THE UNIVERSITY OF EDINBURGH

SCHOOL OF GEOSCIENCES

Investigating the seasonality and environmental drivers of phytoplankton blooms in Loch Leven

BY

Koh Liang Sze Ariel

in partial fulfilment of the requirement for the

Degree of BSc with Honours in

Ecological and Environmental Sciences *(and with Management)*

5 May 2025

Abstract

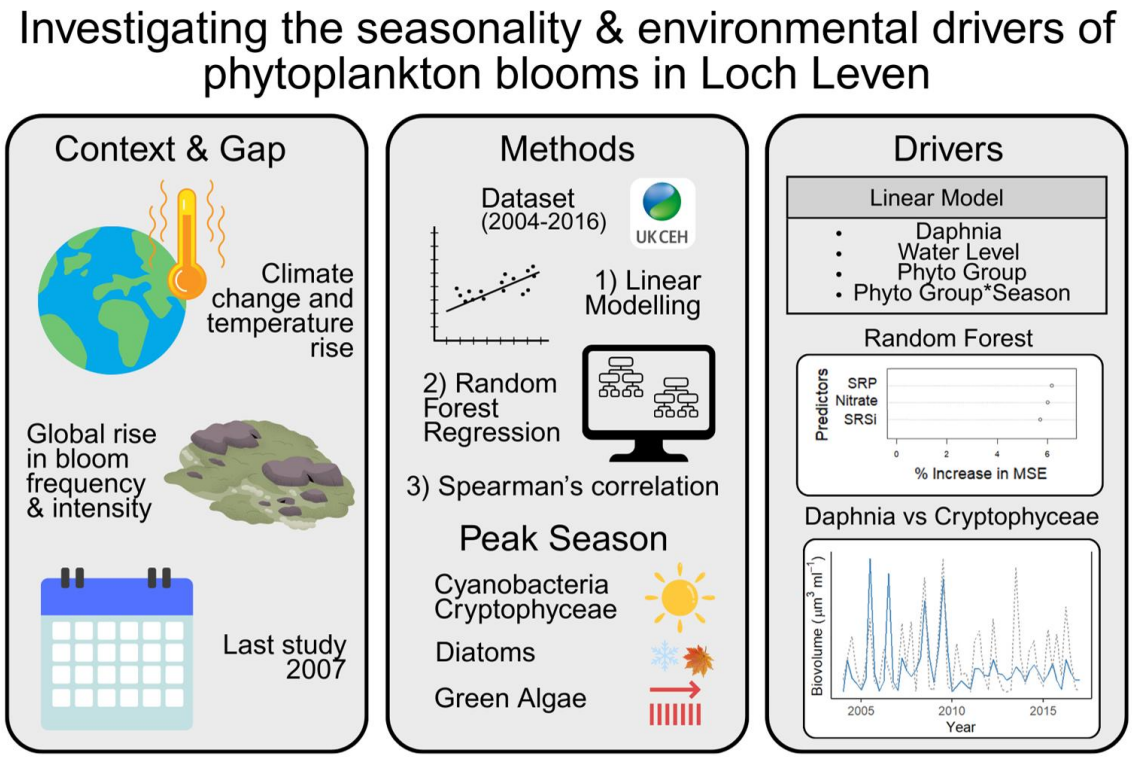
Phytoplankton blooms are intensifying globally due to climate change and anthropogenic pressures, posing risks to water quality, biodiversity, and public health. Despite extensive research on bloom drivers, few studies have integrated both traditional modelling and machine learning approaches. Additionally, the most recent study of phytoplankton dynamics in Loch Leven includes data only up to 2007, leaving a critical gap in understanding recent trends. This study aims to address these gaps by investigating the seasonality and environmental drivers of phytoplankton blooms in Loch Leven, the UK's largest shallow lake.

Long-term monitoring data from the UKCEH, spanning 2004 to 2016, was used. Linear mixed-effects models assessed seasonal variation in phytoplankton biovolume, with post-hoc Tukey tests conducted where appropriate. To identify key environmental drivers, both linear models and random forest regression were employed. Spearman's correlation analysis explored the relationship between Daphnia density and phytoplankton biovolume as a proxy for top-down control.

The results revealed group-specific seasonal biovolume patterns: Cyanobacteria and Cryptophyceae peaked in summer, Diatoms in winter, and Green Algae remained relatively stable. Linear models identified *Daphnia* density, water level, phytoplankton group, and their interaction with season as significant predictors, while random forest models highlighted soluble reactive phosphorus, nitrate, and soluble reactive silica as the most important drivers. Correlations between *Daphnia* and phytoplankton groups were weak and variable, suggesting limited top-down control.

Overall, this study highlights the complex nature of bloom dynamics in Loch Leven, demonstrating the value of combining traditional modelling and machine learning to inform integrated freshwater ecosystem management.

Graphical Abstract



Graphical abstract was designed using Canva (www.canva.com).

Table of Contents

Acknowledgements

A big thank you to my supervisor, Dr James Watt. I am immensely grateful for all your invaluable guidance, support and encouragement throughout my dissertation.

I would like to thank the UK Centre for Ecology & Hydrology (UKCEH) for providing access to long-term monitoring data from Loch Leven. I am especially grateful to Dr Linda May, Philip Taylor, and Dr Toni Dwyer for their patience in answering my many questions.

To my flatmates and friends – thank you for being my support system. For the study sessions, quick lunch breaks, never-ending flat dinners, and pub crawls: I am very lucky to have found lifelong friends. I treasure every single late-night conversation that carried me through the past 4 years.

Mum, Dad, Matt, Po, and Gong – thank you for being my foundation and believing in my potential. Thank you for giving me the opportunity to go to Edinburgh and truly spread my wings. I quite literally would not be where I am now without your constant love and support.

Shoutout to my Kor Kor for paving the way with statistics and teaching me how important they are. Thank you for being there every time Rstudio was making me crash out – and being the only person proofreading this dissertation who actually understood what was going on. I am so grateful to have you as my role model.

To Jack, thank you for your endless support – from my initial dissertation topic software fail to the completion of this one. You kept me grounded when everything seemed impossible and believed in me even when I didn't believe in myself. This journey was far more manageable because of you.

And finally, as my 4 years of university comes to an end, I would like to thank myself. The little girl who once dreamt of studying at a top university overseas has done it. I'm proud of everything I've achieved, and proud of the person I've become.

List of Abbreviations

%IncMSE – Percent Increase in Mean Squared Error

AIC – Akaike's Information Criterion

DBPs – Disinfection By-products

DO – Dissolved Oxygen

DOM – Dissolved Organic Matter

HABs – Harmful Algal Blooms

NO_3 – Nitrate

P – Phosphorus

R^2 – Coefficient of Determination

RMSE – Root Mean Squared Error

SRP – Soluble Reactive Phosphorus

SRSi – Soluble Reactive Silica

TP – Total Phosphorus

UKCEH – UK Centre for Ecology & Hydrology

VIF – Variance Inflation Factor

1. Introduction

Freshwater ecosystems are one of the most vital and vulnerable environments on Earth (Albert et al., 2021). Providing essential both anthropogenic and naturogenic ecosystem services to both humans and wildlife – including food, habitat, drinking water, and recreation – they represent less than 1% of the Earth's surface (Baron et al., 2002; Reid et al., 2020). They contribute a huge amount of biodiversity and are essential to societal well-being (Lynch et al., 2023). However, freshwater resources are increasingly coming under threat from a range of environmental stressors – one of the most pressing being the rise in occurrence and intensity of harmful algal blooms (Brooks et al., 2016). These events not only disrupt aquatic food webs but also pose risks to water quality, human health, and ecosystem function (Kudela, Berdalet and Urban, 2015).

1.1 Ecological Significance and Diversity of Phytoplankton

Phytoplankton are key primary producers that form the foundation of aquatic ecosystems (Field et al., 1998); they are generally microscopic, autotrophic organisms (Falkowski, 2012). Despite constituting less than 1% of the Earth's photosynthetic biomass, they contribute an estimated 50% of global primary production and net oxygen production (Monier et al., 2015). Through their photosynthetic activity, they help regulate atmospheric carbon dioxide and play a major role in global biogeochemical cycles, including carbon, nitrogen, and phosphorus cycling (Williamson, Saros and Schindler, 2009). Forming the base of most aquatic food webs, phytoplankton are a vital energy source for every level of the aquatic food web, from zooplankton to larger invertebrates and fish (Shurin, Gruner and Hillebrand, 2005). As such, their productivity affects food availability at higher trophic levels, which directly influences the distribution, structure and function of aquatic ecosystems (Adrian et al., 2009).

Phytoplankton communities are taxonomically and functionally diverse (Field et al., 1998; Falkowski, 2012). They encompass a range of major groups including Cyanobacteria, Cryptophyceae, Diatoms and Green Algae. These groups differ in key traits such as cell physiologies, biochemical functions and ecological strategies (Beardall et al., 2009). This diversity allows phytoplankton to occupy a wide range of ecological niches and respond differently to environmental fluctuations such as nutrient availability, temperatures and grazing pressure (Vallina et al., 2017).

Phytoplankton are also key indicators of aquatic ecosystem health (Suthers and Rissik, 2009; Zhao, Drakare and Johnson, 2019). Shifts in phytoplankton community structure – such as variation in dominant groups or species – can reflect changes in nutrient dynamics, water quality or broader climate conditions (Jeppesen et al., 2005). Due to their rapid growth and short life cycles, phytoplankton respond quickly to environmental changes, often making them

one of the first biological indicators to reflect ecological disturbances (Rao et al., 2021). This responsiveness, coupled with their ecological importance, makes phytoplankton ideal candidates for long-term monitoring to inform lake management strategies and decisions (Zhao, Drakare and Johnson, 2019).

1.2 Harmful Algal Blooms: Causes and Consequences

Under certain environmental conditions, phytoplankton biovolumes can multiply exponentially, leading to algal blooms (Ebert et al., 2001). While not all algal blooms are harmful, some, termed harmful algal blooms (HABs), have led to negative ecological, economic and public health consequences (Hudnell, 2010; Ekstrom, Moore and Klinger, 2020).

HABs can cause severe ecological damage by creating hypoxic conditions, in which oxygen levels in the water column are severely depleted (Sun et al., 2022). This typically occurs when there is a rapid die-off of dense algal blooms, triggering high rates of decomposition which consumes large amounts of oxygen from the water column creating “dead zones” (Arend et al., 2011). Furthermore, dense blooms can prevent sunlight from penetrating the water column and reaching submerged aquatic plants, reducing photosynthesis and further worsening oxygen depletion (Griffith and Gobler, 2020).

HABs also pose a significant health risk due to the presence of toxin-producing species. Cyanobacteria such as *Microcystis*, *Anabaena*, and *Planktothrix* can release harmful compounds into the water, posing threats not only to aquatic life, but also livestock and humans (Jöhnk et al., 2008; Zanchett and Oliveira-Filho, 2013). When blooms collapse, the rapid die-off of phytoplankton results in the dissolved organic matter (DOM) concentration rapidly increasing (Ji et al., 2024). This is particularly concerning in lakes that serve as sources of drinking water, as during the drinking water treatment the elevated DOM levels lead to increased formation of disinfection by-products (DBPs) like trihalomethanes (Foreman et al., 2021). These compounds are potentially carcinogenic, raising serious concern for drinking water safety and public health (Moreira, Vasconcelos and Antunes, 2022; Tazkiaturrizki, Hartono and Moersidik, 2023).

The frequency and intensity of HABs have risen dramatically over the past decade, largely driven by climate change and increased anthropogenic pressures (Brooks et al., 2016). A rise in global temperatures has been linked to reduced vertical mixing, enhanced thermal stratification, and extended growing seasons – all of which favour bloom-forming taxa, particularly Cyanobacteria (Paerl & Paul, 2012). Wang et al. (2025) conducted a global analysis of large lakes and found a significant increase in algal blooms frequency over the past two decade. The study found that the increase in phytoplankton blooms was closely related with rising surface water temperatures, creating more favourable conditions for bloom development. Even lakes with relatively low nutrient levels are experiencing increased bloom

occurrences, emphasising the impact of climate warming on freshwater ecosystems. This observation is consistent with findings from Lake of the Woods, Canada, where Paterson et al. (2017) reported intensifying Cyanobacterial blooms linked to rising summer temperatures and extended periods of thermal stratification, despite declines in phosphorus inputs from the Rainy River.

1.3 Environmental Drivers of Phytoplankton Blooms

With modelling studies projecting that climate change will continue to increase lake water temperatures and stratification (Mullin et al., 2020), it is important to understand the complex environmental drivers of HABs. Freshwater lakes are ideal environments for bloom formation due to their relatively closed hydrological systems, shallow morphometry, and limited water exchange, which encourage the nutrient accumulation and increased stratification (Paerl, 1996). As sources of drinking water, recreation, and biodiversity, lakes provide many ecosystem services, making managing and understanding phytoplankton dynamics within them extremely important.

There are a range of chemical, physical, and biological drivers of phytoplankton growth in freshwater lakes; they interact in complex ways across seasons and lake systems (Nõges et al., 2010). Many studies have highlighted nutrients as one of the main chemical drivers of phytoplankton blooms (Rigosi et al., 2014). Currently, it is generally agreed that phosphorus (P), is the main nutrient limiting phytoplankton production in most shallow lakes (Schindler, 1977; May et al., 2012). Long-term studies of lakes such as Lake Erie, one of the United States of America's Great Lakes, have demonstrated clear association between P concentration and the occurrence and intensity of blooms (Michalak et al., 2013).

Recent research has also increasingly pointed to the role of nitrate as a driver of algal bloom, and even more so, the possibility of nutrient co-limitation in lakes (Andersen et al., 2020). Nitrate, playing a vital role in protein synthesis and metabolism, can become the key limiting factor determining bloom intensity and composition in systems with high levels of P concentration (Howarth et al., 2021). In the long-term study of Lake Taihu, a large, shallow eutrophic lake in China, Xu et al. (2010) found that both nitrogen and phosphorus availability significantly influenced phytoplankton bloom development.

In addition to nutrient availability, as earlier noted, physical drivers such as temperature and stratification are increasingly recognised as key contributors to algal bloom dynamics (Paerl & Paul, 2012). While these drivers are influenced by climate change, they also exhibit strong seasonality. Warmer temperatures increase the metabolic rates of phytoplankton, encouraging bloom formation (Jöhnk et al., 2008). The increase in temperatures during the summer season also limits vertical mixing, leading to stratification in lakes (Wilkinson, Hondzo and Guala, 2019). This traps nutrients in the upper water layers, which in combination with the penetration

of sunlight, provide ideal environments for phytoplankton growth (Mullin et al., 2020). Such conditions have been shown in studies to lead to more frequent and intense algal blooms during warmer months (Ho and Michalak, 2020).

The role of biological drivers should also not be overlooked. Many studies have demonstrated that zooplankton, particularly large-bodied filter feeders such as *Daphnia*, can exert significant top-down control on phytoplankton biomass through grazing (Jeppesen et al., 1997). In systems where edible phytoplankton dominate, high *Daphnia* densities can suppress bloom formation by consuming phytoplankton faster than they grow, thereby stabilising phytoplankton populations (Muylaert et al., 2010).

These drivers are not temporally constant and do not act in isolation – chemical, physical, and biological drivers often vary seasonally and interact in complex ways (Nõges et al., 2010). These interactions can be synergistic, where the combined effects are greater than the sum of individual drivers, or antagonistic, where one driver counteracts the effects of another (Piggott, Townsend and Matthaei, 2015). For example, Richardson et al. (2018) found that in medium-to-high alkalinity humic lakes, total phosphorus and temperature had a synergistic relationship, resulting in a significant increased Cyanobacteria biovolume. The possibility of these drivers synchronising or desynchronising emphasises the importance of considering temporal overlap in environmental conditions. This temporal variability and complex interplay make it difficult to predict phytoplankton bloom dynamics using single-driver approaches (Spears et al., 2021).

1.4 Modelling Approaches for Phytoplankton Dynamics

Despite advances in conceptual understanding, management efforts often continue to rely on single-driver approaches. These can lead to unexpected ecological outcomes in systems where multiple drivers interact (Spears et al., 2021). Thus, it is important to apply modelling approaches that can account for both direct and interactive effects of environmental variables on phytoplankton communities.

Most traditional studies rely on linear modelling. Linear models are widely used in ecological research due to their transparency and ease of interpretation (Love et al., 2023). They allow researchers to assess the significance and direction of relationships between predictors and response variables (James et al., 2013). It is through the use of linear models that studies have identified factors such as nutrient loading, temperature, or grazing pressure as key environmental drivers of phytoplankton bloom. For example, Jeppensen et al. (2005) used linear and generalised linear models to evaluate nutrient thresholds for algal blooms in European lakes. However, linear models are constrained by several factors – including assumptions of linearity, additivity, and independence, etc – which may lead to difficulties in

capturing the complex, interactive dynamics of aquatic ecosystems (James et al., 2013; Rimpler, Kiers and Van Ravenzwaaij, 2025).

Given recent advances in computational techniques, machine learning algorithms – specifically ensemble methods like random forests – have emerged as powerful tools for analysing ecological data (Liu et al., 2023). Random forests are decision tree-based ensemble models that can account for non-linear, high-dimensional, and collinear datasets without strict parametric assumptions (Breiman, 2001). In recent studies, random forest models have been successfully used to model phytoplankton blooms frequency, intensity, and risk, and evaluate the relative importance of multiple, potentially interacting environmental predictors (Cheng et al., 2021; Liu et al., 2023).

Liu et al. (2023) employed random forest models to predict phytoplankton dynamics in Lake Mjøsa, Norway. Their study found nutrient concentration and physical drivers – such as water temperatures, and lake inflow – to be the most important predictors phytoplankton biovolume. Their results also revealed complex interactive effects that highlight the value of machine learning in analysing the interaction of multiple drivers in aquatic ecosystems.

However, random forest models are not without their weaknesses – they lack explicit coefficients and hypothesis-testing frameworks. Nonetheless, when used in tandem with traditional approaches, random forest can complement linear models providing deeper insights into ecosystem dynamics.

1.5 Loch Leven: a Model System for Freshwater Bloom Research

Loch Leven, the largest shallow freshwater lake in the UK, has long been regarded as a model system for researching nutrient enrichment, restoration, and long-term ecological changes in freshwater lakes (Kirby, 1971; Ferguson et al., 2007). With its extensive monitoring and research record, spanning over five decades, Loch Leven has become a cornerstone for understanding the process and drivers of eutrophication as well as lake management (May and Spears, 2012). Due to its shallow depth and high surface-area-to-volume ratio, Loch Leven has been referred to as a sentinel system, responding rapidly to environmental pressures (Adrian et al., 2009). This makes it a valuable early warning indicator of broader catchment-level and climate-driven stressors.

Having experienced repeated episodes of nutrient-driven phytoplankton blooms since the 1960s, the primary drivers of Loch Leven blooms have historically been identified as excessive phosphorus loading from sewage inputs, agricultural run-offs, and industrial discharges (Bailey-Watts and Kirika, 1987; Holden and Caines, 1974; Spears et al., 2011; May et al., 2012). Beginning in the 1980s, restoration efforts, focused on improvements in catchment management and wastewater treatment, have successfully reduced external P inputs (May et

al., 2012). Despite these reductions, the presence of internal P loading has slowed down the recovery in water quality of the loch (May et al., 2012).

More recent studies on phytoplankton in Loch Leven suggest that interannual weather variation and climate change are exerting increasing influence on bloom dynamics (Carvalho et al., 2012). Furthermore, Gunn et al. (2012) found changes in the structure and timing of phytoplankton blooms, with increasing summer dominance of bloom-forming groups, such as Cyanobacteria and Cryptophyceae. These patterns highlight the influence of multiple, interacting stressors – including nutrient inputs and climatic variables – making Loch Leven an ideal study site for analysing the complex environmental drivers shaping phytoplankton communities in shallow ecosystems.

1.6 Knowledge Gap, Aim and Research Questions

While many studies have investigated the impact of multiple stressors on shallow lakes ecosystems, including nutrient enrichment and climate change, these studies often employ either traditional statistical models or machine learning techniques independently (Rigosi et al., 2014; Liu et al., 2023). There is limited research that integrate both techniques to provide a deeper understanding of drivers influencing phytoplankton biovolumes in shallow lakes.

Furthermore, despite the extensive research on Loch Leven's phytoplankton dynamics, the most recent comprehensive study covers data only up to the year 2007 (Carvalho et al., 2012). This temporal gap highlights the need for updated analysis that incorporates more recent data and novel modelling techniques to better assess ecological conditions and inform lake management strategies and decisions.

Hence, this study aims to investigate the seasonality and environmental drivers of phytoplankton blooms in Loch Leven. This aim will be achieved by answering the following research questions:

(1) How does seasonality influence the occurrence of algal blooms within Loch Leven?

(2) What are the chemical and physical drivers of algal blooms in Loch Leven?

(3) How does environmental drivers influence the composition of phytoplankton communities within Loch Leven?

2. Methods

2.1 Study Site

Loch Leven, situated in Perth and Kinross (56.20019 °N, 3.37963 °W), is the largest shallow lake in the United Kingdom (Fig.1.1). It has an area of 13.3 km^2, mean depth of 3.9 m and maximum depth of 25.5 m (Kirby, 1971). Details on its structure and physical environment are described by Smith (1974). The catchment area, spanning about 145 km^2 and reaching a maximum altitude of 482 m above ordnance datum, is relatively rural but intensively farmed (LLCMP, 1999). There is limited livestock rearing in the area, tending to be further away from the lake, mainly restricted (LLCMP, 1999).

Figure 1.1 Location of Loch Leven within Perth and Kinross Council, Scotland. Map created using Digimap.

The loch has four main inflow sources – North Queich, South Queich, Gairney Water and Green Burn – which provide about two-thirds of the total inflow into the loch, with the remaining coming from smaller tributaries and direct runoff from the surrounding catchment areas (May *et al.*, 2012; May, 2018; Fig. 1.2). In 1850, sluice gates were installed in Loch Leven's only outflow point, allowing them to regulate the outflow from the loch, and thus its overall water level (CEH, no date; Fig 1.3).

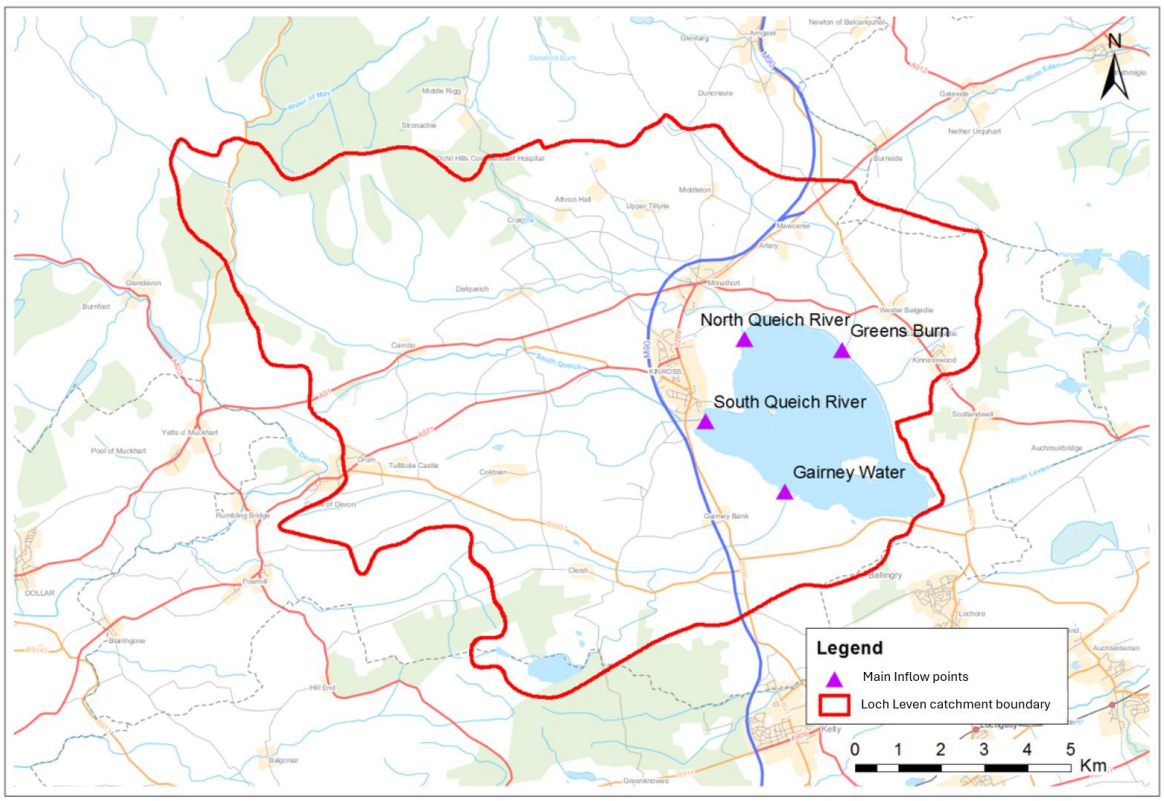


Figure 1.2 Map of the Loch Leven catchment area, showing the four main inflow sources: the North Queich River, South Queich River, Gairney Water, and Greens Burn. The catchment boundary is outlined in red, with inflow points marked in purple triangles. Adapted from May (2018).

The loch has a well-documented history of water quality issues resulting from anthropogenic induced eutrophication (May et al., 2012). Beginning in 1963, the most prominent indicator of this was an increase in the occurrence of Cyanobacteria blooms (Holden and Caines, 1974; Morgan, 1974). By the mid-1980s, the severity of these blooms had escalated to a level that threatened the loch's high conservational, recreational and economic value (May and Spears, 2012). By this time, these blooms were widely recognised to be due to the high inputs of P into the loch (LLCMP, 1999).

The areas of intensive agriculture (Spears et al., 2011) as well as sewage treatment works within the catchment was a major source of P entering the loch (Bailey-Watts and Kirika, 1987). To reduce the nutrient load from these sources and address elevated P concentrations, stricter controls on agriculture run-offs, effluent diversion measures and tertiary treatment was introduced in the 1980s to 1990s (May et al., 2012).

Although there is currently little industry within the catchment area, there have been >1 woollen mills on the lake banks since 1840 (Munro, 1994). These mills have historically been a major source of P-rich effluent into the loch (Holden and Caines, 1974). After peaking in the 1960s and early 1970s, these inputs have been successfully reduced, with the mills ceasing their use of P-based materials in 1989 (May et al., 2012).

These measures have found relative success, with the total phosphorus (TP) loading into the loch decreasing from 20 t P y^{-1} in 1985 to 8 t P y^{-1} in 1995 (Bailey-Watts and Kirika, 1999), remaining at this level in 2005 (May et al., 2012). Despite this, the loch still struggles due to high levels of internal loading, delaying the loch's recovery in terms of in-lake water quality parameters (May et al., 2012).

2.2 Water Sampling, Storage and Analysis

This study's field and laboratory methods were adapted from methods detailed in Carvalho et al (2012), the most recent comprehensive study of phytoplankton dynamics in Loch Leven. Loch Leven hosts one of the longest running lake monitoring programmes in the world, beginning in 1968 (May and Spears, 2012). All environmental and phytoplankton data were obtained from the UKCEH, who runs the long-term monitoring of Loch Leven.

Sampling data for this study was collected between February 2004 and December 2016, this period was selected as it represents the most accurate and reliable recent record of phytoplankton biovolume. Due to sampling consistency and data availability, this study only used data collected from the Reed Bower sampling point (56.1950° N, -3.3945° W; Fig. 1.3).

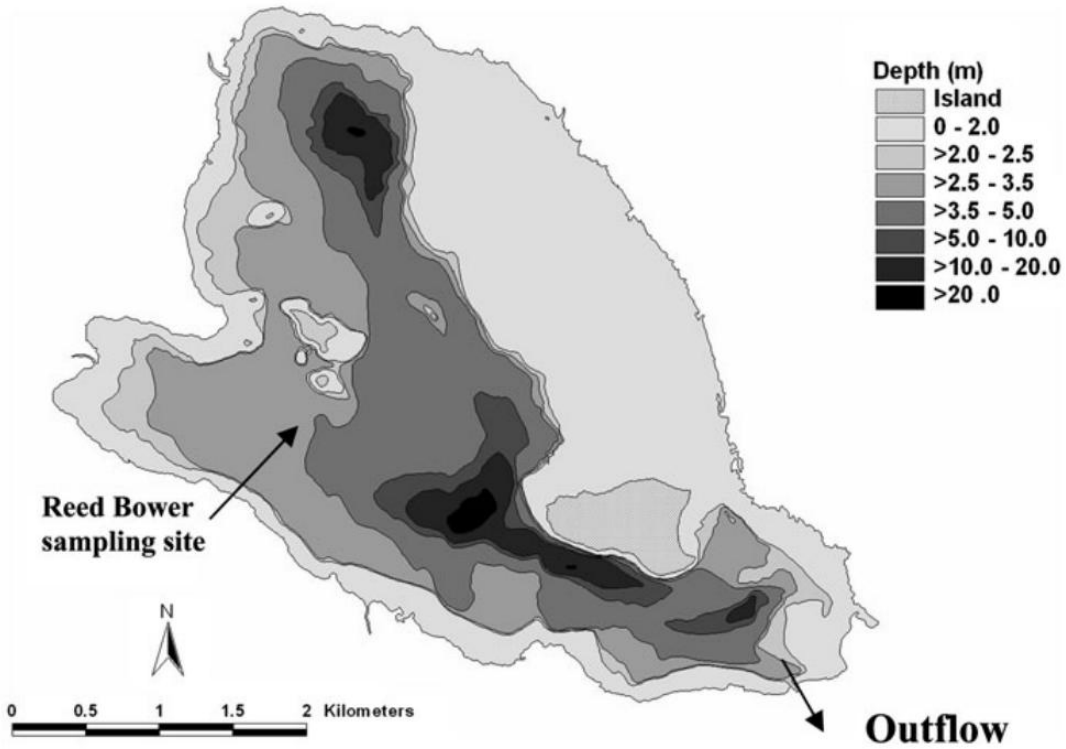


Figure 1.3 Bathymetric map of Loch Leven showing depth gradients (in m), the location of the Reed Bower sampling site, and the lake outflow. Adapted from Spears et al. (2012).

Water sampling is generally conducted on a weekly to fortnightly basis. Water level is measured as the distance from the top of the pier to the surface of the water at Kinross Pier using a metre rule. Water chemistry and phytoplankton sampling was conducted by boat from

a mid-basin station near Reed Bower. Conductivity, pH, temperature and dissolved oxygen (DO) were measured in situ using hand-held electronic probes. Duplicate integrated water samples (between the water surface and about 0.25 m above the lakebed) were collected using a weighted polythene tube. Due to fluctuations in water levels, sampling depths ranged between 3.0 and 3.5 m. A Secchi disk was used to measure water clarity at Reed Bower station.

Phytoplankton surface dips were carried out with a 250 ml wide-mouth screw top bottle attached to the end of an open-end '28-micron' mesh net (210 mm net frame). The net was lowered into the water column (~1m depth) and raised to collect the sample. Excess water was tipped out, before the net was rinsed with deionised water to transfer all material into the bottle. In the laboratory, the samples were divided into two 100 ml glass bottles for preservation. One was preserved with 37% formalin (10 ml per 100ml sample), and the other with 15% Lugol's iodine (typically 8-10 drops per 100ml sample), as per standard protocol for preserving phytoplankton samples (ASTM International, 2012).

Phytoplankton identification and biovolume estimation were carried out using the Utermöhl method with an inverted microscope (Zeiss Axiovert 40 CFL) equipped with phase contrast illumination and a Zeiss Axiocam 305 colour camera, in line with the CEN (WFD Phytoplankton Counting Guidance) recommendations (Brierley *et al.*, 2007). Phytoplankton taxa were grouped into major taxonomic groups: Cryptophyceae, Cyanobacteria, Diatoms, and Green Algae.

Open water crustacean zooplankton samples were collected at Reed Bower and preserved in 4% formaldehyde solution. *Daphnia* densities were estimated using the sampling and counting method detailed in Gunn et al. (2011).

Sub-samples of filtered (Whatman® GF/C) water were taken in the laboratory from each of the duplicate samples. The filtered sub-samples were analysed for soluble reactive phosphorus (SRP) using Murphy's & Riley's (1962) method. A sulphuric acid-potassium persulphate digestion was applied on unfiltered samples, converting all P forms to SRP; total phosphorus (TP) was then measured using a modified Murphy & Riley (1962) method, as outlined by Wetzel and Likens (2000). Nitrate (NO_3) was measured with a SEAL AQ2 analyser (SEAL Analytical Limited), using the sulphanilamide/NEDD reaction, which produces a reddish-purple dye (HMSO, 1981). SRSi concentration was quantified via spectrophotometry with a 10 mm flow-cell, following Golterman et al. (1978).

2.3 Statistical Methods

Monthly averages were calculated by averaging each month's readings for all variables. Each average was assigned a season – winter (D, J, F), spring (M, A, M), summer (J, J, A), autumn (S, O, N) – following the Met Office's meteorological calendar (Met Office, 2022).

2.3.1 Seasonality of Phytoplankton Blooms

To address the research question of assessing the seasonal variation in phytoplankton biovolume, linear mixed-effects models for each phytoplankton group (response variable) with season as a fixed effect were built. Exploratory analysis indicated a violation of the normality assumption by the response variable; therefore, a logarithmic transformation was applied. To account for the temporal dependency of samples collected within the same year, it was included as a random effect. Model diagnostics, including residual distribution and fit, were evaluated to ensure the model assumptions were met and to verify overall model adequacy. Where relevant, when the models returned significant relationships, post-hoc pairwise comparisons between seasons were conducted using Tukey's adjustment for multiple testing.

2.3.2 Chemical and Physical Drivers of Algal Bloom – Linear Modelling

To investigate the relationship between chemical and physical drivers and phytoplankton biovolume, separate additional linear models were built. These models similarly used phytoplankton biovolume as the response variable, with a logarithmic transformation applied, and year was included as a random effect. Fixed effects were selected following an extensive review of literature on phytoplankton dynamics in lakes, prioritising European and shallow lakes. Parameters were chosen based on both the volume of supporting literature and the strength of reported relationships with phytoplankton biovolumes – favouring variables with well-established relationships, such as phosphorus and nitrate concentration (Jeppesen *et al.*, 1997; Elser *et al.*, 2007; Alex Elliott and May, 2008; Carvalho *et al.*, 2012; Sommer *et al.*, 2012; Andersen *et al.*, 2020; Spears *et al.*, 2022).

During model construction, fixed effects were added in a stepwise manner, beginning with the simplest model and incrementally introducing variables such as season, phytoplankton group, and individual water chemistry parameters (e.g. SRP, nitrate, water temperature, etc). An interaction term between season and phytoplankton group was incorporated based on evidence from the literature suggesting that seasonal patterns vary between taxonomic groups (Pálffy and Vörös, 2019; Paltsev et al., 2024).

Model selection was guided by comparisons of Akaike's Information Criterion (AIC) and conditional R^2 values, balancing explanatory power with model parsimony. Multicollinearity amongst predictors was assessed using variance inflation factors (VIFs), all of which were within the acceptable threshold. Model diagnostics, including residual distribution and fit, were

again evaluated. Significance testing of fixed effects helped identify the most influential drivers of phytoplankton dynamics.

A Spearman's correlation analysis was performed to explore the seasonal relationship between the biovolume of each phytoplankton group and overall *Daphnia* densities, a genus of zooplanktons which commonly feed on phytoplankton (Lathrop and Carpenter, 1992). This non-parametric test was selected due to the non-normal distributions of variables. *Daphnia* is a dominant grazer in many freshwater systems; hence it was chosen as a proxy for grazing pressure. Studies has shown that *Daphnia* can exert substantial top-down control on phytoplankton populations (Jeppesen et al., 1997; Muylaert et al., 2010). Additionally, research on Loch Leven has highlighted the potential role that *Daphnia* has to play in controlling phytoplankton dynamics (Carvalho *et al.*, 2012), supporting *Daphnia* suitability as an indicator of grazing pressure in this study.

2.3.3 Chemical and Physical Drivers of Algal Bloom – Random Forest Regression

To further assess variable importance and identify key predictors of phytoplankton biovolume, a random forest regression analysis was performed. This non-parametric ensemble method ranks predictors based on their contribution to model accuracy – offering an alternative perspective on predictor influence without assumptions of linearity, normality, or collinearity. Given that the dataset contained missing values, two modelling strategies were compared to evaluate how best to deal with these gaps: (1) training the random forest model on a dataset with imputed values and (2) on a complete-case dataset with all missing values removed. The full model outputs, including performance metrics and variable importance rankings for both approaches, are presented in *Appendix 7.1.2 and 7.1.3*.

Models' performances were evaluated using root mean squared error (RMSE) and Pearson correlation coefficient (*r*) between predicted and observed values in the testing dataset. RMSE quantifies the average prediction error in the units of the response variable (phytoplankton biovolume), while *r* shows how well the model explains variability in the response variable. Model performances with both datasets were compared, and the complete-case dataset yielded better explanatory power – this version was selected and used for all subsequent analysis.

To ensure reproducibility, a random seed (40) was set. The dataset was then split into training (70%) and testing (30%) sets using stratified sampling to preserve the distribution of key variables, following standard machine learning practices (Nguyen et al., 2021). An initial random forest model was trained using default parameters of 500 trees, with the number of predictors selected at each split (mtry) determined by the square root of the total number of predictors – a commonly used approach in random forest modelling (Scornet, 2017).

The model was then tuned to optimise its performance– specifically the number of trees and mtry were adjusted – and retrained with the optimised model parameters. The optimised model was retrained and used to predict on the testing dataset. Each predictors importance was assessed using the percent increase in mean squared error (%IncMSE), which shows how much predictive error increases when each predictor is excluded (Liaw and Wiener, 2002); a higher %IncMSE indicates that the variable contributes more strongly to accurate prediction of phytoplankton biovolume. A full list of predictor variables used in the random forest model is provided in *Appendix 7.1.*

All statistical analyses were conducted using R version 4.4.2. (R Core Team, 2024). Specific packages utilized include *tidyverse* for data cleaning and data visualisation (Wickham et al., 2019), *lmerTest* for linear mixed-effects models (Kuznetsova, Brockhoff and Christensen, 2017), *DHARMa* for model diagnostics (Hartig, 2024), *emmeans* for post-hoc analyses (Lenth, 2025), *car* for assessing multicollinearity (Fox and Weisberg, 2019), *MuMIn* for calculating R^2 values (Bartoń, 2024), and *randomForest* for random forest regression (Liaw and Wiener, 2002). The full analysis workflow and codebase are available via the project's GitHub repository (https://github.com/arielkoh02/dissertation-loch-leven).

3. Results

3.1 Seasonality in Phytoplankton Groups

Seasonal variation in phytoplankton biovolume varied across phytoplankton taxonomic groups (Fig 2.1). The highest biovolumes values for Cryptophyceae and Cyanobacteria was in summer, while Diatoms peaked in winter and autumn. Green Algae biovolumes were more stable across seasons, with a relatively smaller peak in summer.

When back-transformed from the log scale, Cryptophyceae exhibited the highest mean biovolume in summer – approximately 2.6×10^6 μm^3 ml^{-1} (SE = 1.2×10^6). Cyanobacteria similarly peaked in summer, with an estimated biovolume of 5.6×10^6 μm^3 ml^{-1} (SE = 2.1×10^6). Diatom biovolume was highest in winter, reaching approximately 4.1×10^7 μm^3 ml^{-1} (SE = 1.8×10^7). Green Algae maintained more stable values throughout the year, with a summer peak of ~1.6×10^6 μm^3 ml^{-1} (SE = 0.8×10^6), with the mean seasonal differences being comparatively smaller.

The results of the post-hoc test showed that Cryptophyceae was significantly lower biovolume in winter compared to the other seasons. Winter and summer had the largest seasonal difference (–1.91 ± 0.27, $p < 0.0001$), indicating a distinct increase during the summer months. Summer biovolume was significantly higher than autumn (-0.79 ± 0.26, $p < 0.05$), but not significantly different from spring (–0.57 ± 0.25, $p > 0.05$). No significant difference was found between spring and autumn.

Cyanobacteria showed a similar trend, with winter biovolume significantly lower than in spring (–1.07 ± 0.37, $p < 0.05$), summer (–2.75 ± 0.37, $p < 0.0001$), and autumn (–1.94 ± 0.37, $p < 0.0001$). Summer biovolume was significantly higher than spring (-1.68 ± 0.35, $p < 0.0001$), but not significantly different from autumn (0.81 ± 0.35, $p > 0.05$). Similarly, no significant difference was found between spring and autumn.

In contrast, Diatoms showed an opposing trend, with peak biovolumes in winter and autumn. Summer biovolume was significantly lower than both winter (1.54 ± 0.41, $p < 0.001$) and autumn (1.93 ± 0.39, $p < 0.0001$). Spring biovolume was an intermediate between the colder seasons, autumn and winter, and summer; it did not significantly differ from any other season.

Green Algae showed the least seasonal variation. Biovolume in winter was significantly lower than in summer (–0.66 ± 0.29, $p < 0.05$); no other seasonal differences were observed, indicating a relatively stable pattern throughout the year.

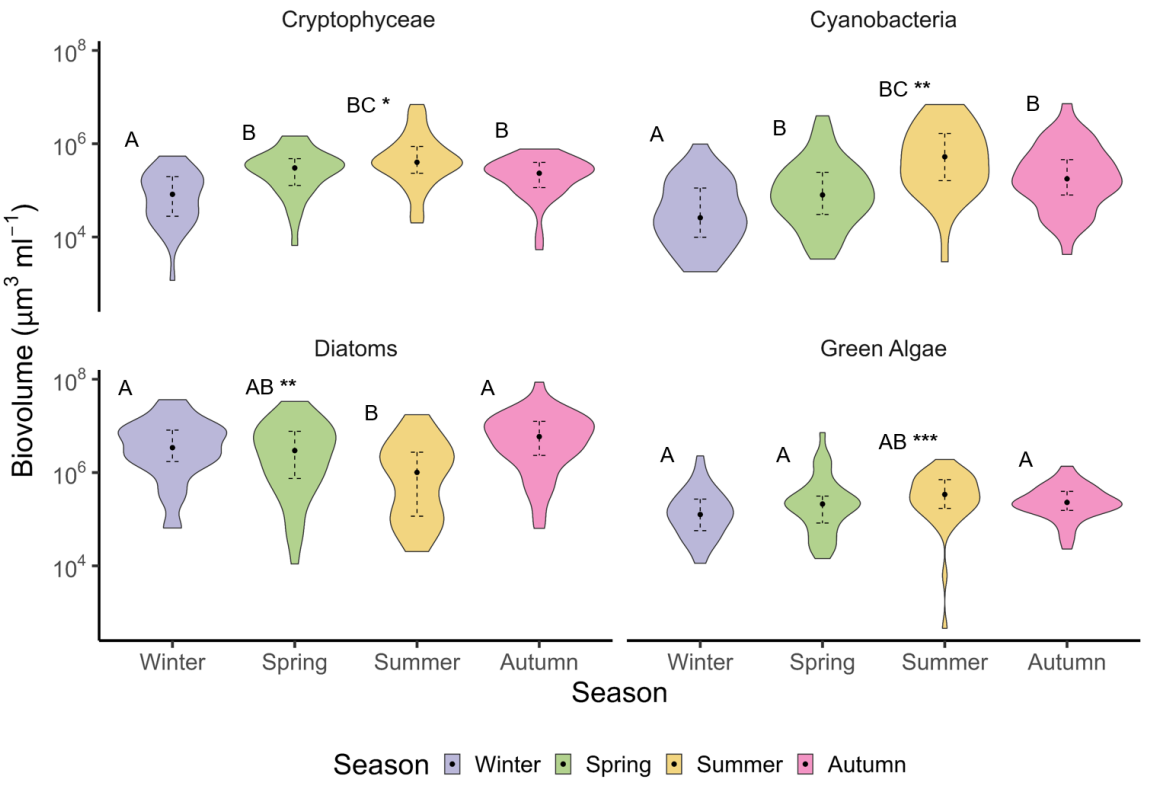


Figure 2.1 Seasonal variation in phytoplankton biovolume across four major taxonomic groups. The violin plots show the distribution, density, and spread of biovolume data across each season, differentiated by colour. Embedded boxplots represent the median (dot) and interquartile range (whiskers) of each season's values. Letters above each violin represent groupings based on post-hoc pairwise comparisons: seasons that do not share a letter are significantly different (p = 0.05). The y-axis is plotted on a logarithmic scale. **Comparisons should be made within each phytoplankton group (facet), rather than across groups.** BC * for Cryptophyceae in summer indicates biovolume is significantly higher than in autumn, but not different from spring. BC ** for Cyanobacteria in summer indicates biovolume is significantly higher than in spring, but not different from autumn. AB ** for Diatoms in spring reflects an intermediate grouping, not significantly different from any other season. AB *** for Green Algae in summer suggests biovolume is only significantly different from winter.

3.2 Chemical and Physical Drivers of Algal Bloom

3.2.1 Linear Modelling

Amongst the models built, Model 7 is the best fitting model, with the lowest AIC and the highest R^2 value (AIC = 287.64; R^2 = 0.050; Table 2.2). This suggests that the model explains the greatest proportion of the variation in phytoplankton biovolume, while still maintaining a high degree of model parsimony (Burnham and Anderson, 2004). Importantly, as fixed effects were incrementally added across models, the AIC generally remained constant or continued to decrease, indicating improved model performance without overfitting.

In Model 7, SRP, season, nitrate, temperature, and dissolved oxygen all had insignificant effects on biovolume. *Daphnia* abundance had a significant negative relationship with biovolume (–0.16 ± 0.051, $p < 0.01$). Similarly, water level had a strong and significant negative effect (–2.03 ± 0.93, $p < 0.05$).

While season on its own was not a significant predictor ($p > 0.05$), phytoplankton group was significant predictor ($p < 0.01$), and the interaction term between season and phytoplankton group was highly significant ($p < 0.001$), emphasizing the seasonal variation in phytoplankton biovolume is dependent on phytoplankton taxonomic groups. Detailed fixed effect results for season, phytoplankton groups, and the interaction between them are presented in *Appendix 7.1*.

Table 2.2 Summary of linear model outputs evaluating the influence of environmental, seasonal, and biotic factors on phytoplankton biovolume. Models (Mod 1 - Mod 7) are presented in order of increasing complexity, with fixed effects added incrementally in a stepwise manner. A dash (–) in a given row indicates that the corresponding fixed effect was not included in that particular model. Significance levels are denoted as follows: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, N.S. = not significant. Detailed results for Season, Phytoplankton Group, and *Season*Phytoplankton Group* are presented in *Appendix 7.2.*

No	Response Variable	Fixed Effect								Interaction Terms	AIC	R^2
		Soluble Reactive Phosphorus	Season	Phyto Group	Nitrate	Daphnia	Temperature	Water Level	Dissolved Oxygen	Season* Phyto Group		
1	Total Biovolume	-0.017±0.0099	-	-	-	-	-	-	-	-	450.76	0.029
2		-0.013±0.0063 *	***	***	-	-	-	-	-	***	2063.55	0.047
3		0.0096±0.014	**	***	-0.043±0.29	-	-	-	-	***	703.45	0.44
4		0.0058±0.014	**	***	-0.17±0.30	-0.040±0.018 *	-	-	-	***	706.83	0.45
5	Biovolume	0.0031±0.014	**	***	-0.056±0.32	-0.046±0.020 *	0.057±0.063	-	-	***	711.71	0.46
6		0.0020±0.014	*	***	0.16±0.34	-0.040±0.019 *	0.069±0.063	-1.21±0.55 *	-	***	693.12	0.47
7		0.017±0.040	N.S.	**	-0.38±1.54	-0.16±0.051 **	0.093±0.15	-2.03±0.93 *	0.019±0.031	***	287.64	0.50

3.2.2 Random Forest Regression

The optimised random forest model, consisting of 500 trees with an mtry value of 4, explained approximately 35.7% of the variance in total phytoplankton biovolume and produced a mean squared residual error of 7.34×10^{13}.

Variable importance, assessed using the percent increase in MSE, identified SRP (6.17%), nitrate (6.02%), and SRSi (5.71%) as the top three most influential predictors of phytoplankton biovolume (Fig 2.5). *Daphnia* abundance also ranked highly (4.83%), partially aligning with findings from the linear model. Conversely, the least influential predictnumors included season (–3.24%), year (–2.16%), and Secchi depth (–1.81%), all of which showed negative or negligible contributions to predictive accuracy.

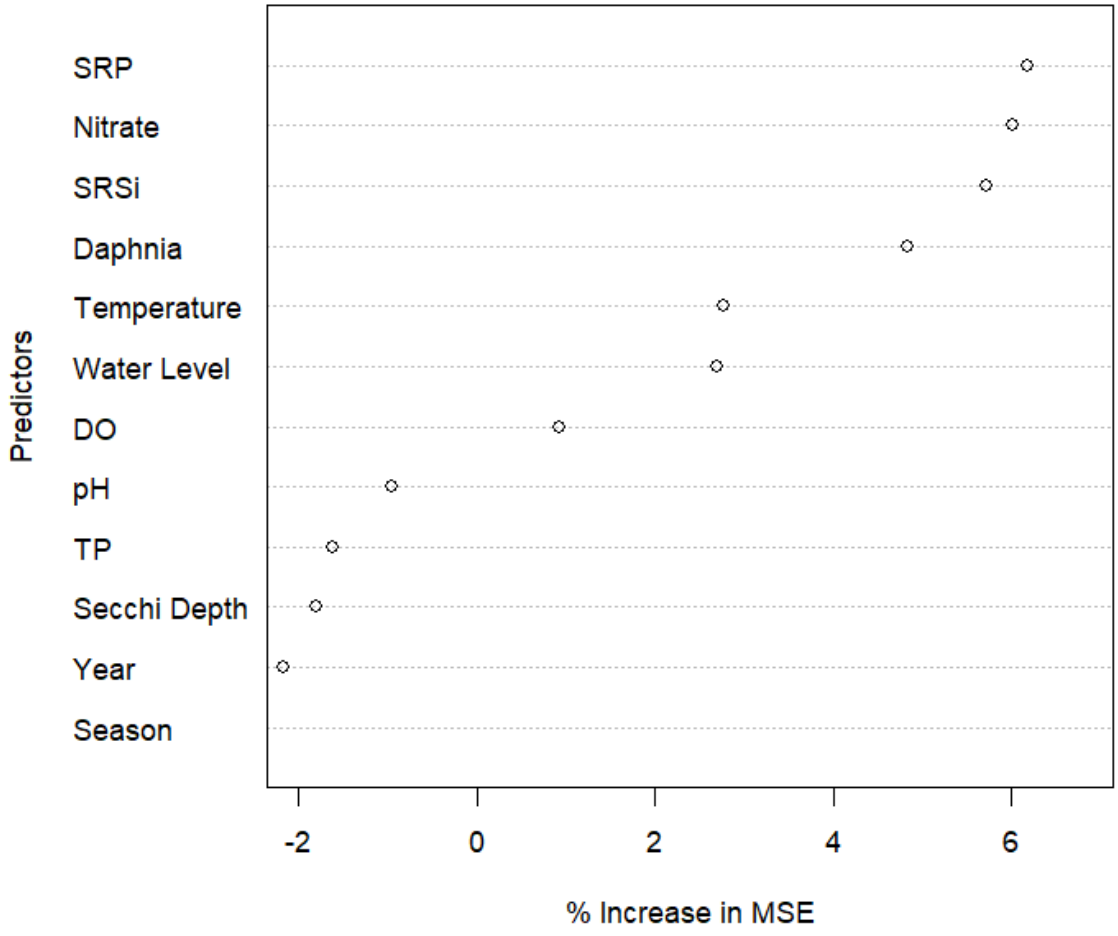


Figure 2.5 Variable importance plot based on the percent increase in mean squared error (%IncMSE) from the random forest regression model predicting total phytoplankton biovolume. Higher %IncMSE values indicate greater predictor importance.

3.2 Daphnia Densities and Phytoplankton Community Composition

A 1:1 plot of observed *Daphnia* densities versus non-Diatom phytoplankton biovolume revealed a visually broadly positive relationship, suggesting that higher Daphnia densities were generally linked with higher non-Diatom phytoplankton biovolume (Fig 2.3). The choice of only using non-Diatoms biovolume is further explained in the discussion section.

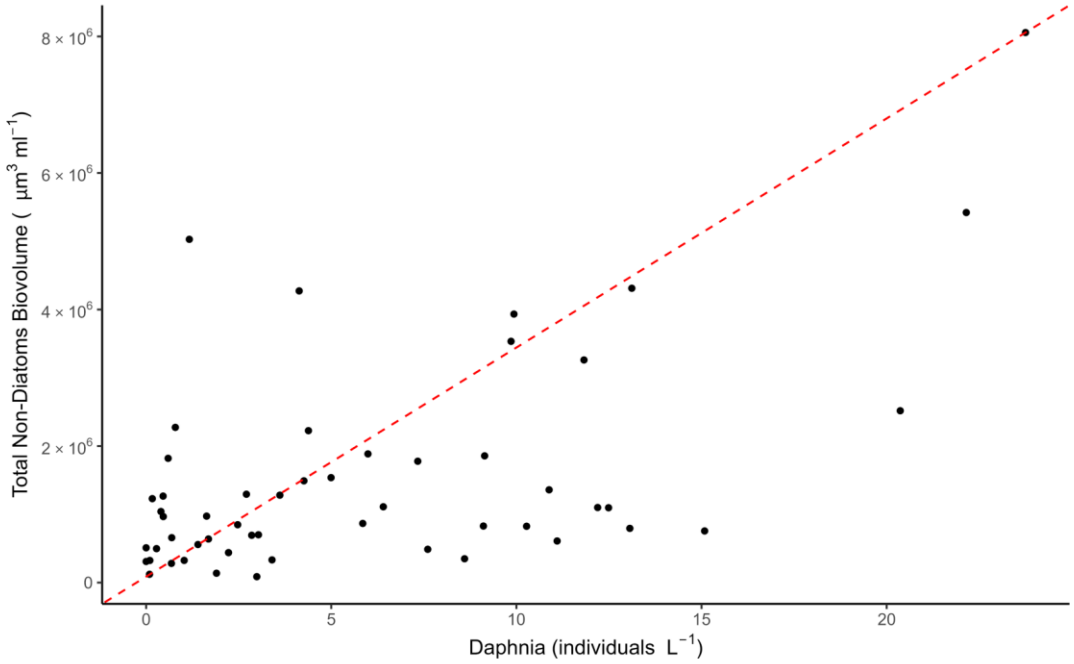


Figure 2.3 Scatter plot showing the relationship between *Daphnia* density and total non-Diatom phytoplankton biovolume (Cryptomonads, Cyanobacteria, and Green Algae combined). The red dashed line represents the 1:1 line.

The seasonal relationship between *Daphnia* densities and the biovolume of each phytoplankton group was plotted – to assess co-variation patterns over time (Fig 2.4). Cryptophyceae showed the clearest alignment with Daphnia patterns. Peaks in Cryptophyceae biovolume frequently coincided with *Daphnia* density peaks, and both declined in parallel during certain seasons. In contrast, Green Algae tended to show an inverse pattern, with higher biovolume occurring during periods of low Daphnia density. Both Cyanobacteria and Diatoms showed no visually consistent alignment.

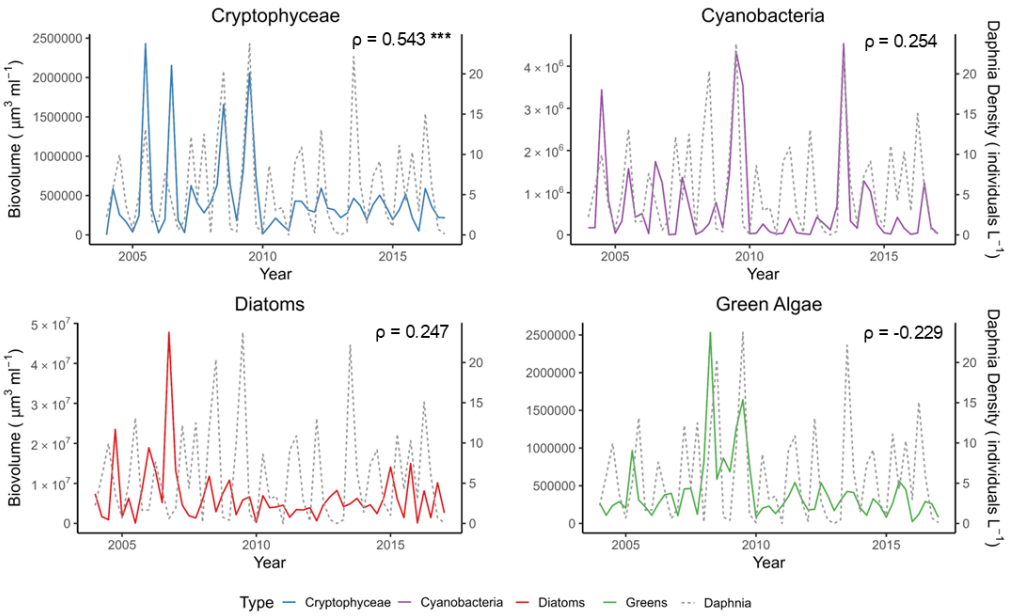


Figure 2.4 Seasonal trends in Daphnia densities (dashed line, right y-axis shared across all panels) and phytoplankton biovolume (solid lines, left y-axis also shared across all panels) for four phytoplankton groups from 2004 to 2016. Spearman's correlation coefficients (ρ) between Daphnia density and each group's biovolume are shown in the top right of each panel. Asterisks indicate statistical significance: *** $p < 0.001$.

Upon conducting Spearman's correlation, among the groups, Cryptophyceae displayed the strongest positive correlation with Daphnia density (ρ = 0.543, $p < 0.001$). Green Algae exhibited a weak negative correlation (ρ = –0.229), while Cyanobacteria and Diatoms showed weak positive correlations (ρ = 0.254 and ρ = 0.247, respectively), with all three groups lacking the statistical power to arrive at a decisive conclusion.

4. Discussion

This study aimed to investigate the seasonality and environmental drivers of phytoplankton blooms in Loch Leven, the largest shallow freshwater lake in the UK (Carvalho et al., 2012). Using a combination of statistical tests, including linear modelling and random forest regression, the results revealed that phytoplankton dynamics are influenced by a complex combination of seasonality, physical drivers, biotic interactions, and nutrient availability.

The study's findings demonstrated seasonal variation in biovolumes across phytoplankton taxon groups – with Cryptophyceae and Cyanobacteria peaking in summer, Diatoms favouring winter and autumn, and Green Algae remaining relatively constant throughout the year. Linear modelling of environmental drivers identified *Daphnia* density and water level as key predictors of phytoplankton biovolume, while random forest regression, a machine learning algorithm, highlighted SRP, nitrate and SRSi as important chemical predictors of algal bloom.

With most recent studies on Loch Leven about this topic being conducted in the early 2000s and 2010s – covering monitoring data up until 2007 (Alex Elliott and May, 2008; Carvalho et al., 2012) – this research paper extends existing knowledge by incorporating an additional decade of data (up to 2016). By understanding phytoplankton-environment relationships using both traditional statistical modelling and machine learning, this study can help to inform management decisions for shallow lakes, especially in the current context of climate change and increased anthropogenic pressures.

4.1 Seasonal Variation and Phytoplankton Community Dynamics

Addressing the first research question, seasonal variation had a significant impact on phytoplankton community composition in Loch Leven. These results align with the well-established concept of seasonal succession of phytoplankton, a recurring ecological process where phytoplankton communities shift over the course of a year (Sommer et al., 2012).

The increased summer dominance of Cyanobacteria is likely due to the higher summer water temperatures, increased light availability and reduced vertical mixing throughout the water column (Paerl and Paul, 2012; Wilkinson, Hondzo and Guala, 2019; Mullin et al., 2020). The higher summer temperatures and strong light conditions can enhance Cyanobacteria photosynthetic rates, further encouraging Cyanobacteria blooms (Jöhnk et al., 2008). Many species of Cyanobacteria can possess gas vesicles which allow them to regulate their buoyancy, therefore

are able to gain access to nutrients while remaining near the well-lit surface of a stratified lake, allowing them to outcompete the other phytoplankton groups (Han et al., 2020). Although no direct wind or mixing data was available in this study, it is possible that intermittent wind-induced upwelling events occurred, temporarily transporting nutrients from the hypolimnion to the surface, further supporting Cyanobacterial growth (Planas and Paquet, 2016). In contrast, winter conditions in Loch Leven are typically characterised by the opposite – lower temperatures and reduced light, likely contributing to the significantly lower biovolumes observed during this season.

The observed summer peak in Cryptophyceae biovolume could potentially reflect their physiological advantages during this season's stratified and nutrient-enriched conditions. Due to limited research available focussing on the Cryptophyceae class, research on *Cryptophytes* and *Cryptomonas* – groups within this class – will be used to provide some ecological context. Previous studies have shown that *Cryptomonas* tend to bloom during summer months, likely due to their motility, rapid growth, and ability to exploit light and nutrient gradients in stratified water columns (Barone and Naselli-Flores, 2003; Knapp et al., 2003). *Cryptophyte's* ability to photosynthesise at lower light intensities enables them to occupy deeper layers of the photic zone, reducing direct competition with Cyanobacteria (Gervais, 1998). These traits may help explain Cryptophyceae's seasonal summer peak in Loch Leven.

Diatoms typically thrive during periods of lower temperatures, high nutrient availability, and increased vertical mixing (Behrenfeld et al., 2021), likely explaining their higher biovolume in winter and autumn. Diatoms have relatively low light requirements and can photosynthesise efficiently under winter's reduced light conditions (Fisher and Halsey, 2016; Zhou et al., 2021). Furthermore, silica concentration tends to peak in the colder conditions of winter, providing the key resource Diatoms need to build their siliceous cell walls and supporting their seasonal dominance (Shatwell, Köhler and Nicklisch, 2013). Moreover, the increased vertical mixing typically observed during these seasons help to suspend Diatoms in the photic zone, counteracting their tendency to sink, maintaining access to light for photosynthesis (Fogg, 1991). This sinking tendency puts Diatoms at a disadvantage in the more stratified summer months, allowing more buoyant phytoplankton groups to outcompete them.

Green Algae's relatively constant biovolume across seasons in Loch Leven may reflect their functional versatility and broad environmental tolerance. Their varied physiological traits allow them to persist under a wide range of irradiance, temperature, and salinity conditions (Taylor, Fletcher and Raven, 2001; Reynolds, 2006). Green Algae are generally less competitive under

extreme conditions that favour specialists like Diatoms in winter or Cyanobacteria in summer, which may prevent them from dominating at any one time but allow them to persist consistently throughout the year (Sommer et al., 2012). This ecological generalism likely explains the observed stable biovolumes of Green Algae in this study.

Overall, comparing this study's seasonal trends in phytoplankton biovolume with historical patterns in Loch Leven reveals both continuity and evidence of long-term ecological change. The dominance of Diatoms during winter and autumn remains consistent with Bailey-Watts' (1978) findings, which reported the genera of *Stephanodiscus* and *Cyclotella* as frequent dominants in cooler seasons. However, in contrast, Bailey-Watts observed Cyanobacteria predominating in late spring and Green Algae in early summer, with Diatoms also maintaining a strong presence in summer. Cryptophyceae were not mentioned in Bailey-Watts' study.

More recent research has reported increasing summer dominance of Cyanobacteria in Loch Leven, suggesting a shift in community composition (Elliott and Defew, 2012). This is likely reflecting broader climate-driven changes such as elevated water temperatures, increased stratification, and altered nutrient cycling (Elliott and Defew, 2012; Planas and Paquet, 2016). These shifts highlight the importance of long-term monitoring and how phytoplankton communities in shallow lakes can reorganise in response to climate and catchment pressures.

Ultimately, these findings highlight that phytoplankton dynamics in Loch Leven are influenced by taxon-specific responses to seasonal environmental conditions. While some patterns remain consistent with historical observations, the rising dominance of groups such as Cyanobacteria and Cryptophyceae suggests a shifting ecological landscape. To further investigate the drivers behind phytoplankton biovolume, the next section explores the relative effect of chemical, physical, and biotic factors on phytoplankton biovolume throughout the year.

4.2 Environmental Drivers of Algal Bloom

In line with the second research question, the combination of both modelling approaches identified key chemical and physical drivers influencing phytoplankton biovolume.

4.2.1 Linear model Findings

The best fitting linear model, Model 7, found that *Daphnia* density and water level were significant predictors of phytoplankton biovolume, suggesting that biotic interactions and physical factors play key roles in shaping bloom intensity.

A statistically significant negative relationship between *Daphnia* density and phytoplankton biovolume was found. This finding aligns with previous research on zooplankton grazing pressure in shallow lakes, supporting the idea that *Daphnia*, as efficient predators of phytoplankton, can exert top-down control by consuming phytoplankton (Jeppesen et al., 1997; Muylaert et al., 2010). Carvalho et al. (2012) noted in their study that warmer spring temperatures in Loch Leven coincided with increased *Daphnia* densities – which may be the cause of reduced chlorophyll *a*, a proxy for phytoplankton biovolume, and an associated improvement in water clarity. However, the relationship found in the present study was relatively small. This may be due to their grazing effect being weakened during periods of high nutrient availability or when the phytoplankton community is dominated by less edible species (Yuan and Pollard, 2018).

Water level was also negatively correlated with phytoplankton biovolume. As noted in the methods section, the water level in Loch Leven is measured as the distance from the top of the pier to the water surface at Kinross Pier. Therefore, a smaller measured value indicates a higher actual water level. Hence, the significant negative relationship between the measured water level values and phytoplankton biovolume suggests that higher actual water levels are associated with higher phytoplankton biovolume (Fig 3.1).

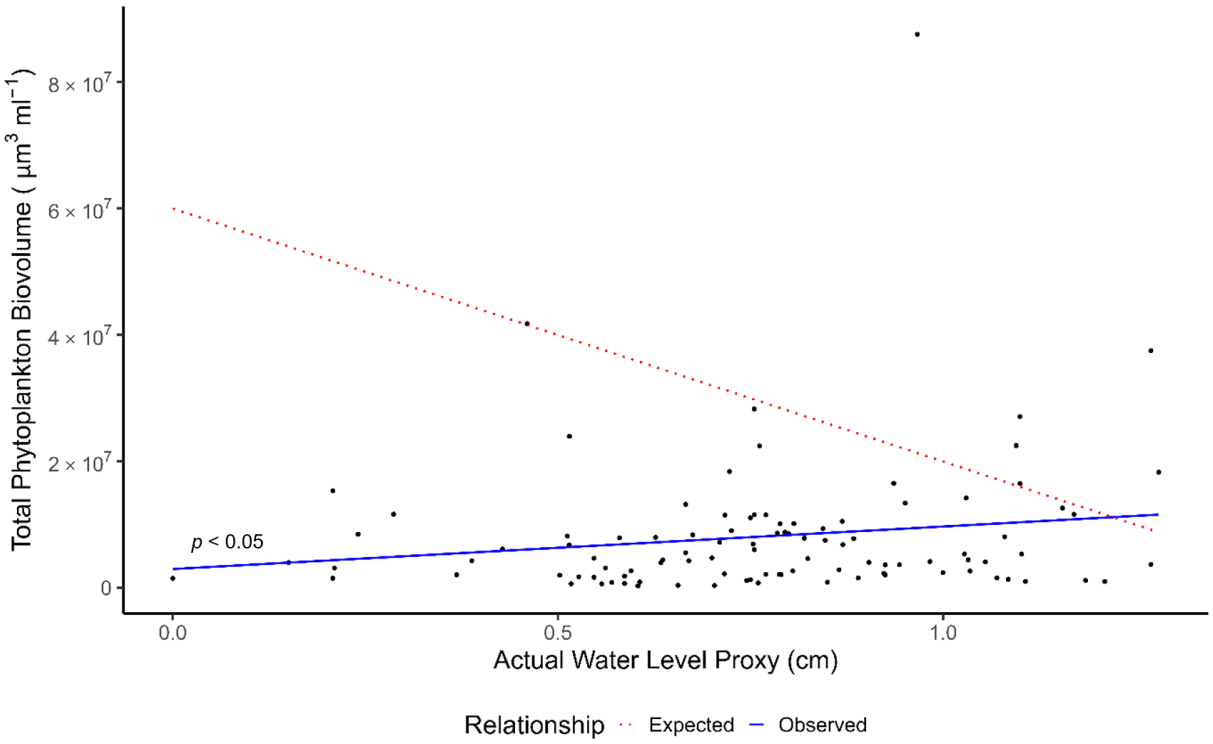


Figure 3.1 Conceptual diagram showing the relationship between phytoplankton biovolume and a proxy for actual water level. The absolute pier height was not available in this study; thus, the furthest measured pier-to-surface distance (107.98 cm) was used as a reference point. A proxy was generated by subtracting this value from each observation, where higher values represent higher actual water levels. The blue line represents the observed significant relationship from the data ($p < 0.05$); the red dotted line indicates the expected relationship based on previous literature.

This observed relationship runs counter to expectations. It opposes previous research that have found higher lake levels can enhance flushing rates, which can dilute algal biomass and export nutrients from the system (Carvalho et al., 2012; Elliott and Defew, 2012). However, it is possible that in shallower conditions, zooplankton and phytoplankton are confined to a reduced volume, increasing encounter rates and potentially leading to higher grazing pressure (Põllumäe and Haberman, 1998). Another possible explanation is that reduced water levels can enhance sediment resuspension rates. Wind and bioturbation can stir up bottom sediments, increasing turbidity and light attenuation – conditions that can suppress phytoplankton productivity (Da Costa, Attayde and Becker, 2016). Regardless of the direction of the relationship, these findings underscore the importance of hydrological management in regulating phytoplankton blooms in shallow lake systems.

While season on its own was not identified as a significant predictor of total phytoplankton biovolume, the phytoplankton group and its interaction with season were both found to be

significant. The significance of the linear model's interaction term supports the findings of the earlier section: Seasonal Variation and Phytoplankton Community Dynamics, where different taxonomic groups showed distinct seasonal biovolume patterns. These findings align with broader ecological theories of phytoplankton seasonal succession, and that dynamics are highly taxon-specific, rather than uniform across all groups (Sommer et al., 2012).

4.2.2 Random Forest Findings

The random forest regression identified chemical variables – SRP, nitrate and SRSi – as the most important predictors of phytoplankton biovolume, differing from the linear model's results. Notably, *Daphnia* density also ranked relatively high in importance, aligning with the linear model results.

Phosphorus is well-established as a key limiting nutrient in freshwater ecosystems (Correll, 1999; Elser et al., 2007; Alex Elliott and May, 2008). Soluble reactive phosphorus, the most readily bioavailable form, supports phytoplankton growth; its presence is often linked to more frequent and intense blooms (Reynolds and Davies, 2001; Smith, King and Williams, 2015). Nitrate, a major source of nitrogen, is similarly important for phytoplankton growth, particularly for Cyanobacteria and Green Algae (Hecky and Kilham, 1988*).* This also points to the potential that Loch Leven experiences nutrient co-limitation, where growth of phytoplankton is limited by the availability of both nitrogen and phosphorus – a pattern observed in other shallow, eutrophic lakes (Andersen et al., 2020; Howarth et al., 2021). As previously noted, SRSi is a crucial component of Diatoms' cell walls (Shatwell, Köhler and Nicklisch, 2013). Fluctuations in SRSi availability can therefore influence Diatom biovolume and, in turn, the broader community composition (Ngupula et al., 2014), especially during autumn and winter when Diatoms tend to dominate.

Interestingly, the random forest model ranked temperature as the most influential physical driver. The warmer temperatures coupled with higher SRP concentrations may have acted synergistically to favour Cyanobacteria dominance during the stratified summer periods. This finding is supported by Richardson et al. (2019), who found that elevated temperatures and nutrient enrichment together markedly increased Cyanobacteria biovolume in shallow lakes.

In contrast, Secchi depth, season, and year were among the lowest-ranked predictors in the model. Secchi depth readings can be strongly influenced by phytoplankton biovolume, potentially making it more of an indicator than a causal driver (Lee et al., 2018). Season may have limited additional explanatory power because nutrient concentrations, strong predictors in the model, already vary predictably with seasonal cycles (Wang et al., 2019). Furthermore, if long-term

changes are more effectively reflected through nutrient availability trends, interannual variation may not be a significant predictor of biovolume. These findings highlight that although these variables are often used as proxies for light availability and temporal variation, their low predictive importance may reflect that their effects are already accounted for in more direct chemical variables.

4.2.3 Traditional Modelling vs Machine Learning

The linear model did not identify SRP and nitrate as significant predictors, which contrasts with previous research that highlighted these variables as key drivers of phytoplankton blooms (Smith, King and Williams, 2015; Andersen et al., 2020). This may be reflective of the model's inability to capture non-linear interactions among these variables. For example, Carvalho et al. (2012) found a highly significant positive relationship between total phosphorus and chlorophyll *a* in Loch Leven – however they noted this relationship levelling off beyond a certain concentration threshold. These threshold effects and nutrient saturation points may reduce the apparent importance of nutrient predictors in linear modelling approaches (James et al., 2013). These findings suggest that the influence of phosphorus and nitrate may be contingent on other environmental conditions and better captured using methods that accommodate non-linear relationships, such as the random forest regression.

Traditional linear models are limited in their ability to detect complex relationships and interactions, unless directly stated in the model structure (Zuur, Ieno and Smith, 2007). While interaction terms can be manually included – such as *season*phytoplankton* group as done in this study – this requires ecological reasoning and prior assumptions on which variables may interact and how (Rimpler, Kiers and Van Ravenzwaaij, 2025). In contrast, random forest inherently captures interactions between variables without the need for explicit specification (Breiman, 2001). This is done through their decision tree-based structure splitting the data in different ways across many variables, allowing the model to learn how each variable works together to influence outcomes (James et al., 2013).

However, this flexibility comes at a cost. Despite their strong predictive power, random forests do not provide clear effect sizes or significance values, which makes them less transparent and harder to interpret (Breiman, 2001; James et al., 2013). They are often referred to as “black boxes” due to their internal structure, comprising of hundreds of decision trees, being difficult to unpack (Breiman, 2001). Additionally, random forests rely on random sampling and tuning of model

parameters causing their outputs to vary slightly between runs (James et al., 2013), unless settings such as the random seed are explicitly fixed – as was done in this study. As a result, random forests can be less reproducible than traditional linear models. In contrast, linear models are highly transparent and reproducible, producing clear, interpretable outputs that are easy to replicate and communicate (Love et al., 2023), making them especially valuable for hypothesis testing in research.

Ultimately, this study combined both approaches to offer complementary perspectives. The linear model highlighted a small set of interpretable, significant drivers – such as *Daphnia* density and water level, while the random forest model captured more complex relationships and interactions – identifying chemical parameters (SRP, nitrate, SRSi) as important drivers of phytoplankton biovolumes. This integrated approach enhances our ability to understand and predict algal blooms in Loch Leven, a shallow freshwater lake.

4.3 Daphnia as a Driver of Phytoplankton Community Composition

The third research question examined how environmental drivers influence phytoplankton community composition; analysis of grazer-prey dynamics provides valuable insight. While the linear model identified that *Daphnia* density had a small but significant influence on total phytoplankton biovolume, the 1:1 plot and group-specific time series suggests a more complex relationship.

4.3.1 Daphnia vs Total Non-Diatom Biovolume

Diatoms were excluded from the 1:1 plot based on ecological reasoning. *Daphnia*, as filter feeders, mainly consume small, suspended phytoplankton such as Cryptophyceae, Green Algae, and some Cyanobacteria (Hiltunen et al., 2017). However, they are generally less effective at grazing on heavier, silica-based Diatoms, which are less accessible in the water column (Fogg, 1991).

In the 1:1 plot, Daphnia density was visually observed to be broadly positively related to the total biovolume of non-Diatom phytoplankton. This opposes the negative relationship found in the linear model, thus challenging the theory that *Daphnia* exert top-down grazing pressure on phytoplankton populations. One possible explanation is favourable environmental conditions may stimulate phytoplankton growth, which in turn supports increases in *Daphnia* populations – a pattern consistent with bottom-up control theory (Li et al., 2020). Alternatively, this relationship might be explained by a time-lag effect, where *Daphnia* densities rise following increases in

phytoplankton biovolume, but do not immediately suppress phytoplankton due to delays in grazing impact.

To better understand the grazer-prey dynamics, a group-specific analysis was conducted to investigate how different phytoplankton taxa responded to changes in *Daphnia* density.

4.3.2 Group-Specific Relationships with Daphnia

Among the phytoplankton groups, Cryptophyceae showed the strongest positive association with *Daphnia* density; the relationship was statistically significant. This relationship suggests potential bottom-up dynamics, where higher Cryptophyceae biovolumes may support increased *Daphnia* populations rather than being suppressed by them. As previously noted, limited research is available on the Cryptophyceae class, thus studies on *Cryptomonas* and *Cryptophytes* will be used to potentially give us some ecological insights. *Cryptophytes* are generally small, fast-growing, and highly nutritious, making them a preferred food source for *Daphnia* (Sarnelle, 1992). Similar relationships have been found in other freshwater systems, where cryptomonas abundance has been positively linked to zooplankton growth and abundance (DeMott, 1986; Enawgaw et al., 2023).

Green Algae showed a contrasting pattern, where biovolumes were inversely related to *Daphnia* density. While this trend was not statistically significant, this may suggest that reduced grazing pressure results in increases in the biovolume of Green Algae, supporting the theory of top-down control. The structural and ecological diversity of Green Algae likely increases their ability to persist under changing grazing pressures throughout the year (Taylor, Fletcher and Raven, 2001; Reynolds, 2006). Traits such as low palatability, colony formation, mucilage production, and larger cell sizes can reduce grazing efficiency, allowing certain taxa to resist predation even during high periods of *Daphnia* density (Lürling, 2003).

Cyanobacteria and Diatoms were both weakly positively correlated with *Daphnia* density, though not statistically significant. This suggests that both groups are generally less affected by direct grazing. Cyanobacteria often form filamentous or colonial structures that interfere with *Daphnia's* filter-feeding apparatus, and many produce toxins, making them a poor-quality food source (Bednarska, Pietrzak and Pijanowska, 2014). Diatoms tend to sink out of the photic zone (Fogg, 1991), where *Daphnia* migrate vertically to feed and avoid predation (Chiapella et al., 2021) – reducing their availability to grazers.

Overall, these findings highlight the complexity of grazer-prey interactions in Loch Leven, a shallow freshwater lake. The study's findings point towards a dynamic interplay between both top-down and bottom-up mechanisms, where grazers' influence varies considerably depending on phytoplankton group traits and environmental conditions.

4.4 Broader Ecological Implications and Management

This study's findings provide valuable insights into how seasonal patterns and environmental drivers – focusing on water levels, nutrient availability, and *Daphnia* grazing – influence phytoplankton community composition in Loch Leven. The presence of both top-down and bottom-up dynamics highlights the complexity of ecosystem function in freshwater shallow lakes. Given the context of climate change, which is expected to increase the frequency and severity of harmful algal blooms (Paerl and Paul, 2012; Planas and Paquet, 2016), understanding these interactions is essential to inform long-term management decisions and strategies.

Understanding the seasonality of phytoplankton groups, knowing which phytoplankton taxonomic group dominate at different times of the year, offers practical implications for monitoring and interventions. Being able to anticipate periods of higher bloom risk allows the targeting of interventions accordingly. For example, the summer peaks of Cyanobacteria biovolumes, is a larger public health concern due to their production of cyanotoxins (Jöhnk et al., 2008). Hence, this season may require increased nutrient monitoring and public health communications to ensure the well-being of loch users.

Water level being highlighted as a significant predictor has emphasised the role hydrological processes has in influencing phytoplankton biovolumes. Loch Leven is equipped with sluice gates at its singular outflow point to regulate water level and discharge (CEH, no date) – this existing infrastructure could play an increasingly important role in algae bloom management. During periods of higher risk, the sluice gates can potentially be used to manipulate water levels, decreasing the occurrence of bloom formation.

While managing zooplankton grazing is more complex, understanding grazer-prey dynamics is still ecologically valuable. Enhancing conditions that support *Daphnia,* such as maintaining clear water conditions and reducing fishing predator pressure (Manca et al., 2008), could help encourage natural top-down control of phytoplankton biovolume.

Chemical drivers – especially SRP and nitrate – have unsurprisingly emerged as important predictors of biovolume. This reinforces the need to continue managing nutrient inputs into Loch Leven. In recent years, internal nutrient loading from sediments have been shown to play a role in sustaining blooms (Spears et al., 2012). Future management would benefit from further investigations of sediment dynamics and exploration of potential solutions.

These findings show that a single intervention strategy is likely to be an insufficient solution. Effective management of phytoplankton blooms in Loch Leven requires an integrated approach that accounts for the interplay of chemical, physical, and biological drivers.

4.5 Limitations and Future Studies

This study was not without its limitations. Although the dataset spanned over a decade, it contained substantial data gaps, particularly in later years. These data gaps may have limited the robustness of the linear model's output and the strength of correlations. In the random forest regression, missing values were excluded entirely, as noted in the methods section, which reduced the overall dataset size. This limited dataset size may have limited the model's ability to fully capture underlying patterns and interactions in the data.

Another limitation is the taxonomic resolution of phytoplankton groupings. In this study, Cyanobacteria and Green Algae were analysed at the phylum level, while Diatoms and Cryptophyceae were analysed at the class level. At such broad taxonomic levels, finer-scale ecological interactions and species dynamics may be overlooked or obscured. Similarly, the simplification of using *Daphnia* as a proxy for zooplankton grazing may ignore the influence of other zooplankton taxa that fill different ecological roles.

Future research could consider applying similar modelling approaches to other shallow lake systems affected by multiple stressors, both within and beyond the UK. This would allow for cross-system comparisons to assess if the patterns observed in Loch Leven are consistent in other lakes. This will help determine if these findings reflect generalisable ecological processes in shallow freshwater ecosystems. Expanding the temporal coverage and improving data continuity would allow for stronger interferences about long-term ecological change. The addition of other environmental factors such as light availability, wind speed, or internal phosphorus release could further improve model accuracy and predictive powers. Furthermore, the use of other advanced modelling techniques such as generalised additive models, could allow for the flexibility of

modelling non-linear trends and interactions, offering an additional avenue to capture the complex relationships between phytoplankton and the environment.

By addressing these limitations and expanding both the spatial and temporal scope of future research, a deeper understanding of shallow freshwater lakes can be achieved, helping to inform adaptive management strategies in the face of a changing climate.

5. Conclusion

This study explored the seasonal and environmental drivers of phytoplankton biovolumes and community composition in Loch Leven using both traditional linear modelling and a machine learning approach, random forest regression. The results revealed a clear seasonality of phytoplankton taxonomic groups, with Cyanobacteria and Cryptophyceae peaking in summer, and Diatoms dominating in autumn and winter. The findings highlighted the multifaceted nature of algal bloom dynamics in Loch Leven. While linear modelling identified physical (water level) and biotic (*Daphnia* density) factors as significant predictors, the random forest model emphasised the importance of chemical variables: SRP, nitrate, and SRSi. When *Daphnia* density was plotted against each phytoplankton group, the study found that grazer-prey interactions are more complex than top-down controls, suggesting the dynamic is influenced by environmental conditions.

Ultimately, this study's findings demonstrate that phytoplankton community dynamics in Loch Leven are influenced by the interaction between physical, chemical and biological drivers. Effective lake management of phytoplankton blooms will therefore require integrated solutions that account for both nutrient control and ecological interactions. As pressure from climate change and anthropogenic influences intensify, developing adaptive, evidence-based approaches to lake management will be essential to preserve freshwater ecosystem health.

6. References

Adrian, R. *et al.* (2009) 'Lakes as sentinels of climate change', *Limnology and Oceanography*, 54(6part2), pp. 2283–2297. Available at: https://doi.org/10.4319/lo.2009.54.6_part_2.2283.

Albert, J.S. *et al.* (2021) 'Scientists' warning to humanity on the freshwater biodiversity crisis', *Ambio*, 50(1), pp. 85–94. Available at: https://doi.org/10.1007/s13280-020-01318-8.

Alex Elliott, J. and May, L. (2008) 'The sensitivity of phytoplankton in Loch Leven (U.K.) to changes in nutrient load and water temperature', *Freshwater Biology*, 53(1), pp. 32–41. Available at: https://doi.org/10.1111/j.1365-2427.2007.01865.x.

Andersen, I.M. *et al.* (2020) 'Nitrate, ammonium, and phosphorus drive seasonal nutrient limitation of chlorophytes, Cyanobacteria, and Diatoms in a hyper-eutrophic reservoir', *Limnology and Oceanography*, 65(5), pp. 962–978. Available at: https://doi.org/10.1002/lno.11363.

Arend, K.K. *et al.* (2011) 'Seasonal and interannual effects of hypoxia on fish habitat quality in central Lake Erie: Hypoxia effects on fish habitat', *Freshwater Biology*, 56(2), pp. 366–383. Available at: https://doi.org/10.1111/j.1365-2427.2010.02504.x.

ASTM International (2012) *ASTM D4137-82(2012): Standard Practice for Preserving Phytoplankton Samples*, *ASTM International*. Available at: https://www.astm.org/d4137-82r12.html (Accessed: 14 April 2025).

Bailey-Watts, A.E. and Kirika, A. (1987) 'A re-assessment of phosphorus inputs to Loch Leven (Kinross, Scotland): rationale and an overview of results on instantaneous loadings with special reference to runoff', *Transactions of the Royal Society of Edinburgh: Earth Sciences*, 78(4), pp. 351–367. Available at: https://doi.org/10.1017/S0263593300011299.

Bailey-Watts, A.E. and Kirika, A. (1999) 'Poor water quality in Loch Leven (Scotland) in 1995 in spite of reduced phosphorus loadings since 1985: the influences of catchment management and inter-annual weather variation', *Hydrobiologia*, 403, pp. 135–151. Available at: https://doi.org/10.1023/A:1003758713050.

Baron, J.S. *et al.* (2002) 'Meeting Ecological and Societal Needs for Freshwater', *Ecological Applications*, 12(5), pp. 1247–1260. Available at: https://doi.org/10.1890/1051-0761(2002)012[1247:MEASNF]2.0.CO;2.

Barone, R. and Naselli-Flores, L. (2003) 'Distribution and seasonal dynamics of Cryptomonads in Sicilian water bodies', *Hydrobiologia*, 502(1–3), pp. 325–329. Available at: https://doi.org/10.1023/B:HYDR.0000004290.22289.c2.

Bartoń, K. (2024). *MuMIn: Multi-Model Inference*. R package version 1.48.4. Available at: https://CRAN.R-project.org/package=MuMIn (Accessed 10 Apr. 2025).

Beardall, J. *et al.* (2009) 'Allometry and stoichiometry of unicellular, colonial and multicellular phytoplankton', *New Phytologist*, 181(2), pp. 295–309. Available at: https://doi.org/10.1111/j.1469-8137.2008.02660.x.

Bednarska, A., Pietrzak, B. and Pijanowska, J. (2014) 'Effect of poor manageability and low nutritional value of Cyanobacteria on Daphnia magna life history performance', *Journal of Plankton Research*, 36(3), pp. 838–847. Available at: https://doi.org/10.1093/plankt/fbu009.

Behrenfeld, M.J. *et al.* (2021) 'Thoughts on the evolution and ecological niche of Diatoms', *Ecological Monographs*, 91(3), p. e01457. Available at: https://doi.org/10.1002/ecm.1457.

Breiman, L. (2001) 'Random Forest', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Brierley, B. *et al.* (2007) 'Guidance on the quantitative analysis of phytoplankton in Freshwater Samples'.

Brooks, B.W. *et al.* (2016) 'Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems?', *Environmental Toxicology and Chemistry*, 35(1), pp. 6–13. Available at: https://doi.org/10.1002/etc.3220.

Burnham, K.P. and Anderson, D.R. (eds) (2004) *Model Selection and Multimodel Inference*. New York, NY: Springer New York. Available at: https://doi.org/10.1007/b97636.

Canva (2013) 'Canva'. Canva Pty Ltd. Available at: https://www.canva.com/ (Accessed: 4 May 2025).

Carvalho, L. *et al.* (2012) 'Water quality of Loch Leven: responses to enrichment, restoration and climate change', *Hydrobiologia*, 681(1), pp. 35–47. Available at: https://doi.org/10.1007/s10750-011-0923-x.

CEH (no date) *About Loch Leven — a case study of global importance*, *UK Centre for Ecology & Hydrology - Loch Leven Portal*. Available at: https://eip.ceh.ac.uk/apps/loch-leven/about (Accessed: 14 April 2025).

Cheng, Y. *et al.* (2021) 'A novel random forest approach to revealing interactions and controls on chlorophyll concentration and bacterial communities during coastal phytoplankton blooms', *Scientific Reports*, 11(1), p. 19944. Available at: https://doi.org/10.1038/s41598-021-98110-9.

Chiapella, A.M. *et al.* (2021) 'A day in the life of winter plankton: under-ice community dynamics during 24 h in a eutrophic lake', *Journal of Plankton Research*. Edited by B.E. Beisner, 43(6), pp. 865–883. Available at: https://doi.org/10.1093/plankt/fbab061.

Correll, D. (1999) 'Phosphorus: a rate limiting nutrient in surface waters', *Poultry Science*, 78(5), pp. 674–682. Available at: https://doi.org/10.1093/ps/78.5.674.

Da Costa, M.R.A., Attayde, J.L. and Becker, V. (2016) 'Effects of water level reduction on the dynamics of phytoplankton functional groups in tropical semi-arid shallow lakes', *Hydrobiologia*, 778(1), pp. 75–89. Available at: https://doi.org/10.1007/s10750-015-2593-6.

DeMott, W.R. (1986) 'The role of taste in food selection by freshwater zooplankton', *Oecologia*, 69(3), pp. 334–340. Available at: https://doi.org/10.1007/BF00377053.

Ebert, U. *et al.* (2001) 'Critical Conditions for Phytoplankton Blooms', *Bulletin of Mathematical Biology*, 63(6), pp. 1095–1124. Available at: https://doi.org/10.1006/bulm.2001.0261.

Ekstrom, J.A., Moore, S.K. and Klinger, T. (2020) 'Examining harmful algal blooms through a disaster risk management lens: A case study of the 2015 U.S. West Coast domoic acid event', *Harmful Algae*, 94, p. 101740. Available at: https://doi.org/10.1016/j.hal.2020.101740.

Elliott, J.A. and Defew, L. (2012) 'Modelling the response of phytoplankton in a shallow lake (Loch Leven, UK) to changes in lake retention time and water temperature', *Hydrobiologia*, 681(1), pp. 105–116. Available at: https://doi.org/10.1007/s10750-011-0930-y.

Elser, J.J. *et al.* (2007) 'Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems', *Ecology Letters*, 10(12), pp. 1135–1142. Available at: https://doi.org/10.1111/j.1461-0248.2007.01113.x.

Enawgaw, Y. *et al.* (2023) 'Zooplankton as ecosystem indicators and their effects on eutrophication in Lake Arekit (Ethiopia) – implication for freshwater habitat management', *Journal of Freshwater Ecology*, 38(1), p. 2287433. Available at: https://doi.org/10.1080/02705060.2023.2287433.

Falkowski, P. (2012) 'Ocean Science: The power of plankton', *Nature*, 483(7387), pp. S17–S20. Available at: https://doi.org/10.1038/483S17a.

Ferguson, C.A. *et al.* (2007) 'Model Comparison for a Complex Ecological System', *Journal of the Royal Statistical Society Series A: Statistics in Society*, 170(3), pp. 691–711. Available at: https://doi.org/10.1111/j.1467-985X.2006.00462.x.

Field, C.B. *et al.* (1998) 'Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components', *Science*, 281(5374), pp. 237–240. Available at: https://doi.org/10.1126/science.281.5374.237.

Fisher, N.L. and Halsey, K.H. (2016) 'Mechanisms that increase the growth efficiency of Diatoms in low light', *Photosynthesis Research*, 129(2), pp. 183–197. Available at: https://doi.org/10.1007/s11120-016-0282-6.

Fogg, G.E. (1991) 'The phytoplanktonic ways of life', *New Phytologist*, 118(2), pp. 191–232. Available at: https://doi.org/10.1111/j.1469-8137.1991.tb00974.x.

Foreman, K. *et al.* (2021) 'Effects of harmful algal blooms on regulated disinfection byproducts: Findings from five utility case studies', *AWWA Water Science*, 3(3), p. e1223. Available at: https://doi.org/10.1002/aws2.1223.

Fox, J. and Weisberg, S., 2019. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks, CA: Sage. Available at: https://www.john-fox.ca/Companion/ [Accessed 10 Apr. 2025].

Gervais, F. (1998) 'Ecology of cryptophytes coexisting near a freshwater chemocline', *Freshwater Biology*, 39(1), pp. 61–78. Available at: https://doi.org/10.1046/j.1365-2427.1998.00260.x.

Griffith, A.W. and Gobler, C.J. (2020) 'Harmful algal blooms: A climate change co-stressor in marine and freshwater ecosystems', *Harmful Algae*, 91, p. 101590. Available at: https://doi.org/10.1016/j.hal.2019.03.008.

Gunn, I.D.M. *et al.* (2012) 'Long-term trends in Loch Leven invertebrate communities', *Hydrobiologia*, 681(1), pp. 59–72. Available at: https://doi.org/10.1007/s10750-011-0926-7.

Han, Y. *et al.* (2020) 'Assessing vertical diffusion and Cyanobacteria bloom potential in a shallow eutrophic reservoir', *Lake and Reservoir Management*, 36(2), pp. 169–185. Available at: https://doi.org/10.1080/10402381.2019.1697402.

Hartig, F. (2024). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.7. Available at: https://CRAN.R-project.org/package=DHARMa (Accessed 10 Apr. 2025).

Hecky, R.E. and Kilham, P. (1988) 'Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on the effects of enrichment1', *Limnology and Oceanography*, 33(4part2), pp. 796–822. Available at: https://doi.org/10.4319/lo.1988.33.4part2.0796.

Hiltunen, M. *et al.* (2017) 'Trophic upgrading via the microbial food web may link terrestrial dissolved organic matter to Daphnia', *Journal of Plankton Research*, 39(6), pp. 861–869. Available at: https://doi.org/10.1093/plankt/fbx050.

Ho, J.C. and Michalak, A.M. (2020) 'Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes', *Limnology and Oceanography*, 65(5), pp. 992–1009. Available at: https://doi.org/10.1002/lno.11365.

Holden, A.V. and Caines, L.A. (1974) 'Nutrient Chemistry of Loch Leven, Kinross', *Proceedings of the Royal Society of Edinburgh. Section B. Biology*, 74, pp. 101–121. Available at: https://doi.org/10.1017/S0080455X00012340.

Howarth, R.W. *et al.* (2021) 'Role of external inputs of nutrients to aquatic ecosystems in determining prevalence of nitrogen vs. phosphorus limitation of net primary productivity', *Biogeochemistry*, 154(2), pp. 293–306. Available at: https://doi.org/10.1007/s10533-021-00765-z.

Hudnell, H.K. (2010) 'The state of U.S. freshwater harmful algal blooms assessments, policy and legislation', *Toxicon*, 55(5), pp. 1024–1034. Available at: https://doi.org/10.1016/j.toxicon.2009.07.021.

James, G. *et al.* (2013) *An Introduction to Statistical Learning*. New York, NY: Springer New York (Springer Texts in Statistics). Available at: https://doi.org/10.1007/978-1-4614-7138-7.

Jeppesen, E. *et al.* (1997) 'Top-down control in freshwater lakes: the role of nutrient state, submerged macrophytes and water depth', *Hydrobiologia*, 342/343, pp. 151–164. Available at: https://doi.org/10.1023/A:1017046130329.

Jeppesen, E. *et al.* (2005) 'Lake responses to reduced nutrient loading – an analysis of contemporary long-term data from 35 case studies', *Freshwater Biology*, 50(10), pp. 1747–1771. Available at: https://doi.org/10.1111/j.1365-2427.2005.01415.x.

Ji, G. *et al.* (2024) 'Response of dissolved organic matter and disinfection by-product precursors to algal blooms and thermal stratification in deep reservoirs', *Chemosphere*, 368, p. 143757. Available at: https://doi.org/10.1016/j.chemosphere.2024.143757.

Jöhnk, K.D. *et al.* (2008) 'Summer heatwaves promote blooms of harmful Cyanobacteria', *Global Change Biology*, 14(3), pp. 495–512. Available at: https://doi.org/10.1111/j.1365-2486.2007.01510.x.

Kirby, R.P. (1971) 'The Bathymetrical Resurvey of Loch Leven, Kinross', *The Geographical Journal*, 137(3), p. 372. Available at: https://doi.org/10.2307/1797274.

Knapp, C.W. *et al.* (2003) 'PHYSICAL AND CHEMICAL CONDITIONS SURROUNDING THE DIURNAL VERTICAL MIGRATION OF *CRYPTOMONAS* SPP. (CRYPTOPHYCEAE) IN A SEASONALLY STRATIFIED MIDWESTERN RESERVIOR (USA)', *Journal of Phycology*, 39(5), pp. 855–861. Available at: https://doi.org/10.1046/j.1529-8817.2003.02139.x.

Kudela, R., Berdalet, E. and Urban, E. (2015) *Harmful Algal Blooms : A scientific summary for policy makers*, *SIDALC*. Unesco. Available at: https://sidalc.net/search/Record/dig-icm-es-10261-141712 (Accessed: 30 April 2025).

Kuznetsova, A., Brockhoff, P.B. and Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), pp.1–26. doi:10.18637/jss.v082.i13. Available at: https://doi.org/10.18637/jss.v082.i13 (Accessed 10 Apr. 2025).

Lathrop, R.C. and Carpenter, S.R. (1992) 'Zooplankton and Their Relationship to Phytoplankton', in J.F. Kitchell (ed.) *Food Web Management*. New York, NY: Springer New York (Springer Series on Environmental Management), pp. 127–150. Available at: https://doi.org/10.1007/978-1-4612-4410-3_8.

Lee, Z. *et al.* (2018) 'Resolving the long-standing puzzles about the observed Secchi depth relationships', *Limnology and Oceanography*, 63(6), pp. 2321–2336. Available at: https://doi.org/10.1002/lno.10940.

Lenth, R. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.0. Available at: https://CRAN.R-project.org/package=emmeans (Accessed 10 Apr. 2025).

Li, Y. *et al.* (2020) 'Bottom-up and top-down effects on phytoplankton communities in two freshwater lakes', *PLOS ONE*. Edited by X. Guo, 15(4), p. e0231357. Available at: https://doi.org/10.1371/journal.pone.0231357.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), pp.18–22. Available at: https://CRAN.R-project.org/doc/Rnews/ (Accessed 10 Apr. 2025).

Liu, M. *et al.* (2023) 'Algal community structure prediction by machine learning', *Environmental Science and Ecotechnology*, 14, p. 100233. Available at: https://doi.org/10.1016/j.ese.2022.100233.

LLCMP (1999) *Loch Leven Catchment Management Plan: The Report of the Loch Leven Area Management Advisory Group*, p. 99.

Love, P.E.D. *et al.* (2023) 'Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction', *Advanced Engineering Informatics*, 57, p. 102024. Available at: https://doi.org/10.1016/j.aei.2023.102024.

Lürling, M. (2003) 'Phenotypic plasticity in the Green Algae Desmodesmus and Scenedesmus with special reference to the induction of defensive morphology', *Annales de Limnologie - International Journal of Limnology*, 39(2), pp. 85–101. Available at: https://doi.org/10.1051/limn/2003014.

Lynch, A.J. *et al.* (2023) 'People need freshwater biodiversity', *WIREs Water*, 10(3), p. e1633. Available at: https://doi.org/10.1002/wat2.1633.

Manca, M. *et al.* (2008) 'Daphnia body size and population dynamics under predation by invertebrate and fish predators in Lago Maggiore: an approach based on contribution analysis', *Journal of Limnology*, 67(1), p. 15. Available at: https://doi.org/10.4081/jlimnol.2008.15.

May, L. (2018) *Water governance at Loch Leven, Scotland*. C05846. NERC/Centre for Ecology & Hydrology, p. 10. Available at: https://nora.nerc.ac.uk/id/eprint/527064/.

May, L. and Spears, B.M. (2012) 'A history of scientific research at Loch Leven, Kinross, Scotland', *Hydrobiologia*, 681(1), pp. 3–9. Available at: https://doi.org/10.1007/s10750-011-0929-4.

May, L. *et al.* (2012) 'Historical changes (1905–2005) in external phosphorus loads to Loch Leven, Scotland, UK', *Hydrobiologia*, 681(1), pp. 11–21. Available at: https://doi.org/10.1007/s10750-011-0922-y.

Met Office (2022) *When does spring start?*, *Met Office*. Available at: https://weather.metoffice.gov.uk/learn-about/weather/seasons/spring/when-does-spring-start (Accessed: 14 April 2025).

Michalak, A.M. *et al.* (2013) 'Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions', *Proceedings of the National Academy of Sciences*, 110(16), pp. 6448–6452. Available at: https://doi.org/10.1073/pnas.1216006110.

Monier, A. *et al.* (2015) 'Oceanographic structure drives the assembly processes of microbial eukaryotic communities', *The ISME Journal*, 9(4), pp. 990–1002. Available at: https://doi.org/10.1038/ismej.2014.197.

Moreira, C., Vasconcelos, V. and Antunes, A. (2022) 'Cyanobacterial Blooms: Current Knowledge and New Perspectives', *Earth*, 3(1), pp. 127–135. Available at: https://doi.org/10.3390/earth3010010.

Morgan, N.C. (1974) 'Historical Background to the International Biological Programme Project at Loch Leven, Kinross', *Proceedings of the Royal Society of Edinburgh. Section B. Biology*, 74, pp. 45–55. Available at: https://doi.org/10.1017/S0080455X00012303.

Mullin, C.A. *et al.* (2020) 'Future Projections of Water Temperature and Thermal Stratification in Connecticut Reservoirs and Possible Implications for Cyanobacteria', *Water Resources Research*, 56(11), p. e2020WR027185. Available at: https://doi.org/10.1029/2020WR027185.

Munro, D. (1994) *Loch Leven and the River Leven: a landscape transformed*. Markinch: River Leven Trust.

Muylaert, K. *et al.* (2010) 'Influence of nutrients, submerged macrophytes and zooplankton grazing on phytoplankton biomass and diversity along a latitudinal gradient in Europe', *Hydrobiologia*, 653(1), pp. 79–90. Available at: https://doi.org/10.1007/s10750-010-0345-1.

Ngupula, G. *et al.* (2014) 'Spatial distribution of soluble reactive silica (SRSi) in the Tanzanian waters of Lake Victoria and its implications for Diatom productivity', *African Journal of Aquatic Science*, 39(1), pp. 109–116. Available at: https://doi.org/10.2989/16085914.2014.888330.

Nguyen, Q.H. *et al.* (2021) 'Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil', *Mathematical Problems in Engineering*. Edited by Y.-S. Shen, 2021, pp. 1–15. Available at: https://doi.org/10.1155/2021/4832864.

Nõges, P. *et al.* (2010) 'The Impact of Variations in the Climate on Seasonal Dynamics of Phytoplankton', in G. George (ed.) *The Impact of Climate Change on European Lakes*. Dordrecht: Springer Netherlands, pp. 253–274. Available at: https://doi.org/10.1007/978-90-481-2945-4_14.

Paerl, H.W. (1996) 'A comparison of Cyanobacterial bloom dynamics in freshwater, estuarine and marine environments', *Phycologia*, 35(sup6), pp. 25–35. Available at: https://doi.org/10.2216/i0031-8884-35-6S-25.1.

Paerl, H.W. and Paul, V.J. (2012) 'Climate change: Links to global expansion of harmful Cyanobacteria', *Water Research*, 46(5), pp. 1349–1363. Available at: https://doi.org/10.1016/j.watres.2011.08.002.

Pálffy, K. and Vörös, L. (2019) 'Phytoplankton functional composition shows higher seasonal variability in a large shallow lake after a eutrophic past', *Ecosphere*, 10(5), p. e02684. Available at: https://doi.org/10.1002/ecs2.2684.

Paltsev, A. *et al.* (2024) 'Phytoplankton biomass in northern lakes reveals a complex response to global change', *Science of The Total Environment*, 940, p. 173570. Available at: https://doi.org/10.1016/j.scitotenv.2024.173570.

Paterson, A.M. *et al.* (2017) 'Climate as a driver of increasing algal production in Lake of the Woods, Ontario, Canada', *Lake and Reservoir Management*, 33(4), pp. 403–414. Available at: https://doi.org/10.1080/10402381.2017.1379574.

Piggott, J.J., Townsend, C.R. and Matthaei, C.D. (2015) 'Reconceptualizing synergism and antagonism among multiple stressors', *Ecology and Evolution*, 5(7), pp. 1538–1547. Available at: https://doi.org/10.1002/ece3.1465.

Planas, D. and Paquet, S. (2016) 'Importance of climate change-physical forcing on the increase of Cyanobacterial blooms in a small, stratified lake', *Journal of Limnology*, 75(s1). Available at: https://doi.org/10.4081/jlimnol.2016.1371.

Põllumäe, A. and Haberman, J. (1998) 'THE EFFECT OF FLUCTUATING WATERLEVEL ON THE ZOOPLANKTON OF LAKE VÕRTSJÄRV, CENTRAL ESTONIA', *Proceedings of the Estonian Academy of Sciences. Biology. Ecology*, 47(4), p. 259. Available at: https://doi.org/10.3176/biol.ecol.1998.4.03.

Rao, K. *et al.* (2021) 'The relative importance of environmental factors in predicting phytoplankton shifting and Cyanobacteria abundance in regulated shallow lakes', *Environmental Pollution*, 286, p. 117555. Available at: https://doi.org/10.1016/j.envpol.2021.117555.

Reid, A.J. *et al.* (2020) 'Conservation Challenges to Freshwater Ecosystems', in M.I. Goldstein and D.A. DellaSala (eds) *Encyclopedia of the World's Biomes*. Oxford: Elsevier, pp. 270–278. Available at: https://doi.org/10.1016/B978-0-12-409548-9.11937-2.

Reynolds, C.S. (2006) *The Ecology of Phytoplankton*. 1st edn. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511542145.

Reynolds, C.S. and Davies, P.S. (2001) 'Sources and bioavailability of phosphorus fractions in freshwaters: a British perspective', *Biological Reviews*, 76(1), pp. 27–64. Available at: https://doi.org/10.1111/j.1469-185X.2000.tb00058.x.

Richardson, J. *et al.* (2018) 'Effects of multiple stressors on Cyanobacteria abundance vary with lake type', *Global Change Biology*, 24(11), pp. 5044–5055. Available at: https://doi.org/10.1111/gcb.14396.

Rigosi, A. *et al.* (2014) 'The interaction between climate warming and eutrophication to promote Cyanobacteria is dependent on trophic state and varies among taxa', *Limnology and Oceanography*, 59(1), pp. 99–114. Available at: https://doi.org/10.4319/lo.2014.59.1.0099.

Rimpler, A., Kiers, H.A.L. and Van Ravenzwaaij, D. (2025) 'To interact or not to interact: The pros and cons of including interactions in linear regression models', *Behavior Research Methods*, 57(3), p. 92. Available at: https://doi.org/10.3758/s13428-025-02613-6.

Sarnelle, O. (1992) 'Nutrient Enrichment and Grazer Effects on Phytoplankton in Lakes', *Ecology*, 73(2), pp. 551–560. Available at: https://doi.org/10.2307/1940761.

Schindler, D.W. (1977) 'Evolution of Phosphorus Limitation in Lakes: Natural mechanisms compensate for deficiencies of nitrogen and carbon in eutrophied lakes.', *Science*, 195(4275), pp. 260–262. Available at: https://doi.org/10.1126/science.195.4275.260.

Scornet, E. (2017) 'Tuning parameters in random forests', *ESAIM: Proceedings and Surveys*. Edited by J.-F. Coeurjolly and A. Leclercq-Samson, 60, pp. 144–162. Available at: https://doi.org/10.1051/proc/201760144.

Shatwell, T., Köhler, J. and Nicklisch, A. (2013) 'Temperature and photoperiod interactions with silicon-limited growth and competition of two Diatoms', *Journal of Plankton Research*, 35(5), pp. 957–971. Available at: https://doi.org/10.1093/plankt/fbt058.

Shurin, J.B., Gruner, D.S. and Hillebrand, H. (2005) 'All wet or dried up? Real differences between aquatic and terrestrial food webs', *Proceedings of the Royal Society B: Biological Sciences*, 273(1582), pp. 1–9. Available at: https://doi.org/10.1098/rspb.2005.3377.

Smith, D.R., King, K.W. and Williams, M.R. (2015) 'What is causing the harmful algal blooms in Lake Erie?', *Journal of Soil and Water Conservation*, 70(2), pp. 27A-29A. Available at: https://doi.org/10.2489/jswc.70.2.27A.

Smith, I.R. (1974) 'The Structure and Physical Environment of Loch Leven, Scotland', *Proceedings of the Royal Society of Edinburgh. Section B. Biology*, 74, pp. 81–100. Available at: https://doi.org/10.1017/S0080455X00012339.

Sommer, U. *et al.* (2012) 'Beyond the Plankton Ecology Group (PEG) Model: Mechanisms Driving Plankton Succession', *Annual Review of Ecology, Evolution, and Systematics*, 43(1), pp. 429–448. Available at: https://doi.org/10.1146/annurev-ecolsys-110411-160251.

Spears, B.M. *et al.* (2012) 'Long-term variation and regulation of internal phosphorus loading in Loch Leven', *Hydrobiologia*, 681(1), pp. 23–33. Available at: https://doi.org/10.1007/s10750-011-0921-z.

Spears, B.M. *et al.* (2021) 'Making waves. Bridging theory and practice towards multiple stressor management in freshwater ecosystems', *Water Research*, 196, p. 116981. Available at: https://doi.org/10.1016/j.watres.2021.116981.

Spears, B.M. *et al.* (2022) 'Assessing multiple stressor effects to inform climate change management responses in three European catchments', *Inland Waters*, 12(1), pp. 94–106. Available at: https://doi.org/10.1080/20442041.2020.1827891.

Sun, X. *et al.* (2022) 'Effects of Algal Blooms on Phytoplankton Composition and Hypoxia in Coastal Waters of the Northern Yellow Sea, China', *Frontiers in Marine Science*, 9, p. 897418. Available at: https://doi.org/10.3389/fmars.2022.897418.

Suthers, I.M. and Rissik, D. (eds) (2009) *Plankton: a guide to their ecology and monitoring for water quality*. Collingwood, Vic London: CSIRO.

Taylor, R., Fletcher, R.L. and Raven, J.A. (2001) 'Preliminary Studies on the Growth of Selected "Green Tide" Algae in Laboratory Culture: Effects of Irradiance, Temperature, Salinity and Nutrients on Growth Rate', *Botanica Marina*, 44(4). Available at: https://doi.org/10.1515/BOT.2001.042.

Tazkiaturrizki, T., Hartono, D.M. and Moersidik, S.S. (2023) 'A critical analysis of potential formation and health risk of disinfection by products in drinking water', *IOP Conference Series: Earth and Environmental Science*, 1239(1), p. 012027. Available at: https://doi.org/10.1088/1755-1315/1239/1/012027.

Vallina, S.M. *et al.* (2017) 'Phytoplankton functional diversity increases ecosystem productivity and stability', *Ecological Modelling*, 361, pp. 184–196. Available at: https://doi.org/10.1016/j.ecolmodel.2017.06.020.

Wang, M. *et al.* (2019) 'Seasonal Pattern of Nutrient Limitation in a Eutrophic Lake and Quantitative Analysis of the Impacts from Internal Nutrient Cycling', *Environmental Science & Technology*, 53(23), pp. 13675–13686. Available at: https://doi.org/10.1021/acs.est.9b04266.

Wang, Y. *et al.* (2025) 'Global elevation of algal bloom frequency in large lakes over the past two decades', *National Science Review*, 12(3), p. nwaf011. Available at: https://doi.org/10.1093/nsr/nwaf011.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K.,

Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), p.1686. doi:10.21105/joss.01686. Available at: https://doi.org/10.21105/joss.01686 (Accessed 10 Apr. 2025).

Wilkinson, A.A., Hondzo, M. and Guala, M. (2019) 'Investigating Abiotic Drivers for Vertical and Temporal Heterogeneities of Cyanobacteria Concentrations in Lakes Using a Seasonal In Situ Monitoring Station', *Water Resources Research*, 55(2), pp. 954–972. Available at: https://doi.org/10.1029/2018WR024228.

Williamson, C.E., Saros, J.E. and Schindler, D.W. (2009) 'Sentinels of Change', *Science*, 323(5916), pp. 887–888. Available at: https://doi.org/10.1126/science.1169443.

Yuan, L.L. and Pollard, A.I. (2018) 'Changes in the relationship between zooplankton and phytoplankton biomasses across a eutrophication gradient', *Limnology and Oceanography*, 63(6), pp. 2493–2507. Available at: https://doi.org/10.1002/lno.10955.

Zanchett, G. and Oliveira-Filho, E. (2013) 'Cyanobacteria and Cyanotoxins: From Impacts on Aquatic Ecosystems and Human Health to Anticarcinogenic Effects', *Toxins*, 5(10), pp. 1896–1917. Available at: https://doi.org/10.3390/toxins5101896.

Zhao, X., Drakare, S. and Johnson, R.K. (2019) 'Use of taxon-specific models of phytoplankton assemblage composition and biomass for detecting impact', *Ecological Indicators*, 97, pp. 447–456. Available at: https://doi.org/10.1016/j.ecolind.2018.10.026.

Zhou, L. *et al.* (2021) 'Photosynthesis acclimation under severely fluctuating light conditions allows faster growth of Diatoms compared with dinoflagellates', *BMC Plant Biology*, 21(1), p. 164. Available at: https://doi.org/10.1186/s12870-021-02902-0.

Zuur, A.F., Ieno, E.N. and Smith, G.M. (2007) *Analysing Ecological Data*. New York, NY: Springer New York (Statistics for Biology and Health). Available at: https://doi.org/10.1007/978-0-387-45972-1.

7. Appendix

7.1 Random Forest

7.1.1 Full list of Parameters

Table 4.1 Variables included in the random forest regression model predicting total phytoplankton biovolume.

Variable	Description	Units
Year	Calendar year of observation	-
Season	Meteorological season (Winter–Autumn)	-
Water Level	Measured from the top of Kinross pier to the water surface	cm
Daphnia	Zooplankton density	individuals/L
Nitrate	Nitrate concentration	mg/L
SRP	Soluble reactive phosphorus	µg/L
SRSi	Soluble reactive silica	mg/L
TP	Total phosphorus	µg/L
SD	Secchi disk depth	m
Temperature	Water temperature	°C
pH	pH level of water	-
DO	Dissolved oxygen concentration	% saturation

7.1.2 Imputed Dataset vs Complete Case Dataset – Comparison of Performance Metrics

To determine the most appropriate dataset for the final model, random forest models were trained using both an imputed dataset and a complete case dataset. After fine-tuning of both models, the complete case model achieved a higher Pearson correlation coefficient value ($r = 0.42$) compared to the imputed model ($r = 0.043$), indicating a much better explanatory power of the complete case model. Although RSME was higher in absolute terms for the complete case model (RSME = 37×10^6) compared to the imputed model (RSME = 14×10^6), this likely reflected the reduced variability introduced by imputation, which tends to fill in missing values with median or model estimates – smoothing the data and narrowing the range of observed values. Most notably, the percentage of variance explained by the models improved from 13.43% to 37.04% when moving from imputed to complete case data. Thus, the decision was made to proceed with the complete-case dataset for the final model results and interpretation.

Table 4.2 Comparison of model performance metrics for fine-tuned random forest models using both imputed and complete case data. While imputation allowed for more data inclusion, the complete case model showed better performance and was selected to proceed with data analysis.

Dataset Used	**mtry**	***r***	**RMSE**	**% Variance Explained**
Imputed	6	0.043	14×10^6	13.43
Complete Case	4	0.42	37×10^6	37.04

7.1.3 Imputed Dataset vs Complete Case Dataset – %IncMSE Variable Importance Scores

Table 4.3 Comparison of variable importance scores (%IncMSE) from imputed and complete-case datasets random forest models. Higher values indicate greater contribution to predictive accuracy of phytoplankton biovolume. Negative values suggest negligible or counterproductive influence on model performance.

Variables	%IncMSE	
	Imputed Dataset	Complete Case Dataset
Year	2.83	-0.91
Season	3.98	-3.01
Water Level	2.92	4.20
Daphnia	1.10	5.42
Nitrate	6.51	6.45
SRP	0.31	6.32
SRSi	5.76	5.25
TP	0.14	-1.60
Secchi Depth	15.57	-0.94
Temperature	3.66	2.87
pH	2.98	-0.92
Dissolved Oxygen	2.38	0.76

7.2 Detailed Model Output – Seasonal and Taxonomic Effects on Phytoplankton Biovolume

Table 4.4 details the results of the constructed linear models, focusing on the fixed effects of season and phytoplankton group, and their interaction term. As done in the results section, focusing on model 7, the best fitting model, only Diatoms showed a statistically significant pattern. Diatoms had a significantly higher overall biovolume compared to Cryptophyceae (5.81 ± 1.83, $p < 0.01$), particularly in summer which was significantly lower than Cryptophyceae in winter (-6.42 ± 2.11, $p < 0.01$), suggesting a strong seasonal effect specific to this group. No other seasonal or group-level effects reached statistical significance.

Table 4.4 Fixed effects estimates (± standard error) from the linear mixed-effected models constructed predicting the log-transformed phytoplankton biovolume – focusing on season, phytoplankton group, and their interaction (*season*phytoplankton group*). Model numbers align with those in Table 2.2, indicating the stepwise inclusion of fixed effects. Significance levels are denoted as follows: *p < 0.05, **p < 0.01, ***p < 0.001. Estimates are relative to the reference category: *Cryptophyceae in winter*.

Mod No	Intercept	Season			Phytoplankton Group			Season*Phytoplankton Group								
	Crypto : Winter	Spring	Summer	Autumn	Cyano	Diatoms	Greens	Cyano : Spring	Diatoms: Spring	Greens: Spring	Cyano : Summer	Diatoms : Summer	Greens : Summer	Cyano : Autumn	Diatoms : Autumn	Greens : Autumn
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	11.23± 0.29 ***	1.26± 0.36 ***	1.86± 0.36 ***	1.08± 0.36 **	-0.94± 0.38 *	3.82± 0.38 ***	0.58± 0.38	-0.13± 0.51	-1.78± 0.50 ***	-0.80± 0.50	0.94± 0.51	-3.44± 0.50 ***	-1.07± 0.50 *	0.92± 0.51	-0.64± 0.51	-0.40± 0.51
3	10.11± 0.78 ***	2.04± 0.74 **	2.60± 0.84 **	1.01± 1.08	-0.14± 0.78	5.20± 0.76 ***	1.39± 0.76	-1.02± 1.06	-3.06± 1.03 **	-1.46± 1.03	0.40± 1.06	-5.05± 1.03 ***	-2.06± 1.03 *	1.01± 1.08	-1.20± 1.06	-0.87± 1.06
4	10.44± 0.79 ***	2.24± 0.74 **	2.68± 0.83 **	1.38± 0.88	-0.12± 0.77	5.20± 0.75 ***	1.39± 0.75	-1.01± 1.05	-3.07± 1.02 **	-1.46± 1.02	0.34± 1.04	-5.05± 1.02 ***	-2.06± 1.02 *	0.99± 1.06	-1.19± 1.05	-0.86± 1.05
5	10.06± 0.89 ***	2.03± 0.77 **	2.12± 1.03 *	1.13± 0.92	-0.13± 0.77	5.20± 0.75 ***	1.39± 0.75	-0.99± 1.06	-3.07± 1.02 **	-1.46± 1.02	0.35± 1.04	-5.05± 1.02 ***	-2.06± 1.02 *	1.02± 1.07	-1.18± 1.05	-0.85± 1.05
6	139.42± 58.95 *	1.97± 0.76 *	2.32± 1.04 *	0.97± 0.91	-0.10± 0.76	5.20± 0.74 ***	1.39± 0.74	-1.06± 1.05	-3.07± 1.01 **	-1.46± 1.01	0.12± 1.06	-5.41± 1.03 ***	-2.31± 1.03 *	1.02± 1.05	-1.14± -1.04	-0.82± 1.04
7	226.14± 100.19 *	2.17± 1.73	1.64± 1.95	-0.88± 2.20	0.56± 1.83	5.81± 1.83 **	0.34± 1.83	-1.91± 2.24	-3.70± 2.24	-0.005± 2.24	-1.43± 2.11	-6.42± 2.11 **	-1.91± 2.11	0.52± 2.16	-1.20± 2.16	0.56± 2.16

7.3 Code

7.3.1 Load and Clean Datasets

1. Monitoring Data

Loads RB5 monitoring data, filters for the target period, and drops unused variables.

```
RB5_monitoring_data <-
read.csv("../data/csv/RB5_monitoring_data_monthly_average_1980-2023.csv")
 RB5_monitoring_data <- RB5_monitoring_data %>%
   filter(between(year, 2004, 2016)) %>%
   select(-RB5.Cond, -RB5.NO2, -RB5.NH4, -RB5.PP, -RB5.PSiO2, -RB5.SURP, -RB5.TSi,
-RB5.TSP, -RB5.TON) %>%
   mutate(YearMonth = as.Date(YearMonth))
```

2. Biovolume Data

Processes monthly biovolume values and standardizes date format.

```
biovol_fulldata <- read.csv("../data/csv/biovolfulldata_2004-2016.csv") %>%
   mutate(YearMonth = as.Date(YearMonth))
```

3. Predator (Daphnia) Data

Selects and processes zooplankton data, aggregates monthly means, and filters for the target timeframe.

```
predator_data <- read.csv("../data/csv/crustaceanzooplankton_1980-2023.csv") %>%
   select(Date, Daphnia.hyalina.ind.L) %>%
   mutate(Date = dmy(Date), year = year(Date), month = month(Date)) %>%
   filter(!is.na(Date)) %>%
   mutate(Daphnia = as.numeric(Daphnia.hyalina.ind.L)) %>%
   select(-Daphnia.hyalina.ind.L) %>%
   group_by(year, month) %>%
   summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
   mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"))) %>%
   filter(between(YearMonth, as.Date("2004-01-01"), as.Date("2016-12-01")))
```

4. Water Level Data

Combines two water level datasets and computes monthly means.

```
waterlevelp2 <- read.csv("../data/csv/waterlevels_1993-2007.csv") %>%
   rename(waterlevel.masl = Harbour.masl) %>%
   mutate(Date = dmy(Date)) %>%
   select(Date, waterlevel.masl)

waterlevelp3 <- read.csv("../data/csv/waterlevels_2008-2013.csv") %>%
   rename(Date = SAMPLE_DATE, waterlevel.masl = Level.maod) %>%
   mutate(Date = dmy(Date)) %>%
   select(Date, waterlevel.masl)

fullwaterlevel <- bind_rows(waterlevelp2, waterlevelp3) %>%
   mutate(year = year(Date), month = month(Date)) %>%
   group_by(year, month) %>%
   summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
   mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"))) %>%
   filter(between(year, 2004, 2016))
```

5. Temperature and SD Data

Processes temperature and standard deviation data, summarising to monthly means.

```
tempsddata <- read.csv("../data/csv/Temp_SD_2004-2016.csv") %>%
   mutate(Date = dmy(Date), RB5.SD = as.numeric(RB5.SD), RB5.Temp =
as.numeric(RB5.Temp),
          year = year(Date), month = month(Date)) %>%
   group_by(year, month) %>%
   summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
   mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))
```

6. pH Data (2010–2016)

Processes pH values, available only for 2010–2016, into monthly means.

```
pHdata <- read.csv("../data/csv/pH_2010-2016.csv") %>%
   select(Date, Value) %>%
   mutate(Date = dmy(Date), RB5.pH = Value,
          year = year(Date), month = month(Date)) %>%
   group_by(year, month) %>%
   summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
   mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))
```

7. DO Data (2010–2016)

Filters and processes dissolved oxygen (% saturation) into monthly averages.

```
DOdata <- read.csv("../data/csv/DO_2010-2016.csv") %>%
   filter(Units == "%") %>%
   select(Date, Value) %>%
   mutate(Date = dmy(Date), RB5.DO = Value,
          year = year(Date), month = month(Date)) %>%
   group_by(year, month) %>%
   summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop') %>%
   mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-")))
```

8. Merge All Data

This step incrementally merges all cleaned datasets into one by matching on the `YearMonth` column.

```
all_data1 <- full_join(RB5_monitoring_data, select(predator_data, YearMonth,
Daphnia), by = "YearMonth")
 all_data2 <- full_join(biovol_fulldata, select(fullwaterlevel, YearMonth,
waterlevel.masl), by = "YearMonth")
 all_data <- full_join(all_data2, all_data1, by = "YearMonth")

all_data <- all_data %>%
   left_join(tempsddata, by = "YearMonth") %>%
   mutate(RB5.SD = coalesce(RB5.SD.x, RB5.SD.y),
          RB5.Temp = coalesce(RB5.Temp.x, RB5.Temp.y)) %>%
   select(-RB5.SD.x, -RB5.SD.y, -RB5.Temp.x, -RB5.Temp.y) %>%
   left_join(pHdata, by = "YearMonth") %>%
   mutate(RB5.pH = coalesce(RB5.pH.x, RB5.pH.y)) %>%
   select(-RB5.pH.x, -RB5.pH.y) %>%
   left_join(DOdata, by = "YearMonth") %>%
   mutate(RB5.DO = coalesce(RB5.DO.x, RB5.DO.y)) %>%
   select(-RB5.DO.x, -RB5.DO.y)
```

9. Final Cleaning

Refines merged dataset by selecting relevant columns, deriving seasonal categories, and calculating total biovolume.

```
all_data <- all_data %>%
   distinct(YearMonth, .keep_all = TRUE) %>%
   select(YearMonth, year.y, month.y, Cryto.Biovolume, Cyano.Biovolume,
          Diatoms.Biovolume, Greens.Biovolume, waterlevel.masl, Daphnia,
          starts_with("RB5")) %>%
   rename(Year = year.y, Month = month.y) %>%
   mutate(Season = case_when(
     Month %in% c(12, 1, 2) ~ "Winter",
     Month %in% c(3, 4, 5) ~ "Spring",
     Month %in% c(6, 7, 8) ~ "Summer",
     Month %in% c(9, 10, 11) ~ "Autumn"
   )) %>%
   mutate(Total.Biovolume = Cryto.Biovolume + Cyano.Biovolume +
                            Diatoms.Biovolume + Greens.Biovolume) %>%
   relocate(Season, .after = Month) %>%
   relocate(Total.Biovolume, .after = Greens.Biovolume)
```

10. Save Final Dataset

Exports the cleaned and merged dataset as CSV files for further analysis.

```
write.csv(all_data, "../data/csv/alldata-2004-2016v2.csv", row.names = FALSE)
```

7.3.2 Random Forest Regression

1. Load and Prepare Data

```
alldata <- read.csv("../data/csv/alldata-2004-2016v2.csv") %>%
   filter(!is.na(Total.Biovolume)) %>%
   select(-YearMonth, -Month, -Cryto.Biovolume, -Cyano.Biovolume, -
Diatoms.Biovolume, -Greens.Biovolume, -RB5.ChlA) %>%
   mutate(Year = as.numeric(Year), Season = as.factor(Season))
 summary(alldata)
```

2. Random Forest with Imputed Data

2.1 Impute Missing Values

```
data_impute <- rfImpute(Total.Biovolume ~ ., data = alldata, ntree = 500)

##      |      Out-of-bag   |
 ## Tree |      MSE  %Var(y) |
 ##  500 | 8.908e+13    90.04 |
 ##      |      Out-of-bag   |
 ## Tree |      MSE  %Var(y) |
 ##  500 | 8.853e+13    89.48 |
 ##      |      Out-of-bag   |
 ## Tree |      MSE  %Var(y) |
 ##  500 | 9.256e+13    93.56 |
 ##      |      Out-of-bag   |
 ## Tree |      MSE  %Var(y) |
 ##  500 | 9.231e+13    93.30 |
```

```
##      |      Out-of-bag    |
## Tree |      MSE  %Var(y) |
##  500 | 9.126e+13    92.24 |
```

2.2. Train-Test Split

```
trainIndex1 <- createDataPartition(data_impute$Total.Biovolume, p = 0.7, list = FALSE)
 trainData1 <- data_impute[trainIndex1, ]
 testData1 <- data_impute[-trainIndex1, ]
```

2.3. Fit Random Forest

```
rfmodel1 <- randomForest(Total.Biovolume ~ ., data = trainData1, ntree = 500, mtry = sqrt(ncol(trainData1)-1), importance = TRUE)
 print(rfmodel1)

##
## Call:
 ##  randomForest(formula = Total.Biovolume ~ ., data = trainData1,      ntree = 500, mtry = sqrt(ncol(trainData1) - 1), importance = TRUE)
##                Type of random forest: regression
 ##                      Number of trees: 500
 ## No. of variables tried at each split: 3
 ##
##           Mean of squared residuals: 4.594635e+13
 ##                     % Var explained: 14.54

rfmodel1$importance

##                      %IncMSE IncNodePurity
 ## Year            4.197329e+11  2.289577e+14
 ## Season          1.471044e+12  1.707812e+14
 ## waterlevel.masl 9.991508e+11  6.201864e+14
 ## Daphnia         2.409907e+12  4.231888e+14
 ## RB5.NO3         4.790721e+12  6.029073e+14
 ## RB5.SRP         8.322700e+11  2.951942e+14
 ## RB5.SRSi        2.935821e+12  6.515032e+14
 ## RB5.TP          1.045831e+12  3.243275e+14
 ## RB5.SD          9.020735e+12  6.990395e+14
 ## RB5.Temp        1.713124e+12  2.887353e+14
 ## RB5.pH          1.060274e+12  3.521908e+14
 ## RB5.DO          1.052345e+12  3.056314e+14

varImpPlot(rfmodel1)
```

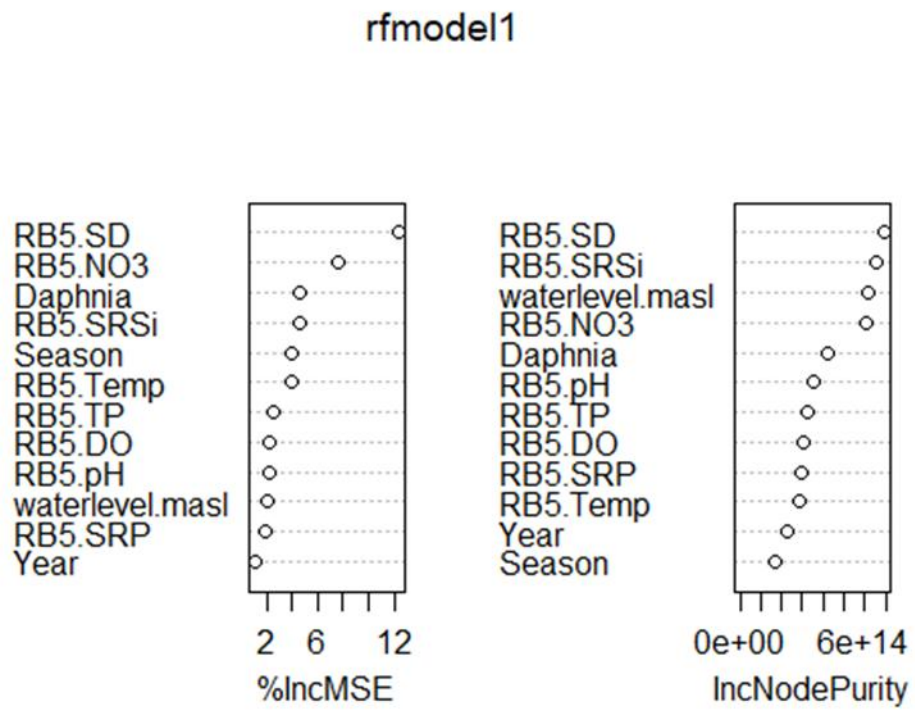


Figure 5.1 Variable Importance Plot for the initial random forest model trained on imputed data – based on %IncMSE and IncNodePurity.

2.4. Evaluate Model

```
predictions <- predict(rfmodel1, testData1)
 rmse <- sqrt(mean((predictions - testData1$Total.Biovolume)^2))
 rsq <- cor(predictions, testData1$Total.Biovolume)^2
 cat("R²:", rsq, "\nRMSE:", rmse)

## R²: 0.04814227
## RMSE: 14318130
```

2.5. Tune Model Hyperparameters

```
tuned_model <- tuneRF(trainData1[, -which(names(trainData1) ==
"Total.Biovolume")], trainData1$Total.Biovolume, stepFactor = 1.5, improve = 0.01,
ntreeTry = 500, trace = TRUE)

## mtry = 4  OOB error = 4.596766e+13
## Searching left ...
 ## mtry = 3     OOB error = 4.568297e+13
## 0.006193347 0.01
## Searching right ...
 ## mtry = 6     OOB error = 4.684678e+13
## -0.0191247 0.01
```

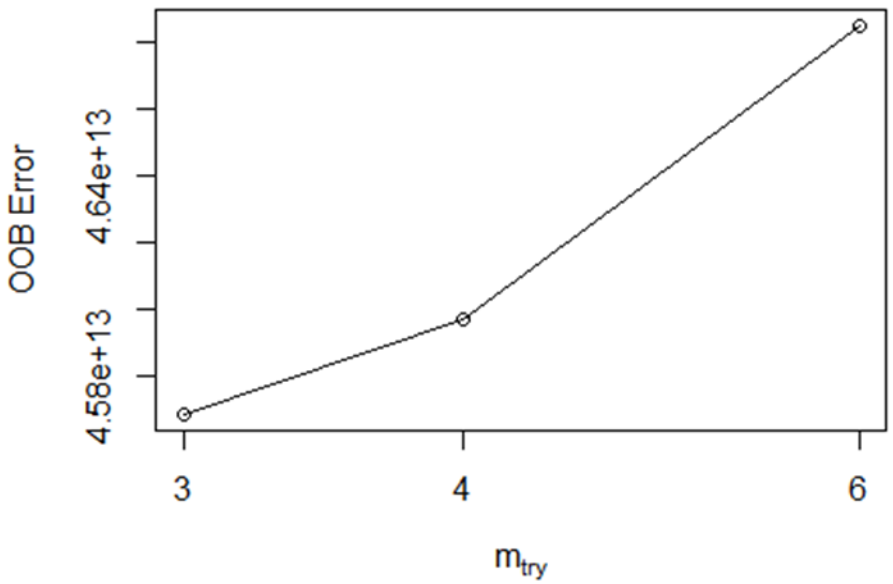


Figure 5.2 Out-of-bag (OOB) error across different mtry values for the initial random forest trained on imputed data.

2.6. Final Tuned Model

```
rfmodel1opt <- randomForest(Total.Biovolume ~ ., data = trainData1, ntree = 500,
mtry = 6, importance = TRUE)
 print(rfmodel1opt)

##
## Call:
 ##  randomForest(formula = Total.Biovolume ~ ., data = trainData1,      ntree =
500, mtry = 6, importance = TRUE)
##               Type of random forest: regression
 ##                     Number of trees: 500
 ## No. of variables tried at each split: 6
 ##
##           Mean of squared residuals: 4.654127e+13
 ##                     % Var explained: 13.43

rfmodel1opt$importance

##                     %IncMSE IncNodePurity
 ## Year            1.059677e+12  2.314507e+14
 ## Season          1.489809e+12  1.435579e+14
 ## waterlevel.masl 1.481133e+12  7.167117e+14
 ## Daphnia         5.039160e+11  2.988729e+14
 ## RB5.NO3         4.391759e+12  6.564095e+14
 ## RB5.SRP         1.197435e+11  2.685283e+14
 ## RB5.SRSi        3.446938e+12  7.289899e+14
 ## RB5.TP          5.630557e+10  2.737771e+14
 ## RB5.SD          1.534417e+13  9.065972e+14
 ## RB5.Temp        1.632840e+12  2.557006e+14
```

```
 ## RB5.pH          1.221846e+12  3.033555e+14
 ## RB5.DO          1.156705e+12  2.667971e+14

predictions1opt <- predict(rfmodel1opt, testData1)
 rmse1opt <- sqrt(mean((predictions1opt - testData1$Total.Biovolume)^2))
 rsq1opt <- cor(predictions1opt, testData1$Total.Biovolume)^2
 cat("R²:", rsq1opt, "\nRMSE:", rmse1opt)

## R²: 0.04315796
## RMSE: 14361269

varImpPlot(rfmodel1opt)
```

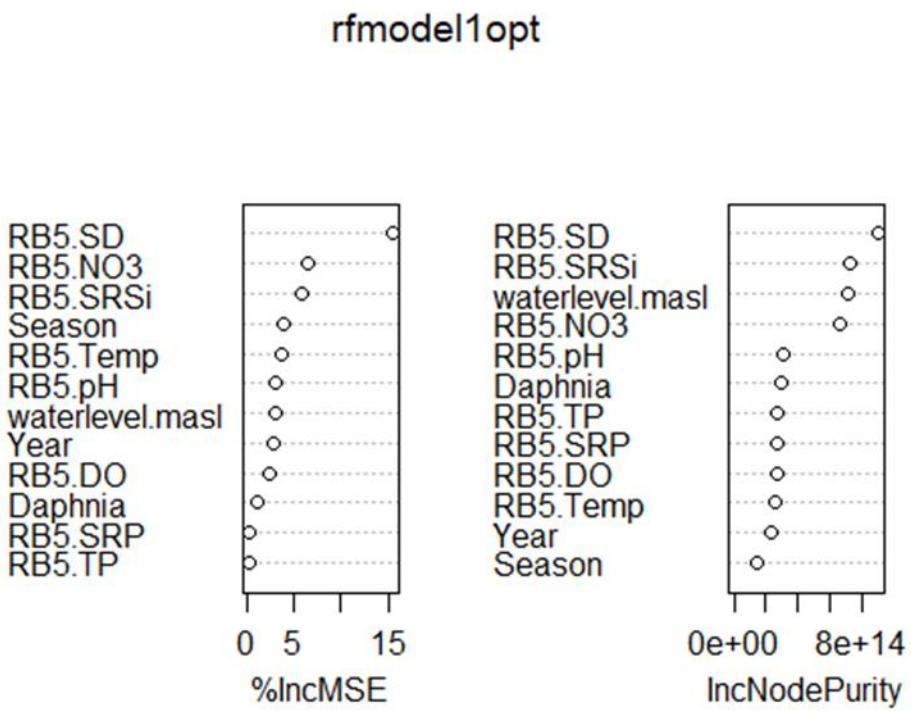


Figure 5.3 Variable Importance Plot for the fine-tuned random forest model trained on imputed data – based on %IncMSE and IncNodePurity.

3. Random Forest on Complete Cases Only

```
set.seed(40)
 noNAdata <- alldata %>% drop_na()
```

3.1. Train-Test Split

```
trainIndex2 <- createDataPartition(noNAdata$Total.Biovolume, p = 0.7, list =
FALSE)
 trainData2 <- noNAdata[trainIndex2, ]
 testData2 <- noNAdata[-trainIndex2, ]
```

3.2. Fit Model

```
rfmodel2 <- randomForest(Total.Biovolume ~ ., data = trainData2, ntree = 500, mtry
= sqrt(ncol(trainData2)-1), importance = TRUE)
 print(rfmodel2)
```

```
##
## Call:
 ##  randomForest(formula = Total.Biovolume ~ ., data = trainData2,      ntree =
500, mtry = sqrt(ncol(trainData2) - 1), importance = TRUE)
##                Type of random forest: regression
 ##                      Number of trees: 500
 ## No. of variables tried at each split: 3
 ##
##           Mean of squared residuals: 7.809481e+13
 ##                     % Var explained: 31.54

rfmodel2$importance

##                       %IncMSE IncNodePurity
 ## Year            -2.621286e+10  1.668237e+12
 ## Season          -2.325194e+12  4.117518e+13
 ## waterlevel.masl  1.466149e+12  1.524454e+14
 ## Daphnia          1.277275e+13  1.042920e+14
 ## RB5.NO3          1.463132e+13  1.611270e+14
 ## RB5.SRP          2.504472e+13  2.669766e+14
 ## RB5.SRSi         1.406638e+13  2.040490e+14
 ## RB5.TP           4.451574e+11  4.491608e+13
 ## RB5.SD          -1.951062e+12  4.412091e+13
 ## RB5.Temp         4.492040e+12  1.309090e+14
 ## RB5.pH          -7.652727e+11  3.509055e+13
 ## RB5.DO           1.349827e+12  3.827267e+13

predictions2 <- predict(rfmodel2, testData2)
 rmse2 <- sqrt(mean((predictions2 - testData2$Total.Biovolume)^2))
 rsq2 <- cor(predictions2, testData2$Total.Biovolume)^2
 cat("R²:", rsq2, "\nRMSE:", rmse2)

## R²: 0.4345314
## RMSE: 37353204

varImpPlot(rfmodel2)
```

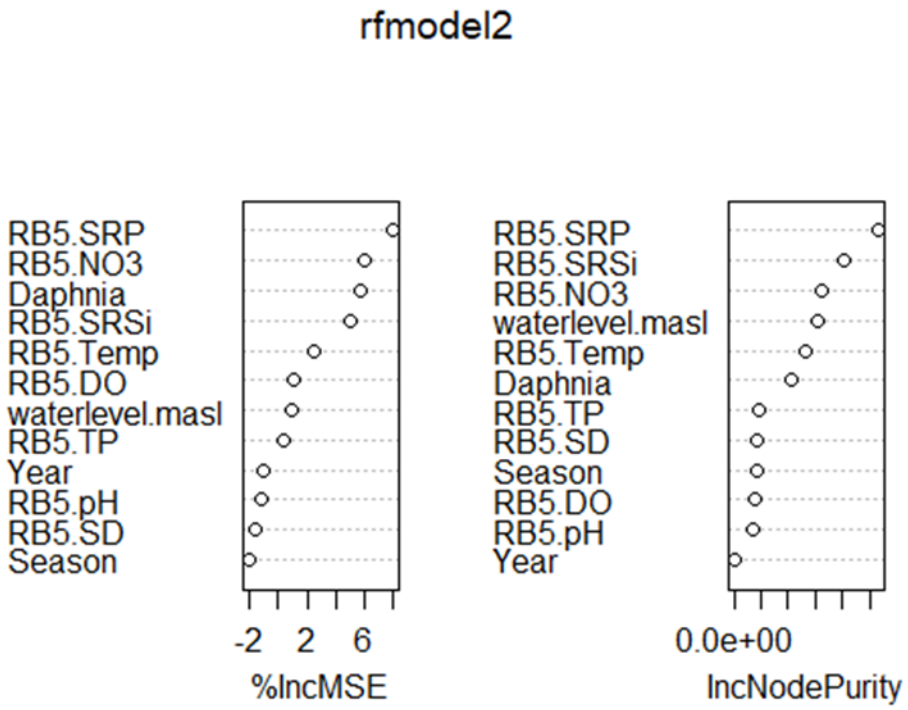


Figure 5.4 Variable Importance Plot for the initial random forest model trained on complete case data – based on %IncMSE and IncNodePurity.

3.3. Tune Final Model

```
tuned_model2 <- tuneRF(trainData2[, -which(names(trainData2) ==
"Total.Biovolume")], trainData2$Total.Biovolume, stepFactor = 1.5, improve = 0.01,
ntreeTry = 500, trace = TRUE)

## mtry = 4  OOB error = 7.89978e+13
## Searching left ...
 ## mtry = 3     OOB error = 7.632914e+13
## 0.03378144 0.01
## mtry = 2     OOB error = 8.372919e+13
## -0.09694922 0.01
## Searching right ...
 ## mtry = 6     OOB error = 7.14342e+13
## 0.06412938 0.01
## mtry = 9     OOB error = 8.129322e+13
## -0.1380154 0.01
```

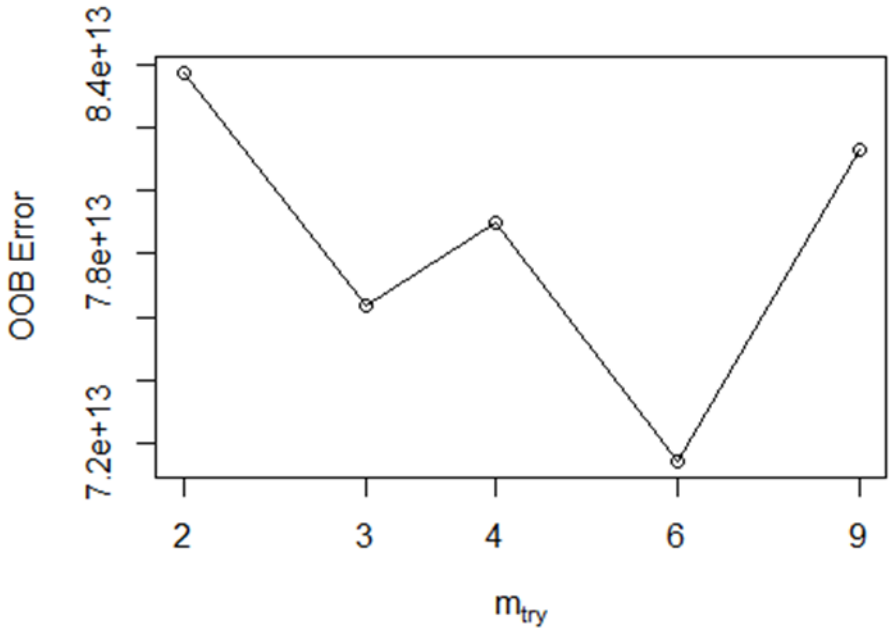


Figure 5.5 Out-of-bag (OOB) error across different mtry values for the initial random forest trained on complete case data.

```
rfmodel2opt <- randomForest(Total.Biovolume ~ ., data = trainData2, ntree = 500,
mtry = 4, importance = TRUE)
 print(rfmodel2opt)

##
## Call:
 ##  randomForest(formula = Total.Biovolume ~ ., data = trainData2,      ntree =
500, mtry = 4, importance = TRUE)
##                Type of random forest: regression
 ##                      Number of trees: 500
 ## No. of variables tried at each split: 4
 ##
##           Mean of squared residuals: 7.18227e+13
 ##                     % Var explained: 37.04

rfmodel2opt$importance

##                        %IncMSE IncNodePurity
 ## Year            -8.465859e+10  2.618577e+12
 ## Season          -3.653437e+12  6.577670e+13
 ## waterlevel.masl  9.305259e+12  1.619684e+14
 ## Daphnia          1.078503e+13  8.345826e+13
 ## RB5.NO3          1.891470e+13  2.167845e+14
 ## RB5.SRP          1.972447e+13  2.645145e+14
 ## RB5.SRSi         1.365172e+13  2.014512e+14
 ## RB5.TP          -1.467488e+12  3.168691e+13
 ## RB5.SD          -9.548901e+11  2.833717e+13
 ## RB5.Temp         5.520234e+12  9.475459e+13
 ## RB5.pH          -6.353382e+11  1.565802e+13
 ## RB5.DO           9.002710e+11  3.550268e+13
```

```r
predictions2opt <- predict(rfmodel2opt, testData2)
 rmse2opt <- sqrt(mean((predictions2opt - testData2$Total.Biovolume)^2))
 rsq2opt <- cor(predictions2opt, testData2$Total.Biovolume)^2
 cat("R²:", rsq2opt, "\nRMSE:", rmse2opt)

## R²: 0.4212883
## RMSE: 37758025
```

4. Variable Importance Plot (Renamed)

```r
imp_df <- as.data.frame(importance(rfmodel2opt, type = 1))
 imp_df$vars <- rownames(imp_df)
 imp_df$vars <- recode(imp_df$vars,
                       "RB5.SRP" = "SRP",
                       "RB5.NO3" = "Nitrate",
                       "RB5.SRSi" = "SRSi",
                       "RB5.Temp" = "Temperature",
                       "waterlevel.masl" = "Water Level",
                       "RB5.DO" = "DO",
                       "RB5.pH" = "pH",
                       "RB5.TP" = "TP",
                       "RB5.SD" = "Secchi Depth")
 imp_df <- imp_df[order(imp_df$`%IncMSE`), ]
 dotchart(imp_df$`%IncMSE`, labels = imp_df$vars, xlim = c(-2,
max(imp_df$`%IncMSE`, na.rm = TRUE) * 1.1),
          pch = 1, xlab = "% Increase in MSE", ylab = "Predictors")
```

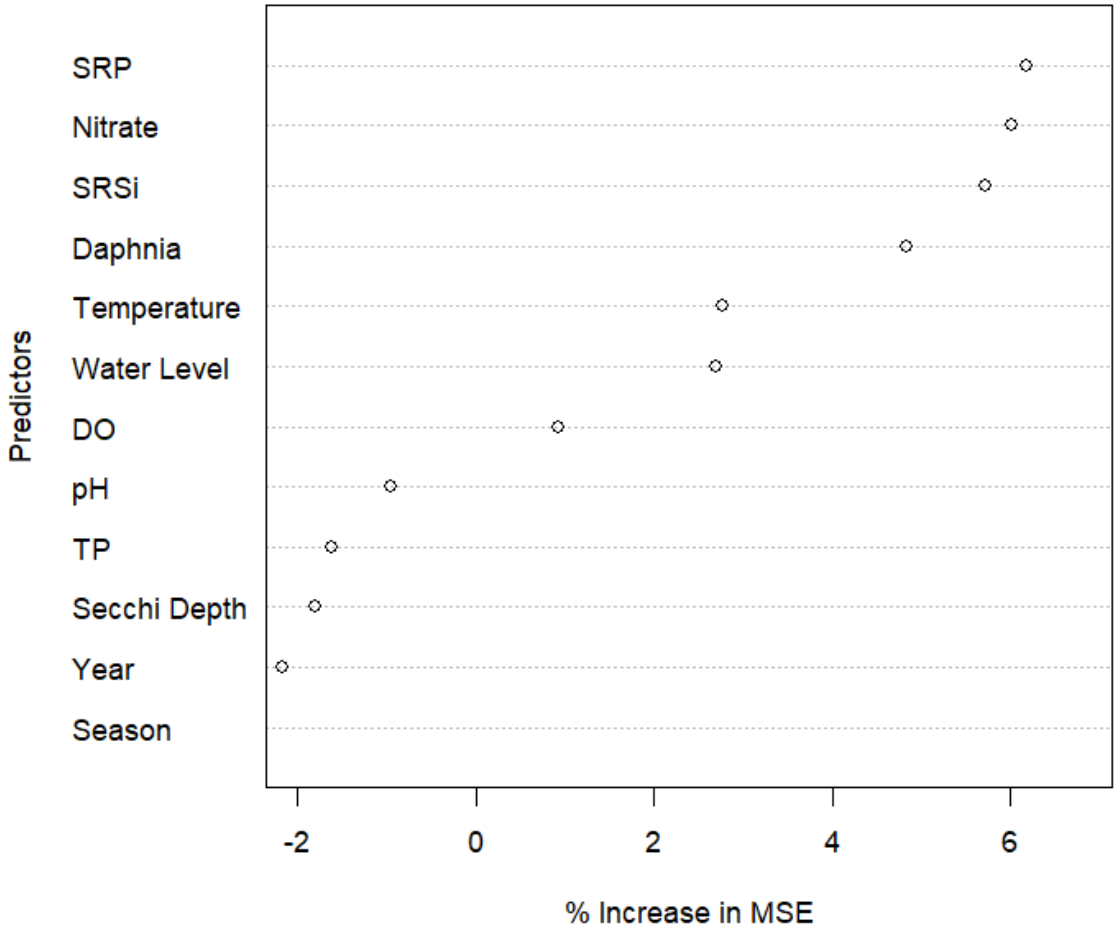


Figure 5.6 Ranked variable importance for the final random forest model using complete-case data, shown by %IncMSE.