Surface solar radiation and lake productivity: Investigating a global relationship

by

Jane Ye

20652790

Coauthors:

Hannah Adams, Steph Slowinski, Bhaleka Persaud, Ishit Ranjan, Rahim Barzegar, Homa Kheyrollah Pour, Philippe van Cappellen

Earth 436B Undergraduate Thesis

April 14, 2021

Bachelor of Science in Environmental Science (Geoscience Specialization)

Supervised by Homa Kheyrollah Pour and Philippe Van Cappellen

University of Waterloo

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Jane Ye conducted data collection, processing, and exploratory analysis of *in situ* SSR data and processing of satellite data. Hannah Adams conducted data collection, processing and exploratory analysis of *in situ* chlorophyll-*a* data and processing of satellite data. Ishit Ranjan extracted satellite data and performed pre-processing and validation. Rahim Barzegar provided the code template for the random forest model and helped with troubleshooting. Jane Ye selected, processed, and compiled the input data for the random forest model, ran and troubleshooted the model, and assessed model outputs. Bhaleka Persaud, Steph Slowinski, Homa Kheyrollah Pour, and Philippe van Cappellen offered help, advice and guidance throughout the timeline of the project.

Abstract

In recent decades, the intensity and frequency of lake algal blooms have been increasing worldwide. In addition to potentially toxic effects for humans and wildlife, intense algal blooms negatively impact recreation and economy. Therefore, it is crucial to understand the underlying mechanisms controlling the blooms. Most algal bloom management programs focus on limiting nutrient input; however, other controlling factors such as solar radiation are often not considered in practice. This study examined time series of surface solar radiation (SSR), chlorophyll-*a*, temperature and other factors controlling algal growth since the 1990's, using a combination of *in situ* and satellite data. A random forest regression model was used to qualitatively investigate the importance of different controlling factors on chlorophyll-*a* rates of increase during algal blooms. Results of the modelling support that temperature and SSR – both during and immediately before periods of rapid growth – were important predictive factors in seasonal chlorophyll-*a* rates overall. This study joins recent literature in successfully demonstrating the feasibility of using satellite data for global scale lake monitoring, using a widely applicable supervised machine learning tool. The results of this study, and further research taking advantage of satellite data for lake monitoring, will increase our understanding of factors controlling algal bloom intensification and improve our ability to evaluate best management practices.

Acknowledgements

Thank you to Cole Warling for additional assistance with satellite data extraction.

Thank you to Kaden McCulloch, Steph Slowinski, and Bhaleka Persaud for their thoughtful and constructive feedback on the draft of this document.

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund Global Water Futures Project. Thanks also to the University of Waterloo for their generosity in funding this project.

Thank you to the many organizations that provided free and easily accessible data for use in this project, including but not limited to: Environment and Climate Change Canada, Alberta Agriculture and Forestry, UK Environment Agency, Baseline Surface Radiation Network, Global Energy Balance Archive, IISD Experimental Lakes Area, and UK Centre for Ecology and Hydrology.

Finally, I would like to thank all of my coauthors and supervisors (Dr. Homa Kheyrollah Pour and Dr. Philippe van Cappellen) for their support, advice, and patience. Most of all I would like to thank them for taking me on as a thesis student and always believing in me.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

In recent decades, algal blooms in lakes worldwide have been increasing in intensity, duration, and frequency (Ho et al., 2019; Kudela et al., 2015). One accepted criterion for measuring the intensity of algal blooms is the measurement of chlorophyll-*a* concentration in lake waters (Huot et al., 2007). Chlorophyll-*a* is the molecule used by photosynthetic primary producers to harness incoming light energy for carbon fixation (Melkozernov & Blankenship, 2007). Chlorophyll-*a* concentration, which is measurable *in situ* by fluorescent probes or remotely by optical satellites, is commonly used as a proxy for primary productivity in aquatic ecosystems (Ho et al., 2019; Huot et al., 2007).

Primary productivity is the net rate of organic carbon fixation for biomass production conducted by primary producers (Ito, 2011). In most lake ecosystems, approximately half of primary productivity results from photosynthetic algae (Vadeboncoeur et al., 2002). Algal blooms occur when growth of algae occurs so suddenly and explosively that lake waters may become discolored, harmful toxins may be released, and light penetration may become limited (Anderson, 2009). When the bloom is over, the subsequent die-off and aerobic decay of biomass may cause oxygen depletion in the lake (Anderson, 2009). In addition to potentially toxic effects for humans and wildlife, intense algal blooms negatively impact ecosystems, fishing, tourism, property values, and drinking water quality (Anderson, 2009; Kudela et al., 2015; Vadeboncoeur et al., 2002).

The focus of algal bloom management programs is often on nutrient input reduction, due to a well-established connection between anthropogenic eutrophication of lakes and an increase in algal bloom occurrence (Heisler et al., 2008; Kudela et al., 2015). One example, Lake Erie, underwent a phosphorus loading reduction program from the 1960s to 1980s, and was touted as a North American binational success story for algal bloom management through nutrient input reduction (Makarewicz & Bertram, 1991).

However, the contemporary story of Lake Erie is one of re-eutrophication and increasing awareness of multivariate algal bloom drivers (Levy, 2017; Mohamed et al., 2019; Scavia et al., 2014). Other lakes around the world have undergone similar re-eutrophication since the 1990s, such as Lake St. Clair in Canada, Lake Alexandrina in Australia, and Lake Khanka in China (Ho et al., 2019). Studies have shown that although nutrient input is a strong contributor, there are many complex factors which interact to trigger harmful algal blooms (Mohamed et al., 2019; Shuvo et al., 2021). Figure 1, below, summarizes algal bloom drivers considered in this study.

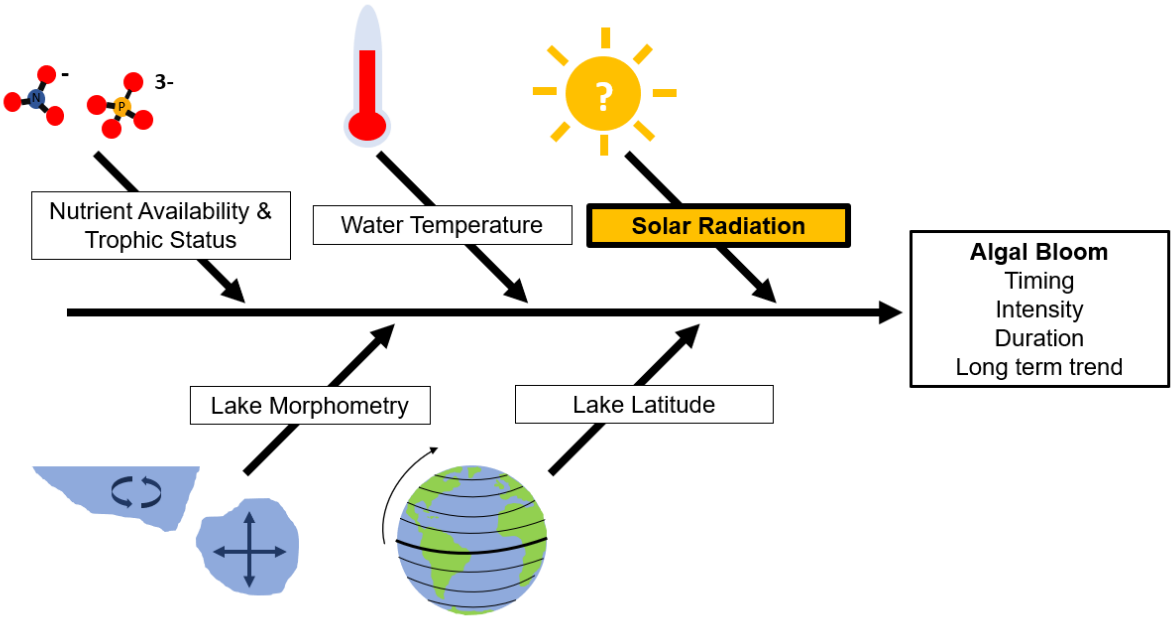

Figure 1: Summary of factors contributing to algal blooms.

One under-addressed factor that contributes to algal growth rates is UV radiation (Inomura et al., 2020). UV radiation, measured at the earth's surface as *surface solar radiation* (SSR), is used directly by photosynthetic organisms for carbon fixation (Melkozernov & Blankenship, 2007). SSR is also a strong control on lake water temperature (Jakkila et al., 2009) and ice breakup timing in seasonally ice-covered lakes (Kirillin et al., 2012), both of which influence primary productivity in lakes. In turn, the global distribution of mean annual SSR is controlled by latitude

(Kirillin et al., 2012) but varies between regions, as well as over time due to atmospheric controls (Alpert & Kishcha, 2008; Cutforth & Judiesch, 2007; Wild, 2009).

Decadal-scale increasing and decreasing trends in SSR have been coined in the literature as solar "brightening" and "dimming", respectively (Wild, 2009). It is understood that these trends are likely caused by changes in atmospheric aerosol concentrations as well as cloud cover, and therefore vary regionally (Cutforth & Judiesch, 2007; Wild, 2009). For example, SSR in the Canadian Prairies has been undergoing a dimming trend (Cutforth & Judiesch, 2007) since the 1990s, while the opposite trend appears to be occurring in most of Europe (Wild, 2009). Globally, urban areas may also be experiencing more dimming than rural areas due to higher aerosol production rates from the burning of fossil fuels (Alpert & Kishcha, 2008). It is important to note that, SSR is not related to or affected directly by the phenomenon of global warming (Kirillin et al., 2012), nor is it controlled by cycles in the sun's energy output (Wild, 2009). Unfortunately, the sparsity or inaccessibility of continuous, long term monitoring networks outside of Western Europe and North America limits much of the study on historical SSR trends to these two regions.

At the lake-basin-scale, a recent study by Tian et al. (2017) showed that SSR is an important predictor of growing-season chlorophyll-*a* concentrations in the Western Basin of Lake Erie. A recent paleolimnological study of Lake Tanganyika fishery productivity, also showed evidence for correlation between multi-centennial oscillations of higher SSR and increased diatom production, dating back to ~1000 CE (McGlue et al., 2020). However, the link between SSR and lake chlorophyll-*a* is less well-understood at a regional or global scale. This study aims to advance our understanding in this area by using satellite data to chlorophyll-*a* in lakes above 40°N.

The comprehensive spatiotemporal coverage of satellite data makes it a useful tool in large scale lake monitoring. Taking advantage of this resource has allowed many studies in a variety of disciplines to investigate global scale phenomena, including increasingly in environmental and climate science fields (examples include Crespo Cuaresma et al., 2017; Ho et al., 2019; Kaufman et al., 2005; Liu et al., 2009; Pilla et al., 2020; Shuvo et al., 2021). However, appropriate tools and techniques must be used to interpret the large amount of data generated from long term satellite monitoring. This study applies one such tool, a machine learning model, to *in situ* and satellite data

to investigate the influence of growing season SSR on chlorophyll-*a* growth rates of temperate Northern Hemisphere lakes. The results of this study increase scientific understanding of the various factors that influence temperate lake algal blooms. This work joins several recent studies demonstrating the feasibility of using satellite data for global scale lake monitoring (examples include Chipman, 2019; Ho et al., 2019; Li & Li, 2004; Philipson et al., 2016; Pilla et al., 2020; Shuvo et al., 2021).

1.2 Hypothesis

The hypothesis behind this research is that SSR is an important factor contributing to seasonal lake algal growth. If all other controlling factors are constant, higher SSR will contribute to higher chlorophyll-*a* growth rates during the spring and fall seasonal periods of peak algal biomass growth (*i.e.*, spring and fall algal blooms). It is also hypothesized that the effect of SSR can be differentiated by trophic status, being stronger in eutrophic > mesotrophic > oligotrophic lakes. This order also reflects the degree to which lake productivity is limited by nutrient availability.

1.3 Research Objective

To investigate relative contributions of SSR, and other environmental factors, on controlling chlorophyll-*a* growth rates during seasonal algal blooms in lakes above 40°N, by applying a machine learning random forest model.

1.4 Research Approach

This study aims to understand the drivers of algal blooms in lakes above 40°N. To account for the variability and interactions between the different controlling variables and their contributions to chlorophyll-*a* growth rates, a random forest regression modeling approach was used. This approach generates a quantitative model that can predict chlorophyll-*a* growth rates as

a function of changes in environmental variables, including SSR, water temperature, trophic status, and lake morphological characteristics. The model and its input variables are described in detail in Section 2.2. The results of the model output for in situ and remotely sensed data are compared.

Many studies have previously demonstrated the role of SSR in controlling algal growth at a variety of spatial scales, including laboratory experiments, individual lake basins, and individual lakes (examples include Deng et al., 2019; Dubourg et al., 2015; Inomura et al., 2020; McGlue et al., 2020; Tian et al., 2017). However, only a few studies have examined the SSR-chlorophyll-a relationship at a global scale (e.g., Shuvo et al., 2021). Many studies that investigate chlorophyll-a simplify time series data by calculating annual mean (Shuvo et al., 2021) or annual maximum (Ho et al., 2019) concentrations. Here, spring and fall growth windows of each year in each lake were isolated and the chlorophyll-a rates of increase were examined (as a proxy for algal growth rate) during these windows.

2. Methods

2.1 Data collection and processing

In situ chlorophyll-*a* and SSR time series from latitudes ≥ 40° N from 1950-2020 were collected from online, open-access, international, federal, and regional datasets to form a compiled database, as outlined in Appendix I.

The restriction of the data to north of 40° N latitude partially controls the study for different mixing and ice cover regimes that dominate at different latitudes. At mid- to high- latitudes, lakes are more likely to be dimictic and have seasonal ice cover (Woolway & Merchant, 2019), while low-latitude lakes are more likely to be meromictic or polymictic, and do not often experience ice cover (except for those at high elevations). The relationship between lake chlorophyll-*a* and other environmental variables, including SSR, may be different in low-latitude lakes and the results described here may be irrelevant.

In situ water samples for chlorophyll-*a* measurements were collected at various depths by lake monitoring agencies (summarized in Appendix I). To ensure consistency across sampling sites, only measurements taken at a depth of 3 meters or less were included. Since lake water temperature data were also required, chlorophyll-*a* measurements without contemporaneous water temperature measurements were removed from the compiled database.

Because of the large variability in consistency and temporal coverage between SSR data from different sources, only records with 15 years of data or more from 1990 to the present were included. This criterion was set so that individual lake and SSR records would have greater temporal overlap when paired. Due to irregular periods of missing or incomplete data from some SSR records, years with more than 30% of SSR data missing were also removed from the compiled database. Figure 2 below shows the spatial distribution of SSR stations in the compiled database after these criteria were applied.

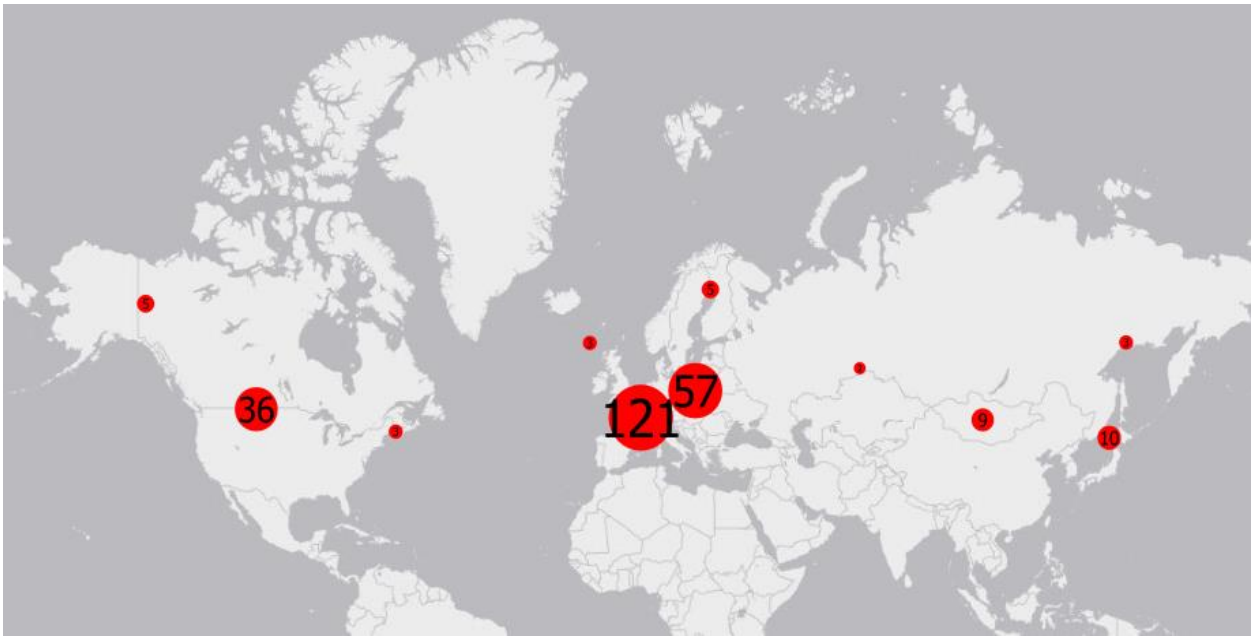

Figure 2: Distribution of solar radiation *in situ* measurement stations in compiled project database

Satellite time series of chlorophyll-*a*, and lake surface water temperature were extracted from the European Space Agency's Lakes_cci data products (Crétaux et al., 2020) for selected lake centroids. Lake centroids were calculated in QGIS by finding the coordinate of the central point of each lake shape polygon. The Lakes_cci data products are homogenized datasets from multiple instruments and satellites (Crétaux et al., 2020). Daily data is available from 1990-2020 for temperature, and from 2002-2020 for chlorophyll-*a*.

The coordinates chosen for satellite data extraction consisted of all lakes in the HydroLAKES database (Messager et al., 2016) with surface area ≥ 100 km^2 and latitude ≥ 40° N, for which the centroid coordinates were calculated. Out of the 1.4 million lake polygons in the global HydroLAKES database, 1,034 lakes satisfied these conditions. The surface area cutoff was chosen as a compromise between spatial coverage of selected lakes, and computational expense of the satellite data extraction.

Z-scores (sometimes called normalized anomalies) were calculated for satellite chlorophyll-*a* data, where the Z-score for a datapoint represents the difference between the datapoint and the sample mean in units of standard deviation (Mendenhall & Sincich, 2007). From visual inspection, all points with a Z-score greater than 5 were identified as suspected outliers and removed. Although these points may have represented valid chlorophyll-*a* measurements, their removal was required to homogenize the data for the performance of the model.

Even after the Z-score filtering was applied, the satellite chlorophyll-*a* data still contained some much higher values compared to the *in situ* measurements. The comparative homogeneity of the *in situ* data was likely due to the infrequency of water sampling relative to the daily temporal resolution of the satellite data. Sudden high-intensity bloom events over several days, which would be captured by satellite data, would likely be missed by *in situ* sampling. *In situ* chlorophyll-*a* measurements, while they may be affected by human measurement and analysis mistakes, are also not subject to optical sensor or imagery interpretation errors that may affect satellite data products. However, due to the pre-publication quality control processes undergone by the Lakes_cci product used here, these types of errors are assumed to be minimal.

All data processing, visualization, and analysis for this project was done in Python (ver. 3.7.6), using the modules NumPy (ver. 1.18.1) (Harris et al., 2020), pandas (ver. 1.0.1) (Mckinney, 2010; The pandas development team, 2020), matplotlib (ver. 3.1.3) (Hunter, 2007), and scikit-learn (ver. 0.23.2) (Pedregosa et al., 2011). Spatial data analysis and visualization were done using QGIS/PYQGIS (ver. 3.14) (QGIS.org, 2021).

Representative radius of SSR point observations

To examine the relationship between chlorophyll-*a* and SSR, satellite and *in situ* lake records were paired with a representative *in situ* SSR record. It is acknowledged in the literature (Schwarz et al., 2018) that *in situ* point measurements of SSR are generally considered spatially representative within a 1° radius. Therefore, *in situ* and satellite lake records were paired with the closest SSR station based on geodesic distance. A maximum radius of 1° was used for *in situ* data.

Maximum pairing radii of 1°, 2°, and 3° were used for the satellite data for a sensitivity analysis of the representative SSR radius. In total, four sets of paired lakes and SSR stations were generated (using *in situ* data at a 1° pairing radius, as well as using satellite data at 1°, 2°, and 3° pairing radii). The model, described below in Section 2.2, was run separately on all four datasets and the results were compared.

2.2 Random forest model

Random forest and cross validation

Random forest is a type of supervised machine learning algorithm. The input to the algorithm is comprised of many *instances* (i.e., rows) of data. Each row is associated with a set of *predictor parameters* (i.e., columns) and one *target parameter*. The algorithm randomly subsamples the input dataset with replacement (i.e., *bootstrap aggregation*) and passes the subsampled data instances onto *K* number of decision trees (Liaw & Wiener, 2002). Each decision tree uses the predictor parameters to predict the target parameter independent of other trees (Breiman, 2001). The final model prediction is taken as an average of all decision tree predictions (Breiman, 2001). This process is illustrated in Figure 3. By testing the accuracy of predictions against actual target parameter values, the algorithm identifies the relative importance of each predictor parameter in governing the target parameter (Liaw & Wiener, 2002).

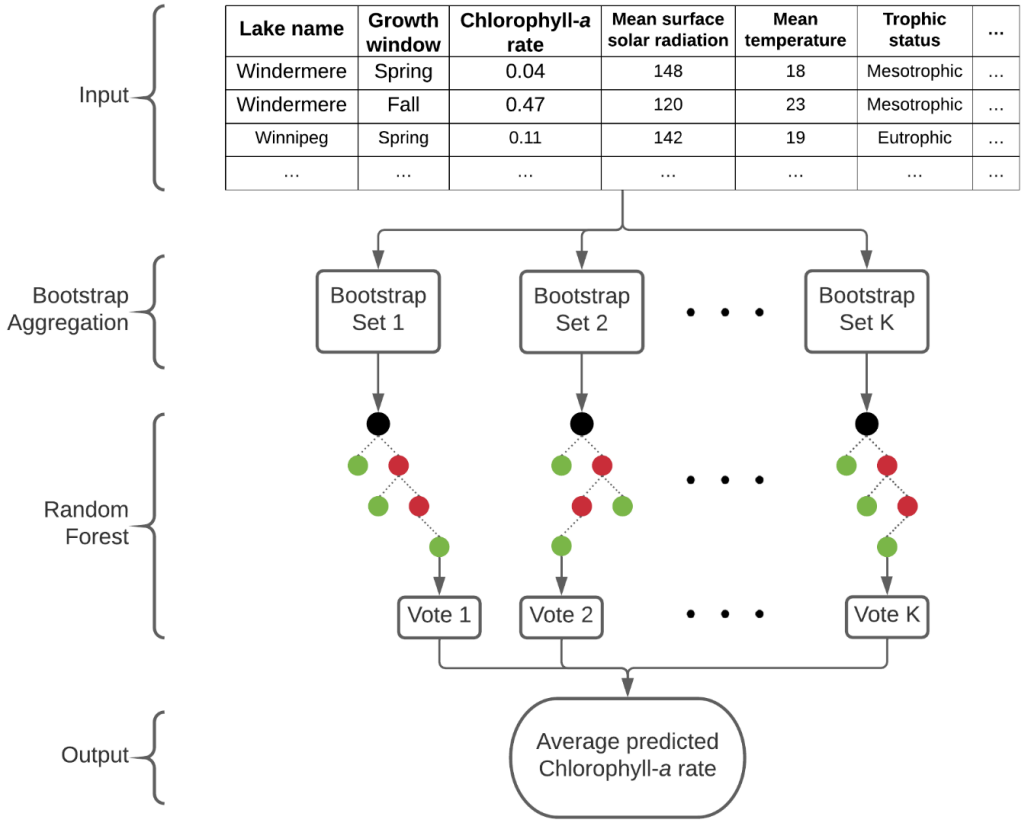

Figure 3: A graphical representation of a random forest algorithm. Figure adapted from Denisko & Hoffman (2018) and Li et al. (2017).

A cross validation was used to evaluate the model performance. *Model parameter* ranges were first defined for the random forest regressor, such as the maximum allowed depth of an individual tree, minimum allowed number of samples per tree leaf, *etc.* A grid-search cross validation was applied, which exhaustively searches through all combinations of parameters to determine the *estimator* (the optimized parameters that return a model with the most accurate prediction) (scikit-learn developers, 2019).

The cross validation applied here was 10-fold, meaning that the data was split into 10 groups. The data underwent 10 iterations of the random forest regressor utilizing all possible parameter combinations. Each iteration used 9 groups as a test set, and 1 group as a training set. With each

subsequent iteration, the group of data assigned for training rotated. Within each iteration and for each combination of parameters, a random forest of decision trees is generated. The optimized parameters are those which generate the most accurate prediction (scikit-learn developers, 2019). Code for the model can be found in Appendix II.

Figure 4 summarizes the workflow from lake-SSR pairing to model input described throughout this section.

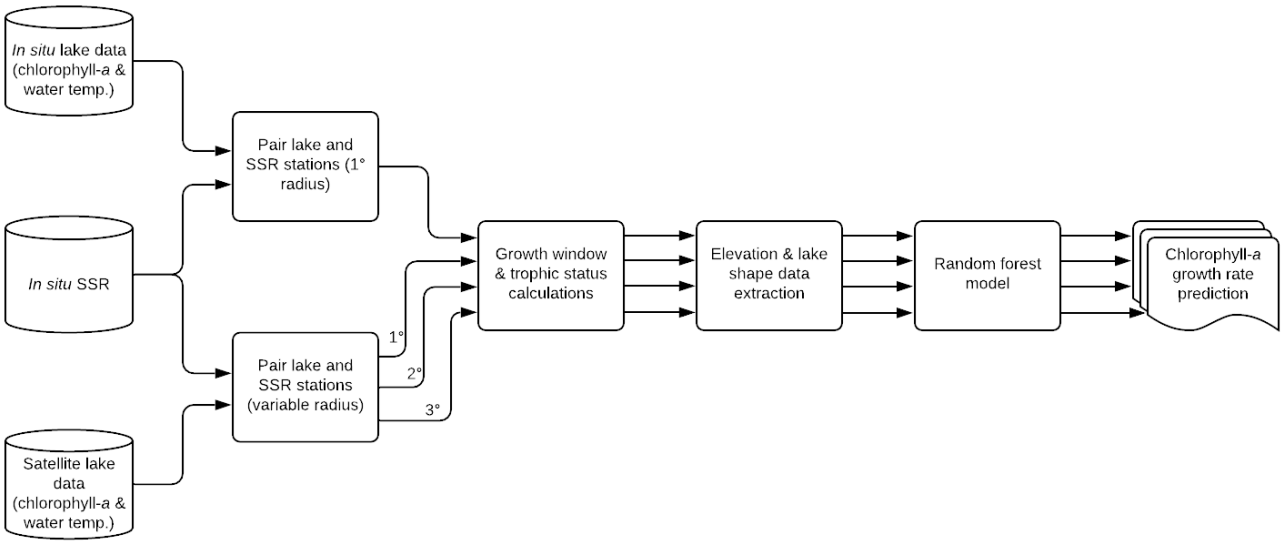

Figure 4: Workflow followed for compiling input datasets for random forest model.

Predictor parameter selection and collinearity testing

Based on literature review and exploratory analysis of relationships between some environmental variables, the parameters in Table 1 were selected as target and input parameters for the random forest model.

Table 1: Target and predictor input parameters for random forest model

	Description	Data source
Target parameter:		
Chlorophyll-*a*	Growth window rate ($\mu g\ L^{-1}\ day^{-1}$)	Global open-access *in situ* databases and ESA Lakes_cci data[1]
Predictor parameters:		
SSR	During- and pre- growth window mean ($W\ m^{-2}$)	Global open-access *in situ* databases
Water temperature	During- and pre- growth window mean (°C)	ESA Lakes_cci data[1]
Trophic status	Categorical (eutrophic, mesotrophic, oligotrophic)	Calculated using NALMS guideline[2]
Lake morphometry	Mean depth (m) and volume ($0.001\ km^3$)	HYDROLakes database[3]
Elevation difference	Between lake and paired SSR station (m)	USGS GMTED2010 DEM[4]
Distance	Between lake and paired SSR station (km)	Calculated using QGIS
Lake latitude	(°N)	Global open-access *in situ* databases and HYDROLakes database[3]
Growth window length	(number of days)	Calculated based on chlorophyll-*a* time series

[1]Crétaux et al., 2020

[2]Carlson & Simpson, 1996

[3]Messager et al., 2016

[4]Danielson, J.J., Gesch, 2011

Using a chlorophyll-*a* rate of growth as the target parameter allowed direct investigation of the effects of various parameters on the growth, rather than the absolute amount, of algae. This parameter was also useful for direct comparison of values between different lakes.

In addition to SSR, lake surface water temperature was included as a predictor. This is due to the understanding that water temperature is closely related to algal growth (Singh & Singh, 2015), as well as being partly influenced by SSR (Jakkila et al., 2009; Schmid & Köster, 2016; Zhong et al., 2016).

Trophic status was also included as a predictor. Trophic status was calculated based on the long-term mean chlorophyll-*a* concentration, one of three methods described in Carlson & Simpson (1996). In the absence of data on nutrient concentrations consistent across lakes in the compiled database, trophic status was used as a proxy measure for lake nutrient availability.

Annual spring and fall "growth windows" were identified for each lake to focus our study on the periods of the year with algal bloom activity. In this study, the start of a spring growth window was defined as the day when the chlorophyll-*a* concentration rate of increase exceeded zero for the first time in a year. The end of the spring growth window was defined as the day when the rate of increase became negative for the first time in year. The fall growth windows were defined similarly, after data from spring growth windows were removed. For years where the rate of change fluctuated, only one growth window was calculated for that year and labelled as a "spring" growth window. These growth windows were defined as the first day of increasing chlorophyll-*a* growth rate until the maximum chlorophyll-*a* concentration in the year was reached. The length and timing of the calculated growth window for each lake therefore varies from year to year. An example of ideal spring and fall growth windows for two different years within the two basins of Lake Windermere in the UK is illustrated in Figure 5 below.

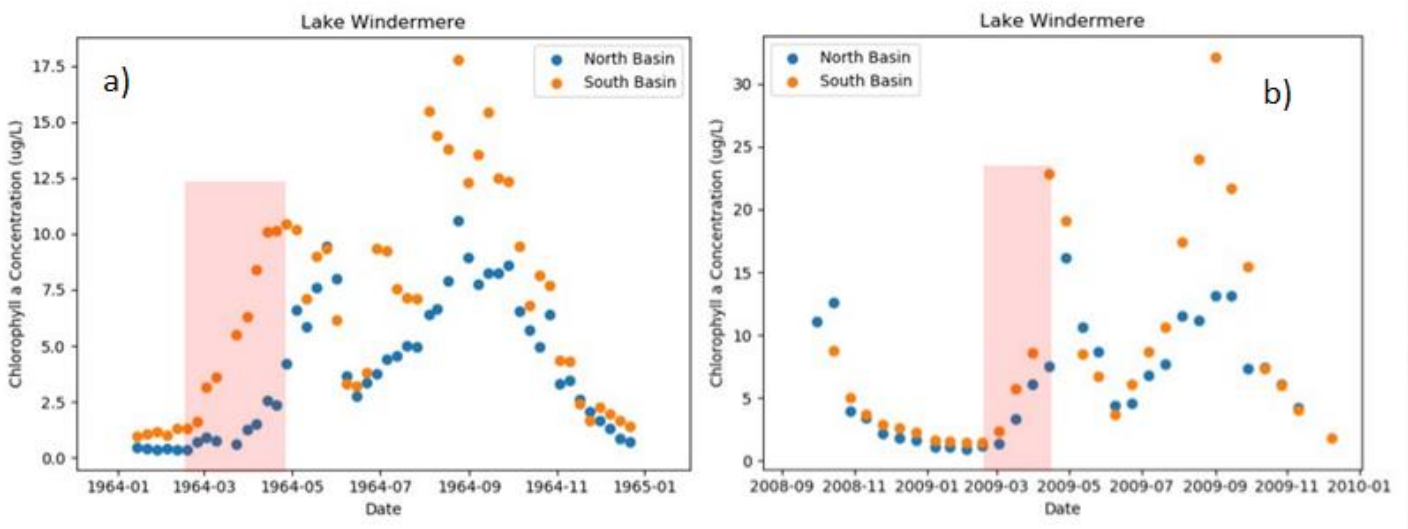

Figure 5: Ideal examples of calculated spring growth windows for North and South basins of Lake Windermere, UK, in a) 1964 and b) 2009. Figure created by Hannah Adams.

For parameters available as time series (SSR, water temperature, and chlorophyll-*a*), mean values within each growth window were calculated and included as predictor parameters for the model. The length of each growth window was also included as a predictor parameter to account for possible effects of extremely long or short calculated growth windows.

Mean lake surface water temperature and mean SSR during the one-week period prior to the start of the growth window were also calculated and included in the model. The inclusion of these predictor parameters allows the investigation of effects that temperature and radiation may have on the growth of algae before any detectable changes in chlorophyll-*a* concentrations appear. These parameters are referred to as the "pre-growth window" SSR and temperature.

The volume and mean depth of lakes, extracted from the HydroLAKES database (Messager et al., 2016) were included as indicators of mixing patterns (Fee et al., 1996). The elevation difference and distances between lakes and paired SSR stations were calculated using QGIS, and were included to account for potential effects of the spatial representativeness of the SSR point measurements (Schwarz et al., 2018).

Collinearity between predictor parameters was tested using Kendall non-parametric correlation testing, as data was not normally distributed. The absolute value of the Kendall correlation coefficient, $|\tau|$, was found to be <0.7 for all predictor parameters, indicating an acceptable level of collinearity between all predictor parameters (Genuer et al., 2010).

Assessment of model performance

The coefficient of determination (R^2) and root mean squared error (RMSE) are used to assess the accuracy of the model predictions. The R^2 value represents the amount of variation in the data that is explained by the model (Neill & Hashemi, 2018). A higher R^2 indicates that the model explains more of the variation in the observed data. The RMSE represents the difference between numerical predictions and observed values. RMSE is a commonly used measure of model accuracy, and a lower RMSE represents a higher model accuracy (Neill & Hashemi, 2018).

3. Results

3.1 Data collection and processing

Four model input dataset scenarios, labelled A through D, were generated from the random forest model using datasets summarized in Table 2. Figure 6 shows the global distribution of lakes included in each of the four input datasets. It is evident that by taking advantage of satellite data, the spatial representativeness of the study results is improved, compared to using only *in situ* data (Figure 6).

Table 2: Summary of four input datasets for random forest model

Scenario label	Dataset	SSR station pairing radius (degrees)	Number of data instances	Number of unique lakes	Number of unique SSR stations
A	*In situ*	1	250	20	4
B	Satellite	1	589	47	14
C	Satellite	2	878	68	22
D	Satellite	3	922	72	24

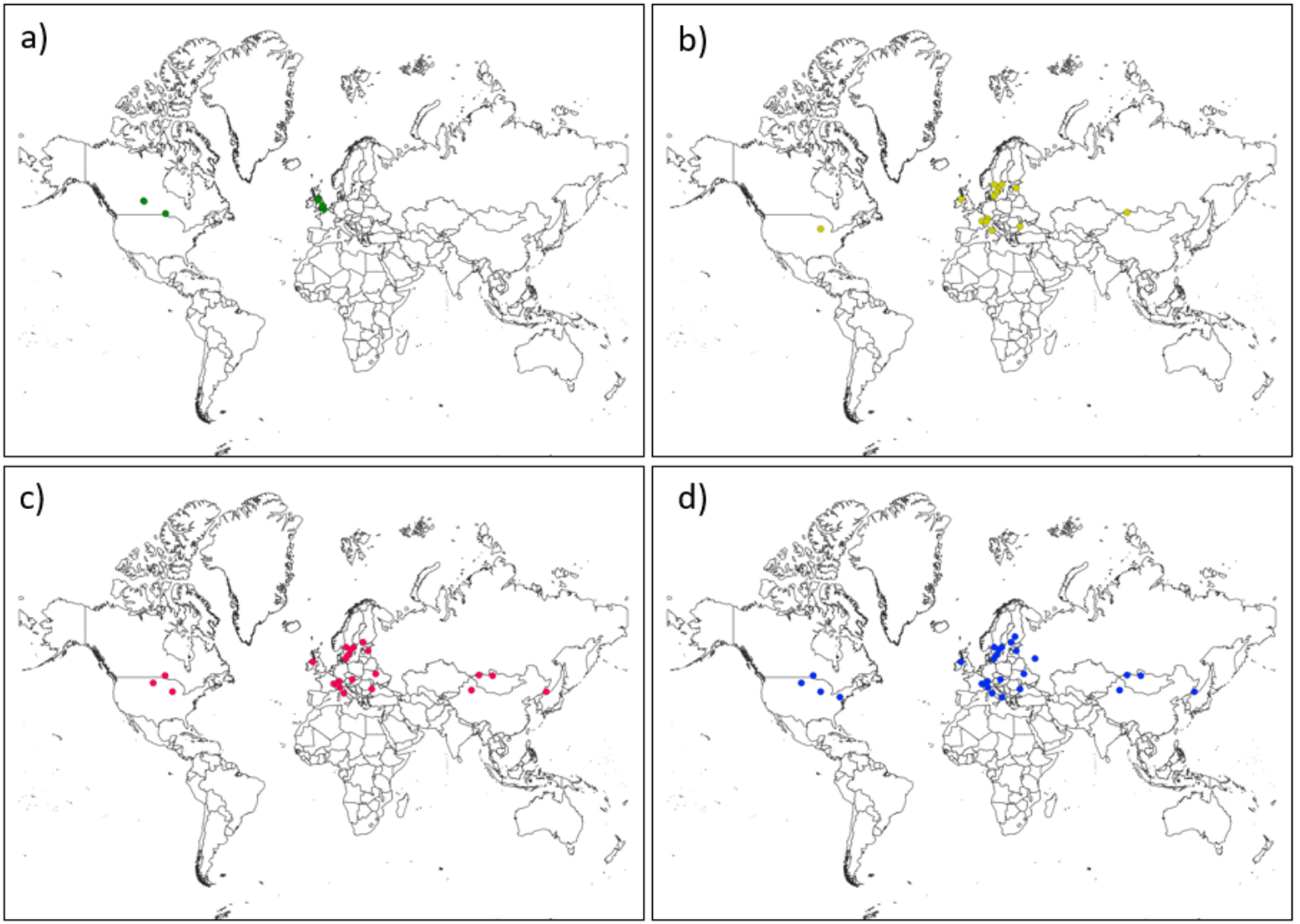

Figure 6: Distribution of lakes included in model input dataset scenarios A-D.

3.2 Random forest model

Assessment of model performance

Figure 7 shows the chlorophyll-*a* rate predicted by the model plotted against the input chlorophyll-*a* values for the four input dataset scenarios. The red line represents the linear regression best fit line, with the shaded red region representing the 90% confidence interval of the linear regression. The R^2 and RMSE values are shown on each plot for the random forest testing and training sets. The highest R^2, as well as the lowest RMSE, for both training and testing sets, is seen in scenario B. Using this dataset, the implemented model explained about 50% of the variability in the chlorophyll-*a* data of the testing set ($R^2 = 0.504$).

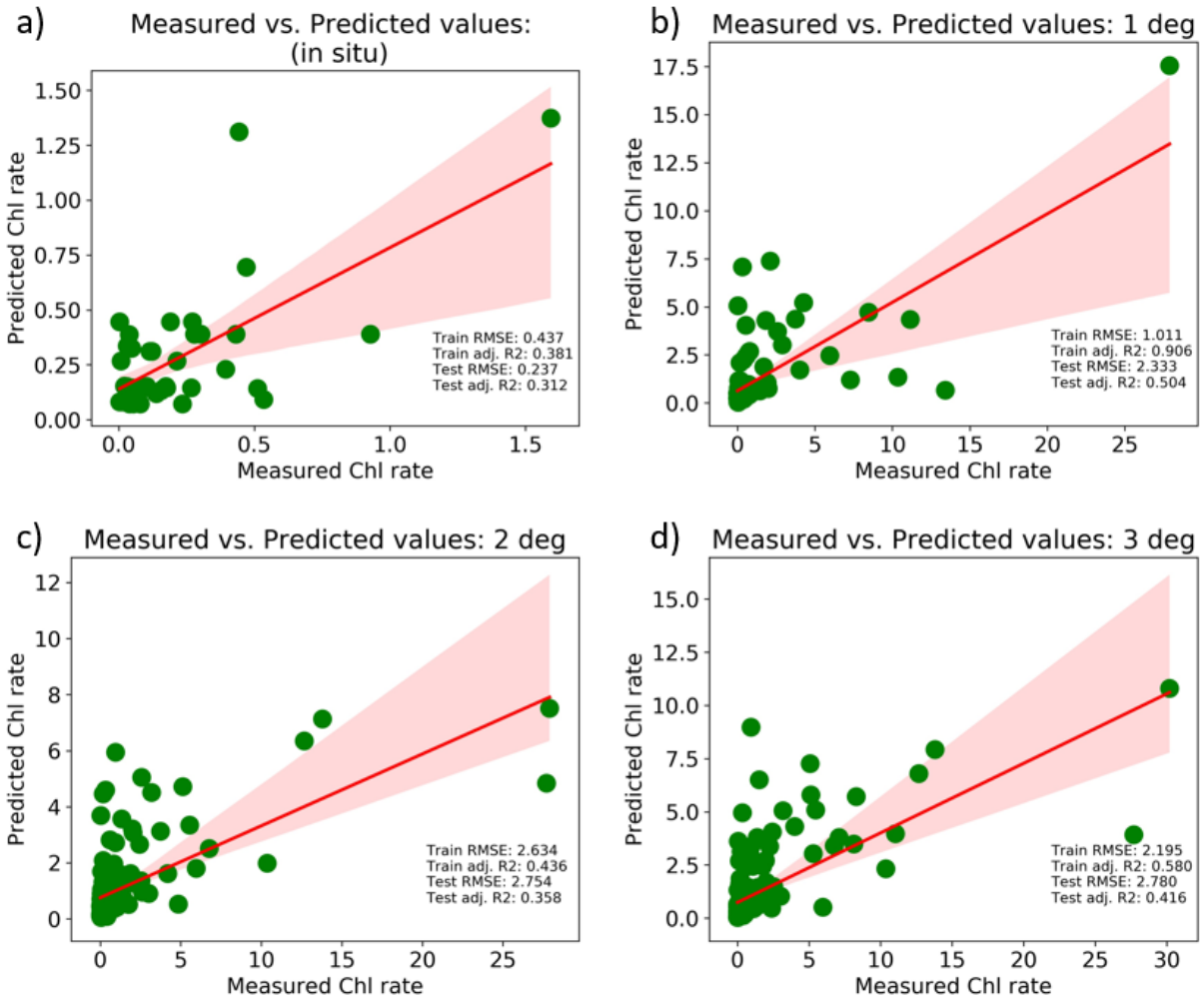

Figure 7: Scatterplots comparing chlorophyll-*a* rate predicted by model with observed chlorophyll-*a* rates for model input dataset scenarios A-D.

Influence of predictor variables

Figures 8 and 9 quantify the importance of each parameter in predicting the chlorophyll-*a* rate, as determined by the random forest model for the four input dataset scenarios. It is important to keep in mind when interpreting this set of figures that a low relative importance does not mean that a predictor parameter is "not important". A low importance implies that given a certain input dataset and model setup, other parameters had more weight in predicting the target value. As such, machine learning models are highly sensitive to the data they are given and should be interpreted with an understanding of the limitations of the input data.

There were noticeable differences in the model results for the satellite datasets (scenarios B-D, Figures 8b, 9c, and 9d) compared to the *in situ* dataset (scenario A, Figure 8a). Lake depth, pre-growth window temperature and during-growth window temperature made up the three most important predictors in scenario A, with a combined representation of ~60% of the model variability (Figure 8a). However, lake depth represented only <2% of the model variability in scenarios B-D (Figures 8b, 9c, and 9d). The three most important parameters for scenario B were instead during-growth window temperature, growth window length, and pre-growth window SSR, with a combined representation of ~54% of the model variability (Figure 8b). Previous studies (Jakkila et al., 2009; Long et al., 2011; Singh & Singh, 2015) have shown the strong influence of water temperature on chlorophyll-*a* concentrations or algal growth rates.

In scenarios C and D (Figures 9c and 9d), the importance of parameters was dominated by the growth window length. The growth window length had a high relative importance for scenarios B-D, with a much lower relative importance for scenario A. In scenarios C and D, growth window length contributed ~40% to the model variability. In scenario B, the growth window length contributed ~17.5% to the model variability. Conversely in scenario A, the growth window length contributed only ~2.5% to the model variability.

In scenario B, which had the highest model accuracy, pre-growth window SSR and during-growth window SSR represented ~17.5 % and ~11% of the model accuracy, respectively (Figure 8b). The combined importance of these two SSR parameters, at ~28.5%, is comparable to the combined importance of pre- and during- growth window temperature, at ~29%.

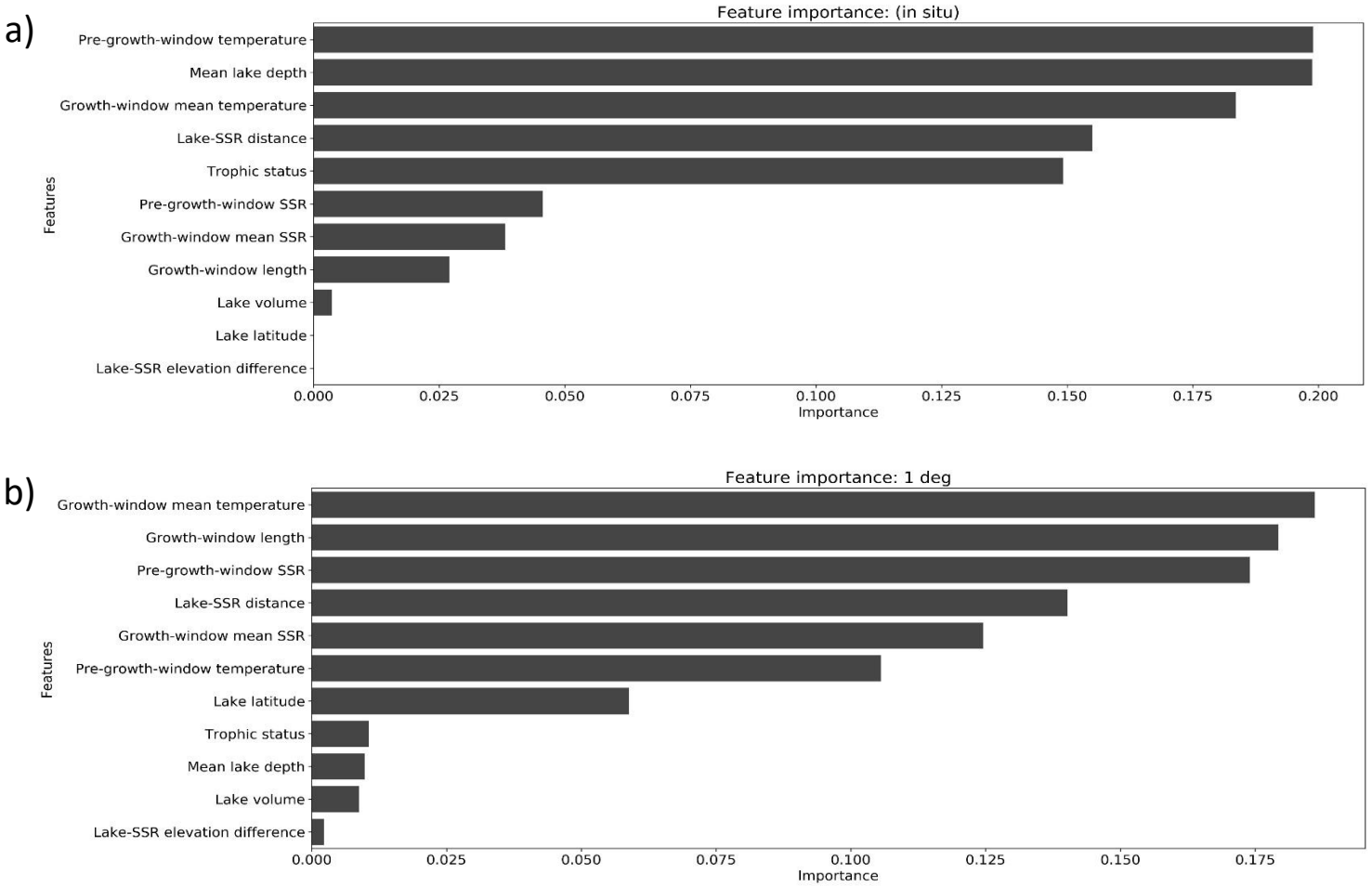

Figure 8: Relative feature importance determined by random forest model for two of four model input dataset scenarios, A and B.

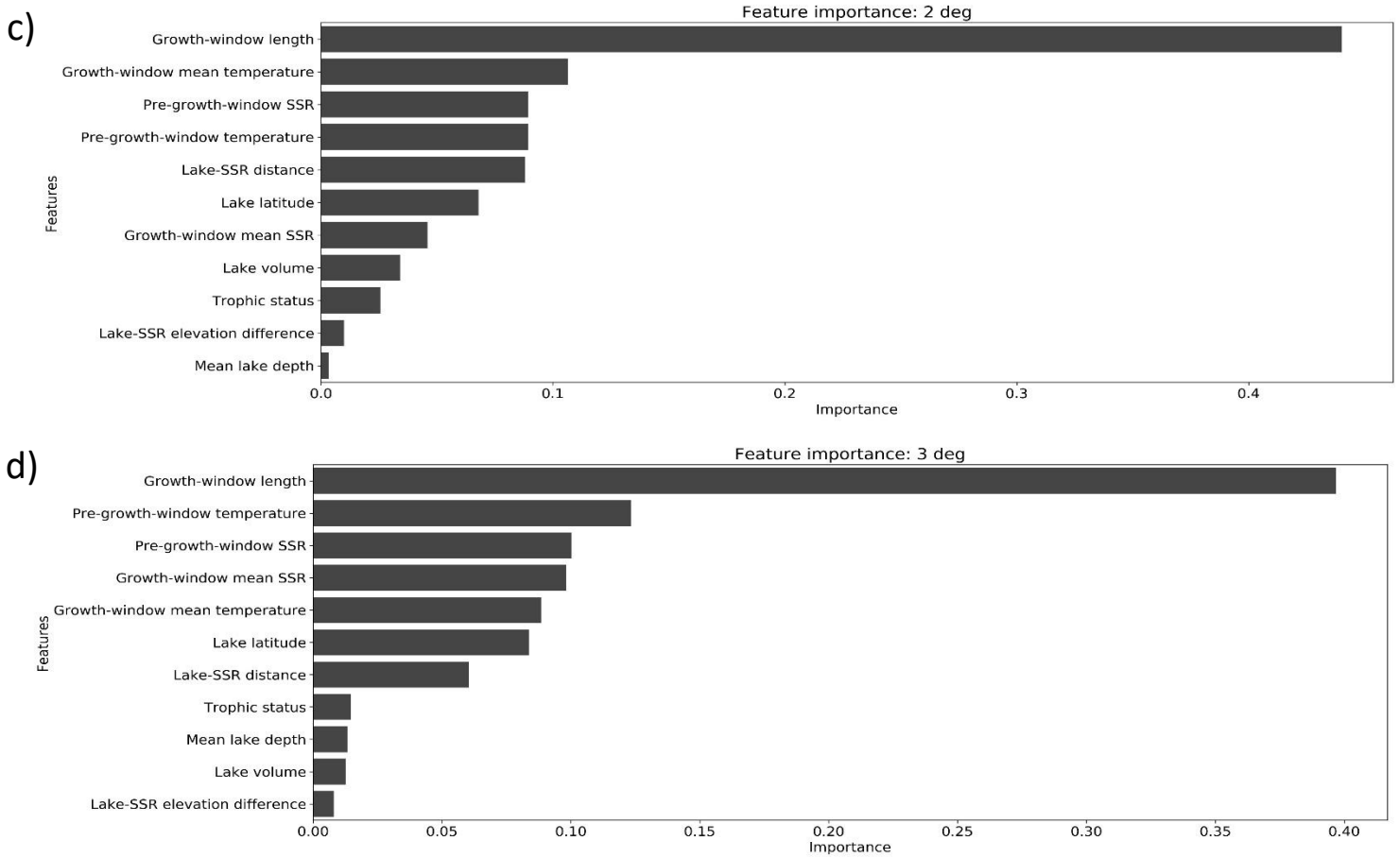

Figure 9: Relative feature importance determined by random forest model for two of four model input dataset scenarios, C and D.

Partial dependence plots are shown below, in Figure 10, for selected important parameters in all four model input dataset scenarios. Partial dependence examines the individual influences of each predictor parameter on the model's prediction of the target parameter. They are a useful tool for understanding the data-driven results of machine learning model predictions. The x-axis shows the range in values of each predictor parameter. The y-axis shows the partial dependence of the chlorophyll-*a* rate target parameter on each predictor parameter. A higher partial dependence indicates a higher likelihood of the model predicting a high chlorophyll-*a* rate for a given predictor value.

In scenario A (Figure 10a), there is a relationship between higher during-growth window temperature, during-growth window SSR, pre-growth window temperature, and pre-growth window SSR with higher chlorophyll-*a* rate. In this scenario, it is also quite evident eutrophic lakes are linked to higher chlorophyll-*a* rates than oligotrophic or mesotrophic lakes. This is consistent with the definition of trophic status defined by Carlson & Simpson (1996).

For scenarios B-D (Figures 10b-d), the relationship between during-growth window temperature and during-growth window SSR with chlorophyll-*a* rate is less clear, and at times appears to be the inverse of the relationship seen in scenario A. However, relationships are seen between higher pre-growth window temperature and pre-growth window SSR with higher chlorophyll-*a* rate. The unexpected relationships between during-growth window parameters and chlorophyll-*a* rates, as well as the strong relationship between short growth windows and high chlorophyll-*a* rates, may be legacies of the growth window length calculation. This is discussed in Section 4.2.

In Figures 10 b-d, it is particularly interesting to note that the greatest response of chlorophyll-*a* rate to SSR occurs around a mean SSR value of ~200 W/m^2. This "optimal SSR" behaviour supports the speculation that photoacclimation or photoinhibition may be taking place at higher SSR exposure. This behaviour is not observed in scenario A (Figure 10a) likely because of the lower temporal resolution of the scenario A *in situ* dataset.

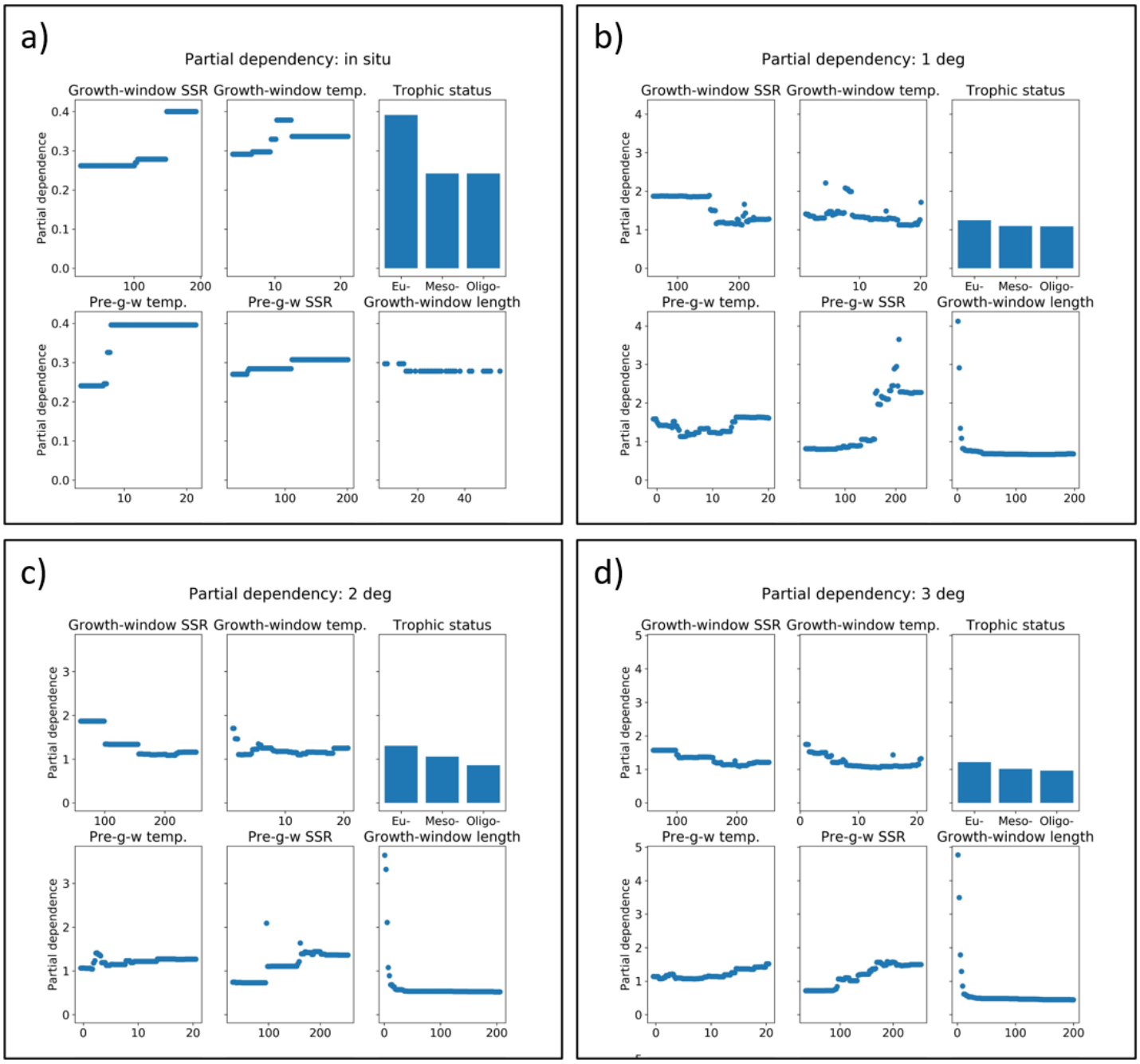

Figure 10: Partial dependence plots for selected predictor parameters in model input dataset scenarios A-D.

Density distribution plots corresponding to the parameters shown in the partial dependence plots are in Figure 11. These plots illustrate the evenness of data distribution in each dataset, and any potential biases that may arise in partial dependence and relative feature importance as a result of uneven data distribution. For example, in the satellite dataset scenarios B-D (Figure 11 b-d), eutrophic lakes are over-represented compared to mesotrophic and oligotrophic lakes. This might result in the importance of trophic status as a predictor being under-estimated. The scenario B-D datasets also show a skew towards short growth windows in the "growth window length" parameter.

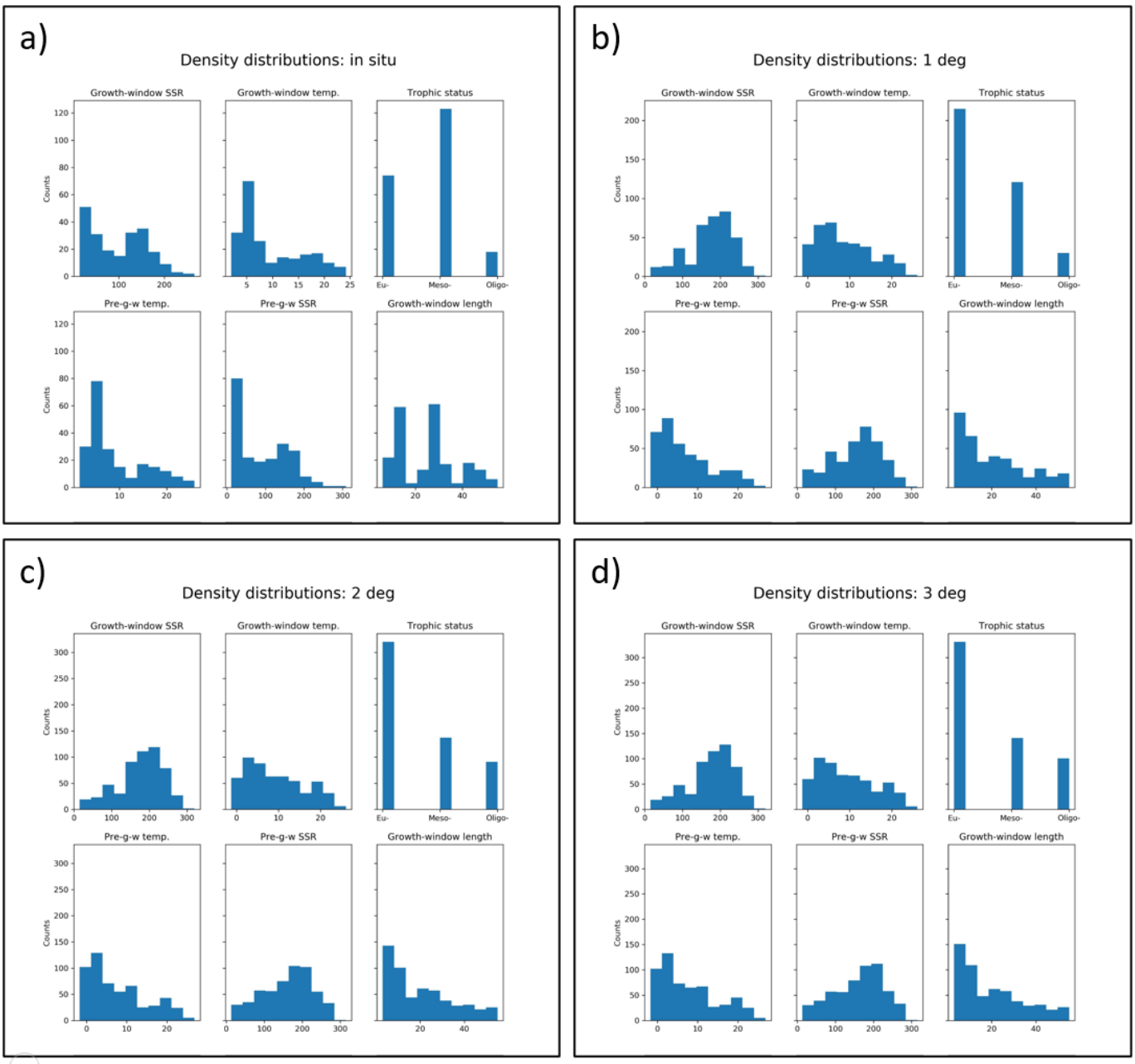

Figure 11: Density distribution plots for selected predictor parameters in model input dataset scenarios A-D.

4. Discussion

4.1 Representative radius of SSR point observations

Representative radii of 1°, 2°, and 3° for SSR point observations (scenarios B-D) were tested using satellite datasets. A radius larger than 1° (*i.e.*, scenarios C and D) resulted in worse model performance based on both R^2 and RMSE (Figure 7). This is in line with findings from Schwarz et al. (2018), who reported that point SSR observations are generally considered representative of a 1° radius or less. The greater distances between the lake and the measured SSR contributed to lower combined importance of SSR parameters (pre-growth window SSR and during-growth window SSR, Figures 8 and 9), as well as worse model performance in scenarios C and D compared to scenario B.

4.2 Influence of predictor variables

The length of each growth window is very highly weighted for the three satellite dataset scenarios B-D (Figures 8 and 9). Partial dependence plots (Figure 10) of the growth window length predictor parameter for the satellite datasets show that shorter growth window lengths result in higher predicted chlorophyll-*a* growth rates. This relationship is caused by the change in chlorophyll-*a* from the start to the end of the growth window being divided over the growth window length to calculate the growth rate. Very long or short calculated growth windows do not actually represent a period of rapid algal growth that would be traditionally defined as an "algal bloom". The challenge of defining the start and end of the growth window is a drawback of the variable growth window length used in this study.

In all three satellite data model scenarios B-D, pre-growth window SSR was weighted higher than mean SSR during the growth window (Figures 8b, 9c, and 9d). It is notable that this pre-growth window SSR, which is calculated for the relatively brief period prior to any detected increase in chlorophyll-*a* concentrations, plays a significant role in controlling the growth rate for the entire growing season overall. For all three satellite data model scenarios, the partial

dependence plots for pre-growth window SSR (Figures 10b-d) show a relationship between higher chlorophyll-*a* growth rate and higher pre-growth window SSR. It is possible that the comparatively low importance of the during-growth window mean SSR (Figure 8b, 9c, and 9d), as well as the unclear partial dependence relationship of chlorophyll-*a* rate on the during-growth window mean SSR (Figures 10b-d), resulted from inaccurately calculated growth window lengths.

Although this study did not distinguish between the fall and spring growth windows in the model, the pre-growth window SSR may be an especially important factor in spring algal growth compared to that in the fall. Several studies (Shatwell & Köhler, 2019; Townsend et al., 1992) have found the driving factors, as well as the dominant algae species, to be distinct between spring and fall algae blooms. For example, spring blooms were found to be phosphorus limited and dominated by diatoms, while summer/fall blooms were found to be nitrogen limited and dominated by cyanobacteria (Kong et al., 2021; Shatwell & Köhler, 2019; Townsend et al., 1992).

The spring growth period is of particular interest from a SSR perspective. Light limitation has been acknowledged in the literature as being an important control on productivity in stratified lakes, particularly in winter and spring (Sommer et al., 1986). More recently, and at a global scale, Shuvo et al. (2021) found that spring SSR was more important than that in the summer when predicting annual average lake chlorophyll-*a* concentrations. Tian et al. (2017) found similarly strong controls on SSR at a basin scale when predicting concurrent chlorophyll-*a* concentrations in the spring compared to other seasons. Additionally, Kirillin et al. (2012) found SSR to be an important factor in ice-break-up timing. They also point out that the timing and intensity of springtime algal blooms are heavily influenced by ice-break-up timing. Therefore, early-spring SSR may be an indirect as well as direct influence on spring algal blooms. The difference in the chlorophyll-*a* growth rate response to SSR (and other environmental factors) between the fall and spring growth periods is an area that warrants further study.

The moderate to high relative importance of lake-SSR distance, even for a 1° lake-SSR pairing radius (Figure 8b), was a symptom of the imperfect lake-SSR pairing methodology used in this study. Schwarz et al. (2018) found that, in general, point SSR observations are

representative of a 1° radius. However, they add that the representative radius becomes much smaller in regions with mountainous terrain and near coastlines. Since these terrain effects were not included in the model, the distance between paired lakes and SSR stations may have been detrimental to identifying an SSR-to-chlorophyll-*a* relationship using the model.

Using gridded SSR time series data products would allow us to estimate SSR exposure better collocated with the lakes in our study. Gridded SSR time series data products derived from reanalysis have been produced for specific regions of the world as well as globally. These products might also be applied to supplement, but not replace, *in situ* SSR measurements in regions where these records are sparse. Some examples of these data products include ERA-Interim (Simmons et al., 2007) and MERRA-2 (Gelaro et al., 2017) with global coverage, and Daymet with North American coverage (Thornton et al., 2020). The benefits and drawbacks of these gridded data products are discussed below.

Lake morphometry parameters (depth and volume) were found to be relatively unimportant at the scale of our study using the satellite datasets (Figures 8b, 9c, and 9d). This is in line with other regional and global scale studies (Shuvo et al., 2021 and citations within), indicating a lower importance of lake morphometry when compared to climate and nutrient input variables. The relative importance of lake depth in the *in situ* scenario A (Figure 8a) was likely because of the small number of unique lakes included in this dataset (see Figure 6 and Table 2, above). 17 of the 20 lakes included in scenario A were located in the UK, and their similarities in climate and environmental conditions may have caused the model to attribute more predictive importance to the lake depth.

As a result of their close proximity, the *in situ* lakes included in scenario A also had similar latitudes; therefore, latitude was found to have a negligible importance in scenario A (Figure 8a). Additionally, the SSR received by the 17 UK lakes in scenario A was represented by only three unique SSR stations (with the fourth SSR station located in Canada). Due to the physical closeness of the UK lakes, there was little variation in lake elevation, and, therefore, little variation in lake-

SSR elevation difference between the lakes. Likely as a result of this, the lake-SSR elevation difference was also found to have negligible importance in scenario A (Figure 8a).

Since trophic status was used here as a proxy for nutrient availability, it was surprising that in all model scenarios, it had relatively low importance. Prior analysis on regional *in situ* chlorophyll-*a* data as part of this study had shown a differentiation by trophic status in chlorophyll-*a* response to SSR. It is possible that this was due to an uneven distribution of different trophic statuses in our input datasets, with over half of the rows of data representing eutrophic lakes in the scenario B dataset (Figure 11b).

Except for trophic status, the results from our model generally seem to agree with previously published investigations into various drivers of spring and fall seasonal chlorophyll-*a* growth rates (*i.e.*, algal blooms), as discussed above. Using a unique growth window approach, it is of particular interest to note the importance of SSR up to one week prior to the start of the growing phase in predicting the overall growth rate.

4.3 SSR controls and predictions

Anthropogenic activities, including climate change, are major drivers of changes in cloud cover, water vapour, and aerosol concentrations. Atmospheric water vapour content and global air temperature are projected to increase with climate change (Intergovernmental Panel on Climate Change, 2001; Lavers et al., 2015), with warmer air temperatures potentially leading to increases in cloud cover (Croke et al., 1999). Deforestation, which modifies the seasonality of cloud cover patterns relative to pristine forest (Durieux et al., 2003), is expected to increase with population growth (Pahari & Murai, 1999). In tropical regions, atmospheric aerosol content is anticipated to be higher on average in the thirty-year period between 2021 and 2050 compared to 1961-1990, but will likely be lower in temperate and (sub)arctic regions (Stier et al., 2006).

All these factors – water vapour, cloud cover, and atmospheric aerosols – act as attenuators of SSR (Ångström, 1962; McCormick & Ludwig, 1967; Renner et al., 2019; Yu et al., 2021).

Inevitably, with increasing anthropogenic impacts on the climate and environment, there will be effects on cloud cover, water vapour content, and aerosol concentrations, and thus on SSR around the globe.

However, predictions of future SSR trends are currently limited in the literature. One study applying aerosol-forcing models for SSR predictions found that forecast results varied greatly depending on the model that was implemented (Gutiérrez et al., 2020). This study acknowledged the limitations of their predictions given that cloud cover and water vapour impacts were not included in their model. Overall, the global interactions between cloud cover, water vapour, and aerosols with climate change are not fully understood (Rosenfeld et al., 2014). These interactions occur at varying spatial and temporal scales, making SSR predictions difficult.

4.4 Implications

Algal blooms are increasing in frequency, duration, and intensity worldwide, and will likely continue to do so with global climate change. Most algal bloom management programs focus on the reduction of nutrient inputs. However, investigating other factors that contribute to lake productivity will increase our understanding of why algal blooms occur, contributing to our ability to assess the effectiveness of best management practices.

By using a machine learning model to analyze long term, global scale datasets, a simple technique is demonstrated by which satellite data, and other large-scale environmental data, can be used to supplement *in situ* environmental and lake monitoring efforts.

Additionally, the hemisphere-scale dataset of long term SSR and chlorophyll-*a* records compiled as part of this work may be a valuable resource for future studies looking to incorporate SSR effects into research on lake productivity.

4.5 Limitations and future directions

As machine learning models are inherently data-driven, this study is limited in scope by the data-derived predictive parameters available at the scale of our study. Thus, this study is unable to completely account for all known or suspected influencing factors on chlorophyll-*a* growth rate.

For example, photoinhibition occurs when algae are exposed to harmful levels of solar radiation and begin to suffer instead of benefit from the dosage (Staehr et al., 2016). Photoacclimation occurs when algae become adapted to higher solar radiation levels and begin to produce less chlorophyll-*a* relative to their biomass (Westberry et al., 2008). It is hypothesized that photoinhibition would cause an observed effect of decreasing chlorophyll-*a* growth rate with increasing SSR, at SSR values higher than some optimal level. It is also expected that photoacclimation would result in decreasing chlorophyll-*a* growth rate (relative to biomass) with increasing SSR, past a certain point.

This study is unable to examine these effects in detail with the current model and compiled dataset due to limitations in biomass and turbidity data at the scale of the study. These factors warrant further study from a physiological modelling approach, and preliminary work on such a model has shown promise.

In addition, the growth window calculations are an aspect of this study's methodology that warrants further attention. The methodology was designed and tested on *in situ* data and may not be directly applicable for use on satellite datasets. The detection and calculation of algal bloom growth windows using satellite data could be improved in future work by implementing aspects of the bloom-detection methodologies tested in a 2020 study by Germán et al. The authors of this study compared techniques such as threshold values, gap-filling and instantaneous slope estimation to identify periods of rapid growth from satellite chlorophyll-*a* data (Germán et al., 2020).

The results of this study should also be viewed with an understanding of the spatial limitations of the data. As Figure 2 shows, long-term *in situ* SSR records outside of Europe and some parts of North America are sparse. Additionally, it was found during the data collection phase

of the study that the quantity and quality of many SSR records around the world have declined since the 1990s. This trend is especially notable in North American SSR records and is well acknowledged in the literature (Wild, 2009). Alpert & Kishcha (2008), Cutforth & Judiesch (2007), and various related publications emphasize acknowledgement of limitations on conclusions about SSR patterns at a purportedly "global scale." These conclusions are not truly global, given that the majority of recent continuous *in situ* SSR records are in Europe, with urban areas especially overrepresented.

Some recent studies investigating SSR as a factor in chlorophyll-*a* concentrations have used reanalysis rather than *in situ* data (Shuvo et al., 2021; Tian et al., 2017). Undoubtedly the improved homogeneity, spatial and temporal representation, and accessibility of reanalysis data products over *in situ* data lends itself to greater ease of use. However, as reanalysis time series are based on numerical simulation using historical climate data (Bengtsson, 2004), there will be deviations from the direct *in situ* SSR measurements used in this study. Zhang et al. (2020) conducted a study evaluating ERA-Interim and MERRA-2 SSR reanalysis products against surface measurements in China. Some deviations from the surface measurements were identified, and these were found to be caused by incorrect estimations of local cloud cover and aerosol optical depth in the reanalysis products (Zhang et al., 2020). Therefore, SSR reanalysis products would probably be most useful in regional or local studies where enough *in situ* SSR records exist for evaluation of the accuracy of the gridded data. Future improvements to, and more validation of, reanalysis products would likely open doors to greater solar radiation and algal bloom modelling capabilities.

Due to the data limitations and scale effects of studying SSR trends, a regional approach incorporating reanalysis data might be a promising direction for this research. Regional trends and interactions between cloud cover, air temperature, aerosols, and water vapour might be better understood at a sub-global scale and more accurate models could be developed. The contributions of SSR to algal blooms could then be better quantified. This increased modelling capability would greatly contribute to improved accuracy of algal bloom and lake health forecasts

5. Conclusions

With the increasing availability of high resolution, large scale satellite data in all fields of environmental science, there is increasing opportunity to take advantage of the data by way of global scale studies not possible at any other time in human history. While the challenges at these scales of study, and the tools to overcome them, may differ from traditional *in situ* environmental and lake research, there is much value to be gained in utilizing satellite data resources. It is promising to see comparable results between *in situ* and satellite data using a widely applicable methodology.

Several representative SSR radii were tested (1°, 2°, and 3°) in this study for pairing SSR data with lake locations. A 1° radius, represented by scenario B, was found to result in the highest model accuracy (represented by R^2 and RMSE scores).

When model accuracy was optimized (scenario B), during-growth window temperature was the most important predictor, explaining ~18% of the chlorophyll-*a* rate variation. The scenario B model also showed that the combined importance of pre- and during- growth window SSR, explaining ~17.5 % and ~11% respectively, was comparable to the combined importance of pre- and during- growth window temperature, at ~29%. In all four model input dataset scenarios, pre-growth window SSR explained more variation than during-growth window SSR.

Partial dependency plots of satellite data model input scenarios B-D suggested that algal photoacclimation or photoinhibition behaviours in response to SSR above ~200 W/m^2 may have been represented by the model.

This study joins recent literature in successfully demonstrating the feasibility of using satellite data for modelling global-scale quantitative relationships between environmental variables and chlorophyll-a growth rates, using a widely applicable supervised machine learning technique. In addition to global scale lake monitoring, this technique could be useful in other fields of environmental science where satellite data or other large-scale datasets are employed.

Satellite data and data driven models are becoming more prevalent in modern day environmental science. However, they are not replacements for vital *in situ* monitoring and

physiological understanding of systems. Rather, they are additional tools that must be validated against high quality, direct measurements to be of use. In the unfolding age of environmental “big” data, “traditional” data collection and modelling for long-term environmental monitoring and forecasting are as important as ever.

6. References

Alpert, P., & Kishcha, P. (2008). Quantification of the effect of urbanization on solar dimming. *Geophysical Research Letters*, *35*(8), 1–5. https://doi.org/10.1029/2007GL033012

Anderson, D. M. (2009). Approaches to monitoring, control and management of harmful algal blooms (HABs). *Ocean and Coastal Management*, *52*(7), 342–347. https://doi.org/10.1016/j.ocecoaman.2009.04.006

Ångström, A. (1962). Atmospheric turbidity, global illumination and planetary albedo of the earth. *Tellus*, *14*(4), 435–450. https://doi.org/10.1111/j.2153-3490.1962.tb01356.x

Bengtsson, L. (2004). Can climate trends be calculated from reanalysis data? *Journal of Geophysical Research*, *109*(D11), D11111. https://doi.org/10.1029/2004JD004536

Binding, C. E., Greenberg, T. A., McCullough, G., Watson, S. B., & Page, E. (2018). An analysis of satellite-derived chlorophyll and algal bloom indices on Lake Winnipeg. *Journal of Great Lakes Research*, *44*(3), 436–446. https://doi.org/10.1016/j.jglr.2018.04.001

Breiman, L. (2001). *Random Forests* (Vol. 45).

Carlson, R., & Simpson, J. (1996). *A Coordinator's Guide to Volunteer Monitoring (digital download) – North American Lake Management Society (NALMS)*. North American Lake Management Society (NALMS). https://www.nalms.org/product/a-coordinators-guide-to-volunteer-monitoring/

Carpenter, S., Kitchell, J., Cole, J., & Pace, M. (2017). *Cascade Project at North Temperate Lakes LTER Core Data Process Data 1984 - 2016 ver 4*. Environmental Data Initiative. https://doi.org/https://doi.org/10.6073/pasta/6a658526e313dbcecbc0331a1f343c01

Chipman, J. (2019). A Multisensor Approach to Satellite Monitoring of Trends in Lake Area, Water Level, and Volume. *Remote Sensing*, *11*(2), 158. https://doi.org/10.3390/rs11020158

Clear Lake Water Quality - Riding Mountain - Open Government Portal. (n.d.). Retrieved April 12, 2021, from https://open.canada.ca/data/en/dataset/2a55313f-26fc-4872-9a57-2a7bf2a4cc38

Crespo Cuaresma, J., Danylo, O., Fritz, S., McCallum, I., Obersteiner, M., See, L., & Walsh, B. (2017). Economic development and forest cover: Evidence from satellite data. *Scientific Reports*, *7*(1), 1–8. https://doi.org/10.1038/srep40678

Crétaux, J.-F., Yésou, H., Merchant, C. J., Duguay, C. R., Simis, S., & Calmettes, B. (2020). *ESA Lakes Climate Change Initiative (Lakes CCI): Lake Products, Version 1.0*. https://doi.org/10.5285/3c324bb4ee394d0d876fe2e1db217378

Cutforth, H. W., & Judiesch, D. (2007). Long-term changes to incoming solar energy on the Canadian Prairie. *Agricultural and Forest Meteorology*, *145*(3–4), 167–175. https://doi.org/10.1016/j.agrformet.2007.04.011

Danielson, J.J., Gesch, D. B. (2011). Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010). *U.S. Geological Survey Open-File Report 2011-1073*, *2010*, 26. http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/GMTED2010

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 115, Issue 8, pp. 1690–1692). National Academy of Sciences. https://doi.org/10.1073/pnas.1800256115

Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F. M., Duprat, T., Fukuda, M., Grobe, H., Haeffelin, M., Hodges, G., Hyett, N., Ijima, O., Kallis, A., Knap, W., Kustov, V., Long, C. N., Longenecker, D., … König-Langlo, G. (2018). *Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017)*. Earth Syst. Sci. Data. https://doi.org/doi:10.5194/essd-10-1491-2018

Fee, E. J., Hecky, R. E., M Kasian, S. E., & Cruikshank, D. R. (1996). *Physical and chemical responses of lakes and streams Effects of lake size, water clarity, and climatic variability on mixing depths in Canadian Shield lakes* (Vol. 41, Issue 5).

Freshwater Inventory and Surveillance of Mercury: Water Quality Data. (n.d.). Retrieved April 12, 2021, from http://data.ec.gc.ca/data/substances/monitor/clean-air-regulatory-agenda-freshwater-inventory-and-surveillance-of-mercury-cara-fishg/freshwater-inventory-and-surveillance-of-mercury-water-quality-data/

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A.,

Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., … Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). *Variable selection using Random Forests*. *31*(14), 2225–2236. http://www.r-project.org/

Germán, A., Andreo, V., Tauro, C., Scavuzzo, C. M., & Ferral, A. (2020). A novel method based on time series satellite data analysis to detect algal blooms. *Ecological Informatics*, *59*, 101131. https://doi.org/10.1016/j.ecoinf.2020.101131

Great Lakes Water Quality Monitoring and Surveillance Data - Open Government Portal. (n.d.). Retrieved April 12, 2021, from https://open.canada.ca/data/en/dataset/cfdafa0c-a644-47cc-ad54-460304facf2e

Gutiérrez, C., Somot, S., Nabat, P., Mallet, M., Corre, L., Van Meijgaard, E., Perpiñán, O., & Gaertner, M. (2020). Future evolution of surface solar radiation and photovoltaic potential in Europe: investigating the role of aerosols. *Environmental Research Letters*, *15*(3), 034035. https://doi.org/10.1088/1748-9326/ab6666

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. In *Nature* (Vol. 585, Issue 7825, pp. 357–362). Nature Research. https://doi.org/10.1038/s41586-020-2649-2

Heisler, J., Glibert, P. M., Burkholder, J. M., Anderson, D. M., Cochlan, W., Dennison, W. C., Dortch, Q., Gobler, C. J., Heil, C. A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H. G., Sellner, K., Stockwell, D. A., Stoecker, D. K., & Suddleson, M. (2008). Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae*, *8*(1), 3–13. https://doi.org/10.1016/j.hal.2008.08.006

Hiriart-Baer, V. P., Boyd, D., Long, T., Charlton, M. N., & Milne, J. E. (2016). Hamilton Harbour over the last 25 years: Insights from a long-term comprehensive water quality monitoring program. *Aquatic Ecosystem Health and Management*, *19*(2), 124–133.

https://doi.org/10.1080/14634988.2016.1169686

Ho, J. C., Michalak, A. M., & Pahlevan, N. (2019). Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, *574*(7780), 667–670. https://doi.org/10.1038/s41586-019-1648-7

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huot, Y., Babin, M., Bruyant, F., Grob, C., Twardowski, M. S., & Claustre, H. (2007). Does chlorophyll a provide the best index of phytoplankton biomass for primary productivity studies? *Biogeosciences Discussions*, *4*, 707–745. https://hal.archives-ouvertes.fr/hal-00330232

Inomura, K., Omta, A. W., Talmy, D., Bragg, J., Deutsch, C., & Follows, M. J. (2020). A Mechanistic Model of Macromolecular Allocation, Elemental Stoichiometry, and Growth Rate in Phytoplankton. *Frontiers in Microbiology*, *11*(February), 1–22. https://doi.org/10.3389/fmicb.2020.00086

Intergovernmental Panel on Climate Change. (2001). *Climate Change 2001: Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change*.

Ito, A. (2011). A historical meta-analysis of global terrestrial net primary productivity: are estimates converging? *Global Change Biology*, *17*(10), 3161–3175. https://doi.org/10.1111/j.1365-2486.2011.02450.x

Jakkila, J., Leppäranta, M., Kawamura, T., Shirasawa, K., & Salonen, K. (2009). Radiation transfer and heat budget during the ice season in Lake Pääjärvi, Finland. *Aquatic Ecology*, *43*(3), 681–692. https://doi.org/10.1007/s10452-009-9275-2

Kaufman, Y. J., Boucher, O., Tanré, D., Chin, M., Remer, L. A., & Takemura, T. (2005). Aerosol anthropogenic component estimated from satellite data. *Geophysical Research Letters*, *32*(17), 1–4. https://doi.org/10.1029/2005GL023125

Kirillin, G., Leppäranta, M., Terzhevik, A., Granin, N., Bernhardt, J., Engelhardt, C., Efremova, T., Golosov, S., Palshin, N., Sherstyankin, P., Zdorovennova, G., & Zdorovennov, R. (2012). Physics of seasonally ice-covered lakes: A review. *Aquatic*

Sciences, *74*(4), 659–682. https://doi.org/10.1007/s00027-012-0279-y

Kong, X., Seewald, M., Dadi, T., Friese, K., Mi, C., Boehrer, B., Schultze, M., Rinke, K., & Shatwell, T. (2021). Unravelling winter diatom blooms in temperate lakes using high frequency data and ecological modeling. *Water Research*, *190*, 116681. https://doi.org/10.1016/j.watres.2020.116681

Kudela, R. M., Berdalet, E., Bernard, S., Burford, M., Fernand, L., Lu, S., Roy, S., Tester, P., Usup, G., Magnien, R., Anderson, D. M., Cembella, A., Chinain, M., Hallegraeff, G., Reguera, B., Zingone, A., & Enevoldsen, H. (2015). Harmful Algal Blooms: A scientific summary for policy makers. *Ioc/Unesco, Paris (IOC/INF-1320)*, 20. http://unesdoc.unesco.org/images/0023/002334/233419e.pdf

Lake Winnipeg DataStream. (n.d.). Retrieved April 12, 2021, from https://lakewinnipegdatastream.ca/en/

Lavers, D. A., Ralph, F. M., Waliser, D. E., Gershunov, A., & Dettinger, M. D. (2015). Climate change intensification of horizontal water vapor transport in CMIP5. *Geophysical Research Letters*, *42*(13), 5617–5625. https://doi.org/10.1002/2015GL064672

Lead, N., Magnuson, J., Carpenter, S., & Stanley, E. (2019). *North Temperate Lakes LTER: Chlorophyll - Trout Lake Area 1981 - current ver 30*. Environmental Data Initiative. https://doi.org/https://doi.org/10.6073/pasta/6c8ee65f6876a7274bfe7714ae7c3a70

Levy, S. (2017). Microcystis Rising: Why Phosphorus Reduction Isn't Enough to Stop CyanoHABs. *Environmental Health Perspectives*, *125*(2), A34–A39. https://doi.org/10.1289/ehp.125-A34

Li, K., Yu, N., Li, P., Song, S., Wu, Y., Li, Y., & Liu, M. (2017). Multi-label spacecraft electrical signal classification method based on DBN and random forest. *PLOS ONE*, *12*(5), e0176614. https://doi.org/10.1371/journal.pone.0176614

Li, R., & Li, J. (2004). Satellite Remote Sensing Technology for Lake Water Clarity Monitoring: An Overview. In *Environmental Informatics Archives* (Vol. 2).

Liaw, A., & Wiener, M. (2002). *Classification and Regression by RandomForest*. https://www.researchgate.net/publication/228451484

Liu, Y., Paciorek, C. J., & Koutrakis, P. (2009). Estimating regional spatial and temporal

variability of PM2.5 concentrations using satellite data, meteorology, and land use information. *Environmental Health Perspectives*, *117*(6), 886–892. https://doi.org/10.1289/ehp.0800123

Long, T. Y., Wu, L., Meng, G. H., & Guo, W. H. (2011). Numerical simulation for impacts of hydrodynamic conditions on algae growth in Chongqing Section of Jialing River, China. *Ecological Modelling*, *222*(1), 112–119. https://doi.org/10.1016/j.ecolmodel.2010.09.028

Maberly, S. C., Brierley, B., Carter, H. T., Clarke, M. A., De Ville, M. M., Fletcher, J. M., James, J. B., Keenan, P., Kelly, J. L., Mackay, E. B., Parker, J. E., Patel, M., Pereira, M. G., Rhodes, G., Tanna, B., Thackeray, S. J., Vincent, C. J., & Feuchtmayr, H. (2017). *Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Windermere North Basin, 1945 to 2013*. NERC Environmental Information Data Centre. https://doi.org/10.5285/f385b60a-2a6b-432e-aadd-a9690415a0ca

Makarewicz, J. C., & Bertram, P. (1991). Evidence for the Restoration of the Lake Erie Ecosystem. *BioScience*, *41*(4), 216–223. https://doi.org/10.2307/1311411

McCormick, R. A., & Ludwig, J. H. (1967). Climate modification by atmospheric aerosols. *Science*, *156*(3780), 1358–1359. https://doi.org/10.1126/science.156.3780.1358

McGlue, M. M., Ivory, S. J., Stone, J. R., Cohen, A. S., Kamulali, T. M., Latimer, J. C., Brannon, M. A., Kimirei, I. A., & Soreghan, M. J. (2020). Solar irradiance and ENSO affect food security in Lake Tanganyika, a major African inland fishery. *Science Advances*, *6*(41), 1–9. https://doi.org/10.1126/sciadv.abb2191

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

Melkozernov, A. N., & Blankenship, R. E. (2007). Photosynthetic Functions of Chlorophylls. In *Chlorophylls and Bacteriochlorophylls* (pp. 397–412). Springer Netherlands. https://doi.org/10.1007/1-4020-4516-6_28

Mendenhall, W. M., & Sincich, T. L. (2007). Statistics for engineering and the sciences, sixth edition. In *Statistics for Engineering and the Sciences, Sixth Edition* (Fifth). Pearson / Prentice Hall. https://doi.org/10.1080/00224065.2016.11918168

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume

and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*, 1–11. https://doi.org/10.1038/ncomms13603

Mohamed, M. N., Wellen, C., Parsons, C. T., Taylor, W. D., Arhonditsis, G., Chomicki, K. M., Boyd, D., Weidman, P., Mundle, S. O. C., Van Cappellen, P., Sharpley, A. N., & Haffner, D. G. (2019). Understanding and managing the re-eutrophication of lake erie: Knowledge gaps and research priorities. *Freshwater Science*, *38*(4), 675–691. https://doi.org/10.1086/705915

National Institute for Environmental Studies. (2016). *Lake Kasumigaura Database, National Institute for Environmental Studies, Japan*. https://db.cger.nies.go.jp/gem/moni-e/inter/GEMS/ database / kasumi / index.html

Neill, S. P., & Hashemi, R. M. (2018). Fundamentals of Ocean Renewable Energy - Generating Electricity from the Sea. In *Fundamentals of Ocean Renewable Energy - Generating Electricity from the Sea*. Elsevier. https://app.knovel.com/hotlink/pdf/id:kt011PCJP2/fundamentals-ocean-renewable/validation-metrics

Pahari, K., & Murai, S. (1999). Modelling for prediction of global deforestation based on the growth of human population. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(5–6), 317–324. https://doi.org/10.1016/S0924-2716(99)00032-5

Pedregosa, F., Varoquaux, G., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12, Issue 85). http://scikit-learn.sourceforge.net.

Philipson, P., Kratzer, S., Ben Mustapha, S., Strömbeck, N., & Stelzer, K. (2016). Satellite-based water quality monitoring in Lake Vänern, Sweden. *International Journal of Remote Sensing*, *37*(16), 3938–3960. https://doi.org/10.1080/01431161.2016.1204480

Pilla, R. M., Williamson, C. E., Adamovich, B. V., Adrian, R., Anneville, O., Chandra, S., Colom-Montero, W., Devlin, S. P., Dix, M. A., Dokulil, M. T., Gaiser, E. E., Girdner, S. F., Hambright, K. D., Hamilton, D. P., Havens, K., Hessen, D. O., Higgins, S. N., Huttula, T. H., Huuskonen, H., … Zadereev, E. (2020). Deeper waters are changing less

consistently than surface waters in a global analysis of 102 lakes. *Scientific Reports*, *10*(1), 1–15. https://doi.org/10.1038/s41598-020-76873-x

QGIS.org. (2021). *QGIS Geographic Information System*. QGIS Association. http://www.qgis.org

Renner, M., Wild, M., Schwarz, M., & Kleidon, A. (2019). Estimating Shortwave Clear-Sky Fluxes From Hourly Global Radiation Records by Quantile Regression. *Earth and Space Science*, *6*(8), 1532–1546. https://doi.org/10.1029/2019EA000686

Rosenfeld, D., Sherwood, S., Wood, R., & Donner, L. (2014). Climate effects of aerosol-cloud interactions. In *Science* (Vol. 343, Issue 6169, pp. 379–380). American Association for the Advancement of Science. https://doi.org/10.1126/science.1247490

Ross, M. R. V., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., Simard, M., & Pavelsky, T. M. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research*, *55*(11), 10012–10025. https://doi.org/10.1029/2019WR024883

Rudstam, L. G. (n.d.). *Limnological data and depth profile from Oneida Lake, New York, 1975 to present*. Knowledge Network for Biocomplexity. Retrieved April 12, 2021, from https://knb.ecoinformatics.org/view/kgordon.35.96

Scavia, D., David Allan, J., Arend, K. K., Bartell, S., Beletsky, D., Bosch, N. S., Brandt, S. B., Briland, R. D., Daloğlu, I., DePinto, J. V., Dolan, D. M., Evans, M. A., Farmer, T. M., Goto, D., Han, H., Höök, T. O., Knight, R., Ludsin, S. A., Mason, D., … Zhou, Y. (2014). Assessing and addressing the re-eutrophication of Lake Erie: Central basin hypoxia. In *Journal of Great Lakes Research* (Vol. 40, Issue 2, pp. 226–246). International Association of Great Lakes Research. https://doi.org/10.1016/j.jglr.2014.02.004

Schmid, M., & Köster, O. (2016). Excess warming of a Central European lake driven by solar brightening. *Water Resources Research*, *52*(10), 8103–8116. https://doi.org/10.1002/2016WR018651

Schwarz, M., Folini, D., Hakuba, M. Z., & Wild, M. (2018). From Point to Area: Worldwide Assessment of the Representativeness of Monthly Surface Solar Radiation Records. *Journal of Geophysical Research: Atmospheres*, *123*(24), 13,857-13,874.

https://doi.org/10.1029/2018JD029169

scikit-learn developers. (2019). *3.2 Tuning the hyper-parameters of an estimator*. Online Documentation for Scikit-Learn. https://scikit-learn.org/stable/modules/grid_search.html

Shatwell, T., & Köhler, J. (2019). Decreased nitrogen loading controls summer cyanobacterial blooms without promoting nitrogen-fixing taxa: Long-term response of a shallow lake. *Limnology and Oceanography*, *64*(S1), S166–S178. https://doi.org/10.1002/lno.11002

Shuvo, A., O'reilly, C. M., Blagrave, K., Ewins, C., Filazzola, A., Gray, D., Mahdiyan, O., Moslenko, · Luke, Quinlan, R., & Sharma, S. (2021). Total phosphorus and climate are equally important predictors of water quality in lakes. *Aquatic Sciences*, *83*, 16. https://doi.org/10.1007/s00027-021-00776-w

Simmons, A., Uppala, S., Dee, D., & Kobayashi, S. (2007). ERA-Interim: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter - Meteorology*, *110*, 25–35. https://www.ecmwf.int/en/elibrary/17713-era-interim-new-ecmwf-reanalysis-products-1989-onwards

Singh, S. P., & Singh, P. (2015). Effect of temperature and light on the growth of algae species: A review. In *Renewable and Sustainable Energy Reviews* (Vol. 50, pp. 431–444). Elsevier Ltd. https://doi.org/10.1016/j.rser.2015.05.024

Sommer, U., Gliwicz, Z. M., Lampert, W., & Duncan, A. (1986). The PEG_model of seasonal succession of planktonic events in fresh waters. *Arch. Hydrobiol.*, *106*(4), 433–471.

Staehr, P. A., Brighenti, L. S., Honti, M., Christensen, J., & Rose, K. C. (2016). Global patterns of light saturation and photoinhibition of lake primary production. *Inland Waters*, *6*(4), 593–607. https://doi.org/10.1080/iw-6.4.888

Stier, P., Feichter, J., Roeckner, E., Kloster, S., & Esch, M. (2006). The evolution of the global aerosol system in a transient climate simulation from 1860 to 2100. In *Atmos. Chem. Phys* (Vol. 6). www.atmos-chem-phys.net/6/3059/2006/

The pandas development team. (2020). *pandas-dev/pandas: Pandas 1.0.1*. https://doi.org/10.5281/ZENODO.3715232

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2020). *Daymet: Daily Surface Weather Data on a 1-km Grid for North*

America, Version 4. ORNL DAAC, Oak Ridge. https://doi.org/10.3334/ORNLDAAC/1840

Tian, D., Xie, G., Tian, J., Tseng, K. H., Shum, C. K., Lee, J., & Liang, S. (2017). Spatiotemporal variability and environmental factors of harmful algal blooms (HABs) over western Lake Erie. *PLoS ONE*, *12*(6), 1–16. https://doi.org/10.1371/journal.pone.0179622

Townsend, D. W., Keller, M. D., Sieracki, M. E., & Ackleson, S. G. (1992). Spring phytoplankton blooms in the absence of vertical water column stratification. *Nature*, *360*(6399), 59–62. https://doi.org/10.1038/360059a0

Vadeboncoeur, Y., Vander Zanden, M. J., & Lodge, D. M. (2002). Putting the lake back together: Reintegrating benthic pathways into lake food web models. *BioScience*, *52*(1), 44–54. https://doi.org/10.1641/0006-3568(2002)052[0044:PTLBTR]2.0.CO;2

Water Quality - Prince Albert - Open Government Portal. (n.d.). Retrieved April 12, 2021, from https://open.canada.ca/data/en/dataset/a0096ea6-6ce0-4007-b43e-9d535cd1a32c

Westberry, T., Behrenfeld, M. J., Siegel, D. A., & Boss, E. (2008). Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles*, *22*(2), 1–18. https://doi.org/10.1029/2007GB003078

Wild, M. (2009). Global dimming and brightening: A review. *Journal of Geophysical Research Atmospheres*, *114*(12), 1–31. https://doi.org/10.1029/2008JD011470

Wild, M., Ohmura, A., Schär, C., Müller, G., Folini, D., Schwarz, M., Zyta Hakuba, M., & Sanchez-Lorenzo, A. (2017). The Global Energy Balance Archive (GEBA) version 2017: A database for worldwide measured surface energy fluxes. *Earth System Science Data*, *9*(2), 601–613. https://doi.org/10.5194/essd-9-601-2017

Williamson, C. E. (2020). *Three decades of limnological data from lakes in the Pocono Mountains region, Pennsylvania USA, 1988-2019*. Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/186/3

Woolway, R. I., & Merchant, C. J. (2019). Worldwide alteration of lake mixing regimes in response to climate change. *NatuRe GeoSCIeNCe |*, *12*, 271–276. https://doi.org/10.1038/s41561-019-0322-x

Yu, L., Zhang, M., Wang, L., Lu, Y., & Li, J. (2021). Effects of aerosols and water vapour on spatial-temporal variations of the clear-sky surface solar radiation in China. *Atmospheric Research*, *248*, 105162. https://doi.org/10.1016/j.atmosres.2020.105162

Zhang, X., Lu, N., Jiang, H., & Yao, L. (2020). Evaluation of Reanalysis Surface Incident Solar Radiation Data in China. *Scientific Reports*, *10*(1), 1–20. https://doi.org/10.1038/s41598-020-60460-1

Zhong, Y., Notaro, M., Vavrus, S. J., & Foster, M. J. (2016). Recent accelerated warming of the Laurentian Great Lakes: Physical drivers. *Limnology and Oceanography*, *61*(5), 1762–1786. https://doi.org/10.1002/lno.10331

国家生态系统观测研究网络. (n.d.). Retrieved April 12, 2021, from http://www.cnern.org/index.action

Appendix I:
Summary of compiled *in situ* SSR and lake datasets

Table A1: Summary of all *in situ* SSR datasets used in this project

Data source	Parameter description	Record length	Number of stations	Region	Retrieval
Global Earth Balance Archive (GEBA), ETH Zurich (co-funded by the Federal Office of Meteorology and Climatology MeteoSwiss within the framework of GCOS Switzerland)	Monthly mean SSR, W/m^2	Varies, longest 1922 - 2017	2290	Global	Online download (May 25, 2020) from https://geba.ethz.ch (Wild et al., 2017)
Baseline Surface Radiation Network (BSRN)	Instantaneous SSR every minute, W/m^2	Varies, longest 1992 - 2020	60	Global	Online download (October 19, 2020) from https://bsrn.awi.de/ (Driemel et al., 2018)
IISD - Experimental Lakes Area (IISD-ELA)	Instantaneous photosynthetically active radiation (PAR) every 15 minutes, µmol/m^2/s	Varies, longest 1973 - 2016	5	North-western Ontario, Canada	Data request; received September 9, 2020

Table A2: Summary of all *in situ* lake datasets used in this project

Data source	Parameter description	Record length	Number of lakes	Region	Retrieval
ECCC Freshwater Inventory and Surveillance of Mercury: Water Quality Data (Contains information licensed under the Open Government Licence – Canada)	Chlorophyll-*a*, pH, DIC, heavy metals	2008 - 2016	20	Canada	Online download from http://data.ec.gc.ca/data/substances/monitor/clean-air-regulatory-agenda-freshwater-inventory-and-surveillance-of-mercury-cara-fishg/freshwater-inventory-and-surveillance-of-mercury-water-quality-data/ (*Freshwater Inventory and Surveillance of Mercury: Water Quality Data*, n.d.)
ECCC Lake Erie Satellite-derived Daily Algal Bloom Indices (Contains information licensed under the Open Government Licence – Canada)	Water quality Index, algal bloom % lake coverage and intensity	2002 - 2018	1	Lake Erie, Southern Ontario, Canada	Online download from https://open.canada.ca/data/en/dataset/0e12e786-b199-4716-891d-529b9f43f904 (Binding et al., 2018)
Lake Winnipeg Datastream	Depth, N, ammonia, P, DIC, DOC, chlorophyll-*a*, pH, alkalinity	2002 - 2020	1	Lake Winnipeg, Manitoba, Canada	Online download from https://lakewinnipegdatastream.ca/ (*Lake Winnipeg DataStream*, n.d.)

IISD - ELA	Chlorophyll-*a* (epilimnion composite sample), DOC, temperature	1968 - 2019	5	North-western Ontario, Canada	Data request; received September 9, 2020
ECCC Great Lakes Water Quality Monitoring and Surveillance Data (Contains information licensed under the Open Government Licence – Canada)	Chlorophyll-*a*, water temperature, N, P, Depth	2000 - 2018	5	Southern Ontario, Canada	Online download from https://open.canada.ca/data/en/dataset/cfdafa0c-a644-47cc-ad54-460304facf2e (*Great Lakes Water Quality Monitoring and Surveillance Data - Open Government Portal*, n.d.)
ECCC Hamilton Harbour Water Quality Data (Contains information licensed under the Open Government Licence – Canada)	Depth, chlorophyll-*a*, DIC, DOC, P, N	1987 - 2018	1	Lake Ontario, Southern Ontario, Canada	Online download from https://open.canada.ca/data/en/dataset/c50e3bb8-97f5-48be-a910-a8a7b59f85ff (Hiriart-Baer et al., 2016)
ECCC Clear Lake Water Quality - Riding Mountain (Contains information	Depth, chlorophyll-*a*, DO, pH, Secchi depth, conductivity	1978 - 2018	1	Clear Lake, Alberta, Canada	Online download from https://open.canada.ca/data/en/dataset/2a55313f-26fc-4872-9a57-2a7bf2a4cc38 (*Clear Lake Water Quality*

licensed under the Open Government Licence – Canada)					- *Riding Mountain - Open Government Portal*, n.d.)
ECCC Water Quality - Prince Albert (Contains information licensed under the Open Government Licence – Canada)	Chlorophyll-*a*, DOC, TP, temp, secchi depth	1992 - 2019	2	Saskatchewan, Canada	Online download from https://open.canada.ca/data/en/dataset/a0096ea6-6ce0-4007-b43e-9d535cd1a32c (*Water Quality - Prince Albert - Open Government Portal*, n.d.)
Environmental Protection Agency (EPA) Great Lakes Environmental Database System (GLENDA)	Chlorophyll-*a*	1983 - 2015	1	Lake Erie, Southern Ontario, Canada	Provided by project partners
UK Centre for Ecology and Hydrology	Water temperature, pH, chlorophyll-*a*, P, N, Secchi depth, alkalinity, misc ions	1945 - 2013	7	United Kingdom	Online download from https://catalogue.ceh.ac.uk/eidc/documents (Maberly et al., 2017)
National Ecosystem Research Network of China	Chlorophyll-*a*, temperature, pH, N, P, DO, misc ions, depth	2005 – 2019	1	Lake Taihu, China	Online download from http://www.cnern.org/index.action (国家生态系统观测研究网络, n.d.)

Environmental Information System of the State Institute for the Environment Baden-Württemberg (LUBW)	Chlorophyll-*a*	1998 – 2019	1	Lake Bodensee, Germany	Online download from https://udo.lubw.baden-wuerttemberg.de/public/index.xhtml
Kasumigaura Long-term Monitoring Project of the National Institute for Environmental Studies, Japan	Secchi, OC, N, C/N, Depth, chlorophyll-*a*	1977 - 2018	1	Lake Kasumigaura, Japan	Online download from https://db.cger.nies.go.jp/gem/inter/GEMS/database/kasumi/contents/datalist.html (National Institute for Environmental Studies, 2016)
Norwegian Institute for Water Research (NIVA)	Chlorophyll-*a*, depth	1976 – 2020	1	Lake Mjøsa, Norway	Online download from http://www.aquamonitor.no/Portal/
Water Information System Sweden (VISS)	Chlorophyll-*a*	1945 - 2020	706	Sweden	Data request; received July, 2020
UK Environment Agency Water Quality Archive (Beta)	Chlorophyll-*a*, P, N, water temperature, misc ions	2000 - 2020	550	United Kingdom	Online download from https://environment.data.gov.uk/water-quality/view/download

Aquasat	Chlorophyll-*a*, DOC, Secchi, temperature	1965 - 2018		North America	Online download from https://figshare.com/articles/dataset/AquaSat/8139383 (Ross et al., 2019)
LTER – Cascade Lakes Project	depth, DIC, chlorophyll-*a*	1992 – 2016	6	Michigan, USA	Online download from https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-ntl.354.4 (Carpenter et al., 2017)
Oneida Lake, New York	water temperature, TDS, pH, chlorophyll-*a*, Secchi depth, P, N, alkalinity, conductivity	1975 - 2017	1	Oneida Lake, New York, USA	Online download from https://knb.ecoinformatics.org/view/kgordon.35.96 (Rudstam, n.d.)
LTER – Trout Lake Area	Chlorophyll-*a*, depth	1981 - 2018	1	Trout Lake, Wisconsin, USA	Online download from https://lter.limnology.wisc.edu/node/55078 (Lead et al., 2019)
Poconos Mountains Region limnological data	depth, chlorophyll-*a*, water temperature, N, P, alkalinity, etc	1989 - 2018	3	Pennsylvania, USA	Online download from https://search.dataone.org/view/https%3A%2F%2Fpasta.lternet.edu%2Fpackage%2Fmetadata%2Feml%2Fedi%2F186%2F3 (Williamson, 2020)

Appendix II:
Example Python code for random forest model

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jan 10 01:31:18 2021

@author: Rahim Barzegar and Jane Ye
"""
# Imports
from sklearn.model_selection import train_test_split
import numpy as np
import pandas as pd
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
import pickle
warnings.filterwarnings("ignore")



# Fix random seed for reproducibility
np.random.seed(1234)


# Load data
df =
  pd.read_csv('../../data/processed_data/MLfiles/SSRLakesML_sat_1deg_21040
  7.csv',  header=0)


# Filter by days_in_window to exclude inaccurately calculated growth
  windows
df = df.loc[(df['days_in_wi'] < 56) & (df['days_in_wi'] > 2)]


# Print number of rows, number of unique lakes, number of unique SSR
  stations
print(df.shape[0])
print(df['Lake'].nunique())
```

```
print(df['SSRID'].nunique())


# Define parameter set from input df
df = df[['mean_ssr', 'mean_temp', 'mean_turbi', 'chla_rate', 'trophic_st',
   'lat', 'GeoDistKm', 'Volume', 'MeanDepth', 'LakeSSRElevDiff']]


# drop nan values in mean ssr and 7 days previous mean ssr columns
df = df.dropna()



#########################################################################
# Dealing with non-numerical (categorical) variables:
# The mask identifies object type columns
# label encoder assigns a number to the categories
# (ie, 4 different categories, they get assigned a # from 0-3)


# Define numerical columns
num_cols = ['mean_ssr', 'mean_temp', 'mean_turbi', 'chla_rate', 'lat',
'GeoDistKm', 'Volume', 'MeanDepth', 'LakeSSRElevDiff']


# Define categorical columns
cate_cols = df.columns.drop(num_cols)


# convert numerical data
df[num_cols] = df[num_cols].apply(pd.to_numeric, errors='coerce')


# Define X and y sections of the data
X = df.drop(columns=['chla_rate'])
y= df['chla_rate']


# Apply categorical boolean mask
categorical_feature_mask = X.dtypes==object


# Filter categorical columns using mask and turn it into a list
categorical_cols = X.columns[categorical_feature_mask].tolist()
```

```
# Initiate labelencoder object
le = LabelEncoder()


# Apply le on categorical feature columns
df[categorical_cols] = X[categorical_cols].apply(lambda col:
  le.fit_transform(col))



#############################################################################
# Assign features and target, train and test values


# Define predictor parameter set
features = ['mean_temp', 'mean_ssr', 'mean_turbi', 'trophic_st', 'lat',
  'GeoDistKm',
           'Volume', 'MeanDepth', 'LakeSSRElevDiff']


# Define target parameter
target = ["chla_rate"]


# Define train and test values
X = df[features].values
y = df[target].values.ravel()


train_X, test_X, train_y, test_y = train_test_split(X, y, test_size=0.2,
  random_state = True)


#############################################################################
# Set up cross validation by grid search and do model fitting/prediction


# Set the parameters for our grid search using bagging
params = {'n_estimators': range(1,20,1),
          'max_depth':range(1,10,1),
          'min_samples_split':range(2,10,1),
          'max_features': ['auto', 'sqrt','log2'],
          'min_samples_leaf':range(1,10,1),
         }
```

```
# Initialize a GridSearchCV with 10-fold cross validation for the Bagging
  Decision Tree Classifier
bdtc = RandomForestRegressor(random_state=0)
model = GridSearchCV(bdtc,params,cv=10,n_jobs=6,verbose=1)


model.fit(train_X,train_y)


# Save model estimator, train and test X and Y, to external file
# External files can be read in to create plots and data visualizations
with open('referencefiles/model_sat1deg_210407_duringgwtempssr.pickle',
  'wb') as handle:
    pickle.dump(model, handle, protocol=pickle.HIGHEST_PROTOCOL)


with open('referencefiles/features_sat1deg_210407_duringgwtempssr.pickle',
  'wb') as handle:
    pickle.dump(features, handle, protocol=pickle.HIGHEST_PROTOCOL)


np.savetxt('referencefiles/trainx_sat1deg_210407_duringgwtempssr.csv',
  train_X, delimiter=',')
np.savetxt('referencefiles/trainy_sat1deg_210407_duringgwtempssr.csv',
  train_y, delimiter=',')
np.savetxt('referencefiles/testx_sat1deg_210407_duringgwtempssr.csv',
  test_X, delimiter=',')
np.savetxt('referencefiles/testy_sat1deg_210407_duringgwtempssr.csv',
  test_y, delimiter=',')
```