cleaning of TwoRocks2014

Right: Figure 4.1.13 Concentration of missing value of raw Leeuwin20131017 data

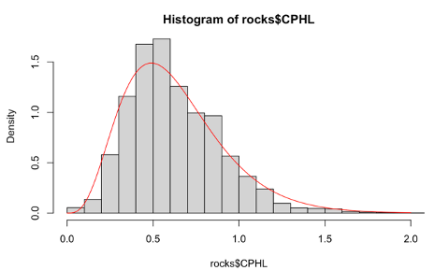
However, not all seven gliders perform precisely the same due to recording differences. Take the glider named Leeuwin20131017 for example, the data related to IRRAD are 100% missing according to the figure 4.1.13.
For reference, IRRAD is a measurement from the optical sensor of how much light from the sun penetrates the water. It is reasonable there are no records regarding the fields, so we deleted all four IRRAD variables and their corresponding quality controls totaling eight variables.
After processing all the features, there are eight features different between the number in TwoRocks2014 at 47 and in Leeuwin at 39, as a result of the IRRAD variables. Overall, all the glider missions can be classified into the above two categories. One, after exploring and processing the raw data, the majority of the features are kept like TwoRocks2014. And two, similar to the Leeuwin glider mission without all the essential IRRAD variables in the processed table. In that case, our modelling process will focus on the above two types.

4.2 Using R

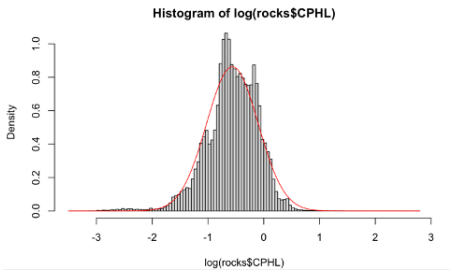
4.2.1 TwoRocks data

We create a brief overview by generating a summary table and find that there are about 47 variables with 21 quality control variables. All quality control variables are

1 and it can be directly concluded that they have no bearing on our modelling process. Hence, we decide to drop quality control variables, producing a new dataset. As seen in the density distribution of the response CPHL, figure 4.2.1, the response is positively distributed and has a right skewness. Therefore, we consider the use of gamma distribution. Based on the diagram we see it fits the density plot well. The red curve indicates the density curve if we assume the response is gamma distributed. We also notice that when we take the logarithm of the response, the density curve is more likely to have a gaussian distribution as shown in figure 4.2.2. Hence, we need to explore gamma family and transformation of response in gaussian family in our generalized linear models.

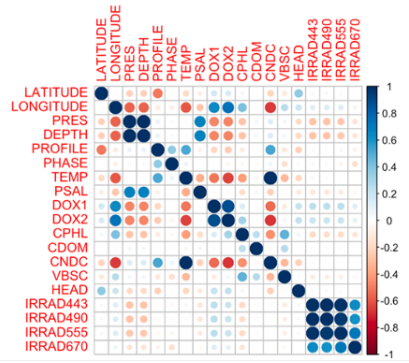


Left:Figure 4.2.1 Response with gamma distribution

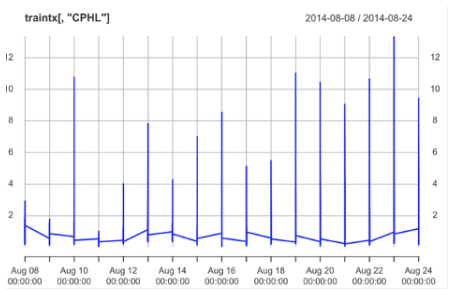


Right:Figure 4.2.2 Logarithm of response with gaussian distribution

By plotting the correlation diagram in figure 4.2.3, we noticed that the response CPHL has tiny correlation with variables such as LATITUDE, PHASE and PROFILE. There are strong correlations pairwise between DEPTH and PHASE as well as the IRRAD group. We also notice that there is a timeline in this dataset. From figure 4.2.4, it clearly shows that this dataset only contains 16 days of information. There is no clear relation in this timeline. So, it is safe to drop this variable.



Left:Figure 4.2.3 Correlation diagram for TwoRocks dataset

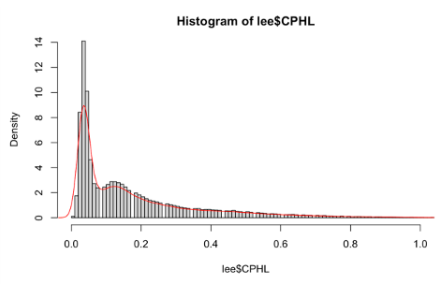


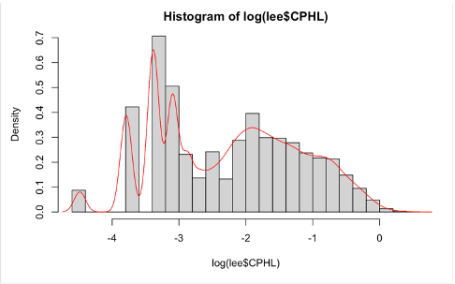
Right:Figure 4.2.4 Timeline for TwoRocks dataset

4.2.2 Leeuwin data

Similarly, we generate a summary table to observe the basic dataset properties. There are 39 variables containing 16 quality control variables. We drop all quality

control variables for the same reason mentioned above and obtain a new dataset. We plot the density distribution of response CPHL for the Leeuwin dataset as shown in figure 4.2.5. We find that the response is positively distributed but with a strong right skewness compared to the normal distribution. Therefore, we take the log of response and use link functions to check whether its distribution could become more concentrated and right tail weaker than those before. However, in figure 4.2.6, we can see that the log of response distribution is more concentrated without skewness distribution. It also shows several small crests compared to the first one. Hence, we decided to explore this dataset and fit the model by comparing the response and log of response separately with generalized linear modelling.

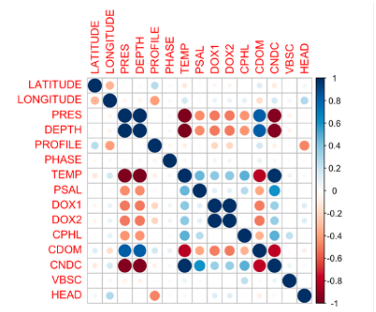


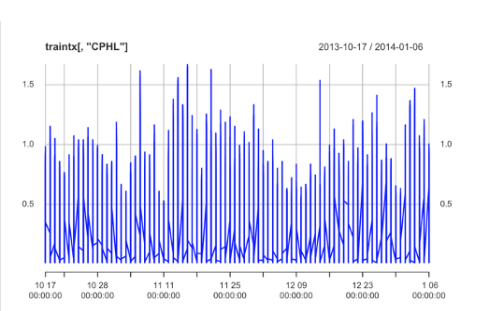


Left:Figure 4.2.5 Response with gamma distribution

Right:Figure 4.2.6 Response take logarithm with gaussian distribution

We then generate the correlation plot of all variables in the Leeuwin dataset as shown in figure 4.2.7 We find that the response CPHL has a small correlation with PRES, DEPTH, and TEMP. We also notice that there is a strong relationship between TEMP and DEPTH, PRES and DEPTH, DOX1 and DOX2, CNDC and DEPTH. These correlations may affect the collinearity of models and needs to be examined during the modelling process.
By plotting this dataset in the timeline as shown in figure 4.2.8, we can see that it only contains data information in two months and a half, it is too short to analyze this dataset in a timeline such as months, seasons, or even years. Moreover, it is obvious that there are no periodical changes for CPHL among times. Hence, we decided to drop the time variables for these data sets.





Left:Figure 4.2.7 Correlation diagram for Leeuwin

Right:Figure 4.2.8 Timeline for Leeuwin

Chapter 5

Analysis and Results

5.1 TwoRocks model

As discussed in part 4.1.1, we observed that the distribution of the response CPHL, is positive and contains a right skewness. We found that the response is either gamma distributed or Gaussian distributed with a logarithm of response. Given this assumption from the data exploration section, it is reasonable to apply a generalised linear model. The models will be produced in R. We will split the whole dataset into 80% for training and 20% for testing.
Note the assumptions for Generalised Linear Model (GLM) are listed below:
1. The data response variable is independently distributed. Errors need to be independently distributed.
2. The distribution of the response variable is from an exponential family
3. GLM assumes a linear relationship between the transformed response in terms of the link function and the explanatory variables

5.1.1 GLM with gamma distribution

GLM with gamma distribution with log link is fitted to the training data to generate the full model. Appendix The summary output suggests that all the predictors are considered significant, the p-value is smaller than 5% significant level.
$R^2 = 1 - D/D_0$where D is the model deviance and D_0 represents the null deviance, which measures how much the model has improved by predictors. The R^2 is 0.63 which means this model is capable in explaining 63% of null deviance.

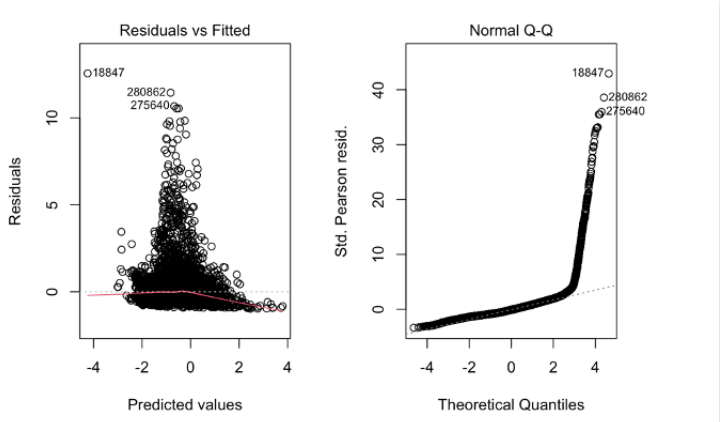


Figure 5.1.1 Residuals diagram and Q-Q plot for Gamma model

From the Residuals vs Fitted diagram (figure 5.1.1), we see that the mean of the residuals is around zero at the range from (-4,0) and decreases after 0. The variance of the residuals is changing especially in the range from (-2,2). There is a peak when the predicted value is at 0.In the Q-Q plot, we see that there is still a heavy right skewness in the diagram, which means that the model may fit the dataset well in the theoretical quantiles from (-4,2).
For all the predictors, the P-value is small, therefore we need to consider multicollinearity. We computed the VIF for each of the predictors in this model to check for the existence of multicollinearity. As shown in the table 5.1.2, there exist 8 predictors with VIF greater than 10 which indicates the existence of multicollinearity.

```
##     LATITUDE    LONGITUDE         PRES        DEPTH      PROFILE        PHASE
## 1.741660e+00 5.751917e+00 1.079367e+08 1.079405e+08 2.985833e+00 1.442119e+00
##         TEMP         PSAL         DOX1         DOX2         CDOM         CNDC
## 5.126674e+04 7.636837e+02 5.582242e+00 7.131016e+00 1.094307e+00 4.792624e+04
##         VBSC         HEAD     IRRAD443     IRRAD490     IRRAD555     IRRAD670
## 1.576649e+00 1.233618e+00 2.525347e+02 1.647447e+02 5.353902e+01 3.532251e+00
```

Table 5.1.2 VIF for full model

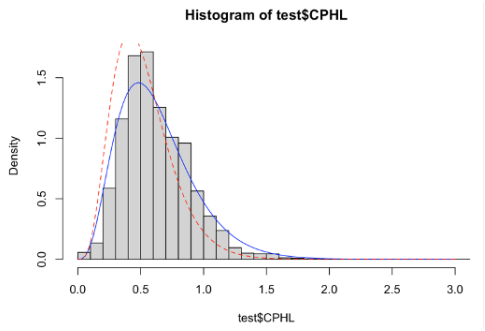


Figure 5.1.3 Fitted value for sub model compare with actual value

We first use all the predictors that have VIF less than 10 and create a new model. In figure 5.1.3, where the red line is the predicted density curve and blue line is the actual density curve for testing data, the red line has higher peak than blue line.
We gradually add each predictor into the model and check the VIF again to ensure there is no multicollinearity.
We first add PRES into the model and all the predictors have small VIF. We then add DEPTH. Both PRES and DEPTH have VIF greater than 10 as shown in Appendix which implies that PRES and DEPTH are highly correlated and therefore, we should only add one predictor into the model.
We use the k-fold-cross-validation method which is a resampling method where we split the dataset into k groups and take the group as a test data set to calculate the prediction error for two models. Since the training model contains around 280,000 observations, we choose k = 10 to impute the value. Both models give us small and close as shown in table 5.1.4, we will choose the one with slightly smaller value. Also, we can easily obtain the value of DEPTH compare to PRES in practice. Hence, we will add DEPTH into our model.

Model with PRES	0.06246883
Model with DEPTH	0.06246632

Table 5.1.4 Cross-validation value for PRES and DEPTH

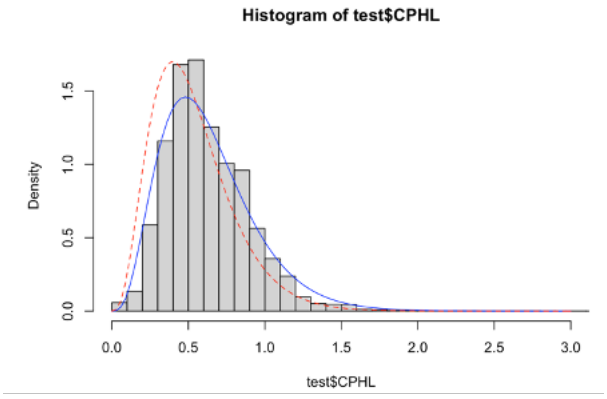


Figure 5.1.5 Fitted value for sub model with depth compare with actual value

After adding DEPTH, figure 5.1.5 clearly shows that the fitted value is getting closer the actual value as red spotted line is getting closer to the actual density.
We apply the same logic and add TEMP and PSAL into our model. However, when we try to put CNDC into our model, the VIF has increased dramatically as shown in the Appendix Based on the correlation diagram in the data exploration part, we see that the CNDC has strong correlation with TEMP. When the model contains CNDC but not TEMP, the VIF tends to be normal which is in Appendix.
Based on the 10-fold cross validation value, the value is similar as shown in table 5.1.6. Yet CNDC means sea water electrical conductivity, which is difficult to measure accurately in practice whereas TEMP is simply the temperature of the sea. Therefore we add TEMP into our model.

Model with CNDC	0.06087
Model with TEMP	0.06111

Table 5.1.6 Cross-validation value for CNDC and TEMP

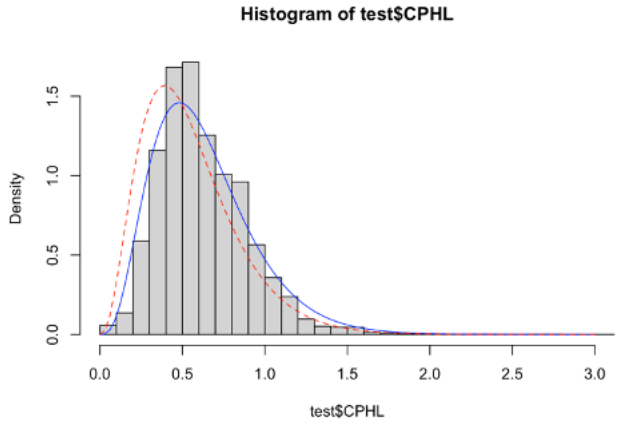


Figure 5.1.7 Fitted value for sub model with temp compare with actual value

Figure 5.1.7 demonstrates that our model is getting better and closer to our aim. Finally, since the IRRAD group is high correlated, we have to decide which IRRAD predictor we should add into our model. By applying the 10-fold cross validation, in table 5.1.8, we see that IRRAD443 has the lowest value.

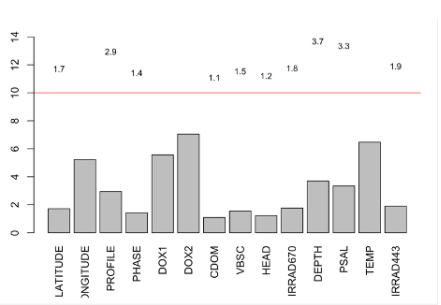
Model with IRRAD443	0.05956
Model with IRRAD555	0.06083
Model with IRRAD490	0.06029

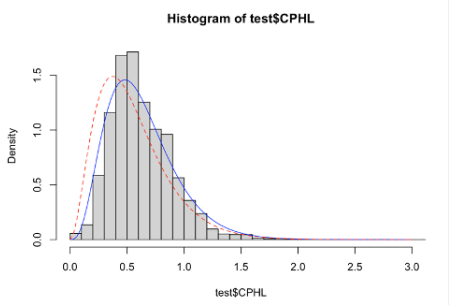
Table 5.1.8 Cross-validation value for IRRAD group

Hence our final model will be

```
glm(CPHL~LATITUDE+LONGITUDE+PROFILE+PHASE+DOX1+DOX2+CDOM+VBSC+HEAD
+IRRAD670+DEPTH+PSAL+TEMP+IRRAD443,family=Gamma(link = "log"))
```

with all the VIF of predictors less than 10, shown in figure 5.1.9, and fitted value density curve close to the actual density plot in figure 5.1.10. The summary table in Appendix shows that all the predictors are significant.





Left: Figure 5.1.9 VIF value for final model

Right: Figure 5.1.10 Fitted value for final model compare with actual value

5.1.2 GLM with gaussian distribution of log of response

As stated in the preliminary analysis, we will apply GLM with gaussian distribution of log of response with identity link to the model. Similar procedure as the previous gamma family, the summary table states Appendix that all predictors are considered significant, the p-value is smaller than 5% significant level.
The R^2 is 0.67 which means this model is capable in explaining 67% of null deviance. From the residual plot in figure 5.1.11, the mean of residuals is around zero, yet the variance of residuals is clearly a non-constant. The Q-Q plot shows the model fails to explain both on the left tail and right tail.

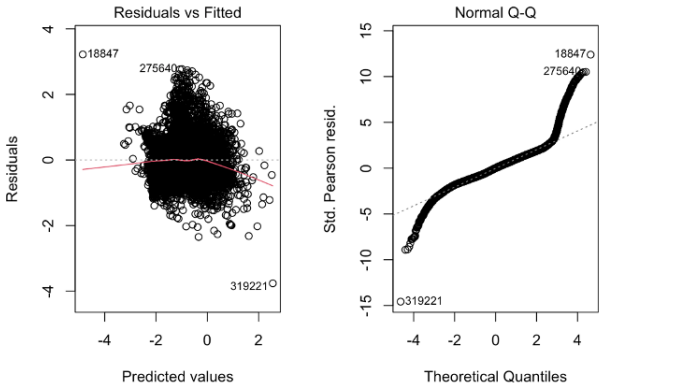


Figure 5.1.11 Residuals diagram and Q-Q plot for Gaussian model

We calculate the VIF for all the predictors as shown in table 5.1.12. There are 8 predictors which contains VIF greater than 10. We repeat the strategy applied before by first gathering all the predictors that have VIF smaller than 10, then adding the rest of them one by one and tracking their VIF.

```
##     LATITUDE    LONGITUDE         PRES        DEPTH      PROFILE        PHASE
## 1.741660e+00 5.751917e+00 1.079367e+08 1.079405e+08 2.985833e+00 1.442119e+00
##         TEMP         PSAL         DOX1         DOX2         CDOM         CNDC
## 5.126674e+04 7.636836e+02 5.582242e+00 7.131016e+00 1.094307e+00 4.792624e+04
##         VBSC         HEAD     IRRAD443     IRRAD490     IRRAD555     IRRAD670
## 1.576649e+00 1.233618e+00 2.525347e+02 1.647447e+02 5.353902e+01 3.532251e+00
```

Table 5.1.12 VIF for full model

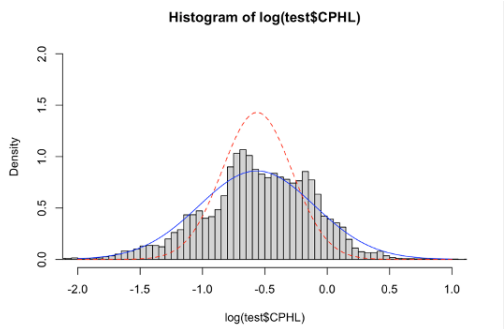
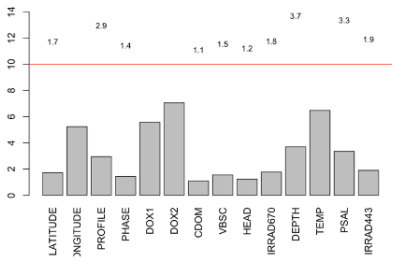
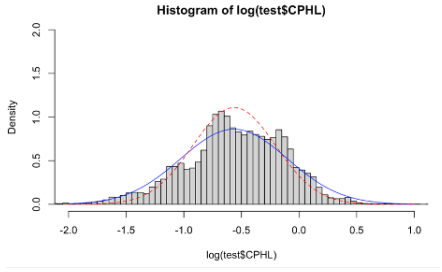


Figure 5.1.13 Fitted value for sub model with all predictors VIF less than 10 compare with actual value

Our first model will be all the predictors that have VIF smaller than 10, the density plot in figure 5.1.13 shows that there are huge difference between fitted value and actual value since the peak is different and the actual value density plot which in blue has a more flattened shape than the predicted density curve, shown in red.
We then need to decide which predictor needs to be added either PRES or DEPTH since they are highly correlated which has been shown in the previous model and in the correlation plot. Based on the 10-fold cross validation results, the values are close as shown in the Appendix, we default to the variable which we can find accurately and easily. Therefore, we will add DEPTH into our model. From the Appendix, the peak of the fitted value is lowering compared to the previous model, moving closer to the actual density cure.
We add TEMP and PSAL into our model since the VIF is smaller than 10. Once we add CNDC in, it will affect PSAL, TEMP and DEPTH. Based on the gamma model and correlation diagram, we need to decide either TEMP or CNDC in our model since they are highly correlated. By conducting the cross-validation value of these two models in Appendix, the values are close. Therefore, we will add TEMP, following the same rationale from the TwoRocks model.
Based on the Appendix, the fitted density curve is getting flatten and the peak value is getting closer to the actual value.
Finally, since the IRRAD443 has the lowest cross-validation value as shown in the Appendix, we decide to add IRRAD443 into our model.





Left: Figure 5.1.14 VIF for final model

Right: Figure 5.1.15 Final model fitted value vs actual value

Hence our final model will be

```
glm(formula = log(CPHL) ~ LATITUDE + LONGITUDE + PROFILE +  PHASE + DOX1+ DOX2 +
+ HEAD + IRRAD670 + DEPTH + TEMP+ PSAL + IRRAD443)
```

with all the VIF of predictors less than 10 which is shown in Figure 5.1.14 and fitted value density close to the actual density plot in figure 5.1.15. .

5.1.3 Model Assessment for TwoRocks model

We split the dataset into training set and testing set. By using the training set to train our model and testing set to calculate the MSE as the criterion for accuracy of the model. This will effectively expose the overfitting of the model. We will also use cross validation to simulate the error.

	RMSE	MSE	Cross Validation
Gamma family with log link	0.54997	0.30246	0.05956
Gaussian family with log of the response	1.28231	1.64433	0.08338

Table 5.1.16 Assessment of TwoRocks 2014 model

As shown in table 5.1.16, RMSE, MSE and cross validation estimate is lower for gamma family with log link. Hence, we can assess the performance of these two models, concluding that the Gamma family with log link performs better.
Furthermore, to assess the accuracy the estimate, we applied a bootstrap of Stepwise Gamma model, since the dataset is huge, we will take the result of 100 bootstrap. The difference of standard error from bootstrap and standard error from summary output did not exceed 15% as shown in Appendix. This implies a good correspondence between bootstrap estimate and standard estimate which shows that our model is not overfitting to the dataset.
Therefore, the final model for TwoRocks dataset is

```
glm(CPHL~ATITUDE+LONGITUDE+PROFILE+PHASE+DOX1+DOX2+CDOM+VBSC+HEAD
+IRRAD670+DEPTH+PSAL+TEMP+IRRAD443,family=Gamma(link = "log"))
```

5.2 Leeuwin model

As discussed in part 4.2.2, we know that the response CPHL, is positive distributed but with heavy right skewness compared to the normal distribution. And we figured out that the response is gaussian distributed with or without logarithm of response. For same reason in TwoRocks model, we would apply generalized linear model by R and splitting the dataset into training and testing dataset using the same percentage from the TwoRocks model. And the assumptions for Generalised Linear Model (GLM) have been mentioned in 5.1

5.2.1 GLM with gaussian distribution and identity link

After fitting the training data into GLM with gaussian distribution and identity link, we obtain the full model with AIC -117941 Appendix. And the summary results show that all parameters are significant, since their p-values are all below the 5% significance level. However, R^2 is 0.32 which means that this model can only explain 32% of null deviance, which is low as we expected.

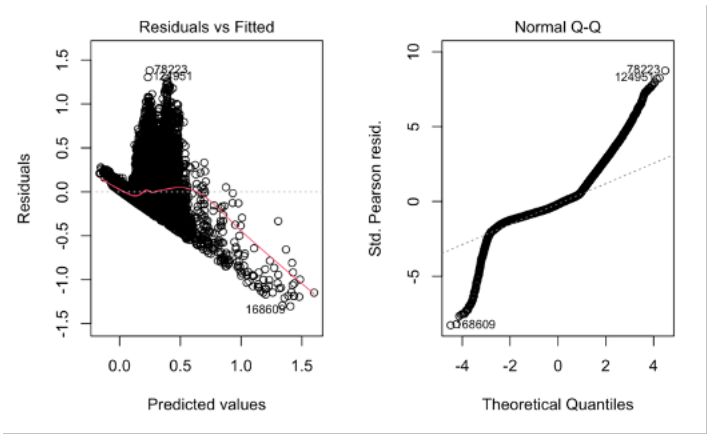


Figure 5.2.1 Residuals diagram and Q-Q plot

Then we plot the Residuals vs Fitted diagram and Q-Q plot of this model (figure 5.2.1). We can see that the mean of residuals is around 0 at the range from (0, 0.5), but extremally goes down after 0.5. And most of the variance of residuals are large. In Q-Q plot we see that it contains two heavy skewness tails both tail, which indicates that this model does not performs well for Leeuwin dataset. Hence, we decided to reject this model directly according to its low R^2, invalid Residuals vs Fitted diagram and Q-Q plot.

5.2.2 GLM with gaussian distribution and log link

As we discussed before, we should try both gaussian distribution with and without log link. Then we fitted the training data into GLM with gaussian distribution and log link, we get a full model with -174411 Appendix. Moreover, we can see that all of parameters with p-value less than 5% level expect TEMP and CNDC. However, TEMP and CNDC both have correlation with response according to correlation plot in 4.2.2. so we cannot drop them. Then we plot the Residuals vs Fitted diagram and Q-Q plot for our model (figure 5.2.2) and R^2 of model is 0.54.

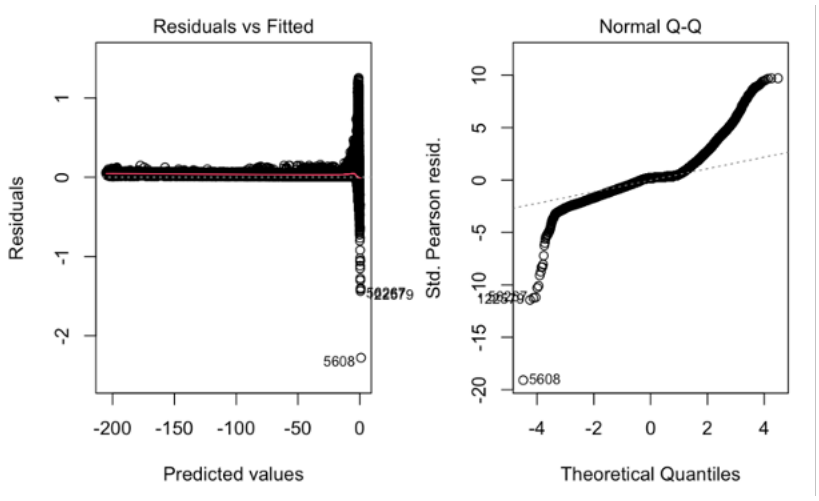


Figure 5.2.2 Residuals diagram and Q-Q plot

It can be seen that the mean of residual is almost zero but variance at 0 is extremally large, which indicates this model does not fits data neither. Considering Q-Q plot , similarly, it has two long tails both sides. Therefore, we reject this model directly neither for same reason with GLM with gaussian distribution and identity link model.

5.2.3 GLM with gaussian distribution of log of response and identity link

Then, we will apply GLM with gaussian distribution of log of response with identity link to Leeuwin dataset. And we obtain the summary table Appendix which shows that variable CNDC might not be significant, but it has correlation with CPHL according to correlation plot, and correlation plot shows that CPHL has no or tiny correlation with PHASE LATITUDE LONGITUDE and HEAD. So, we decide to drop these four variables, PHASE LATITUDE LONGITUDE and HEAD, to get a new model. Then we use anova table to compare these two models. The result shows that the partial model is more significant than the full model Appendix. So, we choose the partial model.

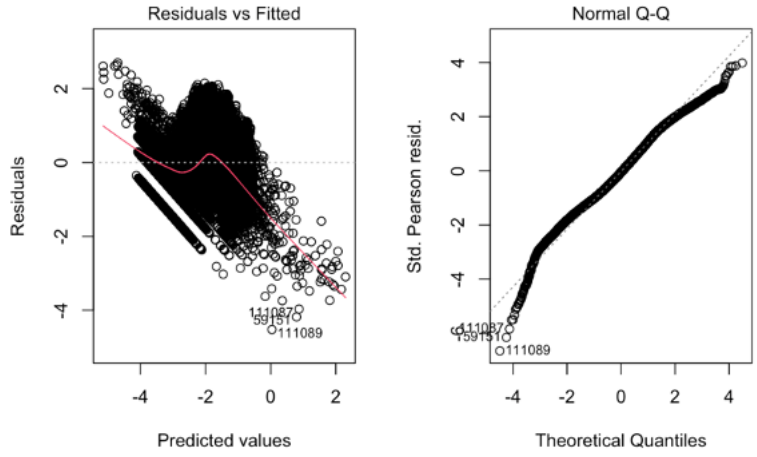


Figure 5.2.3 Residuals diagram and Q-Q plot

The R^2 is 0.59, meaning that this model can explain 59% of data. Although it is not so high, it is the highest one among these four models. The mean of residuals is around zero at the range (-4, 1), but decreasing at the range (1, 0). Obviously, the variance is not constant. However, Q-Q plots performs well without long tails (figure 5.2.3). For the same reason stated in 5.1.1, we calculated VIF for each variable in this model to check whether this model conations multicollinearity. It can be figured out that the VIF of 7 predictors is greater than 10, which indicates that multicollinearity do exist among them (figure 5.2.5).
Therefore, we use the predictors shown on Appendix with VIF less than 10 to create a new GLM with gaussian distribution of log of response and identity link model. For different colors of line in figure 5.2.4, the red one is predicted density curve and the blue one is the density curve of the histogram.

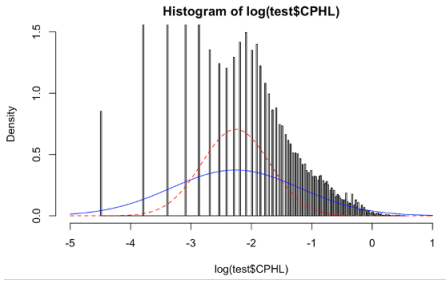


Figure 5.2.4 Fitted value for sub model compares with actual value

We use the same approach as processed before in chapter 5.1.1 to check and ensure the VIF always be less than 10 by adding predictors gradually Appendix.
Firstly, we add DEPTH into this model, and it can be seen on table in Appendix that all of VIF values are less than 10. Then we add PRES to the previous model. The results are shown on Appendix.
We can see that the VIF of DEPTH and PRES are all greater than 10 on table in Appendix, and referring to correlation plot in 4.2, there do exist correlation between DEPTH and PRES. Hence, we need to choose one of them between DEPTH and PRES. We use the same approach, k-fold-cross-validation, to decide which predictor to be stayed, taking k = 10, the value is shown in Appendix. Hence, we decided to choose DEPTH instead of PRES since its smaller value.

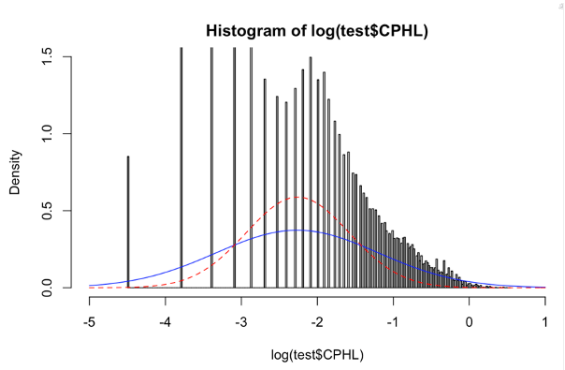


Figure 5.2.5 Fitted value for sub model with depth compares with actual value

Figure 5.2.5 is the histogram after adding DEPTH, it can be seen that the distance between red line and blue has decreased comparing to figure 5.2.4.
Then we add TEMP to previous model and find that VIF of DEPTH and TEMP are all greater than 10 as shown on table in Appendix. Therefore, we use the same method to solve this method. The results in Appendix below shows that we would keep TEMP instead of DEPTH.

The distance between red line and blue has decreased again comparing to figure 5.2.5.

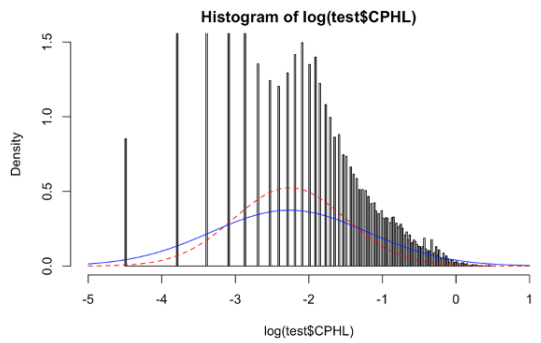


Figure 5.2.6 Fitted value for sub model with TEMP compares with actual value

By using same method, we added PSAL and DOX1 as shown on table in Appendix. However, DOX1 and DOX2 has multicollinearity, which has been shown on Appendix. We keep DOX1 and drop DOX2 according to the table in Appendix. From figure 5.2.7, we can see that the red line and blue line becomes closer than that on figure 2.3.6.

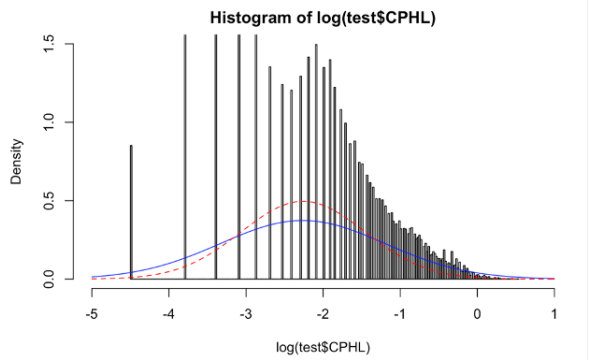
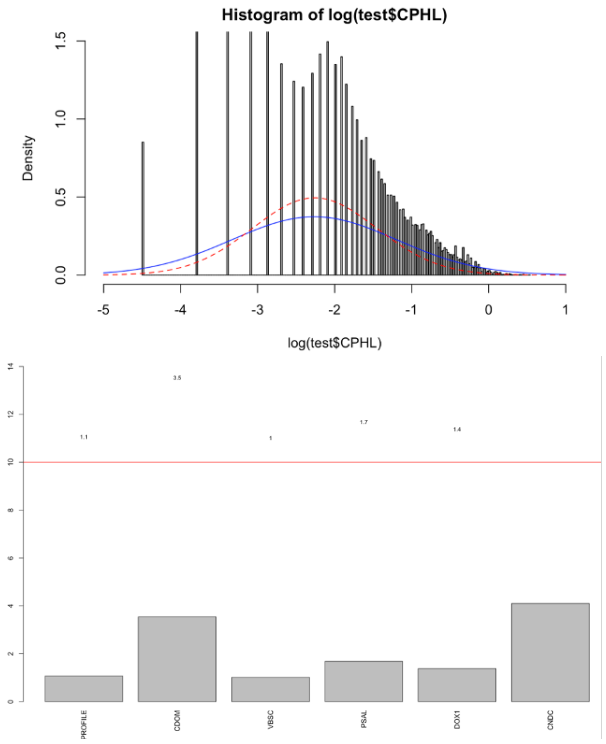


Figure 5.2.7 Fitted value for sub model with TEMP, PSAL and DOX1 compares with actual value

Then we add CNDC to the model and conduct its VIF, the result is shown on Appendix 1.2. We can see that the VIF of both TEMP and CNDC are extremally large and form correlation plot in 4.2, we can also see that there exists correlation between these two variables. We decided to choose CNDC according to the table in Appendix. It can be observed from figure 5.2.8 that these two lines has becomes closer that before, after adding the final predictor.
Hence our final model will be

```
log(CPHL)-PROFILE+DOX1 +CDOM+VBSC+PSAL+CNDC
```

And all of the VIF of these predictors are less than 10 shown on figure 5.2.9.



Up: Figure 5.2.8 Fitted value for final model compares with actual value

Down: Figure 5.2.9 VIF for final model

5.2.4 Model Assessment for Leeuwin model

As mentioned previously, we have split the data randomly into two parts, 90% of are training data and 10% for testing data. We apply test data to this model, we got the RMSE and MAE are 1.57 and 1.44 respectively, which means that our model works well in prediction and we can apply this model for predicting the concentration of phytoplankton in Leeuwin area.

Similarly, we also apply bootstrap of Stepwise gaussian model to obtain the accuracy of estimation by taking the result of 100 bootstrap Appendix. We noticed that the difference between standard error in bootstrap result and standard error in summary table is less than 10%, which implies that our model does not over fitting.

Chapter 6

Discussion

This project aims to find the related features that affect the chlorophyll-a, which can be found in phytoplankton, to model the growth of concentration of phytoplankton (CPHL).

6.1 Discussion for Modelling

We concluded from chapter 5.1 that the final model for TwoRocks data is

```
CPHL~LATITUDE+LONGITUDE+PROFILE+PHASE+DOX1+DOX2+CDOM+VBSC+HEAD
+IRRAD670+DEPTH+PSAL+TEMP+IRRAD443,family=Gamma(link = "log")
```

Since gamma family with log link has a better performance than gaussian family with the log of response, meaning that the predictors like LATITUDE, LONGITUDE, PROFILE, PHASE, DOX1, DOX2, CDOM, VBSC, HEAD, IRRAD670, DEPTH, PSAL, TEMP and IRRAD443 relate to the response with log link. From the summary table in the Appendix, we can determine that all predictors affect the response hugely, including LONGITUDE, PHASE, DOX1, HEAD, IRRAD670, DEPTH, PSAL, TEMP and IRRAD443 all affect the response positively and the rest affect the response negatively.
From the previous chapter 5.2, the final model for Leeuwin data is

```
log(CPHL)~PROFILE+DOX1+CDOM+VBSC+PSAL+CNDC
```

with generalized linear model and identity link, which means that the predictors PROFILE, DOX1, CDOM, VBSC, PSAL and CNDC have relationship with the log of the response. From the summary table in Appendix, VBSC and CNDC affect the response positively and the rest predictors affect the response negatively.
We noticed that there are strong correlations pairwise between DEPTH and PHASE, CNDC and TEMP in both datasets. During the modelling section, we selected one predictor to add to our model as including for instance both DEPTH and PHASE into the model, would cause multicollinearity and weaken our model's accuracy. It indicates that in the further investigation we can record either DEPTH or PHASE and CNDC, or TEMP.
Splitting the dataset according to the glider is required. As in different gliders, data in different variables are missing at random according to the analysis. Therefore, different imputation methods are applied to deal with the missing values. Furthermore, the same model cannot be used in different gliders. As shown in Chapter 5, the log of the response does not perform well in the TwoRocks dataset whereas it gives us the better result in the Leeuwin dataset, and in the TwoRocks dataset, gamma distribution performs well yet in the Leeuwin dataset gaussian distribution

gives us the better performance.

6.2 Discussion for Limitation

Different models have been generated according to each glider since data collected by these gliders contain various missing values and quality of data. Hence, we are required to pre-process data through diverse methods according to the missing value of the original dataset. In this case, we deleted the variables that have a high percentage of missing value which can result in the loss of information. We also used KNN imputation for imputing some missing data, possibly causing bias in the estimation parameters.
The datasets after pre-processing contain different variables. Take the Leeuwin dataset as a example, there are no IRRAD related variables resulting in eight less variables when compared to the TwoRocks dataset.
Further, the limitation of model selection is evident since the GLM method strongly relies on assumptions. It assumes the independence of the variables, whereas in our case, each variable is not entirely independent. By assumption, the response variable is from an exponential distributed family since it is the foundation of the GLM model. However, it does not work the best in this situation since the density diagram doe snot fit perfectly.
Lastly, cross-validation has been chosen and compared in each predictor which is highly time-consuming. Since our dataset is large, it takes around 5 to 10 minutes to run, resulting i long run-times.

6.3 Further improvement

If we were able to work on the project over a long time frame, we would be able to analyse and model glider missions from different areas, as currently both Leeuwin and TwoRocks are located near Perth. We would be able to then apply the models that we have create for Leeuwin and TwoRocks to other glider missions and compare the performance with the final model of that glider. Furthermore, an investigation of the relationship between different locations, the concentration of chlorophyll and the growth of phyotplankton would be possible.
Concerning the residuals vs fitted plot in Chapter 5 where both of the final models show us non-constant variance, if there is was more time, we would have liked to have the chance to understand the existence of the changing variance and try to deal with heteroscedasticity. The final fitted density graph for TwoRocks has a left shift compare to the actual density graph and R^2 value for two models is relatively low which showing us that there are ways still to improve the models.
Just like Desortova did in 1981, exploring the relationship between chlorophyll-a concentration and phytoplankton biomass over a two-year period, we would have liked to investigate datasets from a longer period of time and explore whether the phytoplankton is changing periodically. From the study in Vantrepotte and Melin, 2009, where they use complex time series models to fit the dataset. We can apply the same idea via using different time series models such as ARIMA model and compare the performance with the Generalized Linear Model.

Chapter 7

Conclusion and Further Issues

From the outset of the project our goal was to produce a meaningful model of the growth of oceanic phytoplankton that could help researchers understand the effect of environmental factors upon said growth. Over the course of the project we explored different modeling strategies, informed by the relevant literature we set upon refining the most appropriate models given our particular constraints. Ultimately, we arrived at variations of generalized linear models for both the TwoRocks and Leeuwin data sets as seen in section 6.1.

From these two models we can infer some conclusions about the central interaction, examining the TwoRocks data set model we observe the inclusion of temperature (TEMP) as a positive indicator of phytoplankton growth, in line with much of the similar literature agreeing on the effect of temperature on phytoplankton growth.

Furthermore, from the TwoRocks model we can see the inclusion of multiple IRRAD variables, a measure of radiation intensity (light intensity) as well as the inclusion of VBSC (light penetration) in both the TwoRocks and Leeuwin model. This group of variables encapsulate the effect of light intensity and availability on phytoplankton growth finding positive correlation between the two, this agrees with the academic understanding of light as an energy source for the growth of phytoplankton.

Due to the time limited nature of the project, we were only able to produce models for two glider missions, however, over the course of the current project many interesting questions were uncovered. Avenues for further research include investigations into the effect of water conductivity (CNDC) on chlorophyll concentration, during our modelling process we found that while water conductivity did correlate with the response it often introduced multicollinearity issues. As such research into its effect could be fruitful.

In terms of further modelling, the breadth of missions we modelled over the term were quite limited, both in terms of number and location. Investigations and modelling of mission from the east coast of Australia and comparisons to our current models would be an interesting course of action. Even modelling more glider missions located near the TwoRocks and Leeuwin missions would be helpful in refining our models.

References

[1] B. Desortova, Relationship between chlorophyll-a concentration and phytoplankton biomass in several reservoirs in czechoslovakia, International Review of Hydrobiology 66 (2) (1981) 153–169. doi:10.1002/iroh.19810660202.

[2] Y. Huot, M. Babin, F. Bruyant, C. Grob, M. S. Twardowski, H. Claustre, Does chlorophyll a provide the best index of phytoplankton biomass for primary productivity studies?, Biogeosciences Discussions 4 (2) (2007) 707–745. doi:10.5194/bgd-4-707-2007.

[3] J. K. Moore, M. R. Abbott, Surface chlorophyll concentrations in relation to the antarctic polar front: seasonal and spatial patterns from satellite observations, Journal of Marine Systems 37 (1-3) (2002) 69–86. doi:10.1016/s0924-7963(02)00196-3.

[4] G.-C. Gong, Y.-H. Wen, B.-W. Wang, G.-J. Liu, Seasonal variation of chlorophyll a concentration, primary production and environmental conditions in the subtropical east china sea, Deep Sea Research Part II: Topical Studies in Oceanography 50 (6-7) (2003) 1219–1236. doi:10.1016/s0967-0645(03)00019-5.

[5] C. Gameiro, P. Cartaxana, M. T. Cabrita, V. Brotas, Variability in chlorophyll and phytoplankton composition in an estuarine system, Hydrobiologia 525 (2004) 113–124. doi:10.1023/b:hydr.0000038858.29164.31.

[6] C. Ji, Y. Zhang, Q. Cheng, J. Tsou, T. Jiang, X. S. Liang, Evaluating the impact of sea surface temperature (sst) on spatial distribution of chlorophyll-a concentration in the east china sea, International Journal of Applied Earth Observation and Geoinformation 68 (2018) 252–261. doi:10.1016/j.jag.2018.01.020.

[7] N. Wu, J. Huang, B. Schmalz, N. Fohrer, Modeling daily chlorophyll a dynamics in a german lowland river using artificial neural networks and multiple linear regression approaches, Limnology 15 (2014) 47–56. doi:10.1007/s10201-013-0412-1.

[8] G. Phillips, O.-P. Pietiläinen, L. Carvalho, A. Solimini, A. L. Solheim, A. C. Cardoso, Chlorophyll–nutrient relationships of different lake types using a large european dataset, Aquatic Ecology 42 (2008) 213–226. doi:10.1007/s10452-008-9180-0.

[9] V. Vantrepotte, F. Mélin, Temporal variability of 10-year global seawifs time-series of phytoplankton chlorophyll a concentration, ICES Journal of Marine Science 66 (2009) 1547–1556. doi:10.1093/icesjms/fsp107.

[10] K.-S. Jeong, D.-K. Kim, J.-M. Jung, M.-C. Kim, G.-J. Joo, Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics 211 (2008) 292–300. doi:10.1016/j.ecolmodel.2007.09.029.

[11] Australian Ocean Data Network), OPeNDAP Dataset Access Form (2016).
URL http://thredds.aodn.org.au/thredds/dodsC/IMOS/ANFOG/slocum_glider/Forster20170911/IMOS_ANFOG_BCEOPSTUV_20170911T071056Z_SL287_FV01_timeseries_END-20171002T010328Z.nc.html

[12] J. W. . P. W. Janus Christian Jakobsen, Christian Gluud, When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts, BMC Med Res Methodol (10). doi:https://doi.org/10.1186/s12874-017-0442-1.

[13] J. J. Faraway, Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition, CRC Press LLC, 2016.
URL https://ebookcentral.proquest.com/lib/unsw/detail.action?docID=4711494&query=4711494#

[14] C. B. G.-G. Roman Salmer on Gomez, J. Garcıa-Perez, Overcoming the inconsistences of the variance inflation factor: a redefined vif and a test to detect statistical troubling multicollinearity, Ph.D. thesis (2020).
URL https://arxiv.org/pdf/2005.02245.pdf

[15] Australian National Facility for Ocean Glider, Integrated Marine Observing System (2017).
URL http://imos.org.au/emii.html;http://imos.org.au/anfog.html

[16] Sun, Bin, Ma, Liyao, An Improved k-Nearest Neighbours Method for Traffic Time Series Imputation, 2017 Chinese Automation Congress (CAC), 2017.
URL https://www.researchgate.net/publication/320087317_An_Improved_k-Nearest_Neighbours_Method_for_Traffic_Time_Series_Imputation

Appendix

Codes

See https://github.com/RaymonQ/Ocean-Project-6 and corresponding codes for more details.

Figures

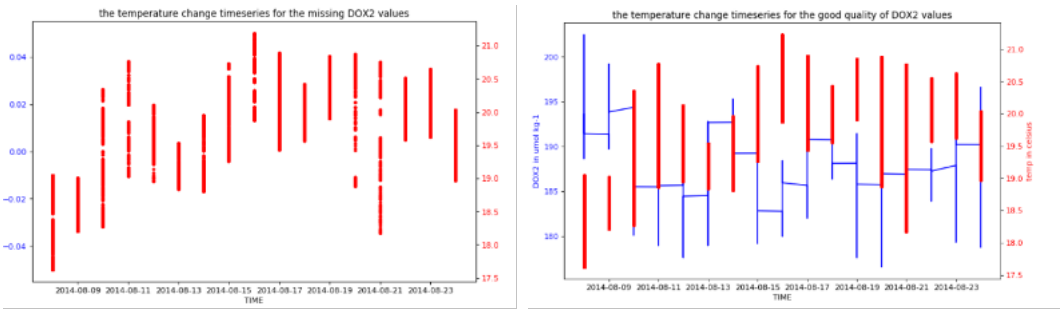
Appendix for Exploratory Data Analysis

```
['TwoRocks20130215' 'Leeuwin20131017' 'AIMS20151127'
 'LizardIsland20131024' 'SpencerGulf20131031' 'TwoRocks20140808'
 'StormBay20141017']
```

The gliders distribution in Australia

```
the quality type of DOX2 are [9. 1. 4.]
the number of bad data for DOX2 8 , and the pecentage is 0.0 %
number of missing value is: 55056 and the pecentage is 8.08 %
```

Number and percentage of bad and missing data for DOX2



Left: Timeseries on the missing DOX2 values Against temperature

Right: Timeseries on the good type DOX2 values Against temperature

```
Call:
glm(formula = CPHL ~ ., family = Gamma(link = "log"), data = train_nt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3222  -0.2153  -0.0318   0.1481   4.4609

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.217e+01  2.675e+00   15.77   <2e-16 ***
LATITUDE     8.017e-02  4.061e-03   19.74   <2e-16 ***
LONGITUDE   -1.082e+00  1.060e-02 -102.05   <2e-16 ***
PRES        -2.199e+01  1.478e-01 -148.79   <2e-16 ***
DEPTH        2.215e+01  1.489e-01  148.79   <2e-16 ***
PROFILE      8.172e-05  6.573e-07  124.33   <2e-16 ***
PHASE       -1.577e-01  3.219e-03  -48.97   <2e-16 ***
TEMP         1.062e+01  1.788e-01   59.37   <2e-16 ***
PSAL         1.106e+01  2.105e-01   52.54   <2e-16 ***
DOX1        -1.912e-02  3.269e-04  -58.50   <2e-16 ***
DOX2         3.526e-02  3.901e-04   90.39   <2e-16 ***
CDOM         1.531e-02  1.860e-04   82.30   <2e-16 ***
CNDC        -1.079e+02  1.757e+00  -61.41   <2e-16 ***
VBSC         1.862e+02  1.127e+00  165.21   <2e-16 ***
HEAD        -2.793e-04  4.693e-06  -59.51   <2e-16 ***
IRRAD443    -8.383e-02  7.928e-04 -105.74   <2e-16 ***
IRRAD490     5.218e-02  5.264e-04   99.13   <2e-16 ***
IRRAD555     9.816e-03  4.568e-04   21.49   <2e-16 ***
IRRAD670     1.427e-02  5.266e-04   27.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.08802653)

    Null deviance: 56863  on 285761  degrees of freedom
Residual deviance: 20530  on 285743  degrees of freedom
AIC: -259634

Number of Fisher Scoring iterations: 8
```

Summary table for full model with Gamma family

```
##
## Call:
## glm(formula = CPHL ~ LATITUDE + LONGITUDE + PROFILE + PHASE +
##     DOX1 + DOX2 + CDOM + VBSC + HEAD + IRRAD670 + DEPTH + PSAL +
##     CNDC + IRRAD443, family = Gamma(link = "log"), data = train_nt
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1241  -0.2362  -0.0358   0.1593   3.6723
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  2.339e+02  1.483e+00  157.669  < 2e-16 ***
## LATITUDE     1.286e-01  4.201e-03   30.606  < 2e-16 ***
## LONGITUDE   -1.119e+00  1.055e-02 -106.080  < 2e-16 ***
## PROFILE      8.625e-05  6.797e-07  126.882  < 2e-16 ***
## PHASE       -1.180e-01  3.339e-03  -35.327  < 2e-16 ***
## DOX1        -2.299e-02  3.399e-04  -67.642  < 2e-16 ***
## DOX2         2.934e-02  4.037e-04   72.687  < 2e-16 ***
## CDOM         1.424e-02  1.936e-04   73.567  < 2e-16 ***
## VBSC         1.825e+02  1.162e+00  156.972  < 2e-16 ***
## HEAD        -3.260e-04  4.867e-06  -66.977  < 2e-16 ***
## IRRAD670    -2.110e-03  3.876e-04   -5.444 5.21e-08 ***
## DEPTH       -3.964e-03  2.853e-05 -138.944  < 2e-16 ***
## PSAL        -2.299e+00  1.345e-02 -170.945  < 2e-16 ***
## CNDC        -4.356e+00  2.056e-02 -211.877  < 2e-16 ***
## IRRAD443    -1.216e-02  7.162e-05 -169.839  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.09542425)
##
##     Null deviance: 56863  on 285761  degrees of freedom
## Residual deviance: 24110  on 285747  degrees of freedom
## AIC: -213114
##
## Number of Fisher Scoring iterations: 7
```

Summary table for final model with Gamma family

```
## Call:
## glm(formula = log(CPHL) ~ ., data = train_nt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7616  -0.1880   0.0016   0.1751   3.2188
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.091e+01  2.376e+00   21.43   <2e-16 ***
## LATITUDE     5.622e-02  3.607e-03   15.59   <2e-16 ***
## LONGITUDE   -1.032e+00  9.413e-03 -109.60   <2e-16 ***
## PRES        -2.344e+01  1.313e-01 -178.58   <2e-16 ***
## DEPTH        2.361e+01  1.322e-01  178.57   <2e-16 ***
## PROFILE      8.305e-05  5.838e-07  142.25   <2e-16 ***
## PHASE       -1.630e-01  2.859e-03  -57.01   <2e-16 ***
## TEMP         1.011e+01  1.589e-01   63.62   <2e-16 ***
## PSAL         1.028e+01  1.870e-01   54.97   <2e-16 ***
## DOX1        -1.896e-02  2.904e-04  -65.29   <2e-16 ***
## DOX2         3.201e-02  3.465e-04   92.38   <2e-16 ***
## CDOM         9.814e-03  1.652e-04   59.40   <2e-16 ***
## CNDC        -1.031e+02  1.560e+00  -66.07   <2e-16 ***
## VBSC         1.538e+02  1.001e+00  153.63   <2e-16 ***
## HEAD        -2.662e-04  4.168e-06  -63.86   <2e-16 ***
## IRRAD443    -8.726e-02  7.042e-04 -123.92   <2e-16 ***
## IRRAD490     5.367e-02  4.676e-04  114.78   <2e-16 ***
## IRRAD555     1.243e-02  4.058e-04   30.64   <2e-16 ***
## IRRAD670     8.989e-03  4.678e-04   19.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.06944416)
##
##     Null deviance: 60778  on 285761  degrees of freedom
## Residual deviance: 19843  on 285743  degrees of freedom
## AIC: 48785
##
## Number of Fisher Scoring iterations: 2
```

Summary table for full model with Gaussian family

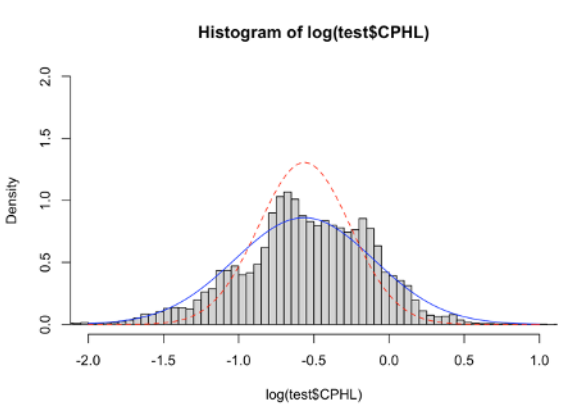
```
## Call:
## glm(formula = log(CPHL) ~ LATITUDE + LONGITUDE + PROFILE + PHASE +
##     DOX1 + DOX2 + CDOM + VBSC + HEAD + IRRAD670 + PRES + CNDC +
##     PSAL + IRRAD443, data = train_nt)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42281  -0.20302   0.00375   0.19359   2.50385
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.472e+02  1.386e+00  178.32   <2e-16 ***
## LATITUDE     1.124e-01  3.926e-03   28.64   <2e-16 ***
## LONGITUDE   -1.132e+00  9.858e-03 -114.80   <2e-16 ***
## PROFILE      9.321e-05  6.354e-07  146.71   <2e-16 ***
## PHASE       -1.358e-01  3.121e-03  -43.52   <2e-16 ***
## DOX1        -2.124e-02  3.177e-04  -66.87   <2e-16 ***
## DOX2         2.525e-02  3.773e-04   66.92   <2e-16 ***
## CDOM         9.343e-03  1.809e-04   51.64   <2e-16 ***
## VBSC         1.587e+02  1.086e+00  146.11   <2e-16 ***
## HEAD        -3.242e-04  4.550e-06  -71.27   <2e-16 ***
## IRRAD670    -3.755e-03  3.623e-04  -10.36   <2e-16 ***
## PRES        -3.839e-03  2.647e-05 -145.04   <2e-16 ***
## CNDC        -4.746e+00  1.922e-02 -247.01   <2e-16 ***
## PSAL        -2.585e+00  1.257e-02 -205.56   <2e-16 ***
## IRRAD443    -1.236e-02  6.695e-05 -184.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08336813)
##
##     Null deviance: 60778  on 285761  degrees of freedom
## Residual deviance: 23822  on 285747  degrees of freedom
## AIC: 101002
##
## Number of Fisher Scoring iterations: 2
```

Summary table for final model with Gaussian family

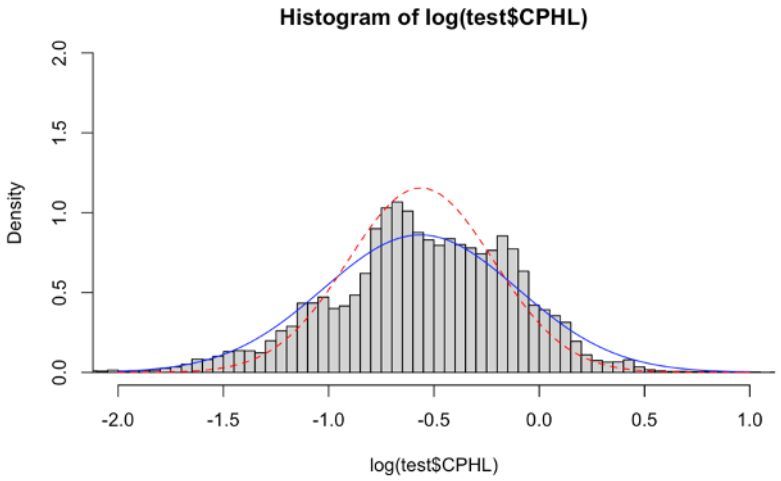
```
Call:
boot(data = train_nt, statistic = boot.gam_pit, R = 100)


Bootstrap Statistics :
          original        bias     std. error
t1*   2.385677e+02  8.560938e-02 1.884788e+00
t2*   1.299703e-01  2.221257e-04 4.964518e-03
t3*  -1.107120e+00 -1.264786e-04 1.248076e-02
t4*   8.604428e-05  2.544127e-08 8.799711e-07
t5*  -1.168333e-01 -8.229374e-04 4.059809e-03
t6*  -2.306732e-02 -5.353556e-05 4.731767e-04
t7*   2.935008e-02  2.073285e-05 5.617572e-04
t8*   1.425113e-02  9.621839e-05 7.918144e-04
t9*   1.822851e+02 -2.669987e-01 3.251797e+00
t10* -3.262128e-04 -9.800868e-07 5.462849e-06
t11* -2.132753e-03 -1.858750e-04 6.923012e-04
t12* -4.150712e-03 -1.628227e-06 3.078622e-05
t13* -2.814099e+00 -1.405088e-03 2.198212e-02
t14* -4.419950e-01 -2.299622e-04 2.866260e-03
t15* -1.215947e-02  1.685249e-05 7.059697e-05
```

Bootstrap for final model



Fitted value for sub model with depth compare with actual value



Fitted value for sub model with temp compare with actual value

```
Call:
glm(formula = CPHL ~ ., data = train_nt)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.30315  -0.09919  -0.03083   0.04566   1.38106

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.031e+00  2.792e-01 -18.019  < 2e-16 ***
LATITUDE     2.471e-02  2.553e-03   9.676  < 2e-16 ***
LONGITUDE    3.186e-02  1.289e-03  24.718  < 2e-16 ***
PRES         1.269e-01  6.543e-03  19.398  < 2e-16 ***
DEPTH       -1.283e-01  6.604e-03 -19.430  < 2e-16 ***
PROFILE      2.370e-05  2.387e-06   9.929  < 2e-16 ***
PHASE       -1.456e-03  2.994e-04  -4.862 1.16e-06 ***
TEMP        -1.556e-02  5.095e-03  -3.053  0.00227 **
PSAL         6.374e-02  9.458e-03   6.739 1.60e-11 ***
DOX1        -3.559e-01  3.252e-02 -10.946  < 2e-16 ***
DOX2         3.638e-01  3.338e-02  10.899  < 2e-16 ***
CDOM         2.275e-02  1.926e-03  11.817  < 2e-16 ***
CNDC         1.103e-01  5.242e-02   2.104  0.03537 *
VBSC         1.502e+03  1.517e+01  98.961  < 2e-16 ***
HEAD         3.092e-05  6.347e-06   4.871 1.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02498948)

    Null deviance: 5182.2  on 138541  degrees of freedom
Residual deviance: 3461.7  on 138527  degrees of freedom
AIC: -117941

Number of Fisher Scoring iterations: 2
```

Summary table for full model with Gaussian distribution

```
Call:
glm(formula = CPHL ~ ., family = gaussian(link = "log"), data = train_nt)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.27659  -0.05069   0.02260   0.04520   1.25183

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  8.199e+00  6.087e+00    1.347 0.177982
LATITUDE    -4.188e-01  9.949e-03  -42.097  < 2e-16 ***
LONGITUDE    9.075e-02  5.099e-03   17.797  < 2e-16 ***
PRES        -9.351e+01  6.007e-01 -155.673  < 2e-16 ***
DEPTH        9.420e+01  6.050e-01  155.687  < 2e-16 ***
PROFILE     -2.032e-04  1.058e-05  -19.202  < 2e-16 ***
PHASE        4.419e-03  1.098e-03    4.024 5.73e-05 ***
TEMP         5.945e-01  4.447e-01    1.337 0.181236
PSAL        -1.810e+00  5.081e-01   -3.562 0.000369 ***
DOX1         1.977e+01  5.173e-01   38.215  < 2e-16 ***
DOX2        -2.028e+01  5.305e-01  -38.223  < 2e-16 ***
CDOM         1.903e-01  7.301e-03   26.068  < 2e-16 ***
CNDC         4.154e+00  4.369e+00    0.951 0.341610
VBSC         1.886e+03  2.001e+01   94.261  < 2e-16 ***
HEAD         1.364e-04  2.362e-05    5.772 7.86e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.01662369)

    Null deviance: 5182.2  on 138541  degrees of freedom
Residual deviance: 2302.9  on 138527  degrees of freedom
AIC: -174411

Number of Fisher Scoring iterations: 15
```

Summary table for full model with Gaussian distribution with log link

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: log(CPHL)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    138541     158940
LATITUDE   1       10    138540     158931 4.515e-06 ***
LONGITUDE  1     2300    138539     156630 < 2.2e-16 ***
PRES       1    58846    138538      97784 < 2.2e-16 ***
DEPTH      1    26519    138537      71265 < 2.2e-16 ***
PROFILE    1      389    138536      70876 < 2.2e-16 ***
PHASE      1       12    138535      70863 2.743e-07 ***
TEMP       1     1376    138534      69487 < 2.2e-16 ***
PSAL       1      713    138533      68774 < 2.2e-16 ***
DOX1       1     1375    138532      67399 < 2.2e-16 ***
DOX2       1     1404    138531      65995 < 2.2e-16 ***
CDOM       1        7    138530      65988 0.0001215 ***
CNDC       1        1    138529      65988 0.2774152
VBSC       1     2116    138528      63872 < 2.2e-16 ***
HEAD       1       18    138527      63853 2.423e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary table for full model of log of response with Gaussian distribution and log link

```
Call:
boot(data = train_nt, statistic = boot.gam_pit, R = 100)


Bootstrap Statistics :
         original        bias     std. error
t1* -6.732033e-01  7.746376e-03 1.998602e-01
t2*  1.088801e-05  1.173312e-06 7.677832e-06
t3*  1.060669e-01  1.153900e-03 7.985080e-03
t4*  4.805510e+03 -2.470250e+00 1.746454e+02
t5* -2.011393e-01 -4.485724e-04 6.104024e-03
t6* -1.813094e-02  6.343550e-06 1.257228e-04
t7*  1.914022e+00  1.177193e-03 8.058367e-03
```

Bootstrap for final model

Tables

Table for TwoRocks2014

```
##     LATITUDE    LONGITUDE      PROFILE        PHASE         DOX1         DOX2
## 1.716249e+00 3.104042e+00 1.796398e+00 1.276264e+00 5.381253e+00 6.209025e+00
##         CDOM         VBSC         HEAD     IRRAD670         PRES        DEPTH
## 1.092950e+00 1.502800e+00 1.209493e+00 1.083788e+00 8.511923e+07 8.512314e+07
```

VIF for sub model contain pressure and depth

Model with PRES	0.11941
Model with DEPTH	0.11940

Cross-validation value for PRES and DEPTH for TwoRocks

```
##     LATITUDE    LONGITUDE      PROFILE        PHASE         DOX1         DOX2
## 1.716249e+00 3.104042e+00 1.796398e+00 1.276264e+00 5.381253e+00 6.209025e+00
##         CDOM         VBSC         HEAD     IRRAD670         PRES        DEPTH
## 1.092950e+00 1.502800e+00 1.209493e+00 1.083788e+00 8.511923e+07 8.512314e+07
```

VIF table for sub model with temp psal and cndc

```
##     LATITUDE    LONGITUDE      PROFILE        PHASE         DOX1         DOX2
##     1.732181     5.632506     2.934021     1.420687     5.557390     7.067269
##         CDOM         VBSC         HEAD     IRRAD670        DEPTH         TEMP
##     1.093629     1.568546     1.225356     1.065166    31.802516 50364.691611
##         PSAL         CNDC
##   747.074111 47094.995703
```

VIF for sub model contain pressure and depth

```
##  LATITUDE LONGITUDE   PROFILE     PHASE      DOX1      DOX2      CDOM      VBSC
##  1.716889  5.254617  2.932393  1.415965  5.547614  7.020983  1.092850  1.546354
##      HEAD  IRRAD670     DEPTH      PSAL      CNDC
##  1.224224  1.065151  3.608991  2.871437  5.961322
```

VIF table for sub model with psal and cndc only

Model with CNDC	0.09334
Model with TEMP	0.09342

cross validation value for CNDC and TEMP for TwoRocks

Model with IRRAD443	0.08338
Model with IRRAD555	0.08491
Model with IRRAD490	0.08501

Cross validation value for IRRAD group for TwoRocks

Table for Leeuwin

```
Analysis of Deviance Table

Model 1: log(CPHL) ~ LATITUDE + LONGITUDE + PRES + DEPTH + PROFILE + PHASE +
    TEMP + PSAL + DOX1 + DOX2 + CDOM + CNDC + VBSC + HEAD
Model 2: log(CPHL) ~ PRES + DEPTH + PROFILE + TEMP + PSAL + DOX1 + DOX2 +
    CDOM + CNDC + VBSC
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    138527      63853
2    138531      64375 -4  -521.52 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA results for two models

```
       PRES        DEPTH      PROFILE         TEMP         PSAL         DOX1         DOX2
1.241681e+07 1.241893e+07 1.339775e+00 3.229619e+03 3.108004e+02 1.355206e+06 1.362629e+06
       CDOM         CNDC         VBSC
3.644228e+00 3.699049e+03 1.016790e+00
```

VIF for full model

```
 PROFILE     CDOM     VBSC
1.032188 1.035022 1.003381
```

predictors with VIF less than 10

```
 PROFILE     CDOM     VBSC    DEPTH
1.089235 3.352770 1.008885 3.264212
```

VIF for sub model contains DEPTH

```
    PROFILE         CDOM         VBSC        DEPTH         PRES
1.091594e+00 3.356594e+00 1.015945e+00 2.763904e+06 2.764044e+06
```

VIF for sub model contains DEPTH and PRES

Model with PRES	0.6881551
Model with DEPTH	0.6878106

CV results for PRES and DEPTH for Leeuwin

PROFILE	CDOM	VBSC	DEPTH	TEMP
1.149645	3.401647	1.011684	15.135472	14.316614

VIF for sub model contains DEPTH and TEMP

Model with TEMP	0.5426375
Model with DEPTH	0.5692239

CV results for TEMP and DEPTH for Leeuwin

PROFILE	CDOM	VBSC	TEMP	PSAL
1.047973	3.140362	1.011647	3.347088	1.263301

VIF for sub model contains PSAL and TEMP

PROFILE	CDOM	VBSC	TEMP	PSAL	DOX1
1.066580	3.552669	1.012535	3.351175	1.269368	1.370803

VIF for sub model contains PSAL, TEMP and DOX1

PROFILE	CDOM	VBSC	TEMP	PSAL	DOX2	DOX1
1.120744e+00	3.624114e+00	1.014119e+00	1.168088e+02	6.636830e+01	7.039376e+05	6.987554e+05

VIF for sub model contains PSAL, TEMP, DOX1 and DOX2

Model with DOX1	0.5007143
Model with DOX2	0.5008438

CV results for DOX1 and DOX2 for Leeuwin

PROFILE	CDOM	VBSC	TEMP	PSAL	DOX1	CNDC
1.131006	3.559244	1.014025	1647.006595	80.255515	2.358578	2013.991686

VIF for sub model contains PSAL, TEMP, DOX1 and CNDC

Model with TEMP	0.5426375
Model with CNDC	0.5007092

CV results for TEMP and CNDC for Leeuwin