Methodology

Role of Student vs. Mentor

The author of this paper was the student, who was mentored by Dr. Kevin Crowthers. From July 2023 to February 2024, the bulk of the work, including idea generation and attainment, research, and model development, validation, and testing, was performed by the former. The latter was responsible for monitoring project progress, as well as offering advice, particularly on potential software usage for parametric testing.

Equipment and Materials

The primary dataset analyzed in this study was the 2018 World Ocean Database (WOD18) provided by the National Oceanic and Atmospheric Administration (NOAA). Both spatially and temporally, this dataset provides a highly cosmopolitan measurement of numerous environmental parameters, including water temperature, micronutrients, pH, salinity, among many others (Boyer et al., 2018). Access to the files of this dataset was attained through the Registry of Open Data provided by Amazon Web Services (AWS). The files used in this dataset were all updated within the AWS S3 explorer system on 17 October 2023 when obtained. Files were organized by year from 1900 to 2023, with pre-1900 data being referred to as 1800 (Amazon Web Services, 2024). The file of each year was systematically downloaded. Regression models were developed using Python code. The main programming interface used was Google Colaboratory, which, when used, was most recently updated on 8 January 2024, supporting Python 3.10.12 (Google Colaboratory, 2024). An online webpage was used to conduct PCA (Statistics Kingdom, 2017). Across all Python programs developed, the Pandas, Xarray, NumPy, SciPy, SciKitLearn, Seaborn, and MatPlotLib packages were utilized. Additionally, to observe the dimensions of the data files more closely, the Panoply software, provided by the National Aeronautic and Space Administration's (NASA's) Goddard Institute for Space Studies (GISS), was downloaded. The most current version, 5.3.1, was used, released 1 January 2024 (NASA GISS, 2024).

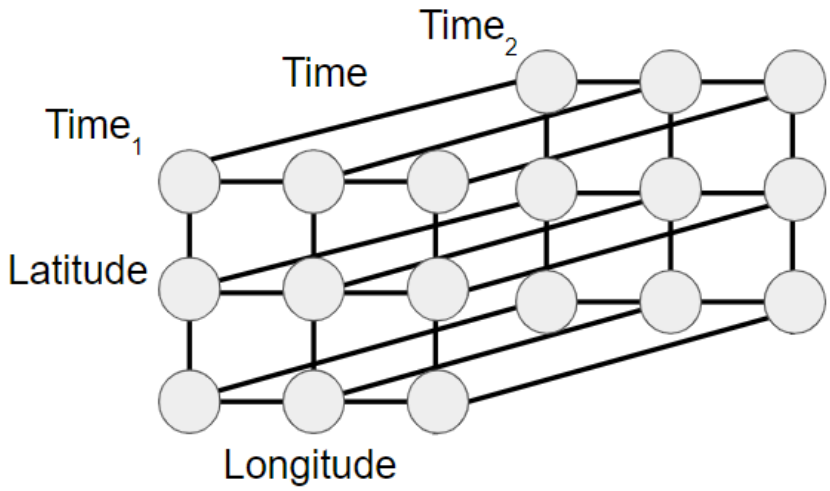
Decisions for Parametric Model Development Based off Data Structure

Two key observations were made during preliminary use and analysis of the dataset that led to two key decisions on the analytical procedure performed. The first observation relates to the data structure of the files used. These were NetCDF (.nc) files, which illustrate parametric measurements at specific points of latitude, longitude, and depth across a series of equal temporal increments (Figure 3). In the metadata attained for the .nc files, these spatiotemporal attributes were referred to as "coordinates." However, it was specifically noted within this directory

that all dimensions, with the exception of the numerical measurements of parametric data, had their latitude, longitude, depth, and time alongside their measurement. Meaning, outside of the year as provided by the file, the spatiotemporal attributes of the primary data on environmental parameters could not be accessed. Therefore, a more holistic approach to analysis was taken, wherein the average global value of each factor was calculated for each year of data. This helped preserve temporal analysis and offered a viable overview of environmental relationships, though at the expense of omitting the nuances and consequences caused by dissimilar spatial attributes among parametric data.

Figure 3

Illustration of the NetCDF Data Structure



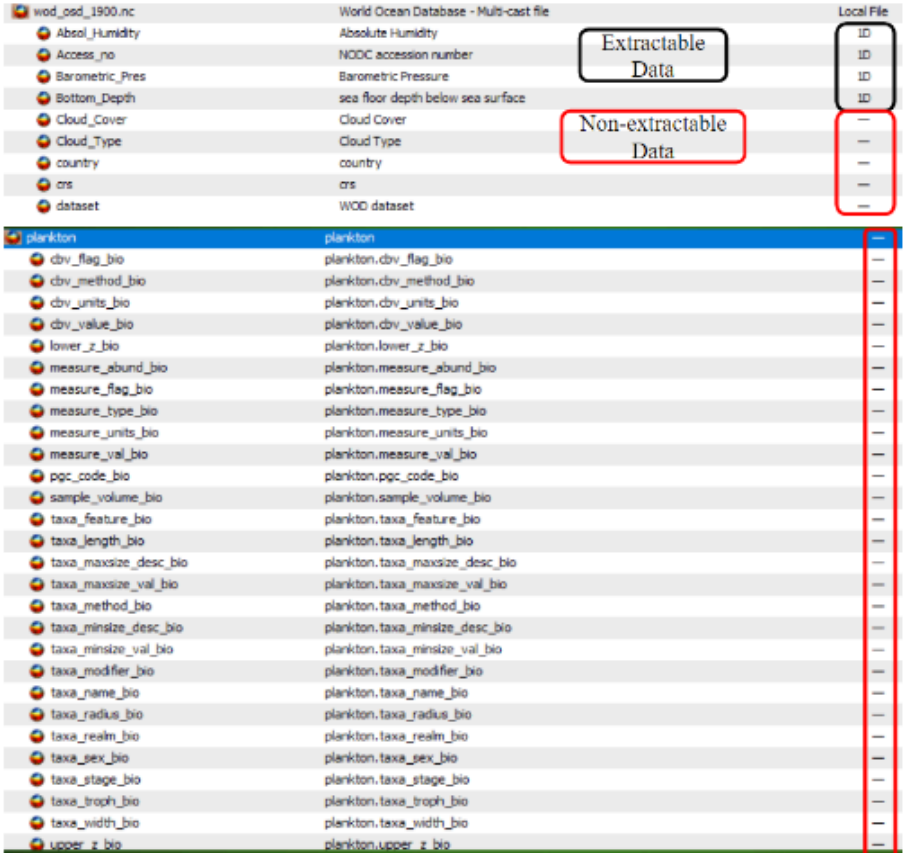
Note. This is a common illustration of the structure of a NetCDF file. The gray circles represent specific points in longitudinal and latitudinal space. Though not illustrated above, each of these coordinates also contains varying levels of depth. At all of these points in space, there exist parametric measurements. These measurements are projected out through a series of time increments, forming a three-dimensional figure.

The second observation made from the metadata was that planktonic data was inaccessible. Within the Panoply software, whereas extractable data was stored within one-dimensional arrays, the values of non-extractable data were unavailable. Figure 4 contains an illustration of this issue. Planktonic data fell under the latter category. Consequently, it was decided to use average global oceanic concentration of total chlorophyll as an indicator of phytoplankton dynamics, namely primary production. This is because chlorophyll is a crucial pigment for carrying out the photosynthetic process, and in turn, all other metabolic processes. As such, a higher concentration of chlorophyll would indicate greater potential for primary production, whereas Though a valuable indicator, it is important to note that it is not a direct measurement of phytoplankton traits. Although data of all available

parameters spanning 1900 to 2019 were downloaded, due to the limited temporal range of measurements for total oceanic chlorophyll, this study focused on data from 1954 to 2017. This also reduced the number of factors assessed.

Figure 4

Inaccessibility of Planktonic Data



Note. The extractable data (black) were stored as one-dimensional arrays, whereas the values of inextricable data could not be obtained. All planktonic data fell under the latter category.

Data Extraction and Cleaning

Using Google Colaboratory, a brief program was written to extract parametric data .nc files and save them as Comma Separated Value (.csv) files. Within each .csv file, the average, standard error, and sample size for each year of each parameter was calculated and compiled into a separate spreadsheet file. Supplementary Files 1 and 2 in the Appendices section (all supplementary files may be found in the Appendices section), provide the exact code used. The main data cleaning involved the removal of non-numerical data within .nc files when converting them to .csv. This was achieved by the Python program written.

However, for total oceanic chlorophyll and alkalinity, the data processing was more complex. When originally creating a time series for the former, it was noted that model strength was inhibited by abnormally high measurements around the early 2000s. Supplementary File 3 (specifically Supplementary Figure 3), provides a visualization of this. Upon further investigation of the .csv files, it was noted that this was due to the abnormally high amount of outliers. In order to maximize model fitness, for chlorophyll data spanning 1998 to 2008, any and all measurements in excess of 20 μg/L were removed, and new averages, standard deviations, and sample sizes were determined. The next iteration of the time series had a stronger fit as a result. For alkalinity, many years had errors in how the data was recorded, in that the decimal place was improperly positioned. This led to values that were orders of magnitude too high for the dataset, and in turn, skewed summary statistics. As such, any such data was eliminated from the set, with summary statistics adjusted accordingly. Besides the processes described, all parametric data from 1954-2017 was preserved when performing data analysis.

Statistical Analysis

A variety of statistical tests and computational tools were used for the three major sets of models developed for this study. The time series for each parameter developed was created primarily using sinusoidal regression. The strength of each regression model was measured using a Pearson's Correlation, including both r and R^2. To assess correlation between each factor and total oceanic chlorophyll, linear regression, in conjunction with a Student's t-test for relationship significance and Pearson's correlation for relationship strength, was used. Finally, driving parameters were identified using PCA, along with supplementary techniques.

Sinusoidal Regression Including Pearson Correlation

Environmental features tend to be periodic in nature. The sine and cosine functions provide an effective way to model cyclical trends. Therefore, this specific type of a regression model was chosen, with R^2 and r measuring model strength and accuracy. This allowed for the evaluation of the validity of the overall computational system as well as projection abilities. Equation 1 represents the template function used for all time series models:

$$f(\kappa) = A\sin((2\pi\gamma)\beta + \varepsilon) + \phi \quad (1)$$

Where $f(\kappa)$ is the function for the total chlorophyll concentration κ, A is the amplitude, $2\pi\gamma$ represents the length of the period (using radians, γ alone being in degrees), β is the given environmental factor (the next section enumerates the variable designation of each parameter), ε is the phase shift, and ϕ is the offset.

Additionally, using the offset as a midline for the sinusoid and the amplitude as a sort of ruler, an interval of all measurements projected by the sinusoid of each parameter was developed. Equation 3 represents the basic construction of the described sinusoidal interval:

$$\phi \pm A \tag{2}$$

Linear Regression Including Pearson Correlation and Student's t-test

For the relationship of every parameter with the indicator, total chlorophyll concentration, a Linear Regression model was developed. On a functional level, assessing each individual parameter's relationship with total chlorophyll acted as a precursor to identifying which of them had a significant influence on chlorophyll when all parameters were considered in tandem. In essence, performing linear regression acted as a prerequisite for then performing PCA. The significance of relationships were determined using a Student's t-test for Linear Regression at $\alpha = 0.05$. Both double- and single- tailed p-values were attained. This was done in conjunction with the use of R^2 and r to measure model accuracy and strength. Similar to the previous set of models, Equation 3 provides a template wherein each parameter's regression model was represented:

$$f(\kappa) = m\beta + \beta_0 \tag{3}$$

Where $f(\kappa)$ is the function for the total chlorophyll concentration κ, m is the predicted slope of the line, β is the given parameter, and β_0 is the y-intercept of the model.

PCA

In order to identify the driving parameters behind total chlorophyll concentrations, PCA was used. PCA is a dimension reduction technique that compresses multiple independent variables into fewer dimensions so as to summarize overarching data patterns and allow for ease of data visualization. Before performing PCA, the data must be standardized so that scale does not impede the accuracy of results. In this study, before PCA was performed, all data of the indicator (chlorophyll) and every parameter were standardized using minimum-maximum normalization. Within the setting of PCA, the total variance of the data is measured. This variance is captured by a finite set of portions of the data known as principal components. In two-dimensional representations, the first two principal components, that is, the two components that account for the highest amount of variance, denoted PC_1 and PC_2, are placed on the horizontal and vertical axes respectively. Since information from the other principal components (PC_3, PC_4, … PC_n) is omitted, it is important that most variance is captured by PC_1 and PC_2. This is measured by each principal component's eigenvector values, which are derived from various matrix operations performed on the data.

Then, to standardize the amount of variation each principal component captures, the eigenvalues are divided by the total variance of the dataset. A scree plot is used to depict the cumulative coverage of variance by all principal components. Along with a PCA plot, a scree plot was used to illustrate these notable properties of the principal components. In two-dimensional PCA, every parameter assessed captures some amount of either principal component, and holds either a positive or negative relationship with the directionality of component variances. This magnitude and directionality is represented by a pair of coordinates that form a vector. The greater the magnitude of variance represented by a parametric vector, the more influential that parameter is relative to the overall data, and in turn, driving the dependent variable. For the study, the magnitude of each parameter was calculated as depicted by Equation 4:

$$M(\beta) = \sqrt{(C_{PC1})^2 \cdot \nu_{PC1} + (C_{PC2})^2 \cdot \nu_{PC2}} \quad (4)$$

Where $M(\beta)$ is the function of the magnitude of parameter β, C_{PC1} is the contribution of the parameter to the variance of PC_1, while C_{PC2} is the contribution of the parameter to the variance of PC_2, and ν_{PC1} is the proportion of the total variance represented by PC_1, and ν_{PC2} is the proportion of the total variance represented by PC_2.

The magnitudes of each parameter were calculated using the above equation, and then subsequently ranked by descending magnitude values. The parameters with the highest calculated magnitude were identified as driving parameters of chlorophyll concentrations. Additionally, to assess the presence of inter-parameter relationships, a covariance matrix was used. A covariance matrix is an intermediate operation performed in the complex matrix calculations involved with PCA wherein all independent variables are arranged in a square array. The cells of this matrix contain the covariance between the row and column parameters. Values vary between 0 and 1, and can be either positive or negative based on the directionality of the relationship. A greater magnitude indicates a stronger relationship between the two parameters. The diagonal cells represent the variance of that individual parameter following dimension reduction procedures. The results from these three sets of computational models were then put into biogeochemical, ecological, and climatic context.