

A changing climate for lake monitoring - harnessing satellite imagery and machine learning for tracking and forecasting algal blooms

Daniel Atton Beckmann

Submitted to:

Biological and Environmental Sciences

University of Stirling, Scotland, United Kingdom

December 2025

For the degree of Doctor of Philosophy

Supervised by:

Ian Jones

Evangelos Spyarakos

Peter Hunter

**UNIVERSITY of
STIRLING**



Abstract

Bloom-forming algae, particularly toxin-producing cyanobacteria, pose significant risks to inland water quality. Consequently, developing algae monitoring and forecasting capabilities is critical for understanding drivers, tracking long-term trends, and delivering early warnings. Satellite remote sensing and machine learning (ML) show considerable promise for addressing these challenges, but at present, operational algal bloom monitoring and forecasting programs are rare, and have only been implemented for large, economically-significant lakes. Therefore, extending these capabilities to smaller, less well-studied lakes is critical. Consequently, it is important to understand data requirements for ML forecasting and explore ways to monitor lakes which have, historically, been too small to study using satellites. Furthermore, climate change may exacerbate algal blooms and diminish the predictability of freshwater ecosystems, and so it is critical to understand and prepare for this.

This thesis addresses these challenges through three novel studies: Chapter two explores data requirements for short-term ML algae forecasts. Subsequently, Chapter three evaluates the impact of several climate change scenarios on ML water temperature forecast performance. Finally, Chapter four evaluates the capabilities of new, high-resolution satellite imagery for monitoring small lakes.

It is found that ML forecasts require approximately five or more years of fortnightly training data to be effective, but that performance improvements show diminishing returns as more training data are used. Furthermore, more extreme climate change scenarios are likely to diminish forecast performance. However, high-resolution satellite imagery may help to address these challenges with increased sampling frequency that can potentially offset climate-induced performance losses; and potential for widespread lake monitoring, which may lead to a step-change in the size of training datasets readily available for forecasting.

Collectively, this work highlights potential future directions for algal monitoring and forecasting practices, and ultimately underlines the importance of harnessing synergies across diverse data sources for addressing present and future water quality challenges.

Acknowledgements

Completing a PhD thesis is a long, difficult, and complex process. During this time I experienced some perhaps typical, but nonetheless taxing challenges, including struggles with fieldwork permissions; sensor failures; a computer fault requiring a factory reset just one month before final submission; various health issues; and some quite serious motivational struggles. Overcoming these challenges was certainly only possible with support from a huge number of people, and so here I will do my best to acknowledge those who played some part in this journey.

My supervisory team: Dr Ian Jones, Dr Evangelos Spyarakos, and Dr Peter Hunter – thank you for the continual support, reassurance, guidance, and kindness throughout this process. I could not have asked for better supervisors, and so feel incredibly lucky to have worked with you on this PhD and other projects.

The various colleagues and friends who have helped in many ways with field, lab, and computer work: particularly David Bryan and Charlie Kenny for covering some critical sampling dates (whilst I was on holiday); Nagendra Jaiganesh and Dalin Jiang for support with lab work and imagery analyses; Maud Siebers, Conor McGlinchey, and Mahtab Yaghouti for many useful (and interesting) conversations about satellite imagery analyses, modelling and PhD procedures.

Those who have co-authored papers published as part of this PhD: Mortimer Werther, Eleanor Mackay, Tom Shatwell - thank you for your insightful comments and great feedback. In particular, I am very grateful to Mortimer for our semi-regular meetings which kept me moving through some quite challenging stages of data analysis.

The technical support staff in BES, particularly Ian Washbourne, Ronnie Balfour and Scott Jackson. I am especially thankful to Scott for your tentative and focused IT support as I faced a serious laptop fault so close to submission.

This PhD was funded by the Hydronation Scholars Programme. I am grateful for the many opportunities which this enabled – various training courses, conference trips, and the excellent Hydronation Scholars field trip. Moreover, through this programme I have been part of a brilliant community of Hydronation PhD students, and have learnt a great deal from this group – thanks to all of you for this.

It would not have been at all possible to complete this PhD without a strong group of friends and family. I am particularly grateful to the wonderful folk that I've shared various houses with during this time – especially Nell, Nick, Connie, Tom, Erin, Jessie, Andre, and Gabi; everyone that I have played music with; and all those who have joined me for journeys on foot, bicycle, and ski in the Scottish hills and mountains. Additionally, I would like to give special thanks to Felix, Seth, and Dougal for providing infrequent, yet essential deep discussion and merriment at some of Edinburgh's finest drinking establishments. Lastly, massive thanks to my parents, Kate and Chris, for providing inspiration, hospitality and neverending support, and to my brother, Jacob, for many much-needed adventures.

Statement on ethics, use of generative artificial intelligence software, and plagiarism

I hereby confirm that all work included in this thesis has been approved by the University of Stirling's ethics review panel under the following application form: Using satellite remote sensing and automated in situ sensors to monitor and predict cyanobacterial blooms in multiple lakes 7021, which was approved on the 2nd of February 2022.

Furthermore, I confirm that whilst I have used generative artificial intelligence software for this research, this has been minimal, and solely as an aid for writing small sections of analysis code more efficiently. I have thoroughly reviewed and tested all AI-generated code, and therefore take full ownership and responsibility for the validity and correctness of all data analyses contained herein. Furthermore, I can confirm that none of the writing of this thesis was done using generative artificial intelligence software.

I confirm that this dissertation is an original piece of work conducted independently by the undersigned, and all work contained herein has not been submitted for any other degree. All research material has been duly acknowledged and cited. I consent to the use of plagiarism detection software for testing this.

All three research chapters in this thesis have been published in peer-reviewed academic journals and have therefore had contributions from several co-authors. However, all analyses, data collection, figure generation, and initial draft writing were carried out by myself. Further details on author contributions are included at the start of each chapter.

Signature of Candidate:

A handwritten signature in dark ink, appearing to read 'D. Beckmann', with a long horizontal flourish extending to the right.

Daniel Atton Beckmann

19th December 2025

Table of Contents

1	<i>Introduction</i>	18
1.1	Overview	18
1.2	Cyanobacteria	19
1.2.1	Cyanobacterial blooms: a costly, global water quality issue	20
1.2.2	Cyanobacterial blooms and climate change	22
1.3	Freshwater lakes & algal blooms	23
1.3.1	Understanding blooms – lake physics	23
1.3.2	Understanding blooms – lake chemistry	25
1.3.3	Understanding blooms – phytoplankton dynamics	26
1.4	Monitoring Algae	27
1.4.1	Manual Monitoring Methods	27
1.4.2	Aquatic Remote Sensing	30
1.4.3	Automated in situ sensors	37
1.4.4	Citizen Science	39
1.5	Forecasting algae	41
1.5.1	Process-based modelling and forecasting	42
1.5.2	Data-driven forecasting	43
1.6	Research Priorities	51
2	<i>Are more data always better? Machine learning forecasting of algae based on long-term observations</i>	53
2.1	Abstract	54
2.2	Introduction	54
2.3	Material and methods	58
2.3.1	Study site	58
2.3.2	Algae forecasting	60
2.3.3	Statistical Analysis of Performance	63
2.3.4	Importance of model features	64
2.3.5	Sampling frequency	65
2.4	Results	66
2.4.1	Historical patterns and non-stationarity	66
2.4.2	Effect of training data on model performance	67
2.4.3	Feature Importance	69
2.4.4	Sampling Frequency	72
2.5	Discussion	74

2.5.1	How many years of data are needed?.....	74
2.5.2	Model Selection	76
2.5.3	Importance of input features	77
2.5.4	Sampling frequency	79
2.6	Conclusion.....	79
3	<i>How climate change erodes short-term lake-temperature predictability: informing climate resilient lake forecasting.....</i>	81
3.1	Abstract.....	82
3.2	Introduction.....	82
3.3	Materials and Methods.....	85
3.3.1	Dataset and study sites	85
3.3.2	Forecasting model and approach.....	88
3.3.3	Statistical analysis	91
3.4	Results	93
3.4.1	Effect of climate scenario on forecast performance.....	93
3.4.2	Interaction between training dataset duration and climate scenario	96
3.4.3	Analysis of relative effect sizes	97
3.4.4	Drivers of performance differences	99
3.5	Discussion.....	104
3.5.1	The impact of climate change on water temperature forecasting	104
3.5.2	Variation across lakes and seasons	105
3.5.3	Relevance to other water quality metrics	106
3.5.4	Forecast model choice implications.....	107
3.5.5	Evaluating forecast skill with synthetic data.....	108
3.5.6	Designing climate-resilient water quality forecasting systems.....	108
3.6	Conclusions	109
4	<i>Widespread phytoplankton monitoring in small lakes: a case study comparing satellite imagery from Planet SuperDoves and ESA Sentinel-2</i>	111
4.1	Abstract.....	112
4.2	Introduction.....	112
4.3	Material and methods.....	116
4.3.1	Study site.....	116
4.3.2	In situ data collection	116
4.3.3	Satellite imagery	119
4.3.4	Bloom detection	122
4.3.5	Spatial resolution analysis.....	123

4.3.6	Bloom detection in other small water bodies.....	123
4.4	Results	124
4.4.1	Atmospheric correction.....	124
4.4.2	Chlorophyll-a algorithm evaluation.....	127
4.4.3	Bloom detection.....	127
4.4.4	Spatial analyses.....	129
4.4.5	Bloom detection in other small water bodies.....	131
4.5	Discussion	132
4.5.1	Atmospheric correction.....	132
4.5.2	Chlorophyll-a retrieval and algal bloom detection	133
4.5.3	Spatial resolution	134
4.5.4	Wider applicability to other small water bodies	135
4.6	Conclusion	137
5	Discussion.....	138
5.1	Rising above the plateau – is it possible to further increase forecast performance?	
	138	
5.1.1	Forecast performance plateaus.....	138
5.1.2	Limited examples of critical training data	140
5.1.3	Improving performance through increasing training dataset size.....	141
5.2	Achieving widespread monitoring in smaller lakes	144
5.2.1	The challenges of small water bodies	144
5.2.2	The costs of high-resolution imagery	146
5.2.3	Sustaining operational funding	148
5.2.4	Accelerating sensor technologies.....	149
5.3	Synergies across diverse data sources	150
5.3.1	Citizen science and satellite imagery.....	150
5.3.2	Combining ML and process-based modelling data streams	152
5.3.3	Manual sampling alongside automated data collection	154
5.4	“Artificial Intelligence” trajectories – Self-Supervised Learning and Foundation models 157	
5.4.1	The recent successes of large language models.....	157
5.4.2	Large pre-trained cross-domain time series forecasters	157
5.4.3	Foundation models for remote sensing.....	159
5.5	Climate change – understanding impacts and preparing for future risks.....	161
6	Conclusions	164
7	Bibliography	166

List of figures

Figure 1.1: Illustration of different approaches to lake sampling.	30
Figure 1.2: An illustration showing the key paths and interactions which sunlight makes to a satellite sensor viewing an inland water body.	32
Figure 1.3: Typical autonomous above water radiance and irradiance sensor.....	39
Figure 1.4: Example diagram of a decision tree with four leaf nodes and a depth of two.	45
Figure 1.5: Fully connected multi-layer perceptron architecture.	46
Figure 1.6: Fundamental units of different recurrent neural network architectures.....	48
Figure 1.7: ML performance curves.....	50
Figure 2.1: Decadal means of (a) Chl-a, (b) nitrate, (c) total phosphorus, and (d) ammonium.....	67
Figure 2.2: Performance metrics, (a) RMSE, and (b) MAPE plotted against number of training years	68
Figure 2.3: Performance comparisons of persistence forecast against (a) MLP, (b) RF, (c) ridge regression, and (d) SVM	69
Figure 2.4: Overall average permutation feature importances	70
Figure 2.5: RMSE (left column) and MAPE (right column) plotted against the number of training years with the number of input features varied.....	71
Figure 2.6: Difference in (a) RMSE, and (b) MAPE between models trained with a reduced feature set and models trained with all 13 features	72
Figure 2.7: RMSE (left column) and MAPE (right column) plotted against the number of training years for two different forecast horizons	73
Figure 3.1: Map illustrating the geographical locations of the lakes used in this study. ...	86
Figure 3.2: Satellite imagery showing the lakes used in this study.....	86
Figure 3.3: Example of measured (black dashed lines) and simulated (blue lines) surface water temperature time series from Windermere (upper plot) and Esthwaite (lower plot)	88
Figure 3.4: MAPE against 2100 anthropogenic radiative forcing separated by season and lake	94
Figure 3.5: Forecast model MAE (top row) and MAPE (bottom row) plotted against number of training years for different sampling periods.....	97
Figure 3.6: Forecast model MAE (a) and MAPE (b), plotted against the 2100 anthropogenic radiative forcing associated with each of the four climate scenarios	98
Figure 3.7: Absolute difference in summer MAE for (a) RF and (b) GRU	99

Figure 3.8: Permutation importance analysis, showing mean increase in MAE due to the selective permutation of each variable	100
Figure 3.9: Linear regression analyses from summer months between climate change scenarios and key feature means (left column) and absolute values of the differenced time series of key features (right column)	101
Figure 3.10: Linear regression analyses from summer months between key feature means (left column), absolute values of the differenced time series of key features (right column) and MAPE	102
Figure 4.1: Aerial photograph of Airthrey Loch	117
Figure 4.2: Photograph showing the DALEC sensor deployed on Airthrey Loch.....	118
Figure 4.3: Atmospheric correction performance metrics: R^2 , slope, RMSE, MAPD, and MAD plotted against wavelength	125
Figure 4.4: SuperDoves remote sensing reflectance plotted against in situ remote sensing reflectance measurements.....	126
Figure 4.5: Sentinel-2 remote sensing reflectance plotted against in situ remote sensing reflectance measurements.....	126
Figure 4.6: Satellite chl-a plotted against in situ measurements	127
Figure 4.7: Median chl-a time series from Western basin of Airthrey loch.....	128
Figure 4.8: F1 Scores from bloom detection task.....	129
Figure 4.9: Satellite chl-a maps	130
Figure 4.10: Comparison of SuperDoves chl-a image from 2023-05-29 at three different spatial resolutions	131
Figure 4.11: Comparison of Superdoves OC3 chl-a images from five small water bodies in central Scotland	132
Figure 5.1: Empirical cumulative distribution function of the Blelham Tarn differenced chl-a time series.	141
Figure 5.2: Mock design for a spatial mapping component of a mobile phone application for citizen science algal bloom monitoring.	151
Figure 5.3: Diagram illustrating synergies between various data sources and modelling approaches used in this thesis.	155

List of tables

Table 1.1: Examples of widely used algorithms for retrieval of algae-related parameters from remote sensing imagery.....	36
Table 2.1: Summary of variables, and sources of data used in the study to train and test ML forecast models.....	59
Table 2.2: Hyperparameters obtained from randomized cross validation search for the models used in this study	63
Table 2.3: Minimum number of years training data needed to outperform the persistence forecast	68
Table 3.1: Summary of the geophysical properties of the lakes used for analysis	85
Table 3.2: Climate change scenarios used for experiments.	87
Table 3.3: Metrics comparing day-to-day and fortnightly variability of surface temperature time series.....	89
Table 3.4: Tuned hyperparameters for the Random Forest (RF) and Gated Recurrent Unit (GRU) network models	90
Table 3.5: Summary of fixed effects from standardised LMM regression	95
Table 3.6: Summary of random effects from standardised LMM regression	95
Table 3.7: Difference in average maximum summer error (°C) between SSP585 and SSP126	96
Table 3.8: Complete results from correlation analyses with key features and MAPE.....	103
Table 3.9: Complete results from correlation analyses with climate change scenario and key features	104
Table 4.1: Number of match-ups obtained for reflectance measurements (DALEC), and chl-a samples for both sensors.	119
Table 4.2: Location, area and mean depth of water bodies used for additional analysis.	124
Table 5.1: Approximate uncompressed storage requirements for an entire year of imagery for every lake in the world for three different satellite sensors.....	147

List of Abbreviations

Abbreviation	Definition
AC	Atmospheric Correction
ACOLITE	Atmospheric Correction for Operational Land Imager ‘lite’
AI	Artificial Intelligence
ANN	Artificial Neural Network
AOI	Area of interest
C2RCC	Case 2 Regional Coast Colour (atmospheric correction)
C2X_COMPLEX_nets	C2RCC processor for complex, extreme case-2 waters
C2X_nets	C2RCC processor for extreme case-2 waters
CDOM	Coloured dissolved organic matter
CZCS	Coastal Zone Colour Scanner
DALEC	Dynamic Above-water Radiance and Irradiance Collector
DO	Dissolved oxygen
ELISA	Enzyme-linked immunosorbent assay
ESA	European Space Agency
FAI	Floating Algae Index
GLM	General Lake Model
GOTM	General Ocean Turbulence Model
GPT	Generative pre-trained transformer
GRU	Gated Recurrent Unit
HPLC	High-Performance Liquid Chromatography
ISIMIP	Inter-Sectoral Impact Model Intercomparison Project
LLM	Large language model

LMM	Linear mixed model
LSTM	Long-short-term memory
MAD	Mean average difference
MADt	Mean of absolute differenced time series
MAE	Mean absolute error
MAPD	Median absolute percentage difference
MAPE	Mean absolute percentage difference
MDN	Mixture density network
MERIS	Medium Resolution Imaging Spectrometer
ML	Machine learning
MLP	Multi-layer perceptron
MODIS	Moderate Resolution Imaging Spectrometer
MOS	Modular Optoelectronic Scanner
MSI	Multispectral imager (Sentinel-2 Sensor)
NAP	Non-algal particles
NASA	United States National Aeronautics and Space Administration
NDCI	Normalised difference chlorophyll index
NDVI	Normalised difference vegetation index
NIR	Near-infrared
NOAA	National Oceanic and Atmospheric Administration
OC2, 3, 4	Ocean colour 2, 3, 4 (chlorophyll-a algorithms)
OLCI	Ocean and Land Colour Imager
OLI	Operational Land Imager (Landsat Sensor)
PB	Petabyte
PBC	Public benefit company

PC	Phycocyanin
PIC	Pre-industrial control
PP2A	Protein phosphatase 2A assay
RBD	Red-band difference algorithm
RBF	Radial basis function
RBR3	Red-band ratio algorithm (3-band version)
RF	Random forest
RMSE	Root mean square error
RNN	Recurrent neural network
S2	Sentinel-2
SABI	Surface algal bloom index
SDDt	Standard deviation of differenced time series
SeaWiFS	Sea-viewing Wide Field-Of-View Sensor
SEPA	Scottish environmental protection agency
SR	Surface reflectance
SRP	Soluble reactive phosphorus
SSI	Surface scum index
SSP	Shared socioeconomic pathway
SVM	Support Vector Machine
SVR	Support Vector Regressor
SWIR	Short-wave infrared
TB	Terabyte
TP	Total phosphorus
UDM	Useable data mask
UK	United Kingdom

UKCEH	United Kingdom Centre for Ecology and Hydrology
UKLO	United Kingdom Lakes Observatory
US / USA	United States of America
USD	United States Dollar
UTC	Coordinated Universal Time
WHO	World Health Organisation

1 Introduction

1.1 Overview

Fresh water is essential to nearly every aspect of human life. We drink it, bathe in it, play in it, grow food with it, and maintain long-standing cultural and spiritual connections to it. Yet, this relatively scarce resource is under increasing pressure. Expanding industry (Akinawo, 2023), intensive agriculture (Moss, 2007), and urbanisation (Stokal *et al.*, 2021) are contributing to rising nutrient levels in freshwater systems, which can lead to a range of water quality issues - including the growing prevalence of harmful algal blooms, large accumulations of potentially toxic algae (Anderson, Glibert and Burkholder, 2002; O'Neil *et al.*, 2012; Brooks *et al.*, 2016). In parallel, our unsustainable reliance on fossil fuels is warming the climate and altering weather patterns, which may further exacerbate algal bloom events (Paerl and Huisman, 2009; Paerl *et al.*, 2020).

Driven by increasing water quality concerns, water monitoring technology has advanced rapidly over the past few decades. In situ sensors can now provide high-frequency, real-time measurements of water quality parameters, and freely available, high-quality satellite imagery enables automated estimation of algal concentrations at scale, from space.

Furthermore, recent advances in computing allow for the processing and even automated interpretation of these very large data streams. Machine learning models, for example, are increasingly being used to forecast harmful algal blooms and can potentially provide early warnings to water managers, public health agencies, and other water users (Rousso *et al.*, 2020; Cruz *et al.*, 2021),.

As it is becoming increasingly critical to predict and track algal blooms in lakes, this thesis is focused on improving monitoring and forecasting capabilities in this field through three novel studies, each presented as a chapter. Chapter 1 investigates how long-term in situ data can support the development of machine learning models for forecasting algal concentrations and explores the data requirements of such models. In Chapter 2, I assess how climate change may impact the reliability of these forecasting tools and evaluate strategies to mitigate these effects. Finally, Chapter 3 explores how new, higher-resolution satellite data may dramatically expand the number of lakes that can be monitored from space, provided adequate ground-truth data are available.

Through this exploration of both data requirements for forecasting, alongside the evaluation of a cutting edge, newly available data source, this thesis aims to explore

potential future directions for algal monitoring and forecasting practices. Therefore, it is possible to evaluate and discuss whether the existing approaches in this field are suitable for tackling future water quality challenges, or whether significant paradigm shifts are required.

Across all chapters, the research highlights not only the opportunities new technologies offer for improving water quality monitoring and management, but also the continued importance of manual, human-led monitoring - through regulatory agencies, scientific research programmes, and citizen science. Ultimately this thesis demonstrates that data from a multitude of sources, including newly developing remote sensing technologies, climate change simulations, citizen scientists, and long-term monitoring programs can all provide critical insights into how we should use, design and plan monitoring and forecasting systems for algal blooms.

What continues now is an overview of key background information and existing literature relevant to this thesis. Broadly, this covers freshwater harmful algal blooms, climate change, lakes, lake-monitoring, remote sensing, forecasting, and machine learning. Note that while this thesis is not exclusively about cyanobacteria, much of the background information provided on algal blooms is focused on cyanobacteria, as they are the most significant toxic and bloom forming taxa found in inland waters.

1.2 Cyanobacteria

Approximately 2.5 billion years ago a sequence of events was set in motion that would dramatically alter the planet, its atmosphere, and the life it supported then and now. Cyanobacteria started producing oxygen as a photosynthesis byproduct, and as this new type of photosynthesis became increasingly common, oxygen started to accumulate in the Earth's atmosphere (Olejarz *et al.*, 2021). The consequences of this were dramatic. The new presence of oxygen would have been lethal to a huge number of anoxic bacteria (Kozik *et al.*, 2022), but nonetheless, this event, known as the great oxygenation event, created the beginnings of the oxygen-rich biosphere we are familiar with, and reliant on today.

The organism responsible, cyanobacteria, are still numerous, although having been shaped by evolutionary processes for the last two or three billion years, are likely quite different now (Demoulin *et al.*, 2019). Today, cyanobacteria can be found in a wide variety of habitats, including extreme environments such as deserts, hot springs, highly saline water, and polar regions, as well as in oceans and freshwaters (Whitton and Potts, 2012).

Recently¹ in fresh waters, which are the focus of this thesis, large accumulations, or blooms, of toxic cyanobacteria have become well-recognised as a public health risk and a threat to the stability and diversity of aquatic ecosystems (Codd *et al.*, 2005; Chorus and Welker, 2021).

Unlike the disruption and mass-extinction caused by cyanobacteria during the great oxygenation event, the risks cyanobacteria pose now are perhaps best thought of as impacts of human-induced water quality and climate issues, rather than hazards caused solely by evolved traits. Specifically, large-scale agriculture, industrialisation, and urbanisation drive eutrophication – the nutrient enrichment of water bodies, enabling large populations of cyanobacteria to grow, potentially forming blooms, isolated scums, or dense biofilms (O’Neil *et al.*, 2012). Furthermore, there is a growing body of evidence suggesting that the increasing temperatures and changing weather patterns associated with human-driven climate change are very likely to contribute towards the further proliferation of cyanobacteria in many lakes across the world (Paerl and Huisman, 2009; Carey *et al.*, 2012; Paerl *et al.*, 2020). Cyanobacteria have clearly played a pivotal role in shaping Earth’s history - from oxygenating the atmosphere to now signalling the environmental impacts of human activity. Therefore, monitoring, anticipating, understanding and limiting their negative impacts is of great importance.

1.2.1 Cyanobacterial blooms: a costly, global water quality issue

Cyanobacteria are a diverse group, well adapted to a plethora of terrestrial and aquatic environments (Whitton and Potts, 2012), and have in recent years received significant research attention, particularly due to the health risks posed by certain genera (Codd *et al.*, 2005). Furthermore, industry, agriculture and urbanisation have resulted in modifications to aquatic environments which cyanobacteria are well placed to take advantage of (O’Neil *et al.*, 2012). These modifications include nutrient over-enrichment, hydrologic alterations, and changes associated with anthropogenic climate change (Carey *et al.*, 2012; Paerl and Otten, 2013).

Cyanobacteria which produce harmful toxins are of particular concern. Cyanobacterial toxins are generally categorised into four principal groups: hepatotoxins, which damage the liver (produced by: *Microcystis*, *Anabaena*, *Nostoc*, *Anabaenopsis*, *Planktothrix*,

¹ It should be noted that, whilst anthropogenic eutrophication and climate change are modern problems which are linked with increased algal bloom occurrence, large blooms of cyanobacteria and other algae have likely occurred throughout history (Emslie *et al.*, 1996; Koenigswald, Braun and Pfeiffer, 2004; Brugger *et al.*, 2021).

Oscillatoria, *Hapalosiphon*); nervous tissue damaging neurotoxins (produced by: *Anabaena*, *Oscillatoria*, *Phormidium*, *Aphanizomenon*, *Rhaphidiopsis*); dermatotoxins and cytotoxins, which target cells and skin (produced by: *Lyngbya*, *Schizothrix*, *Oscillatoria*); and endotoxins which can induce inflammation and fever (Bláha, Babica and Maršálek, 2009; Buratti *et al.*, 2017). Endotoxins are thought to be produced by all toxigenic genera (Carmichael, 2001; Codd *et al.*, 2005). Toxin production varies within genera and between different strains of species. Cyanotoxins have been associated with mass animal deaths (Bengis *et al.*, 2016; Lomeo *et al.*, 2024), alongside a host of negative health outcomes in humans. These include gastroenteritis (Teixeira *et al.*, 1993); visual disturbances, nausea, and liver failure (Azevedo *et al.*, 2002); gastrointestinal and dermatological symptoms (Saadi *et al.*, 1995); primary liver cancer (Fleming *et al.*, 2002); respiratory irritation (Giannuzzi *et al.*, 2011); and even death in children and immuno-compromised individuals (Azevedo *et al.*, 2002; Weirich and Miller, 2014).

Quantifying the monetary cost associated with dealing with harmful cyanobacterial blooms is challenging, as prevention and mitigation strategies are diverse and multifaceted. Furthermore, certain costs are harder to quantify, such as losses related to recreational water use and reduction in the value of waterside properties (Hamilton *et al.*, 2014). Pretty *et al.* (2003) developed a framework to assess the environmental costs of freshwater eutrophication in England and Wales where it was estimated that the annual cost of drinking water treatment related to the removal of algal toxins and decomposition products was £19 million (equivalent to £35 million in 2025), and that sewage treatment (the removal of phosphorus to reduce nutrient over-enrichment) costed a total of £50.3 million (equivalent to £93 million in 2025) annually. A similar study which focused on the costs associated with the eutrophication of freshwater in the USA estimated a total cost of 2.2 billion USD (equivalent to 3.3 billion USD in 2025) annually (Dodds *et al.*, 2009). Whilst the costs associated with eutrophication have many sources and causes, cyanobacteria play a significant role, contributing to costs in a number of areas including taste and odour treatment; surveillance and monitoring; water supply contingency; losses due to the impairment of recreational activities and commercial fisheries; and the management of catchment nutrient loads (Hamilton *et al.*, 2014). Ultimately, despite the considerable finance invested in eutrophication management, this remains a persistent global challenge (Hei *et al.*, 2025).

1.2.2 Cyanobacterial blooms and climate change

An important reason for the increased research attention which cyanobacterial blooms have received in recent years is anthropogenic climate change, which has effects on many systems in the natural world. Global surface temperatures are rising; certain extreme weather events are becoming more frequent; and precipitation patterns are changing on a global scale (Arias *et al.*, 2021). Considering lake systems globally, summer surface water temperatures have been shown to be rising rapidly in recent years (O'Reilly *et al.*, 2015). The magnitude of this effect varies between lakes and has been shown to be greater for colder, deeper lakes (Woolway and Merchant, 2017). Furthermore, modelling lake thermal regions has shown that even for a low emissions scenario, the thermal regime of many lakes (> 12 %) will be significantly altered over the next century (Maberly *et al.* 2020). The ramifications of these kind of changes on inland water systems are varied and difficult to predict, but there is some consensus that certain cyanobacteria will be able to take advantage of these changes (Paerl and Huisman, 2009; Carey *et al.*, 2012; Paerl and Paul, 2012).

For example, climate warming is expected to change lake stratification, the presence of distinct layers of water with different densities and temperatures, in several ways. It is projected that stratification will start earlier in the year, it will be stronger, with greater temperature gradients and will last longer (Woolway *et al.*, 2021; Woolway, Sharma and Smol, 2022). These types of conditions can be exploited by cyanobacteria genera with high optimum growth temperatures that are able to regulate their buoyancy (Carey *et al.*, 2012). However, there is some question surrounding whether algal bloom occurrence has actually increased globally in recent years, as monitoring efforts have also intensified (Hallegraeff *et al.*, 2021). Therefore, remote sensing, which can provide a more continuous global record is very useful for studying this. Recent analyses of global remote sensing datasets have concluded that algal bloom frequency has increased over the last 20 years (Hou *et al.*, 2022; Wang *et al.*, 2025). Furthermore, simulations have demonstrated that this trend is likely to continue in more extreme climate scenarios (Tigli *et al.*, 2025). Despite these emerging trends, there remains a clear need for further research that explicitly investigates long-term changes in global occurrence, severity and causal mechanisms of blooms, given their potential toxicity and severe impacts on freshwater ecosystems. Furthermore, it would be valuable to study how trends in smaller-sized lakes compare with larger-sized lakes which, to date, have been the primary focus of global remote sensing studies due to resolution limitations.

1.3 Freshwater lakes & algal blooms

Although 71 % of the Earth's surface is covered by water, only 3 % of that is freshwater, and of that, only 0.3 % is found in lakes and rivers, which forms the majority of our useable supply (Gleick and Srinivasan, 2011). Freshwater lakes are diverse in their morphology, ecology, and thermal structures; and this diversity is reflected in the great range of ecosystem services which they provide. These include provision of drinking water; fisheries; recreation and tourism; carbon sequestration; and hydrological regulation, for example flood control (Schallenberg *et al.*, 2013). Alongside this, lakes have significant cultural value across the world (Schallenberg *et al.*, 2013). Accordingly, freshwater lakes are most certainly highly precious and therefore, are critical to protect and restore. Furthermore, the physical, chemical, and biological properties of lakes are sensitive to environmental changes, meaning that they can be thought of as sentinels of climate change (Adrian *et al.*, 2009; Schindler, 2009). Unfortunately, this sensitivity puts many of the essential services which lakes provide at risk, illustrated by the fact that freshwater biodiversity decline is significantly greater than even the worst-affected terrestrial systems (Sala *et al.*, 2000; Reid *et al.*, 2019). Illustrating this, the World Wide Fund for Nature's Living Planet Index showed a freshwater population decline of 85 % between 1970 and 2020, which is considerably greater than the corresponding declines for terrestrial (69 %) and marine (56 %) populations (WWF, 2024). Therefore, the need to understand, monitor, and protect these vital ecosystems is of great importance (Carpenter, Stanley and Vander Zanden, 2011). As has already been explained, a major threat to freshwater ecosystems is the rise of harmful algal blooms, especially toxin-producing cyanobacteria, which threaten water quality, ecosystem stability, and human health. Effective monitoring and forecasting are needed to manage these blooms, but such efforts must be rooted in a mechanistic understanding of algal blooms, since their dynamics depend on complex physical, chemical, and biological interactions. This section will therefore explore some critical drivers of cyanobacterial blooms and the conditions that favour their growth.

1.3.1 Understanding blooms – lake physics

Fundamental to understanding lake ecosystems and biological processes, are the physics of the lake. This includes how water temperature, and density varies through the water column; where, when, and how water mixes; how lakes are heated, and lose heat; and how light travels through the water column. Furthermore, algal dynamics are intrinsically linked to lake physics. For example, algal growth rates are temperature dependent, and so

the speed at which an algal bloom may develop is inherently linked with the physical properties of the lake (Lürling *et al.*, 2013).

In temperate climates, during the warmer months, lakes typically stratify into distinct thermal layers: the epilimnion (surface layer), metalimnion (middle layer), and hypolimnion (bottom layer). Stratification inhibits vertical mixing, creating isolated water layers with different temperatures, densities, and nutrient availabilities. The epilimnion, warmed by the sun, is typically well-mixed and oxygen-rich due to wind mixing, while the hypolimnion remains cooler, denser, and can become oxygen-depleted during the summer (Boehrer and Schultze, 2008). In autumn surface cooling reduces the temperature differences between layers. Once the temperature and density across the water column become similar enough, the lake undergoes turnover - a process in which wind mixing redistributes heat, oxygen, and nutrients throughout the lake. This turnover is essential for transporting nutrients that have accumulated in the hypolimnion back to the surface, and can be associated with surface blooms of algae (Fee, 1976; Brooks and Torke, 1977). Lakes that follow this seasonal pattern can be classified based on their mixing regimes. Monomictic lakes stratify during the summer and experience a single mixing period, typically in the autumn. Dimictic lakes, common in temperate regions, stratify in the summer, mix in the autumn, and then inverse stratify during the winter. In this winter state, the coldest water - or ice - rests on the surface, while denser, slightly warmer water (close to 4 °C, the temperature at which water reaches its maximum density) remains at the bottom. This behaviour is due to the unique property of water, which becomes less dense both above and below 4 °C, leading to winter stratification (Boehrer and Schultze, 2008).

Shallow lakes, on the other hand, may not stratify as distinctly and can mix frequently due to wind or inflow events (Scheffer and van Nes, 2007). Lakes with this behaviour are termed polymictic, meaning that they mix multiple times each year. These dynamics make shallow systems particularly sensitive to both nutrient loading and temperature variations, increasing the potential for frequent algal blooms (Nixdorf, 1994). Lake morphology also influences mixing patterns. Deeper lakes tend to stratify more strongly and for longer periods, while shallower lakes are more susceptible to mixing and resuspension of bottom sediments. These physical factors play a critical role in determining the availability of nutrients, which in turn affects the degree to which large populations of algae may grow (Schindler, 1974). However, whilst stratification can effectively establish a barrier between the epilimnion and the often more nutrient-rich hypolimnion, certain buoyancy-

regulating cyanobacteria, such as *Microcystis spp.* can cross this barrier, and therefore obtain additional nutrients when many other phytoplankton are unable to do so (Ganf and Oliver, 1982).

How well light travels through water is also key to understanding algal dynamics. Water clarity, influenced by suspended sediments, dissolved organic matter, and existing algal biomass, affects how deeply light can penetrate and thus where algae can grow (Muhetaer *et al.*, 2020; Bialevich, Zachleder and Bišová, 2022). This feedback mechanism means that once a bloom begins, increased turbidity can further inhibit light penetration, affecting both algal species composition and overall lake productivity (Paerl and Otten, 2013).

Understanding these physical drivers provides the foundation upon which biological and chemical processes operate, setting the stage for more complex interactions that lead to the development and persistence of harmful algal blooms.

1.3.2 Understanding blooms – lake chemistry

The chemical composition of lake water plays a critical role in shaping biological activity and therefore, the development and persistence of algal blooms. Of the many chemical variables in freshwater systems, the ones that have been shown to have greater effect are nutrients (nitrogen and phosphorus), and dissolved oxygen.

Algal blooms are most directly driven by the availability of nutrients, with phosphorus (P) often being the limiting factor in freshwater lakes (Sterner, 2008). In many systems, even small increases in bioavailable phosphorus can lead to significant increases in algal biomass. Phosphorus typically enters lakes through external inputs such as agricultural runoff, wastewater discharge, and atmospheric deposition (Fink *et al.*, 2018). However, it can also be released internally from lake sediments - particularly under low-oxygen and high temperature conditions in the hypolimnion - a process known as internal loading (Nürnberg, 2009; Wu *et al.*, 2014; Gibbons and Bridgeman, 2020). Nitrogen (N) is another key nutrient that influences algal dynamics. While many freshwater systems are phosphorus-limited, it is very common for lakes to be co-limited by both nitrogen and phosphorus, especially during periods of high productivity (Sterner, 2008). In such cases, certain cyanobacteria (e.g., *Aphanizomenon* or *Dolichospermum spp.*) can gain a competitive advantage in some systems by fixing atmospheric nitrogen (Oliver and Ganf, 2002; Wood *et al.*, 2010). Therefore, the N:P ratio can influence not just the magnitude of a bloom but also the composition of the algal community (Vrede *et al.*, 2009; Liu, Lu and Chen, 2011). Importantly, nutrient stoichiometry, and therefore nutrient limitation modes,

can vary considerably over time, which can drive rapid changes in algal biomass and composition. To anticipate such rapid changes in algal biomass requires high frequency nutrient flux monitoring, which is generally challenging due to the high costs associated with this (Page *et al.*, 2017).

Dissolved oxygen (DO) is another key parameter, particularly in stratified lakes. During summer stratification, oxygen in the hypolimnion can become depleted due to microbial decomposition of organic matter (Foley *et al.*, 2012; Snortheim *et al.*, 2017). As oxygen is consumed and not replenished due to lack of mixing with upper layers, anoxic conditions can develop. This can increase the potential for algal blooms in several ways. Firstly, in anoxic conditions, iron-bound phosphorus in sediments can be released into the water column, increasing the internal nutrient load available for algae (Mortimer, 1941, 1942; Tammgeorg *et al.*, 2020). Similarly, ammonium (NH_4^+), a nitrogen source, can be released from the sediment during anoxic conditions and therefore become available to phytoplankton (Beutel, 2006). This is important because ammonium is generally the preferred nitrogen source for cyanobacteria (Herrero, Muro-Pastor and Flores, 2001).

1.3.3 Understanding blooms – phytoplankton dynamics

Phytoplankton (algal) communities are central to lake ecology, as they form the base of the aquatic food web and drive fundamental processes such as primary production and nutrient cycling (Naselli-Flores and Padisák, 2023). Patterns in algal biomass, species composition, and the timing of phytoplankton growth can offer important insights into the ecological state and functioning of lake ecosystems (Paształeniec and Poniewozik, 2010). These attributes are shaped by a complex interplay of physical, chemical, and biological factors, and they respond sensitively to environmental change (Grover and Chrzanowski, 2006; Stockwell *et al.*, 2020; Luong *et al.*, 2024; Varga *et al.*, 2024).

Algal biomass reflects the overall outcome of growth, grazing, and loss processes, and serves as a key metric of lake productivity and metabolism. Often the chlorophyll-a concentration is used as a proxy for algal biomass as this is relatively straightforward to measure and is generally, although not in all cases, well-correlated with biomass (Vörös and Padisák, 1991; Siegel *et al.*, 2013; He, Wang and Xu, 2022). In many lakes, biomass levels are closely linked to nutrient availability. However, biomass alone cannot paint a complete picture of phytoplankton dynamics. Changes to the composition of the phytoplankton community, such as shifts from diatom to cyanobacteria dominance, can have profound implications for food web dynamics, water quality, and biogeochemical

cycling (Zhang *et al.*, 2023; X. Wang *et al.*, 2024). These shifts are not always associated with changes in total biomass, highlighting the importance of understanding community structure.

The seasonal timing of phytoplankton growth, or phenology, is another important dimension of phytoplankton dynamics. In temperate lakes, phytoplankton typically exhibit well-defined seasonal successions driven by temperature, light availability, stratification, and nutrient dynamics (Sommer *et al.*, 2012). For instance, in many productive temperate lakes, diatom blooms typically occur in the spring, followed by summer blooms of buoyant, nitrogen-fixing cyanobacteria (Casterlin and Reynolds, 1977; Sommer *et al.*, 2012; Wei *et al.*, 2020). Climate change is altering these phenological patterns, leading to earlier bloom onsets and longer growing seasons (Thackeray, Jones and Maberly, 2008; Meis, Thackeray and Jones, 2009; Vadadi-Fülöp and Hufnagel, 2014).

Taken together, these dimensions - biomass, community composition, and phenology - are all critical to understanding algal bloom dynamics. While elevated biomass often signals bloom formation, compositional shifts toward harmful cyanobacteria and altered seasonal patterns can occur independently of total biomass increases. Therefore, effective monitoring and forecasting of algal blooms requires a multi-faceted approach that considers not only how much algae is present, but also which species dominate and how this can influence the timing of blooms.

1.4 Monitoring Algae

Monitoring is a cornerstone of modern ecological science. It provides critical insight into how ecosystems respond to both short-term disturbances and long-term environmental change, underpins the development of predictive models, and offers essential evidence to inform policy and management (Lindenmayer and Likens, 2010). In the context of freshwater algal blooms, effective monitoring is particularly important - both for understanding bloom dynamics and for supporting timely, evidence-based responses (Chorus and Welker, 2021; Xiao *et al.*, 2024). Accordingly, a major focus of this thesis is on various aspects of monitoring algae in inland waters: exploring how emerging sensor technologies can expand monitoring capacity and how monitoring systems can be designed to facilitate accurate forecasting and informed decision-making.

1.4.1 Manual Monitoring Methods

Field sampling and laboratory analyses are widely used for measuring algal concentrations, as well as cyanobacteria and cyanotoxin concentrations (Paerl and Otten,

2013). To measure algal cell concentrations, water samples of a known volume are taken from the lake of interest and preserved typically with Lugol's Iodine solution. Then, in a laboratory, sub-samples are analysed under a microscope to identify and count the different phytoplankton taxa present. Following this, appropriate statistical analysis of this cell-count data allows for concentrations of different genera to be obtained (Brierley *et al.*, 2007). More recently, attempts have been made to automate the time-consuming manual element of phytoplankton identification using image classification machine learning approaches, however this technology is largely still in an experimental stage, and has not yet been widely adopted (Schulze *et al.*, 2013; Rivas-Villar *et al.*, 2021; Pitois *et al.*, 2025).

Water samples can also be analysed using spectrophotometry or other analytical methods to obtain pigment concentrations. Chlorophyll-a (chl-a) is one of the most measured pigments, and is generally used as an algal biomass indicator (Vörös and Padisák, 1991). Another pigment of interest is phycocyanin, which is specific to cyanobacteria, and therefore useful for indicating the presence of cyanobacteria (Hunter *et al.* 2010). However, phycocyanin is not always strongly correlated with cyanobacterial biomass, and so approaches which use both phycocyanin and chl-a can be advantageous by enabling estimation of overall algal biomass and cyanobacterial dominance (Matthews, Bernard and Robertson, 2012).

Cell counts and pigment concentrations are useful measures, but do not give an indication of the toxicity of a cyanobacterial bloom. For this reason, it is common to measure cyanotoxins present in water samples. The World Health Organisation (WHO) gives specific guidance for the most prolific cyanotoxin, 'microcystin', providing suggested exposure limits (World Health Organization, 2017). Microcystin concentrations are typically measured using protein phosphatase 2A assay (PP2A) or enzyme-linked immunosorbent assay (ELISA) methods (Paerl and Otten, 2013).

Where cyanotoxin concentrations are not sufficiently high to pose significant health risks, cyanobacterial blooms can still cause issues related to the taste and odour of drinking water. For example, geosmin and 2-Methylisoborneol (MIB) are produced by various cyanobacteria and can cause water to taste earthy or musty. Water companies regularly assess taste and odour of water samples using a variety of methods including lab analyses, and panels of trained tasting experts (Lin *et al.*, 2019).

All these manual approaches are extremely valuable, each providing unique insights into critical aspects of algae and water quality. However, manual sampling does have some limitations. Firstly, collection of samples is often just carried out at a single location for the water body of interest, for example, just one of the locations shown in Figure 1.1b. Even when samples are taken at multiple locations, time and resources still usually limit this to a relatively small number as both collection and processing are time-consuming and costly. This has the consequence that heterogeneity across a water body is either ignored or oversimplified. For example, if sampling exclusively from a shoreline upwind, then large scums of cyanobacteria might be missed, as these would tend to accumulate downwind. Conversely, if one only samples the shoreline scum, then it could be interpreted that the extremely high concentrations found in this sample are representative of the entire lake, when this may only be the case for a very small area (Figure 1.1b). This highlights the importance of recording additional field observations when sampling, as sample data taken out of context can be misleading. Additionally, phytoplankton populations can vary considerably over short time scales (Dubelaar, Geerders and Jonker, 2004). Therefore, important short-term fluctuations can be missed as it is typical for routine manual monitoring schemes to sample only at monthly or quarterly intervals (UK Interagency Freshwater Group, 2015). Furthermore, the majority of lakes, particularly smaller, less economically significant water bodies, are never regularly manually monitored (Biggs, von Fumetti and Kelly-Quinn, 2017). Therefore, whilst manual sampling is essential, its costly and time-consuming nature ultimately limits the temporal and spatial resolution which can be covered. For this reason, automated approaches such as satellite remote sensing have received significant research interest in recent years for inland water quality monitoring.

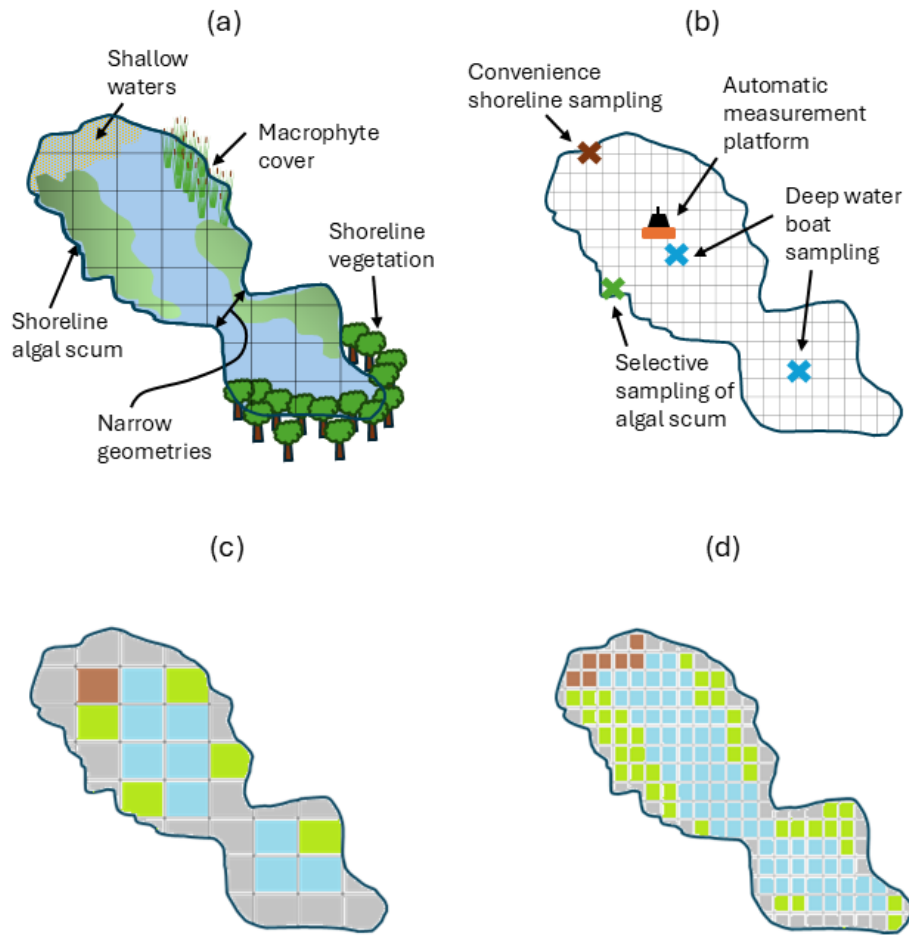


Figure 1.1: Illustration of different approaches to lake sampling. (a) depicts a small, heterogeneous lake; (b) illustrates potential locations for point-sampling; and (c) and (d) show representations of how such a lake might be viewed by medium and high-resolution² satellite imagery respectively. In (c) and (d), blue areas represent relatively clear, deeper waters, green areas are those with a high chlorophyll-a signal, brown areas indicate bottom reflectance effects, and all pixels with a non-negligible land-effect are coloured grey.

1.4.2 Aquatic Remote Sensing

In recent years, satellite remote sensing, which is a key focus of this thesis, has increasingly been used to complement and extend the spatial and temporal range offered by manual algae monitoring approaches. For the purposes of this thesis, the term remote sensing shall be assumed to be synonymous with satellite remote sensing, which is monitoring the Earth using sensors mounted on satellites in orbit. Radiometry, the measurement of electromagnetic radiation, is fundamental to remote sensing. Typically, aquatic remote sensing makes use of radiation from the sun: sunlight reaching the Earth is modified by both the atmosphere and the water it enters, and this modification can be described by radiative transfer theory, which enables quantitative interpretation of satellite

² Note that satellite imagery has multiple resolutions: spatial, temporal, and spectral. However, unless stated otherwise, this thesis uses the term high-resolution to refer to data with relatively high resolutions in only spatial and temporal dimensions. As spectrally high-resolution (hyperspectral) satellite imagery is not a focus of this thesis it is excluded from this terminology.

imagery (Figure 1.2, Gordon, 2019). Specifically, by measuring spectral radiance, the light intensity across a range of wavelengths, it is possible to estimate the optical properties of the water being observed.

For aquatic optical remote sensing, the most widely used reflectance parameter is the remote sensing reflectance (R_{rs}) which is defined, for a given wavelength (λ), as the ratio between the water leaving radiance (L_w) and downwelling irradiance (E_d):

$$R_{rs}(\lambda) = \frac{L_w(\lambda)}{E_d(\lambda)} . \quad (1)$$

Water leaving radiance refers specifically to downwelling light which has entered the water from the air and is then scattered upwards within the water such that it is transferred back into the air (Figure 1.2). However, direct measurement of L_w is not possible as upwelling light from the water's surface also includes light which is reflected by the water's surface without entering the water. Therefore, L_w is often approximated using the measurable total upwelling radiance (L_u) and downwelling sky radiance (L_{sky}) such that:

$$R_{rs}(\lambda) \cong \frac{L_u(\lambda) - \rho L_{sky}(\lambda)}{E_d(\lambda)} , \quad (2)$$

where ρ is the skylight reflectance parameter, which typically varies with windspeed and can be empirically determined (Mobley, 1999). L_u , L_{sky} , and E_d are all ideally measured just above the surface of the water, and so orbiting satellites cannot measure these parameters. Instead, satellites measure the top of atmosphere radiance, which includes atmospheric scattering contributions as well as those from the water's surface.

Accordingly, various atmospheric correction algorithms are used to estimate R_{rs} from top of atmosphere radiance measurements (Section 1.4.2.4).

Generally, oceanic waters are optically less complex than inland waters - their optical properties are primarily determined by phytoplankton (IOCCG, 2000). Conversely, inland, and coastal waters tend to have additional optically active components including suspended particles such as sediments; coloured dissolved organic matter (CDOM) from natural and human sources; and non-algal particles (NAP) like minerals and microbes (Mobley *et al.*, 2015). For this reason, and because these constituents are generally not linearly correlated, aquatic remote sensing of coastal and inland water bodies is generally more challenging and requires a more diverse range of algorithms and approaches (Palmer, Kutser and Hunter, 2015).

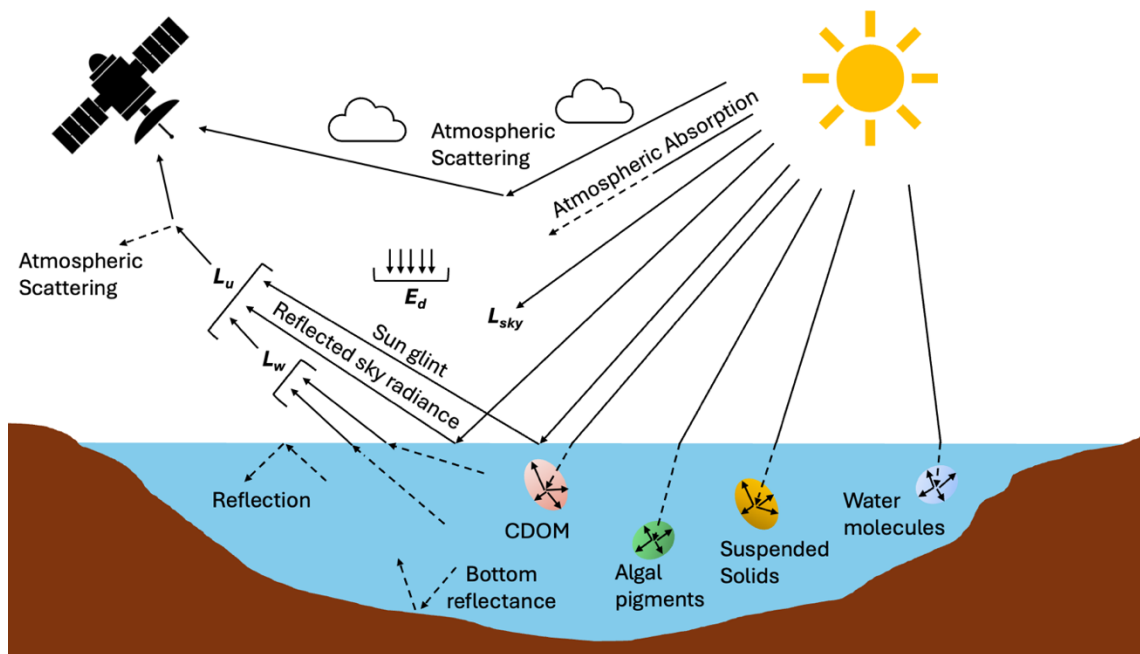


Figure 1.2: An illustration showing the key paths and interactions which sunlight makes to a satellite sensor viewing an inland water body. Also shown are the constituent components of the remote sensing reflectance (R_s): downwelling radiance (L_{sky}), upwelling radiance (L_u), water leaving radiance (L_w), and downwelling irradiance (E_d).

1.4.2.1 A brief history of aquatic remote sensing

Earth observation satellites use remote sensing to monitor the planet's surface, including water bodies. Optical remote sensing is widely used for aquatic applications, and is capable of monitoring characteristic pigments such as chlorophyll-a. Aquatic remote sensing was started in the early 1970s, focusing on water clarity in lakes using Landsat-1 imagery (Yarger *et al.*, 1973). However, the first dedicated satellite ocean colour sensor, the Coastal Zone Colour Scanner (CZCS) was deployed in 1978, and whilst this advanced marine science significantly, there were very few further studies on inland water remote sensing over the next decade (Hovis *et al.*, 1980). Then, in the 1990s and early 2000s further inland water quality studies were undertaken using sensors like the Moderate Resolution Imaging Spectrometer (MODIS) (Pagano and Durham, 1993), the Medium Resolution Imaging Spectrometer (MERIS) (Bezy *et al.*, 1999), the Sea-viewing Wide Field-Of-View Sensor (SeaWiFS) (Hooker and Esaias, 1993), and the Modular Optoelectronic Scanner (MOS) (Zimmermann *et al.*, 1993). Generally, these sensors had relatively coarse spatial resolutions (≥ 300 m pixels), and so could only resolve larger water bodies, but were still pivotal for exploring algorithms suitable for studying inland waters which are optically more diverse and complex than their oceanic counterparts (Palmer, Kutser and Hunter, 2015).

To overcome the limited spatial resolution of ocean colour sensors, higher resolution satellite sensors such as the Landsat series, which were not specifically designed for water

monitoring have also been used extensively for inland water quality monitoring (Lavanderos *et al.*, 1990; Tebbs, Remedios and Harper, 2013; Karimi, Hashemi and Aghighi, 2024). Recently, the European Space Agency's (ESA) Copernicus programme, particularly the Sentinel-3 Ocean and Land Colour Imager (OLCI) and the Sentinel-2 Multispectral Imager (S2 MSI), has expanded capabilities for monitoring both ocean and inland waters. OLCI offers 300 m resolution across 21 spectral bands, whilst MSI has 13 spectral bands and can resolve finer spatial details, particularly in the red, blue, green and near infrared bands which are 10 m resolution (Donlon *et al.*, 2012; Drusch *et al.*, 2012). Importantly, the successful use of higher spatial resolution sensors, such as Sentinel-2 and Landsat for water quality monitoring is a critical step for remote sensing of inland waters, as the vast majority of lakes have an area of only a few square kilometres or less (Downing *et al.*, 2006; Cael and Seekell, 2016). Therefore, this potentially enables developing methodologies for monitoring a huge number of smaller water bodies.

However, this requires addressing various challenges which are more pronounced for smaller water bodies. For example, the mixing of light reflected from the water column with light reflected from land, known as the land-adjacency effect, is particularly consequential for smaller waters where most pixels are relatively close to the shoreline (Figure 1c, d, Jiang *et al.*, 2023; Paulino *et al.*, 2022). Therefore, furthering monitoring capabilities for smaller water bodies using satellite imagery not only increases the number of monitorable water bodies dramatically, it also necessitates solving key problems in aquatic remote sensing and therefore is likely to further the science and technology significantly. Now, as new higher resolution satellite imagery becomes available, this potentially facilitates remote monitoring of even more lakes than is possible with the likes of Sentinel-2 MSI and Landsat OLI. Therefore, it is critical to evaluate this potential and understand how this may further aquatic remote sensing.

1.4.2.2 Cubesat constellations: a paradigm shift?

In 2019, the company Planet Labs started launching the first of what would, over the next few years become several hundred micro-satellites named SuperDoves. This large collection, or constellation, of satellites is able to provide daily images of the Earth which are approximately spectrally comparable to Sentinel 2 MSI, with the benefit of 3 m spatial resolution (Planet Team, 2017). These satellites are an example of *CubeSats*, a term used for very small, compact low-cost satellites (Davoli *et al.*, 2019). Whereas traditional space agencies such as the European Space Agency (ESA) and National Aeronautics and Space Administration (NASA) have typically developed and launched a small number of

relatively expensive, precision-calibrated, and well-instrumented satellite sensor units, Planet has adopted an alternative strategy. Through deploying a large constellation of low-cost CubeSats, Planet effectively increases temporal resolution by simultaneously acquiring multiple images with many satellites.

This constellation-based approach enables the use of lower-altitude orbits, which gives higher spatial resolution, while avoiding the trade-off between altitude, and therefore swath width, and revisit frequency. Consequently, this produces imagery with both high temporal frequency and high spatial resolution. This high effective revisit time of Planet SuperDoves may be particularly advantageous for monitoring areas with frequent cloud cover, as this is significantly higher than the 5-day and 8-day effective revisit times offered by Sentinel-2 and Landsat 8/9. However, the low-cost nature of CubeSats does mean that the imagery they produce is, at present, lower quality than their more expensive, and longer lasting counterparts such as Landsat OLI, Sentinel-2 MSI, and Sentinel-3 OLCI (Selva and Krejci, 2012; Valenzuela, Reinke and Jones, 2024). This difference in quality arises largely from lower signal-to-noise ratios, and inter-sensor calibration issues (Kabir *et al.*, 2025).

One key advantage of the higher resolution imagery offered by the SuperDoves is that this potentially enables monitoring water bodies which are smaller than those resolvable with sensors such as Sentinel-2 MSI or Landsat 8/9 OLI. Specifically, whilst many small water bodies have sufficient surface area that this can be covered by several MSI or OLI pixels, it is common for large cyanobacteria populations, including scums, to accumulate along shorelines or sheltered bays, often downwind (Figure 1.1a). Therefore, when resolved with relatively low resolution imagery, many pixels along shorelines which overlap both land and water must be discarded as the signals from the land and water are typically not easily separable for these “mixed pixels” (Figure 1.1c). In contrast, whilst higher resolution imagery does not completely eliminate the problems of mixed pixels and land-adjacency effects, it provides a larger number of potentially useful pixels along shorelines (Figure 1.1d). As such, the SuperDoves constellation is likely more suitable than lower resolution alternatives for monitoring smaller inland water bodies and complex shoreline geometries (Jiang *et al.*, 2023). However, given the relatively new nature of this satellite constellation, there is, at present, only a limited volume of work which evaluates their potential for water quality monitoring, particularly for small inland water bodies (Vanhellemont, 2023; Kabir *et al.*, 2025).

1.4.2.3 Algal parameter retrieval approaches

There are several approaches which are commonly used for monitoring algae and algal blooms with remote sensing (Table 1.1). Perhaps most common is estimation of the chl-a concentration, which can be done with a variety of algorithms, including physics-based ‘analytical’ approaches as well as empirical (O’Reilly, 2000; Gons, Rijkeboer and Ruddick, 2002; Moses *et al.*, 2009; Gilerson *et al.*, 2010; Yang *et al.*, 2010; Matthews, Bernard and Robertson, 2012; Mishra and Mishra, 2012) and, more recently, machine learning methods (Ioannou *et al.*, 2013; Cao *et al.*, 2020; Pahlevan *et al.*, 2020; Werther, Odermatt, Simis, Gurlin, Lehmann, *et al.*, 2022; Woo Kim *et al.*, 2022). In addition to chl-a retrieval, other parameters such as floating algae index (Hu, 2009), normalised difference vegetation index (NDVI) (Zhou *et al.*, 2022), surface scum index (SSI) (Sayers *et al.*, 2016), surface algal bloom index (SABI) (Alawadi, 2010), and phycocyanin (Hunter *et al.*, 2010; Mishra *et al.*, 2013) have been used to evaluate the intensity and spatial distribution of algae and algal blooms.

Table 1.1: Examples of widely used algorithms for retrieval of algae-related parameters from remote sensing imagery. Notation is as follows: $Y(\lambda)$ is used to indicate parameter Y , measured at wavelength λ ; a_Y is the light absorption coefficient of component Y , such that a_w is that for pure water, a_{PC} is that for phycocyanin, and a_{chl} is that for chlorophyll- a ; a_{PC}^* is the specific absorption coefficient of phycocyanin; R_{rc} is the Rayleigh-corrected reflectance (Hu et al., 2004); and RED , NIR and $SWIR$ indicate red, near-infrared and shortwave infrared wavelengths respectively. All other symbols not explicitly defined are empirically derived constants – see citations for further details.

Name / Citation	Description	Equation
Gons et al. (2002; 2008)	Chlorophyll- a semi-analytical NIR / Red ratio algorithm	$chl_a = \left[\left(\frac{R_{rs}(708)}{R_{rs}(665)} \right) \times (0.7 + b_b) - 0.4 - b_b^k \right] / l$ $b_b = \frac{0.6 \times a_w(778) \times R_{rs}(778)}{0.082 - 0.6 \times R_{rs}(778)}$
Ocean Colour 4 (OC4) (O'Reilly, 2000).	Chlorophyll- a empirical blue / green ratio algorithm	$chl_a = 10^{(a+bX+cX^2+dX^3+eX^4)}$ $X = \log_{10} \left(\frac{\max [R_{rs}(443), R_{rs}(490), R_{rs}(510)]}{560} \right)$
Simis et al. (2005)	Phycocyanin semi-analytical algorithm	$PC = \frac{a_{PC}(620)}{a_{PC}^*(620)}$ $a_{PC}(620) = \left(\left\{ [a_w(709) + b_b] \times \frac{R_{rs}(709)}{R_{rs}(620)} \right\} - b_b - a_w(620) \right) \times \delta^{-1} - [\varepsilon \times a_{chl}(665)]$
MDN (Pahlevan et al., 2020; O'Shea et al., 2021)	Chlorophyll- a and phycocyanin empirical ML algorithm	[Mixture density networks ML approach]
Floating Algae Index (FAI) (Hu, 2009)	Originally for detection of floating algae in oceans, later adapted for lacustrine cyanobacteria and algal bloom monitoring (Oyama et al., 2015; Wang et al., 2025)	$FAI = R_{rc}(NIR) - R'_{rc}(NIR)$ $R'_{rc}(NIR) = R_{rc}(RED) + (R_{rc}(SWIR) - R_{rc}(RED)) \times \frac{\lambda(NIR) - \lambda(NIR)}{\lambda(SWIR) - \lambda(RED)}$

1.4.2.4 Atmospheric correction

Prior to the application of parameter retrieval algorithms, it is standard practice to use an atmospheric correction procedure to remove atmospheric signals from the satellite reflectance data. There are various water-specific atmospheric correction approaches, including radiative transfer models (Vermote et al., 1997; Gao et al., 2000), image characteristic algorithms (Ruddick, Ovidio and Rijkeboer, 2000; Wang and Shi, 2005), and machine learning methods (Brockmann et al., 2016; Schroeder et al., 2022; Soriano-

González *et al.*, 2022; Van Nguyen *et al.*, 2025). Typically, atmospheric correction validation requires a range of in situ reflectance measurements to be taken just above the surface of the water. These data, paired with corresponding satellite imagery from approximately the same time, known as “match-ups”, are then used to evaluate the performance of different atmospheric correction procedures. Ideally the in situ reflectance measurements should closely agree with the corrected satellite reflectance data, but significant errors are very common, particularly in the blue and ultraviolet bands, where the contribution of absorbing aerosols can be hard to estimate (Song *et al.*, 2023). Furthermore, while oceanic atmospheric correction models can typically assume that water reflectance in the near-infrared wavebands is negligible, this is not the case for inland or coastal waters where suspended sediments may be present. Therefore, coastal and inland atmospheric correction is particularly challenging, and is an area of remote sensing research with many unsolved problems (Pahlevan *et al.*, 2021). Importantly, atmospheric correction methods are typically calibrated for specific sensors, and therefore are not always easily portable to new sensor configurations, particularly when the band arrangement differs significantly. For this reason, newer satellite missions such as the Planet SuperDoves constellation have considerably fewer options for atmospheric correction than longer-established missions like Sentinel-2 MSI (Warren *et al.*, 2019; Vanhellemont, 2023).

1.4.3 Automated in situ sensors

Over the last two decades, advances in various sensor technologies alongside developments in data communication and storage infrastructure has significantly increased the reliability and therefore ubiquity of automated, high frequency in situ sensors for water quality monitoring (Marcé *et al.*, 2016; Rozemeijer *et al.*, 2025). There is now a plethora of available sensors that will give high frequency estimates of parameters such as water temperature, chlorophyll-a, phycocyanin, or oxygen concentration. Typically, automated high frequency lake sensors are mounted on a buoy or platform, often powered by solar panels, or small wind turbines (Shukla, Matharu and Bhattacharya, 2023; Nogueira *et al.*, 2025).

Sensors capable of measuring physical parameters such as temperature, conductivity, turbidity, and some simple chemical parameters such as dissolved oxygen and pH, have been in use for several decades, dating back to the 1960s (Cleary, 1962; Bilello, 1968; Sylvester *et al.*, 1972). Following this, advances in measurement systems have enabled automated in situ monitoring of pigments such as chlorophyll-a using fluorescence-based

measurement approaches (Lorenzen, 1966; Uehlinger, 1985). Building upon this, more recently, fluorometric sensors have been developed for monitoring phycocyanin, enabling cyanobacteria-specific high frequency monitoring (Asai *et al.*, 2000). Alongside this, flow cytometry, which is instrumentation designed to pass a sample of particles suspended in a fluid one-by-one through sensing apparatus, is now capable of in situ phytoplankton monitoring (Sosik, Olson and Armbrust, 2010). Historically, flow cytometry instrumentation has largely focused on fluorescence analysis, however, in situ flow cytometers with imaging capabilities have also been developed (Olson and Sosik, 2007). This can enable high frequency automated monitoring of phytoplankton at genus level (Tapics, Gregory-Eaves and Huot, 2025). Alongside high frequency sampling, an additional advantage of most in situ sensor technologies is that they can take measurements at multiple depths. One way to achieve this is by fitting sensors to autonomous winched profiling systems (Morrison, Billings and Doherty, 2000; Singleton, De Young and Bachmayer, 2019). However, for lower-cost instruments, such as temperature sensors, it can be most economical to deploy them in series along a vertical string (Feuchtmayr, I.D. Jones, Clarke, *et al.*, 2021a, 2021b).

Critically, automated in situ sensors provide sensing capabilities which are completely different, but highly complementary to manual and remote sensing approaches. Importantly, they can sample at sub-daily time scales, and are therefore able to capture aspects of phytoplankton dynamics which satellite imagery and manual sampling approaches cannot reasonably resolve (Dubelaar, Geerders and Jonker, 2004; Brentrup *et al.*, 2016). Additionally, they can measure certain parameters which are not directly measurable remotely, such as water temperature at multiple depths. However, automated in situ sensing platforms are relatively expensive to install and operate, and require regular calibration and maintenance to ensure reliable, and consistent operation (Seifert-Dähnn *et al.*, 2021). For this reason, such sensors are typically only deployed in a small number of critical water bodies and locations, and so are well complemented by satellite imagery which can provide much broader spatial coverage. Furthermore, automated in situ measurement of many parameters, including photosynthetic pigments, can often suffer from various issues including biofouling, which causes drift; intermittent and continuous noise, which can impede measurement of low concentrations; and malfunctions, which can result in data gaps, flatlines, or other spurious measurements. Therefore, high frequency monitoring platforms are also well complemented by manual sampling and laboratory analyses as these can provide accurate data for both validation and gap filling in case of sensor outages or malfunction.

Additionally, automated in situ radiometers can also provide data for the validation of satellite remote sensing algorithms (Ruddick *et al.*, 2024). Generally, such sensors are not only capable of automatically taking and storing reflectance acquisitions; ideally, they should also be able to autonomously adjust the viewing angles of the downwelling (L_{sky}) and upwelling (L_u) radiance sensors. This ensures that an optimal relative azimuth angle, typically 135° , between these sensors and the sun is maintained (Figure 1.3). In addition to L_{sky} and L_u measurements, it is also necessary for these sensors to capture the downwelling irradiance, E_d . This approach forms the basis of the atmospheric correction validation data collected in the fourth chapter.

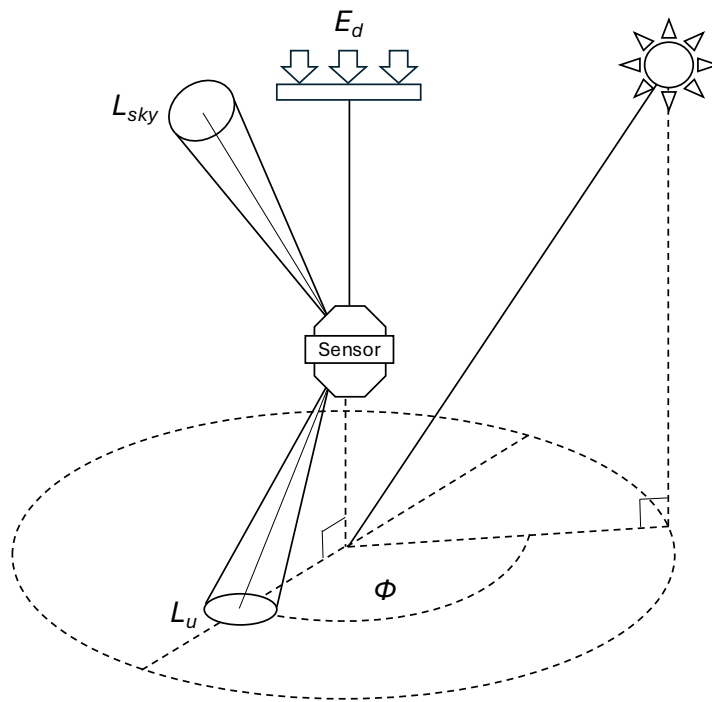


Figure 1.3: Typical autonomous above water radiance and irradiance sensor, which tracks the azimuth angle of the sun, to achieve an optimal relative azimuth angle, Φ .

1.4.4 Citizen Science

Alongside sensor-based approaches and manual monitoring by scientists, government agencies, and water companies, collection of data by members of the public can be a valuable tool for environmental and ecological monitoring. This approach, often called “citizen science”, simultaneously enables collection of diverse and potentially large datasets whilst also fostering a sense of collective environmental stewardship and increased awareness of important ecological issues. Recently, the availability of a diverse range of low-cost sensors has enabled the initiation of a new branch of citizen science, termed citizen sensing (Coulson, Woods and Making Sense EU, 2021). This extends the

already wide breadth of different data types which can now be potentially collected by members of the public for water quality monitoring (Coulson, Woods and Making Sense EU, 2021; Ramírez, van Meerveld and Seibert, 2023). Furthermore, the ubiquity of mobile smartphone devices with internet connectivity and relatively high quality integrated cameras has also revolutionised the collection of citizen science data, enabling users to easily upload information such as water colour indices (Malthus, Ohmsen and Woerd, 2020); reports of algal blooms (Scott, Ramaswamy and Lawson, 2016); wildlife observations and nutrient measurements (Bishop *et al.*, 2025); alongside photographs to centralised repositories. Typically, citizen science data which are uploaded to online repositories can be quality-controlled by experts and viewed by members of the public, thereby providing additional means for engagement with environmental issues.

Importantly, there are several similarities and differences between remote sensing data and citizen science data. Firstly, both have some inherent uncertainties. For example, with remote sensing data, it can be difficult to accurately estimate aerosol contributions over turbid inland waters (Song *et al.*, 2023). Additionally, chlorophyll-a, which is widely used for algae monitoring, is also present in all photosynthetic organisms, and so distinguishing algal blooms from floating macrophytes or leaf litter can be challenging (Liang *et al.*, 2017). Likewise, citizen scientists usually do not have access to expensive, well-calibrated scientific instruments, and therefore cannot make such precise measurements (Dickinson *et al.*, 2012). Additionally, citizen scientists generally do not receive much training, and therefore may report observations incorrectly. Secondly, both remote sensing and citizen science approaches have potential for producing widespread, relatively high frequency data, but this data is prone to gaps or ‘patchiness’. For remote sensing, gaps are typically due to cloud cover, whereas citizen science data can be inherently patchy because measurements and reports are often made on an ad-hoc basis without a regular sampling routine. Whilst both share drawbacks in common, the underlying sources of gaps and error in citizen science data and remote sensing are quite distinct and therefore independent. For this reason, there is potential for synergy between these two data sources, which could produce lower uncertainties, and fewer data gaps (Pattinson *et al.*, 2023). Whilst this potential has been recognised, there are only a handful of studies which have leveraged this for water quality monitoring applications (Putman *et al.*, 2023; Baltodano, Agramont and van Griensven, 2024).

Critically, it is important to recognise that the increasing availability of vast quantities of in situ sensor data, satellite imagery, and data collected through other means such as

citizen science requires embracing new approaches which can process and utilise these data effectively (Smits *et al.*, 2025). With more data available than ever at the fingers of the water quality scientist, it is prudent to look beyond monitoring and explore the additional insights and value this information can generate.

1.5 Forecasting algae

Measures to minimise the negative effects of harmful algal blooms may be broadly divided into 2 categories: (1) long-term mitigation strategies, and (2) reactive harm reduction measures. Strategies which address the root-causes are ultimately most desirable, but interventions like catchment nutrient reduction programmes may take decades to alleviate problems (Ibelings *et al.*, 2007; Janssen *et al.*, 2014; McCrackin *et al.*, 2017). Furthermore, it is now very likely that global warming will exceed 1.5 °C above pre-industrial levels, and that it may be hard to limit warming below 2 °C (IPCC, 2023; Bevacqua, Schleussner and Zscheischler, 2025). Therefore, climate change will continue to exacerbate problems with algal blooms well into the future, even if other remedial measures are taken (Carey *et al.*, 2012). For these reasons, it is very important that short-term harm-reduction and mitigative measures can be taken in a timely and effective way. Such measures include drinking water-supply switching; placement of warning signage or other communication with recreational water users; hypolimnetic siphoning; and activation of artificial mixing systems (Hamilton *et al.*, 2014). Typically, all the actions in this category are most effective if they can be implemented, or at least planned for, prior to the point where algal concentrations become dangerously high. For example, it is best to switch a drinking water supply to an alternative reservoir before toxins in a bloom-affected reservoir reach harmful levels. Therefore, the ability to anticipate, or forecast harmful algal bloom events days to weeks ahead is of great value (Caballero *et al.*, 2025).

Quantitative forecasting generally requires some form of model for the system in question. Such models may be very simple, for example, a “naive” model based on seasonal averages, or much more complex, built upon many equations which either describe various key physical, chemical and biological processes, or more abstract, less-interpretable relationships. This sets two alternative paradigms for algae forecasting approaches: process-based models, which attempt to mathematically synthesise scientific understanding of the system dynamics; and data-driven models which are built upon empirical observations, and normally do not use causal frameworks for prediction (Franks, 2018). Process-based models are most often used for long-term scenario modelling, for example to understand how algal bloom occurrence will change under different climate

change scenarios (Ralston and Moore, 2020). Conversely, for short-term forecasting (days to weeks ahead), data-driven models are generally preferred (Rousso *et al.*, 2020). Furthermore, as long-term datasets and high frequency monitoring platforms are becoming more common, many forecasters are making use of powerful computational systems like cloud computing to apply data-driven machine learning (ML) approaches to algae forecasting problems (Franks, 2018; Cruz *et al.*, 2021). As this thesis is largely focused on short-term forecasting, the forecast models used in this thesis are all data-driven machine learning models. However, the third chapter makes use of data generated by a suite of process-based models, and so it is of importance to understand the fundamentals of the models which underpin this important branch of lake science.

1.5.1 Process-based modelling and forecasting

Fundamentally, process-based models attempt to describe the system in question using equations originating from scientific knowledge of the system. For a lake ecosystem, this requires combining physical laws, such as heat transfer equations with biological models of algal parameters such as growth rates, grazing losses, and flushing. The challenge of using process-based approaches to model such systems is undoubtedly the sheer complexity of the underlying systems and the number of parameters which must be tuned. It is common for lakes to use one-dimensional models of lakes where all variables, including temperature, nutrient concentrations and algal densities are all assumed to be homogenous except in the vertical dimension. This simplification still allows for modelling of what is one of the most important physical processes - stratification. Accordingly, one-dimensional process-based lake models enable modelling of lake physics, chemistry and ecology in a computationally efficient way (Elliott and Thackeray, 2004; Hipsey *et al.*, 2019; Hipsey, 2022). Whilst two- and three- dimensional lake models are also used, the extensive volume of calibration data required means that they are generally only used for extremely well-studied lakes where sufficient data is available (Mooij *et al.*, 2010; Missaghi, Hondzo and Melching, 2013; Hu, 2016; Xue *et al.*, 2017, 2022).

In the field of cyanobacterial and algal bloom modelling, process-based models are most commonly used for longer-term scenario modelling, and although they can be used for making short-term predictions, data-driven models are more commonly used for this application (Rousso *et al.*, 2020). One reason for this is that process-based models typically have a large number of parameters which all need to be calibrated for the lake in question. Accurately measuring some of these parameters can be challenging and therefore

this can result in significant uncertainty or errors, which can be limiting for short-term forecasting tasks (Page *et al.*, 2018). On the other hand, process models have the advantage that they are based upon tested physical, chemical, and biological variable relationships. This means that they should generally be more suitable than data-driven approaches for modelling scenarios which have not yet been measured or recorded (Franks, 2018). For example, the impact of climate change on future algal bloom trends can be modelled using process-based models (Tigli *et al.*, 2025). Similarly, the third chapter of this thesis uses data generated by a suite of one-dimensional process-based lake models for evaluating how short-term ML forecast performance could be affected under different climate scenarios.

1.5.2 Data-driven forecasting

Fundamentally, data-driven forecasting is focused on training computer systems to produce an output based on some inputs, following patterns or trends that have been calculated, or learnt, from some historical training data. In this sense, ‘learning’ can be seen as a process, which is often iterative, where weights in a set of equations are optimised to fit a training dataset such that the resultant equations are useful for making predictions of the output variable of interest. Ideally, this learning process should result in a set of equations or algorithm which is not *overfitted* - it should generalise well to unseen data.

From this, two learning categories can be defined: supervised and unsupervised. In supervised learning, models are trained using data which includes both inputs and the desired corresponding output values. Conversely, in the case of unsupervised learning, only input data is provided, without any desired outputs or labels (Almuqati *et al.*, 2024). Unsupervised learning is most suited to applications involving revealing patterns in data, such as clustering, whereas for forecasting there is typically a target parameter, and therefore supervised learning is more commonly used (Bojer, 2022).

Linear regression can be considered a simple example of supervised learning (Qu, 2024). Using least-squares, or another line fitting approach, the gradient, m , and the y-intercept, c , are calculated from the training dataset consisting of input data, x , and output data, y . Following this training step, the model can make predictions for y when given new, unseen x values using the well-known equation for a straight line:

$$y = mx + c . \quad (3)$$

After training a data-driven or ML model, it is normal to evaluate performance of the model on unseen test data, using metrics such as root mean square error (RMSE) or mean absolute percentage error (MAPE). Whilst in a strictly statistical framework, the validity of using linear regression would be based on meeting a set of stringent assumptions regarding the data, this is not the case for ML (Poole and O’Farrell, 1971). Conversely, if adopting a ML or data-driven approach, then the validity of the model is purely assessed by its performance on a previously unseen test dataset. Under this paradigm, many algorithms have been developed which seek to optimise the process of what is, in effect, fitting equations to data such that they generalise well to unseen data. In this thesis, the terms *data-driven modelling* and *machine learning* are used almost interchangeably, although it should be noted that ML is probably most correctly considered a subset of data-driven modelling, alongside other approaches such as statistical modelling, which is certainly data-driven in nature, but applies different assumptions in order to generate models. What now follows is a brief overview of some of the key ML approaches used in this thesis.

1.5.2.1 *Decision Tree-based approaches*

A decision tree has a structure much like that of a flowchart, where nodes represent a condition check, the outcome of which determines which child node is evaluated next (Figure 1.4). The nodes located at the end of each chain, known as *leaf nodes*, contain the final predictions. For regression problems, decision trees are constructed by recursively partitioning the feature space to minimise the residual sum of squares at each split (Mienye and Jere, 2024). In the example shown in Figure 1.4, the algorithm splits the training data into four groups (the number of leaf nodes), and the average of each of these determines the predictions which are made.

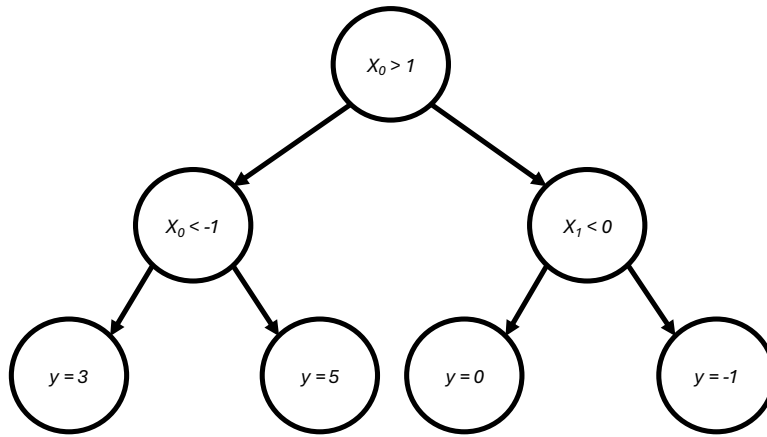


Figure 1.4: Example diagram of a decision tree with four leaf nodes and a depth of two.

Individual decision trees do not generalise well - if made, or grown, too deep they will tend to overfit, whereas if grown shallow, they will tend to oversimplify the problem and will be unable to capture intricacies in the data (Halabaku and Bytyçi, 2024). To overcome this, ensembles of decision trees are commonly used. One such example of a decision tree ensemble is the random forest (RF), which grows multiple trees which are each trained on a random subset of the training data and feature space, known as bootstrapping (Breiman, 2001). Predictions are made by aggregating the output of each individual tree. This bootstrap aggregation process, known as bagging, is effective because although each individual tree may be overfitted, the average of their outputs will not be, as random sampling ensures that trees are not correlated (Breiman, 1996).

1.5.2.2 Support Vector Machines

Support Vector Machines (SVM) were initially invented to solve classification problems but have also been adapted and used extensively for regression problems, where they are usually referred to as Support Vector Regressors (SVR) (Cortes and Vapnik, 1995; Drucker *et al.*, 1996; Vapnik, Golowich and Smola, 1996). Unlike ordinary regression methods that aim to minimize the total prediction error, SVR includes an error-insensitive margin or tube, ε , and attempts to find a linear function that fits the training data, ignoring errors smaller than ε . Predictions that fall outside the ε -tube are penalised using slack variables, whilst controlling for over-fitting using a regularisation parameter, C . SVM and SVR usually make use of the *kernel trick*, which enables modelling of non-linear relationships by effectively projecting input data into a higher-dimensional feature space where linear SVM regression is possible (Boser, Guyon and Vapnik, 1992; Vapnik, 1995). Instead of explicitly evaluating this projection, a kernel function, usually either polynomial or radial basis function (RBF), is used to determine inner products in the

transformed space directly. This allows efficient training without increasing computational cost significantly.

1.5.2.3 Feedforward Artificial Neural Networks

Artificial neural networks (ANN) take a form inspired by the neural structure of the human brain (Rosenblatt, 1962). They generally consist of nodes, which are artificial neurons, connected in layers. Each neuron has a weight which is fitted to the data, allowing for some connections in the network to effectively be stronger than others. Typically, an ANN consists of an input layer which receives the input data, several hidden layers which extract features and perform calculations on the data, and then an output layer which produces the final prediction or result. The simplest modern ANN architecture is the multi-layer perceptron (MLP), which is a fully connected feedforward neural network – that is, data only travels forwards through the network, there is no feedback, and every node in each layer is connected to all other nodes in adjacent layers (Figure 1.5) (Murtagh, 1991). The number of neurons in the input layer is determined by the number of available input variables or features, and the number of output nodes is determined by how many output variables there are. Conversely, the number and size of hidden layers is fully adjustable and should be tuned for the specific problem and size of training dataset (Yu and Zhu, 2020).

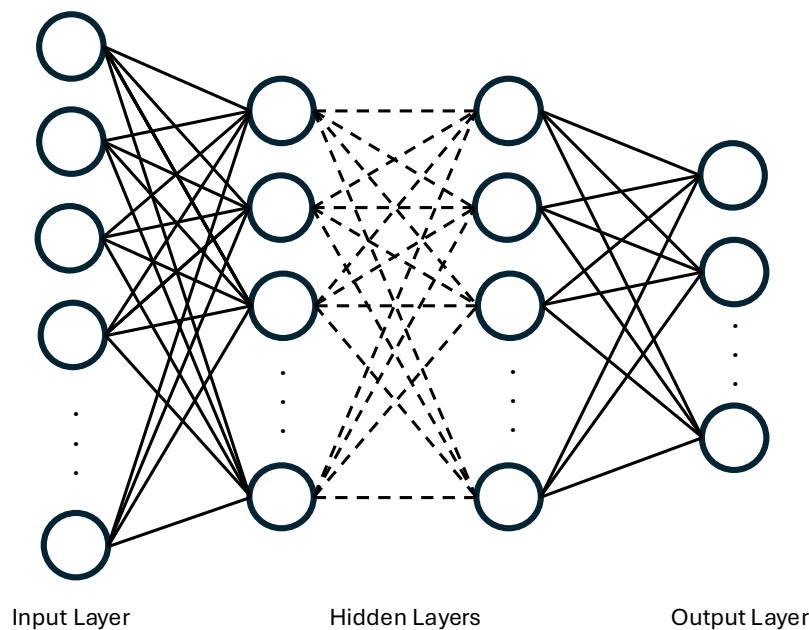


Figure 1.5: Fully connected multi-layer perceptron architecture.

1.5.2.4 Recurrent Neural Networks

For forecasting tasks, a potential disadvantage of ML models like MLP, SVM and RF is that these do not take advantage of any inherent order in data with a clear structure such as time series data. Whilst such models can learn to make use of structured patterns in data, when fed a time series as input, they will initially treat this as if there is no inherent order. This potentially increases the volume of training data required for meaningful learning of important patterns. To overcome this shortcoming, various recurrent neural network (RNN) architectures have been designed for sequential data such as time series or text.

In their simplest form, RNN process sequential data, by maintaining a hidden state that retains information from previous time steps (Figure 1.6a) (Rumelhart and McClelland, 1987; Sherstinsky, 2020). This allows RNNs to use patterns in time series data more effectively than traditional machine learning models. However, standard RNNs often suffer from vanishing or exploding gradients which can hinder convergence during training and make it difficult to learn long-term dependencies (Pascanu, Mikolov and Bengio, 2013). To address this, Long Short-Term Memory (LSTM) networks were introduced (Hochreiter and Schmidhuber, 1997). LSTM networks use units called memory cells, along with input, forget, and output gates that regulate information flow (Figure 1.6b). This enables them to retain relevant information over longer sequences and “forget” less important data. Therefore, LSTM networks are well-suited for time series forecasting where it is important for a model to be able to learn how past events influence future values. More recently, gated recurrent units (GRU) were devised as a simplified alternative to the LSTM which combines the forget and input gates into a single update gate, therefore requiring fewer parameters to be learnt (Figure 1.6c) (Cho *et al.*, 2014). Although simpler, GRU networks can often perform comparably to LSTM networks and therefore are highly suited for applications with limited training data or computational resources (Chung *et al.*, 2014).

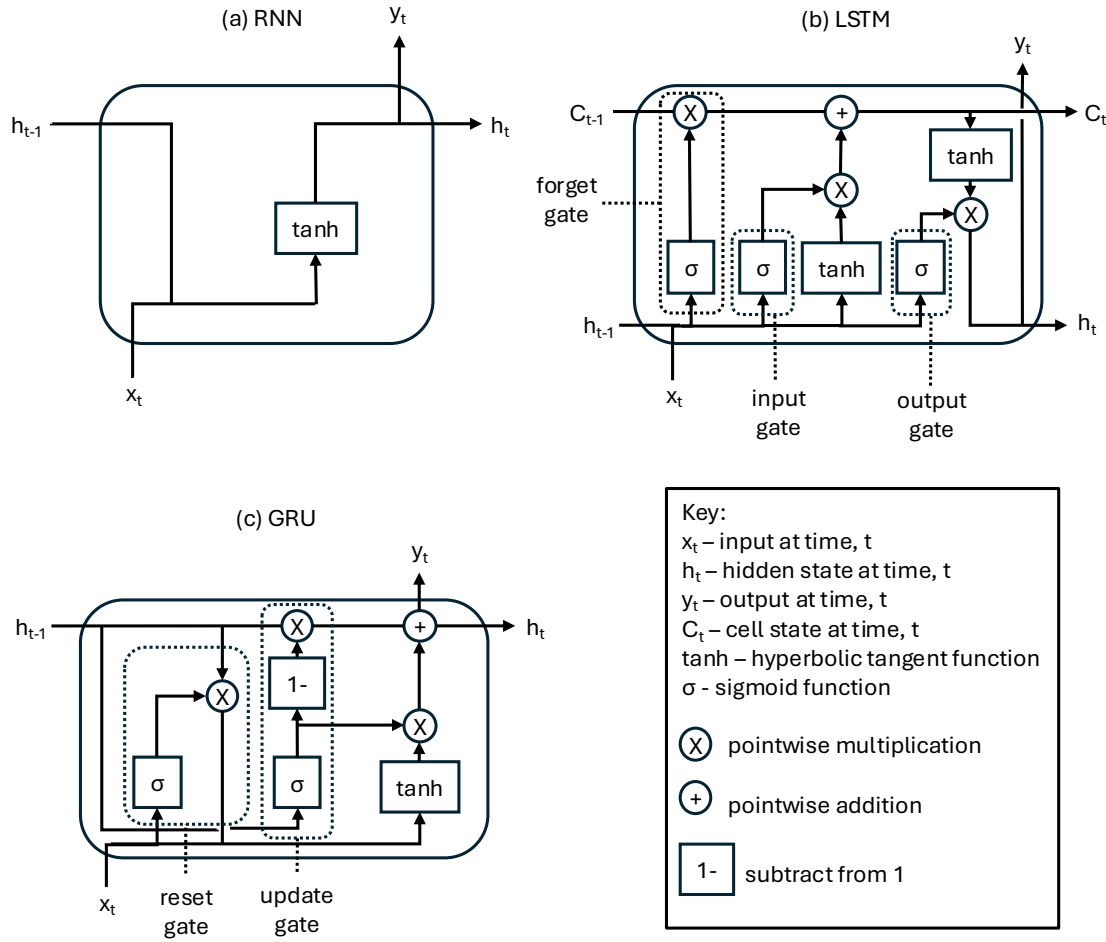


Figure 1.6: Fundamental units of different recurrent neural network architectures – (a) simple RNN, (b) LSTM, (c) GRU.

1.5.2.5 Machine learning for forecasting algal blooms

In recent years, there has been an increasing volume of research carried out which investigates the use of machine learning for algal forecasting (Rousso *et al.*, 2020; Cruz *et al.*, 2021). This has included use of models like RF and SVM/R, which are often used alongside new, more complex models for comparison (Ai *et al.*, 2023; Gupta, Hantush and Govindaraju, 2023; Zhu *et al.*, 2023; Chen *et al.*, 2024; Medina *et al.*, 2024; Schaeffer *et al.*, 2024; Amara *et al.*, 2025; Nguyen *et al.*, 2025), as well as time series specific models like LSTM and GRU (Ahn, Kim and Kim, 2023; Wu *et al.*, 2023; Zamani *et al.*, 2023; Fournier *et al.*, 2024; Zhang, Ding and Zhang, 2024; Busari *et al.*, 2025; Loussaief, Ktari and Hadj Kacem, 2025). There is no strong consensus on which models are most effective - this appears to change depending on the problem and dataset. For example, whilst more complex models such as LSTM sometimes outperform classical ML approaches such as SVM or RF (Chen *et al.*, 2024), there are also examples of simpler models outperforming more complex models (Vincent *et al.*, 2023). To-date, most of the work in this field has been oriented towards the application of ML models to algal forecasting tasks, often with

a focus on how new model formulations or structures can enhance predictive performance (Rousso *et al.*, 2020; Zahir *et al.*, 2024). However, whilst ML algal forecasting is reliant on sufficient training data, both in terms of volume and quality, there is little work which evaluates the parameters of this reliance specifically. Therefore, there is a need for research which seeks to understand how monitoring practices should be optimised or focused in order to facilitate effective ML algae forecasting (Park, Patel and Lee, 2024).

1.5.2.6 Machine learning data requirements

A core part of this thesis concerns data requirements for short-term algal bloom forecasting systems. Assuming that a ML approach is used, then a fundamental question which arises is: how large a dataset is required to train algal forecasting systems which perform adequately? There is some degree of subjectivity here because defining what is meant by *adequate performance* will vary depending on the management aims, and how the forecast is intended to be used. Nonetheless, even without a universal definition of adequate performance, it is still very much possible to experimentally measure how machine learning performance varies as the training dataset size is increased.

Understanding this for algal bloom forecasting may be critical to guiding monitoring strategies and importantly would allow for evaluation of whether data-driven forecasting is feasible using existing monitoring datasets. Furthermore, this understanding should be helpful in setting expectations for how forecast performance may improve in the future as more training data is collected over several years.

ML performance is wholly reliant on a training dataset of sufficient size and quality. A large volume of data which has no relevance to the predictions being made will be unhelpful. Similarly, if only a very small number of observations of variables which are useful exist, then ML is likely a poor modelling choice. On average, machine learning performance typically scales quite predictably with training dataset size – error reductions tend to be correlated with the volume of training data according to an inverse power law: $1/D^\alpha$ (Figure 1.7), where D is the training dataset size and α the scaling exponent, which varies depending on the problem and model constraints (Mahmood *et al.*, 2022; Bahri *et al.*, 2024). One important consequence of this is that over relatively small scales performance improvements may appear to show increasingly diminishing returns as more training data are added (Figure 1.7a). In other words, performance appears to plateau.

Understanding at what training dataset size this plateau appears for algal forecasting problems is of great importance, as this potentially sets an upper bound on the performance improvements that can be made through collecting additional training data.

This is critical because collecting additional data may take several years and carries significant costs. Therefore, understanding how effective this will be in advance is key to setting management expectations. Beyond the apparent plateau it is still possible to improve performance by increasing the training dataset size, but only with increases of several orders of magnitude (Figure 1.7b). Furthermore, for problems where the training dataset size is limited, known as resolution-limited regimes, it has been shown that $0 < \alpha < 1$ (Bahri *et al.*, 2024). However, where D is very large, Bahri *et al.* (2024) show that α approaches 1, which is referred to as a variance-limited regime. This indicates that the rate at which performance improves as more training data are added may increase if sufficient data become available to enable a shift from resolution- to variance- limited regimes. Such regime shifts have been empirically observed: as training dataset size is scaled over several orders of magnitude, the gradient at which performance improves may dramatically increase (Caballero *et al.*, 2023).

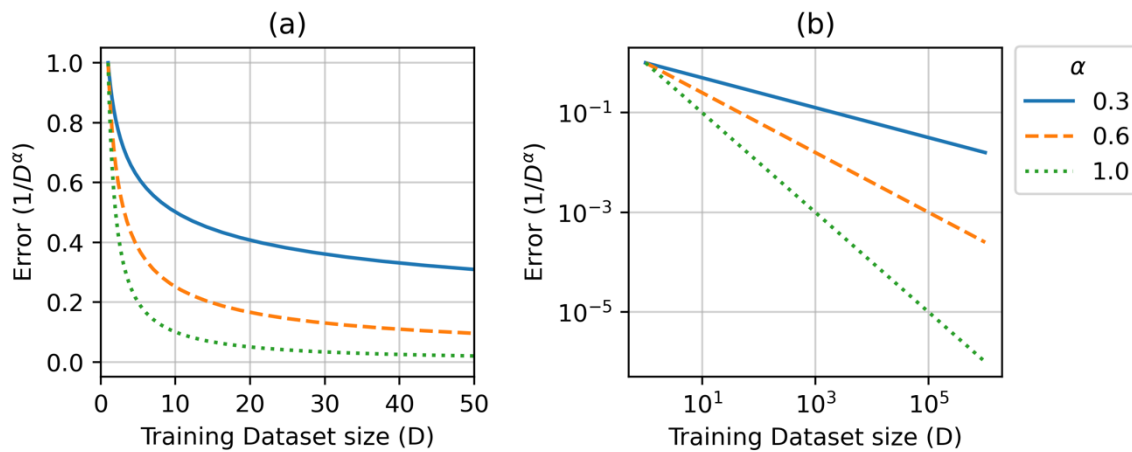


Figure 1.7: ML performance curves showing how error decreases as the training dataset size (D) is increased according to an inverse power law ($1/D^\alpha$) in both linear space (a), and log-space (b). Three different values of the scaling parameter, α , are shown: $\alpha = 0.3$ (blue solid lines), $\alpha = 0.6$ (orange dashed lines), and $\alpha = 1.0$ (green dotted lines).

In the context of short-term algal forecasting, the main consequence of this typical training behaviour is as follows. It is expected that ML forecast performance will effectively plateau after some number of years' training data is exceeded. If at this point, the performance is not adequate, then simply continuing to collect more years' training data from the lake of interest will not be an effective way to improve forecast performance – different approaches must be explored. Accordingly, alternative approaches for improving performance might include: increasing the sampling frequency to better capture critical components of the algal dynamics (Dubelaar, Geerders and Jonker, 2004; Pobel, Robin

and Humbert, 2011), taking additional measurements of potentially useful predictor variables (Ahn, Oh and Park, 2011; Roussio *et al.*, 2020), or exploring alternative models (Harris and Graham, 2017; Ahn, Kim and Kim, 2023). Alternatively, training with data from multiple lakes may be a way to effectively increase the training dataset size by several orders of magnitude and therefore increase performance despite the plateau (Mahmood *et al.*, 2022; Schaeffer *et al.*, 2024).

1.6 Research Priorities

The dual pressures of climate change and human-induced eutrophication suggest that harmful algal blooms will remain a persistent and potentially escalating water management challenge. In particular, cyanobacterial blooms in lakes are likely to become more widespread in many regions due to climate change. However, at present, the vast majority of water bodies are not actively monitored. Therefore, pre-emptive harm mitigation and prevention measures for algal blooms have currently only been implemented for a very small number of lakes where there has been sufficient investment in monitoring and forecasting infrastructure to facilitate this. However, at the same time, our ability to monitor algal dynamics is advancing rapidly. Remote sensing technologies now enable the estimation of algal biomass in inland waters from space, and new CubeSat constellations may offer the potential to monitor and track algal blooms at an unprecedented spatial and temporal resolution. Additionally, the volume of potentially useful water quality data being collected by scientists, regulatory agencies, and members of the public is continually increasing. Furthermore, with improvements in computational power and ML methods, forecasting harmful algal blooms is becoming increasingly feasible and accurate. However, for forecasting tools to be truly effective, future monitoring strategies must be designed with ML requirements in mind. In addition, it may be critical for these strategies to account for the heightened uncertainty introduced by climate change, particularly in short-term predictive models. Therefore, these challenges and opportunities highlight several key questions which this thesis is focused upon:

1. What are the data requirements in terms of monitored parameters, dataset duration, and sampling frequency for short-term data-driven algal forecasting to outperform naïve benchmark models?
2. Will climate change affect short-term lake forecast performance or data requirements?
3. How can we track and ultimately forecast algal blooms in the large number of typically smaller-sized lakes which are presently unmonitored?

4. Where, across the diverse selection of data sources available, are there untapped synergies which can improve algal bloom monitoring and forecasting capabilities?

This thesis addresses these questions through three key investigations, each presented as a research paper. The first of these (Chapter 2) investigates the data requirements for effective ML-based algal bloom forecasting, focusing on the influence of dataset duration, sampling frequency, and parameter selection. Following this, Chapter 3 assesses how water quality forecast skill may be affected under different climate change scenarios and explores adaptive monitoring strategies to mitigate potential declines in model performance. Finally, Chapter 4 explores the potential of a new source of algal monitoring data – the recently launched Planet SuperDoves satellite constellation. Throughout, data from a wide variety of sources is used and evaluated, highlighting potential synergies and novel uses. By pairing a study of the potential of new monitoring technology with an in-depth critical evaluation of data requirements for algal bloom forecasting, this thesis aims to shed light on the future of algal monitoring and forecasting practices. Ultimately, this enables evaluating if the existing paradigms in this field are truly suitable for tackling future water quality challenges.

2 Are more data always better? Machine learning forecasting of algae based on long-term observations

This chapter was published as:

Atton Beckmann, D., Werther, M., Mackay, E.B., Spyrakos, E., Hunter, P., Jones, I.D., 2025. Are more data always better? – Machine learning forecasting of algae based on long-term observations. J. Environ. Manage. 373, 123478.

<https://doi.org/10.1016/j.jenvman.2024.123478>

CRedit authorship contribution statement:

D. Atton Beckmann: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **M. Werther:** Writing – review & editing, Supervision. **E.B. Mackay:** Writing – review & editing. **E. Spyrakos:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **P. Hunter:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **I.D. Jones:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

2.1 Abstract

Bloom-forming algae present a unique challenge to water managers as they can significantly impair provision of important ecosystem services and cause health risks to humans and animals. Consequently, effective short-term algae forecasts are important as they provide early warnings and enable implementation of mitigation strategies. In this context, machine learning (ML) emerges as a promising forecasting tool. However, the performance of ML models is heavily dependent on the availability of appropriate training data. Consequently, it is essential to determine the volume of data necessary to develop reliable ML forecasts. Understanding this will guide future monitoring strategies, optimize resource allocation, and set realistic expectations for management outcomes. In this study, we used 30 years of fortnightly measurements of 13 different parameters from a lake in the English Lake District (UK) to examine the impact of training data duration on the performance of ML models for forecasting chlorophyll-a two weeks in advance. Once training data availability exceeded four years, a Random Forest model was found to consistently outperform naive benchmarks (mean absolute percentage error 16.4 % lower than the best-performing benchmark). With more than five years of training data, model performance generally continued to improve, but with diminishing returns. Furthermore, it was found that equivalent and, in some cases, better performance could be achieved by only using a subset of the most important input features. Additionally, it was found that reducing the sampling frequency had negative impacts on performance, both due to the reduced number of training observations available, and increased forecast horizon. Our findings demonstrate that for lakes ecologically similar to the study site, a consistent and regular sampling programme focused on monitoring a limited number of key parameters can provide sufficient observations for generating short-term algae forecasts after approximately five years of data collection. Importantly, this result provides justification for the initiation of new monitoring programmes for sites where algal blooms are a concern, and suggests that there are likely many pre-existing monitoring datasets which would be suitable for training algae forecast models.

2.2 Introduction

Excessive growths of algae in freshwater, commonly referred to as algal blooms, compromise the safety of drinking water sources (Brooks *et al.*, 2016; Igwaran *et al.*, 2024), endanger recreational water activities (Carvalho *et al.*, 2013; Wolf, Georgic and Klaiber, 2017), and threaten the stability and diversity of aquatic ecosystems (Dolah, Roelke and Greene, 2001; Amorim and Moura, 2021). Blooms can also impose significant

economic impacts, affecting tourism, the fish industry, and even property prices (Hoagland and Scatasta, 2006; Hamilton *et al.*, 2014). The far-reaching effects of this issue underscores the urgency for effective management and mitigation strategies to protect public health and preserve ecological integrity (Codd *et al.*, 2005; D'Anglada *et al.*, 2016; Chorus and Welker, 2021).

Measures such as drinking water supply switching, chemical treatment, hypolimnetic syphoning, ultrasonic control, and use of artificial mixing systems can be effective to mitigate some of the negative effects of harmful algal blooms (Stroom and Kardinaal, 2016). However, many of these short-term measures are best used for only a brief period, due to cost, efficacy, or water resource management implications. For example, a drinking water supply may be switched to an alternate reservoir whilst there is an algal bloom in the main supply reservoir, but the secondary reservoir may only have limited water supply capacity. Therefore, the implementation of these management measures may benefit from early warning systems that enable implementation in a timely way. Short-term algae forecasting (days to weeks ahead) is therefore needed to enable effective management of algal blooms (Hamilton *et al.*, 2014; Ibelings *et al.*, 2016; Stroom and Kardinaal, 2016; Rouso *et al.*, 2020; Cruz *et al.*, 2021).

In recent years, machine learning (ML) approaches have become popular for algae forecasting due to their ability to model non-linear dynamics using complex, and large datasets (Franks, 2018; Rouso *et al.*, 2020; Cruz *et al.*, 2021; Xiao *et al.*, 2024). ML models vary in complexity and form, and include linear regression models, decision-tree based approaches, and artificial neural networks (Sarker, 2021b). Appropriate training data are crucial for ML algae forecasts. Selection of input variables, data frequency, and data collection methods have an effect on performance and this has been explored to some extent for algae forecasts (Muttill and Chau, 2006; Bai *et al.*, 2017; Bertani *et al.*, 2017; Xiao *et al.*, 2017; Thomas *et al.*, 2018; Rouso *et al.*, 2020; Lin, Pierson and Mesman, 2023). In contrast, there are very few studies which explicitly explore the effect of the length of training dataset on forecast performance. Understanding this would guide future monitoring strategies and set realistic expectations for management outcomes, and therefore be of great value to water managers considering developing algae forecast systems.

Climate change, variations in weather, and factors such as changes to land-use or management practices present a further challenge when forecasting water quality. That is, we cannot expect that water systems will behave in an unchanging way year after year:

they exhibit non-stationarity (Milly *et al.*, 2008, 2015). Therefore, whilst it might be assumed that longer training datasets will always give better performance, it may be that if older data represent conditions which are no longer relevant to the present, this could hinder forecasts rather than improve them.

The length of time required to collect sufficient samples for training ML forecasts is likely to be a more considerable constraint for manually collected data than for sub-daily data generated by automated monitoring platforms (Rousso *et al.*, 2020). Consequently, it is particularly important to understand how the length of training data affects forecast performance for manually collected data. Importantly, sampling frequency influences forecasts in two ways. First, it limits how many observations can be collected over time, directly affecting how quickly sufficient training data is accumulated for accurate ML forecasts. Secondly, it establishes the lower boundary for the forecast horizon, how far ahead predictions can be made. The further ahead the prediction, the greater the uncertainty and errors (Derot, Yajima and Schmitt, 2020).

In previous algae forecasting studies that used data sampled on weekly or longer intervals, the training dataset length has varied considerably, from less than four years (Teles *et al.*, 2006; Torres *et al.*, 2011), to more than ten years training data (Recknagel *et al.*, 1997; Talib *et al.*, 2008; Welk *et al.*, 2008; Lin, Pierson and Mesman, 2023). Whilst there is considerable variation in performance between studies using different length datasets (Rousso *et al.*, 2020), attributing this specifically to the length of training dataset is difficult due to differences in the performance metrics, forecast horizons, models, and study sites used.

Understanding how many years of data are required for effective forecasting is only useful if coupled with an understanding of which input variables (features) should be measured. This is particularly important for managers in the early stages of initiating new water quality monitoring schemes for algal bloom management. There are some studies which found that blooms could be forecasted using only previous measurements of chlorophyll-a (chl-a), a common proxy for algal biomass (Muttill and Chau, 2006; Xiao *et al.*, 2017). However, algal blooms are known to be caused by multiple drivers, for example water temperature, stratification, residence time and nutrient availability (Sellner, Doucette and Kirkpatrick, 2003; Paerl and Otten, 2013). Consequently, for some systems, these and perhaps other drivers will be necessary for training effective forecasts (Rousso *et al.*, 2020). Therefore, understanding which features are most important, and how this couples

with the length of training dataset available would be of great utility to water managers interested in forecasting algae.

Exploring how the performance of ML algae forecast models is influenced by the length of training data is crucial. This knowledge would enable evaluation of the applicability of ML models to existing data, thus guiding the development of algae forecasting systems for water bodies where monitoring is already being undertaken. Additionally, this understanding would inform the design of new monitoring programmes where forecasting algae is a specific objective. This would give an estimation of the number of years' data required for viable ML forecasts, enabling better planning. This is particularly pertinent as many water managers may soon wish to implement more regular and thorough monitoring programs given that harmful algal blooms may increase in intensity and frequency due to anthropogenic climate change (Paerl and Huisman, 2009; Paerl and Paul, 2012; Ho, Michalak and Pahlevan, 2019).

Evidently, there is a need to understand the data requirements for ML forecasts of algae. In particular, the question of how long data must be collected for before they become useful for forecasting has not been investigated explicitly. Therefore, we study the following questions using 30 years of fortnightly monitoring data from a small, meso-eutrophic lake:

1. What is the minimum number of years' training data required to surpass the performance of naive benchmark models in forecasting?
2. As we extend the dataset over more years, do we observe diminishing returns in forecast performance improvements?
3. How does the impact of dataset duration on forecast performance vary across different ML models?
4. What is the minimum number of features required to achieve satisfactory performance from a model?
5. How does a reduction in the sampling frequency affect the impact of dataset duration on performance?

2.3 Material and methods

2.3.1 Study site

Blelham Tarn is a small, sheltered lake in the English Lake District. Located at 54° 23' 44" N, 2° 58' 41" W, it has a mean depth of 6.8 m, maximum depth of 14.5 m, surface area of 0.102 km², and volume of 693000 m³ (Ramsbottom, 1976). The lake has several significant external nutrient sources, including a sewage works for the nearby village of Outgate, and sheep grazing in the catchment. The lake is monomictic and is usually stratified from spring to late autumn (Atkinson, 1999). It is on the meso-eutrophic/eutrophic boundary and has diatom blooms (currently dominated by *Asterionella formosa*) in spring followed by algal blooms in the summer that presently include cyanobacteria (*Dolichospermum sp.*), green algae (*Paulschulzia sp.*), and cryptophytes (*Cryptomonas sp.*). The summer blooms in Blelham Tarn have been associated with hypolimnetic anoxia (Elliott and Thackeray, 2004; Foley *et al.*, 2012).

Thirty years of data (1987–2017) were collated including measurements from in-lake biogeochemical monitoring, nearby meteorological stations and river flow stations. The year 2001 was excluded from all analyses as there was no in-lake monitoring for most of that year due to an outbreak of foot and mouth disease.

The in-situ lake data used were from the extensive long-term monitoring programme that was started by the Freshwater Biological Association in 1945 and continued by the UK Centre for Ecology and Hydrology (UKCEH) since 1989. Details of the data collection procedure can be found in the dataset documentation (Maberly *et al.*, 2017; Feuchtmayr, M.A Clarke, De Ville, *et al.*, 2021). The fortnightly monitoring data used includes a large variety of variables, such as surface temperature, surface oxygen, several water chemistry parameters, and phytoplankton chl-a. Initially, all the available variables were used, excluding those with continuous data gaps of more than two months (56 days) during the study period (see Table 2.1).

Daily meteorological data from the UK Meteorological Office (Met Office, 2019) were obtained from four relatively nearby weather stations in the English Lake District (Walney Island, Shap, Newton Rigg, and Keswick). These data were down-sampled to fortnightly intervals by taking the mean from the two-week period of interest, and then the mean value from all four stations was used. These stations were chosen as they are all approximately equidistant (~ 30 km) from Blelham Tarn and therefore the mean should be representative of conditions at the lake itself. Additionally, this approach has the

advantage of ensuring a weather data time series with very few missing values as it is rare for missing values to occur simultaneously at all four stations.

For catchment-specific rainfall measurements, catchment rainfall data from the two nearest stations to Blelham Tarn (Brathay and Rothay) were provided by the UK National River Flow Archive (NRFA) (Tanguy *et al.*, 2021; NRFA, 2023b, 2023a). The average from these two stations was used for analysis.

All data were processed by first linearly interpolating to daily samples, and then down-sampling to generate data at regular fortnightly intervals with no missing values. All variables used are summarised in Table 2.1.

Table 2.1: Summary of variables, and sources of data used in the study to train and test ML forecast models.

Variable	Units	Source
Ammonium	$\mu\text{g L}^{-1}$	UKCEH
Nitrate	$\mu\text{g L}^{-1}$	UKCEH
Surface oxygen saturation	% Air Saturation	UKCEH
Soluble reactive phosphate (SRP)	$\mu\text{g L}^{-1}$	UKCEH
Dissolved reactive silica	$\mu\text{g L}^{-1}$	UKCEH
Surface water temperature	$^{\circ}\text{C}$	UKCEH
Phytoplankton chlorophyll-a	$\mu\text{g L}^{-1}$	UKCEH
Total phosphorus (TP)	$\mu\text{g L}^{-1}$	UKCEH
Mean daily wind speed	Knots	UK Met Office
Mean daily relative humidity	%	UK Met Office
Mean daily cloud amount	Oktas	UK Met Office
Mean daily air temperature	$^{\circ}\text{C}$	UK Met Office
Mean daily rainfall (average from Brathay and Rothay catchments)	mm	NRFA

To explore any consistent changes that occurred in Blelham Tarn over the 30-year study period, we examined decadal trends in the input variables. This was done by linearly interpolating each variable to daily values and then averaging these over three ten-year periods (1987-1996, 1997-2007, 2008-2017), noting that 1997-2007 excluded the year 2001 because of the missing data then. This allowed for identification of variables that have changed either in magnitude or seasonal timing during the study period.

2.3.2 Algae forecasting

Chl-a was used as the forecast target, as it is commonly used as a proxy for phytoplankton biomass, and was the focus of many previous algae forecast studies (Lin et al., 2023; Luo et al., 2017; Mellios et al., 2020). Chl-a forecasts were made two weeks ahead to match the sampling frequency. The forecasting models developed in this study incorporate measurements from the four weeks leading up to the forecast date, equating to two sampling dates, or observations. This selection was strategic, prioritising recent changes in input features while avoiding an overly dimensional feature space. Such a constraint is crucial because a ML model's effectiveness hinges on training with more observations than features (Koutroumbas and Theodoridis, 2008). Considering the fortnightly frequency of chl-a data collection, one year yields 26 observations. Thus, employing 26 features, derived from 13 variables, across two separate observations, represents an upper limit to maintain model trainability with limited annual data.

A simple approach was used to simulate the integration of forecasted weather observations into the models, as these data would likely be readily available for operational forecasting. Specifically, the five meteorological variables were adjusted to reflect a two-week forward shift, meaning that for a given observation date, the meteorological input features correspond to the average conditions of the subsequent two weeks. Forecasting models are usually more effective at making predictions of a stationary time series, one that does not exhibit a trend or seasonality (Granger and Newbold, 1974). Therefore, the forecast target should have a constant mean and variance over time, and low autocorrelation. To convert a non-stationary time series into a stationary one, a common technique is to calculate the differences between consecutive data points, a process known as ‘differencing’ (Hyndman and Athanasopoulos, 2021). Preliminary research demonstrated that applying first-order differencing - subtracting the previous observation from the current one in a time series - successfully eliminated both the seasonal patterns and the overall trend in chl-a. Consequently, we trained our forecasting models to predict the changes in chl-a levels, rather than the absolute values..

2.3.2.1 Machine Learning Models

It is anticipated that models with different complexities and architectures will require varying amounts of training data to perform optimally (Lones, 2024). Accordingly, several models were employed in this study to encompass a broad range of computational approaches. Specifically, we used a Random forest (RF), Neural Network (multilayer perceptron: MLP), recurrent neural network with gated units (GRU), and a Support Vector

Machine (SVM), all implemented through the Python package *scikit-learn* (Pedregosa *et al.*, 2011). These models were chosen as they all use different model architectures and have been shown in previous studies to be effective for algae forecasting (Recknagel *et al.*, 1997; Velo-Suárez and Gutiérrez-Estrada, 2007; Talib *et al.*, 2008; González Vilas *et al.*, 2014; Park *et al.*, 2015; Harris and Graham, 2017; Hill *et al.*, 2020; Mellios, Moe and Laspidou, 2020; Aláez *et al.*, 2021; Izadi *et al.*, 2021; T. Kim *et al.*, 2022). Alongside these four non-linear machine learning models, a multiple linear regression model with L2 regularisation (ridge regression) was also used to provide a comparison against a more simplistic approach (Soranno, 1997; Onderka, 2007; Peretyatko *et al.*, 2010; Ahn, Oh and Park, 2011; Fornarelli *et al.*, 2013). Given that these approaches have different underlying statistical models, they can be considered broadly representative of the classical ML learning landscape. The MLP and GRU served as an initial test to determine the potential benefits of exploring more sophisticated architectures, such as long-short-term-memory (LSTM) neural networks. Advanced architectures, as noted by LeCun *et al.* (2015) and Sarker (2021a), demand extensive datasets for effective training. Therefore, the decision to expand our research to include more advanced models was contingent on the MLP and GRU's performance. Specifically, we looked to determine whether the MLP and GRU's effectiveness was constrained by the size of the available data. If these models demonstrated robust performance despite potential data limitations, this would justify further exploration into more sophisticated network architectures. Our stepwise approach ensures that the extension of research into advanced models is predicated on empirical evidence of their potential efficacy under constrained, real-world conditions.

Recurrent neural networks are designed to take advantage of data with inherent order, such as time series (Schmidt, 2019). Therefore, to take advantage of this, the GRU model was configured to use a longer series of inputs than the other models. An input series of four observations (data from the last two months) was chosen to balance providing sufficient data for the GRU to learn time series patterns whilst ensuring that the ratio of features to observations was not excessive.

Prior to training, input features to the GRU, MLP, SVM, and ridge regression models were scaled to improve convergence, and reduce bias towards larger magnitude variables. All data were standardized by removing the mean and scaling to unit variance using the following equation:

$$z = \frac{x - u}{s} , \quad (4)$$

where x is the sample of interest, u the mean of the training sample, and s the standard deviation of the training sample. RF models generally do not require scaling of the input features as decision thresholds for each feature are learnt independently of other features (Werther, Odermatt, Simis, Gurlin, Jorge, *et al.*, 2022; de Amorim, Cavalcanti and Cruz, 2023). As there was potential for models to predict negative values of chl-a, for example if the test data contained values outside the typical ranges of the training data, any negative values were set to a very small positive number (1×10^{-12} mg/m³). This is a reasonable theory-guided adjustment, as a negative value of chl-a is meaningless (Karpatne *et al.*, 2017).

Machine learning models were implemented in Python (v3.11) using the scikit learn (v1.4) and tensorflow (v2.15) packages (Pedregosa *et al.*, 2011; Martín Abadi *et al.*, 2015).

2.3.2.2 Hyperparameter Optimization

The selected ML models expose several hyperparameters that were optimized using a randomized parameter search with 5-fold cross validation (Table 2.2). Optimization was undertaken once for each model using 20 years of training data from 1987-2007, excluding 2001 due to the lack of data from that year. The data from 2008 onwards were set aside as test data and therefore not used for hyperparameter optimization. In the random search, the layer sizes for both the MLP and GRU were constrained to a maximum of 100 to avoid excessive over-parameterisation.

Table 2.2: Hyperparameters obtained from randomized cross validation search for the models used in this study. Parameter naming follows the scikit learn v1.3.2 convention (Pedregosa et al., 2011).

Random Forest		Support Vector Machine		Multilayer Perceptron		Ridge Regression		Gated Recurrent Neural Network	
<i>max depth</i>	84	<i>C</i>	43	<i>activation</i>	<i>relu</i>	<i>alpha</i>	117	<i>units</i>	4
<i>max features</i>	0.43	<i>epsilon</i>	0.21	<i>alpha</i>	1.76e-3			<i>activation</i>	<i>relu</i>
<i>min samples leaf</i>	3	<i>kernel</i>	<i>linear</i>	<i>Hidden layer sizes</i>	[16, 36]			<i>dropout</i>	7.4e-5
<i>min samples split</i>	7			<i>Learning rate</i>	<i>adaptive</i>			<i>recurrent activation</i>	<i>sigmoid</i>
<i>n estimators</i>	1845			<i>solver</i>	<i>sgd</i>			<i>batch size</i>	100
<i>oob score</i>	<i>True</i>			<i>Early stopping</i>	<i>True</i>				
				<i>validation fraction</i>	0.10				

2.3.2.3 Benchmark Models

Naive models are important for benchmarking more complex forecast models, as they provide a minimum baseline level of performance that should be achieved (McLaughlin, 1983; Hyndman and Athanasopoulos, 2021). For this study, two different benchmarks were used, chosen for their ubiquity in the forecasting literature, and previous use in algae and other water management forecasting studies (Page *et al.*, 2018; Thomas *et al.*, 2020; Hyndman and Athanasopoulos, 2021; Jackson-Blake *et al.*, 2022):

1. *Persistence Forecast* – chl-a persists over time and does not change since the last time it was measured
2. *Seasonal Naive Forecast* – chl-a measurements are linearly interpolated to daily values and the value from the same day and month, but previous year to the forecast date used as the prediction.

2.3.3 Statistical Analysis of Performance

For testing purposes, the dataset's most recent decade, spanning from 2008 to 2017, was selected. To assess how the duration of the training dataset impacts model performance, we conducted experiments varying the number of training years from 1 to 20 for each year under test. This setup was designed to mimic a real-world forecasting context, where training is performed using the most recent data available before the year being forecasted. For instance, in forecasting the year 2008, we developed 20 separate models using the RF algorithm. These models began with a minimal approach, utilising only 2007 as the single year of training data, and incrementally added more years, culminating in a model trained

on a 20-year span from 1987 to 2007 (excluding 2001). For each subsequent test year, models were retrained using the years directly prior to the test year. For example, forecast models for 2009 were at first trained only with data from 2008, then 2008 and 2007, and so on. Considering each number of training years separately, every test year was used as a fold in a non-randomized ten-fold cross validation, ensuring a systematic evaluation of model performance. To identify significant performance variations across the models, we applied a Friedman test (Demšar, 2006; Derrac *et al.*, 2011). Following the Friedman test, a post-hoc test using Holm's procedure for p -value adjustment was used to make multiple comparisons between the benchmark and machine learning models (Holm, 1979).

The criterion for determining the minimal requisite training data duration was established at the juncture where all model-specific p -values for a given performance metric fell below 0.1, with over half also under 0.05, indicating statistical significance. This procedure was repeated for two different performance metrics: the root mean square error (RMSE) and mean absolute percentage error (MAPE). These metrics were chosen for their prevalent use in forecasting tasks, and for the differences in their properties: RMSE penalises errors equally regardless of the magnitude of the true value, whereas MAPE scales errors according to the magnitude of the true value (Bowerman, O'Connell and Koehler, 2005; Chai and Draxler, 2014). Performance metrics were calculated excluding data from the winter months (December, January, and February). This adjustment accounts for the natural absence of algal blooms during these periods in Blelham Tarn, documented by Atkinson (1999). This prevents over-ranking of any conservative models that tend to predict minimal changes in chl-a, something which is pertinent in the context of algal bloom forecasting, a key application of this work.

All statistical analyses were carried out using R Statistical Software version 4.3.2 (R Core Team, 2021).

2.3.4 Importance of model features

Following the initial investigation, which used all available input features, the models were re-trained using the same train-test splits. However, this time, we systematically varied the number of input features. This approach allowed us to assess the impact of feature selection on model performance explicitly, providing insights into the optimal set of features for effective predictions (Breiman, 2001). To avoid the need for a fully exhaustive evaluation of all possible feature combinations, the relative importance of the input features was first established using the permutation feature importance. This method

evaluates a trained model's dependency on specific features by randomly shuffling the values of each feature in isolation and observing the resultant decrease in model performance. To capture some of the variability in feature importance as the number of training years varied, feature importances from two scenarios were evaluated – “low-data availability” with five years training data, and “high-data availability” with fifteen years training data. For both scenarios, the permutation importance of all ML models was evaluated for the ten test years 2008-2017, using the years directly prior to these for training. Given that we were most interested in the importance of specific variables rather than the associated measurement lag of two or four weeks, the permutation importances were averaged for each variable to give a hierarchy of 13 permutation importances. This hierarchy was subsequently used to assess scenarios in which only the top n most critical variables were accessible for analysis. Therefore, the permutation importances for all models were averaged, and an overall order of feature importance was established by ranking the importances from the averages of the two scenarios. The RMSE metric was used for this ranking as it is more stable than MAPE for evaluating performance where the target is close to zero (Makridakis, 1993). This characteristic of RMSE minimises the likelihood of generating misleading feature importances.

Subsequently, the training and testing process, which varied the training duration from 1 to 20 years for the test period spanning 2008 to 2017, was applied to scenarios where the ML models had access to only the n most critical variables. Here, n was adjusted from one, representing a single variable, up to the full set of thirteen variables.

Additionally, scatter plots of the input variables were examined to assess the nature of their pairwise relationships. Where apparent relationships existed, these appeared broadly linear, supporting the use of Pearson correlation as a measure of association. Following this, the Pearson correlation between all input variables was calculated to identify any strongly correlated features which might be redundant. Then, pairs of highly correlated features (correlation coefficient > 0.7) were selected for further testing. To check these for redundancy the train-test routine was run first with all features excluding one of the pair, and then ran again with the exclusion of the other.

2.3.5 Sampling frequency

Adequate machine learning performance is contingent on there being a sufficient number of training observations available. Therefore, the frequency at which data are collected can be expected to influence how long data would need to be collected for before the desired

performance is reached. Additionally, sampling frequency dictates the minimum forecast horizon (number of days ahead) that a model can be trained to predict over, which is expected to have a strong effect on performance. For these reasons, the effect of reducing the sampling frequency from fortnightly to monthly (28 days) was investigated. Four different scenarios were compared. Firstly, the standard train test procedure was used as a baseline (fortnightly observations used to make predictions two weeks into the future). Then to isolate the effect of increasing the forecast horizon, the same fortnightly data was used to make predictions 28 days into the future. Following this, the data was down sampled to monthly (28 day) observations by removing every other observation. This monthly data was used to make forecasts 28 days into the future. Finally, 14 day ahead forecasts were trained using input data sampled at 28 day intervals. This isolates the effect of reducing the sampling frequency without changing the forecast horizon or number of observations available in a given training year. For each of these sampling scenarios, the number of training years made available to the models was varied from 1-20 years, and performance tested for each of the 10 training years (2008-2017).

2.4 Results

2.4.1 Historical patterns and non-stationarity

There have been some shifts in the timing and behaviour of chl-a blooms. In the two most recent decades, the summer chl-a bloom generally starts in July, but in the earliest decade (1987-1996) this occurs approximately one month later. Furthermore, this earliest decade shows a more distinct “clearwater” phase, where the chl-a levels decrease between the spring and summer blooms (Figure 2.1a). Additionally, nutrient levels (nitrate, total phosphorus, ammonium) in the most recent decade are generally lower than in the two previous decades (Figure 2.1b, c, d).

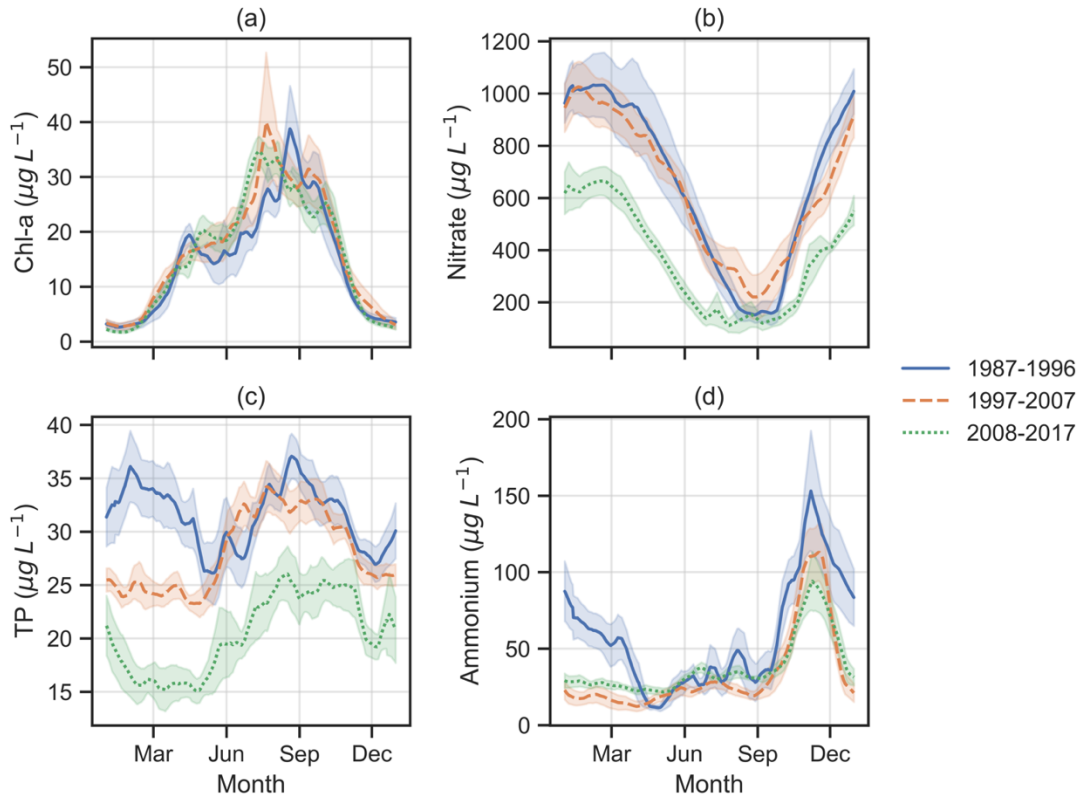


Figure 2.1: Decadal means of (a) Chl-a, (b) nitrate, (c) total phosphorus, and (d) ammonium obtained by interpolating the time-series to daily values and averaging these over three ten-year periods: 1987-1996, 1997-2007 (excluding 2001), 2008-2017. The shaded margins indicate standard error.

2.4.2 Effect of training data on model performance

With less than ten years training data, the RF and ridge regression models outperformed all other models. With just five years training data the RF had an RMSE of $7.9 \mu\text{g L}^{-1}$, and MAPE of 39.4 %. This is $1.0 \mu\text{g L}^{-1}$ (11.4 %) and 16.4 % lower than the respective values for the better performing of the two benchmarks, the persistence forecast. With sufficient training data, all ML models achieved similar performance, but the ridge regression achieved the lowest RMSE of $7.6 \mu\text{g L}^{-1}$ (which represents a 15.5 % improvement over the better of the benchmarking estimates) with fourteen years of training data, and the SVM achieved the lowest MAPE of 36.8 % (which represents a 22.0 % improvement over the better of the benchmarking estimates) with nineteen years training data. The MLP was generally not competitive with the best performing models.

The bulk of the improvement in the performance of the most competitive models was obtained with the first five years of training data, and whilst the addition of more years of training data did further improve performance, this was generally with diminishing returns (Figure 2.2).

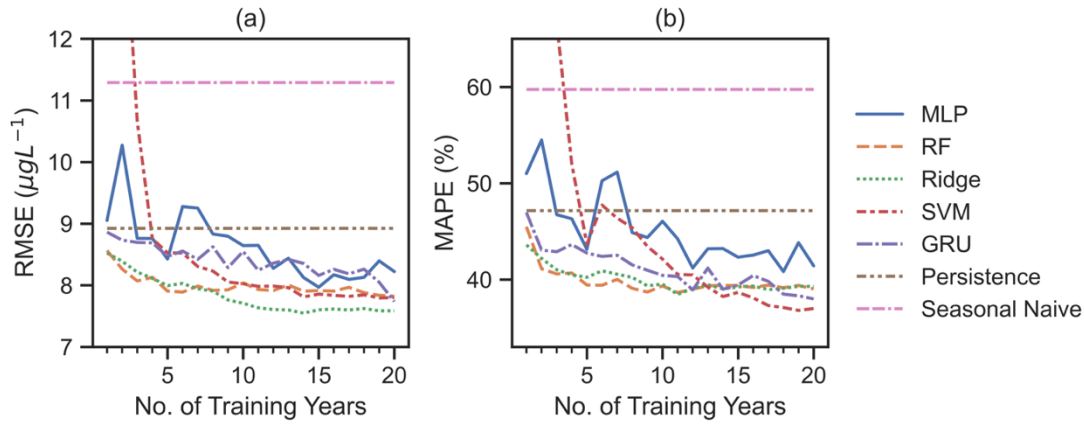


Figure 2.2: Performance metrics, (a) RMSE, and (b) MAPE plotted against number of training years. Lines indicate mean values. Blue solid line is MLP, orange dashed line is RF, green dotted line is ridge regression, red dash-dot line is SVM, purple horizontal dash-dot line is persistence forecast, and brown horizontal dash-dot line is seasonal naive forecast.

For a given number of training years, the Holm adjusted p -values from the post-hoc tests indicate if the differences in performance between the machine learning models and benchmarks are statistically significant. It was found that the RF and ridge regression models had much lower requirements in terms of the number of training years for statistical significance than the other models (Table 2.3, Figure 2.3b, Figure 2.3c). The MLP never consistently outperformed the benchmarks in a statistically significant way.

Table 2.3: Minimum number of years training data needed to outperform the persistence forecast in a consistent and statistically significant way, defined as the number of years training data beyond which all p -values were < 0.1 , and over half of these < 0.05 .

Minimum number of years for consistent performance difference		
Model	RMSE	MAPE
MLP	-	-
RF	4	3
Ridge	4	3
SVM	10	12
GRU	20	8

With a low number of training years (< 3), SVM performance is close to being statistically significantly worse than the persistence forecast, which is why the p -values are initially low (Figure 2.3d). However, all other models achieved similar performance to the persistence forecast even with only one year of training data (Figure 2.2, Figure 2.3).

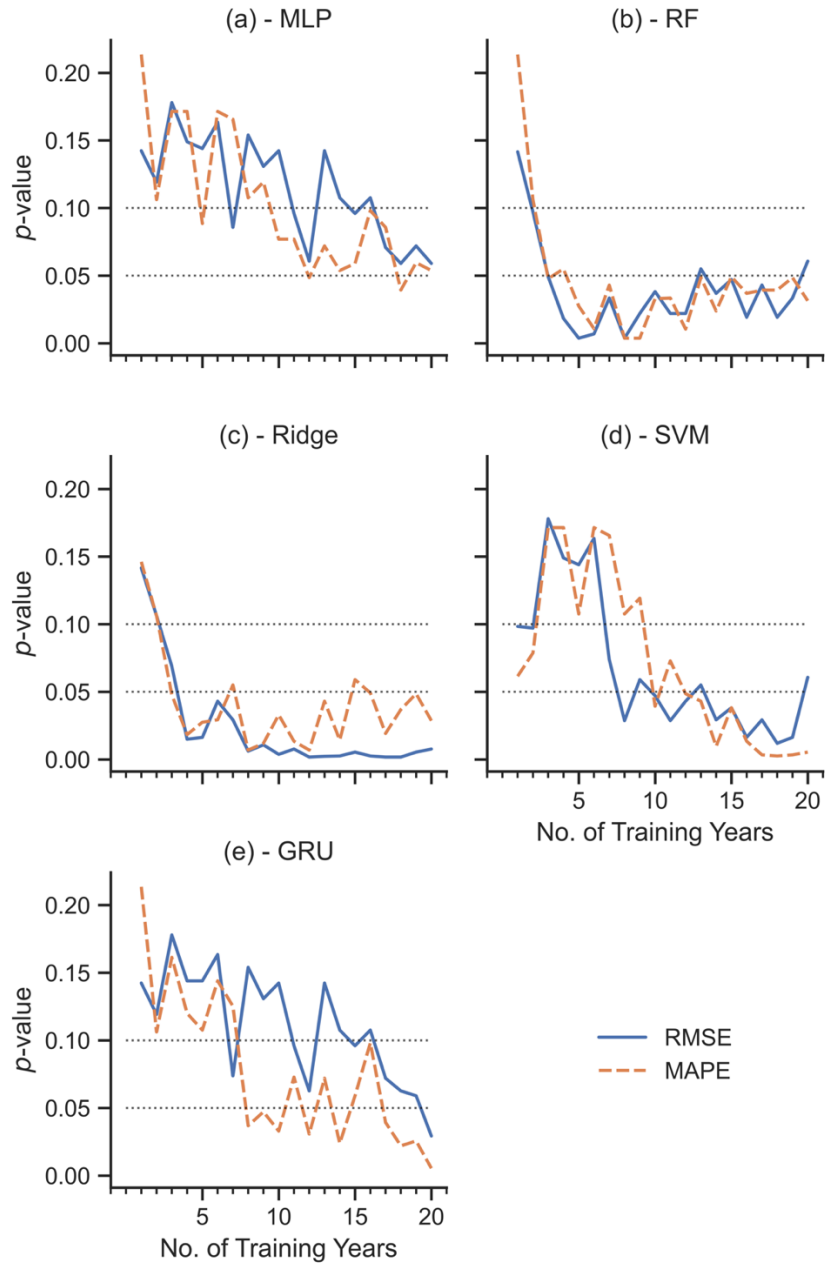


Figure 2.3: Performance comparisons of persistence forecast against (a) MLP, (b) RF, (c) ridge regression, and (d) SVM. All plots show p -values from Holm multiple comparisons procedure for both RMSE (solid blue lines) and MAPE (dashed orange lines). Dotted horizontal lines indicate 0.05 and 0.1 significance levels.

The seasonal naive forecast had the poorest performance overall, and accordingly the number of years training data needed for statistically significant differences was lower for all models.

2.4.3 Feature Importance

In both the five- and fifteen-year training scenarios, chl-a was the most important feature, followed by surface water temperature and air temperature (Figure 2.4). Although there was some variation in the order of feature importance between the two scenarios, they follow very similar patterns. Notably, the most obvious difference is that the average

change in RMSE from shuffling chl-a was $0.6 \mu\text{g L}^{-1}$ higher than the next most important feature, water temperature, in the higher data availability scenario. Here, "shuffling" refers to the random reordering of a feature's values, which breaks its relationship with the target variable while leaving all other inputs unchanged. Comparing this with the lower data availability scenario, where the difference was $0.3 \mu\text{g L}^{-1}$, suggests that on average, given more training data, models were less reliant on lower ranking features.

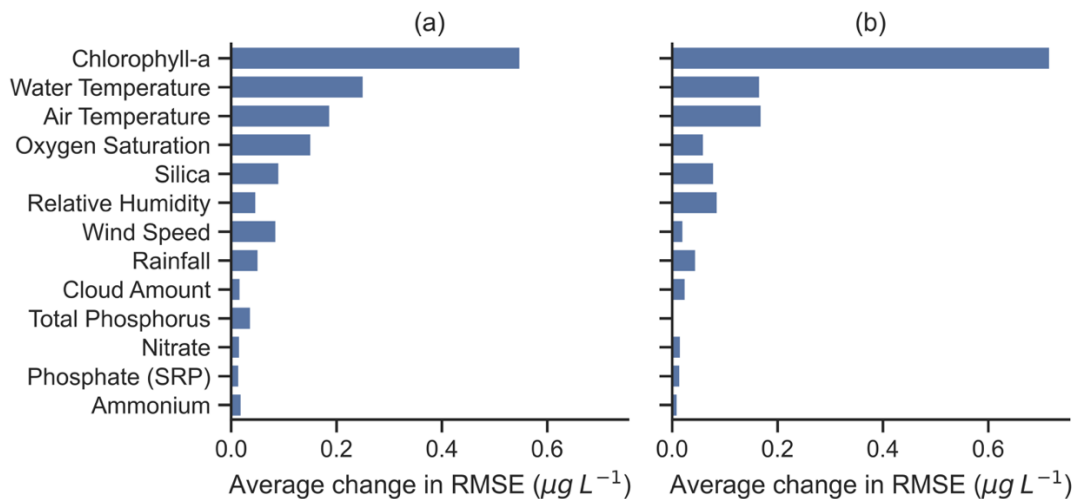


Figure 2.4: Overall average permutation feature importances (evaluated through a performance change in RMSE) with (a) five years, and (b) fifteen years training data.

Models trained exclusively with the four most significant parameters or fewer exhibited reduced performance, despite being trained over the maximum duration of years available. Generally, using a larger number of input features had a positive effect on the performance of most models (Figure 2.5), but there was little benefit to including more than six features (Figure 2.6). For example, all ridge regression models trained with more than four input features had very similar RMSE curves, indicating that for this model, the addition of the least important features (total phosphorus, ammonium, SRP, cloud amount, nitrate) had little benefit to performance (Figure 2.5e). The SVM models trained with just the six most important features reached lower RMSE values with fewer training years than the same model trained with the maximum number of features (Figure 2.5g, Figure 2.6). The benefit of including the five least important features is more apparent in the MAPE curves but is still minimal (Figure 2.5b, d, f, h).

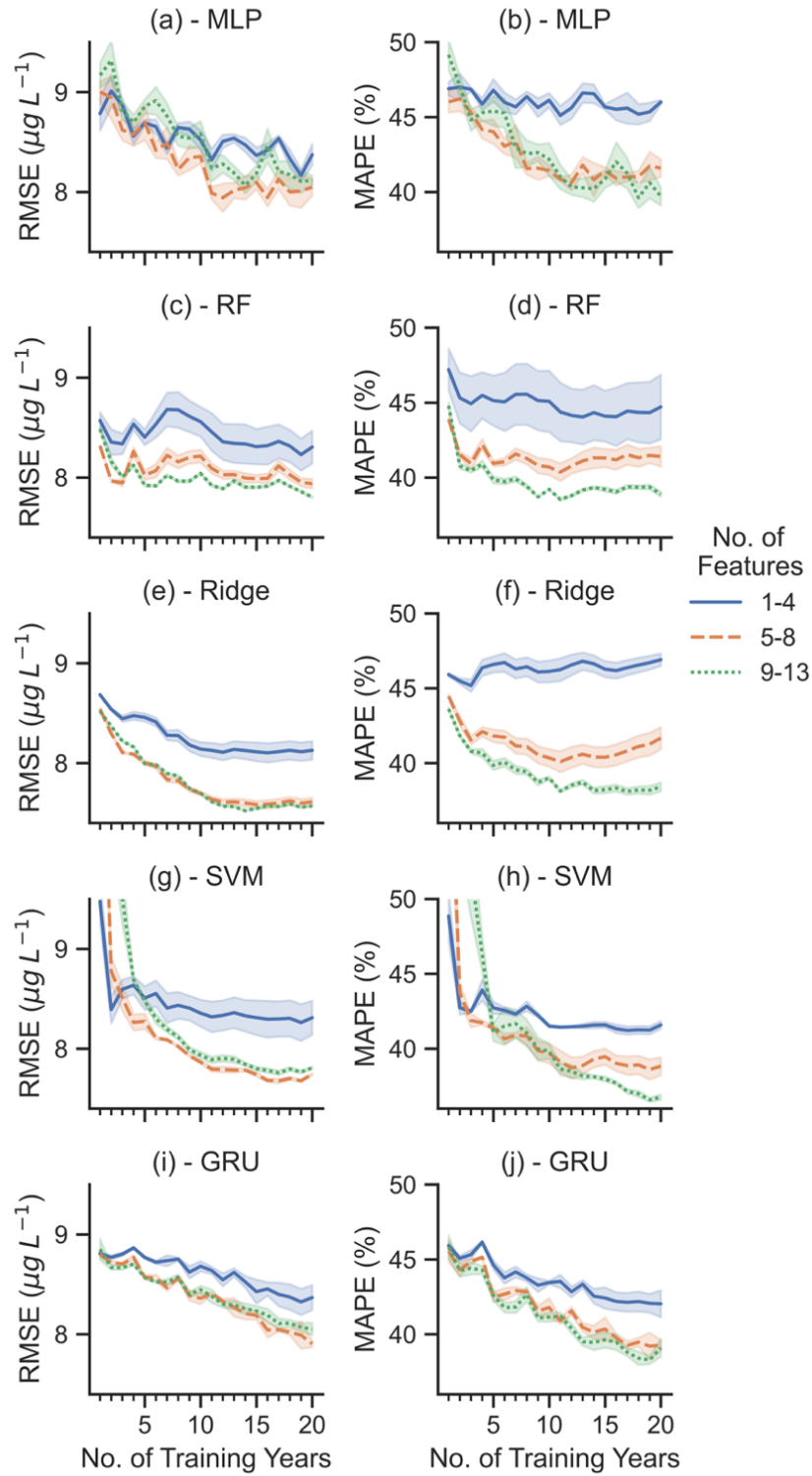


Figure 2.5: RMSE (left column) and MAPE (right column) plotted against the number of training years with the number of input features varied for MLP: (a), (b); RF: (c), (d); ridge regression: (e), (f); SVM: (g), (h); and GRU: (i), (j). Blue lines indicate the mean performance of models trained with the 1-4 most important features (i.e. 1: only chl-a; 2: chl-a and water-temperature; 3: chl-a, water temperature and air temperature; 4: chl-a, water temperature, air temperature and oxygen saturation). Orange dashed lines indicate models trained with the 5-8 most important features, and green dotted lines indicate models trained with the 9-13 most important features. Shaded margins indicate standard error around the mean centre line.

The only two input features with a high correlation (> 0.7) were air temperature and surface water temperature, which had a correlation coefficient of 0.93. For the best

performing model with five years of training data, the RF, the removal of air temperature reduced the RMSE by $0.05 \mu\text{g L}^{-1}$ (0.60 %) and increased the MAPE by 0.40 %. Similarly, the removal of surface water temperature increased the RMSE by $0.01 \mu\text{g L}^{-1}$ (0.14 %) and increased the MAPE by 0.33 %. For the best performing model with 20 years training data, the SVM, the removal of air temperature increased the RMSE by $0.05 \mu\text{g L}^{-1}$ (0.69 %) and increased the MAPE by 1.2 %. Likewise, the removal of surface water temperature increased the RMSE by $0.02 \mu\text{g L}^{-1}$ (0.2 %) and increased the MAPE by 0.28 %. Therefore, given the low performance changes observed, there is a degree of redundancy in these two variables.

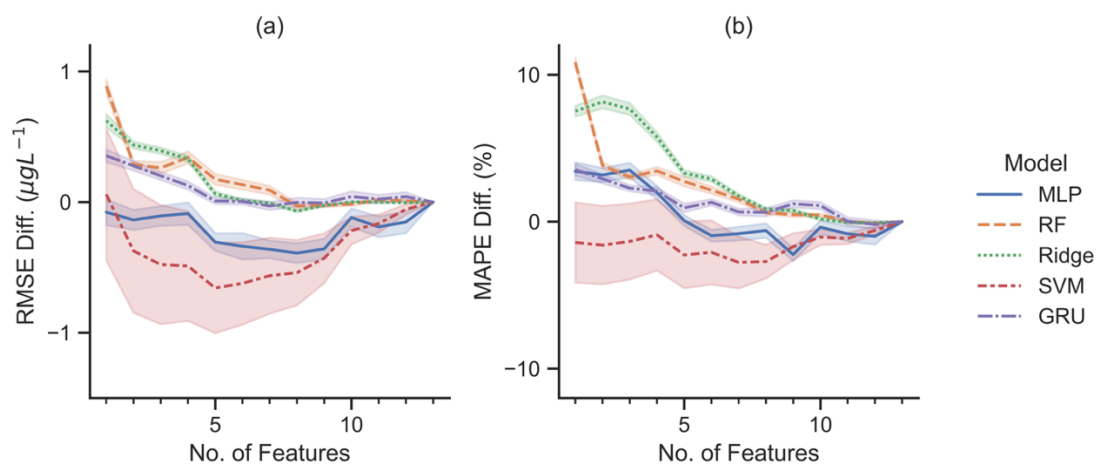


Figure 2.6: Difference in (a) RMSE, and (b) MAPE between models trained with a reduced feature set and models trained with all 13 features. Positive differences indicate that the reduced feature set performance is worse than the full feature set. Differences were averaged over all the different data availability scenarios (number of training years from 1-20), with shaded margins showing the standard error around the average. The blue solid line is for the MLP; orange dashed line for the RF; green dotted line for the ridge regression; red dash-dot line for the SVM; and purple dash-dot line for the GRU.

2.4.4 Sampling Frequency

For all models, increasing the forecast horizon from 14 to 28 days incurred a larger performance reduction than reducing the sampling frequency alone (Figure 2.7). The performance of 14 day ahead forecasts made using data sampled at 28 day intervals were largely very similar to those made with a 14 day sampling period. In general, 28 day ahead forecasts with 14 day samples had lower RMSE than those made with 28 day samples. However, there were only two models (SVM and GRU) for which the MAPE of 28 day ahead forecasts made with 14 day samples was generally lower than those made with 28 day samples (Figure 2.7 h, j). For the SVM model, the performance of 28 day ahead forecasts made with more than 15 years' training data were approximately equivalent in performance regardless of the sampling frequency (Figure 2.7 g, h). In contrast, with less

than 10 years' training data, 28 day ahead SVM forecasts made with 14 day samples performed better than those made with 28 day samples (Figure 2.7 g, h).

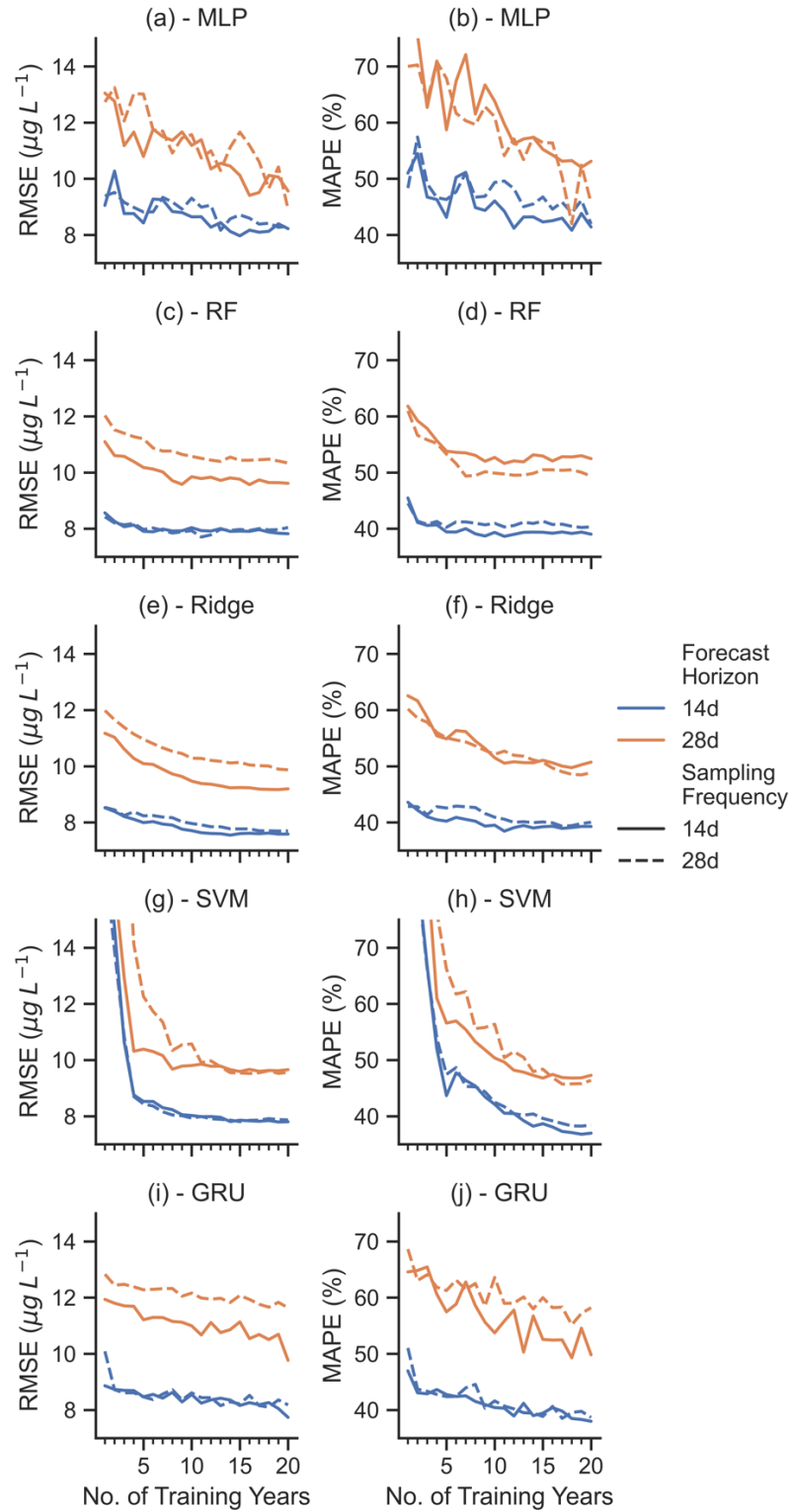


Figure 2.7: RMSE (left column) and MAPE (right column) plotted against the number of training years for two different forecast horizons of 14 (blue lines) and 28 (orange lines) days. Sampling frequencies of 14 days are indicated by solid lines, and 28 days dashed lines. Results are shown for all five ML models: MLP: (a), (b); RF: (c), (d); ridge regression: (e), (f); SVM: (g), (h); and GRU: (i), (j).

2.5 Discussion

2.5.1 How many years of data are needed?

With the full input parameter set and four or more years' training data, we found that the RF and ridge models were able to consistently outperform both benchmarks in terms of RMSE and MAPE. This suggests that ML can be effective for algae forecasting even when the number of years' training data available are limited, with the obvious caveat that some models may be more suitable than others. Furthermore, the RF, GRU and ridge regression models always had mean RMSE and MAPE lower than both benchmarks. Therefore, it can be argued that certain ML approaches would be effective, or at least no-less effective than naive models, even with only one year of training data. However, the best RMSE (ridge regression with fourteen years training data) was only 4.6 % lower than the RF with five years training data, and the best MAPE (SVM with nineteen years training data) was only 2.6 % lower than the RF with five years training data. Ultimately, we have shown that for this study site, only a low number of training years are required for ML to be useful for forecasting, but that there is not very much to be gained from the inclusion of additional years' training data. For the management of fortnightly sampled lakes which are ecologically-similar to Blelham Tarn, we would therefore encourage managers to implement ML forecasts even when the length of the available data is limited, and to continue data collection for at least five years before making judgements about the utility of the data for forecasting. For lakes with a significantly different sampling approach or ecology from Blelham Tarn, we would encourage repetition of the workflow that has been presented. In short, this would involve examining the forecast performance of multiple models over several test years whilst varying the number of training years.

In a review of forecasting models for cyanobacterial blooms in freshwater lakes, Rousso et al. (2020) found that of the reviewed 90 data-driven studies using similar input data to the present study (i.e. algal concentrations estimated through microscopy or pigment analysis), 90.0 % used at least one year of monitoring data to develop data-driven prediction models. Of the same group of studies, 55.6% used more than four years, 26.7 % used more than 10 years, and 8.9 % used 15 or more years of data. Importantly, this distribution is likely to reflect the general availability of long-term monitoring data: datasets of less than 10 years are much more common than those with longer time series than this. Therefore, placed alongside our findings, this suggests that numerous existing datasets are suitable for ML forecasts of algal concentrations, if a similar level of

predictability and sampling frequency to Blelham Tarn is assumed. It is likely that some systems, for example those which do not experience algal blooms every year, would require more training years' training data than this (Shyalika, Wickramarachchi and Sheth, 2023). However, Blelham Tarn can be considered a typical example of a fertile temperate lake. It consistently has spring diatom blooms succeeded by cyanobacteria in the summer, and therefore our findings are likely widely applicable to the many lakes that share similar seasonal behaviour (e.g. Bailey-Watts, 1981; Rose et al., 2021; Wei et al., 2020).

Furthermore, for monitoring routines where automated high frequency sensors are not used, fortnightly or monthly lake sampling is common (Dubelaar, Geerders and Jonker, 2004; Marcé *et al.*, 2016; Spaulding *et al.*, 2024), and so there are likely many existing datasets for which these findings are applicable.

Rousso et al. (2020) also found that almost half (44.4 %) of the reviewed studies with similar input data to the present study used less than five years data. This could be an indication that these studies were able to achieve good performance with fewer training years than in the present study, or it might indicate that some of these models were not performing significantly better than naive benchmarks would have. Whilst benchmarking against naive models is not a new concept in the wider field of forecasting (McLaughlin, 1983), the authors of the present study are aware of only a handful of studies which use naive benchmarks for algae forecasts (Page *et al.*, 2018; Jackson-Blake *et al.*, 2022; Matthews, 2023). Our results demonstrate that some relatively sophisticated forecasts, such as the MLP, may not be guaranteed to be significantly better than a naive model, unless provided with much more training data than is available for most lakes. Therefore, we strongly suggest the use of naive benchmarking for algae forecasting.

Many studies have found that ML performance improves with diminishing returns as the training dataset size increases (Frey and Fisher, 1999; Last, 2007; Mahmood *et al.*, 2022). In the present study, this was likely caused by several factors. Changes in the sampling approach and improvements to measurement methods over the duration of the data collection period may affect the utility of older training data. For example, there have been organizational changes, sampling frequency variations, and upgrades to the lab equipment used for analysis of samples (Maberly et al., 2017). Alongside this, nutrient loads and associated phytoplankton dynamics in Blelham Tarn have changed over the 30-year study period (Figure 2.1). These have been accompanied by many other changes including an earlier onset and longer duration of stratification; increased mean surface water temperature; and increased occurrence of hypolimnetic anoxia (Foley et al., 2012).

Evidently this is a non-stationary system, which likely means that older training years are less representative of the conditions being modelled in a given test year. Globally, lakes are experiencing rapid warming and changes to mixing patterns due to anthropogenic climate change (O'Reilly *et al.*, 2015; Woolway and Merchant, 2019; Maberly *et al.*, 2020). This is likely to have a profound effect on phytoplankton species composition and biomass, and is expected to lead to increased cyanobacterial dominance in many water bodies (Carey *et al.*, 2012; Huisman *et al.*, 2018). Therefore, if warming rates further increase, it may be that historic data becomes increasingly less relevant to modelling short-term fluctuations. Investigating this explicitly would be of great value in understanding how the performance of near-term ML algae forecasts could be affected by a rapidly changing climate.

The use of ML approaches does improve our ability for modelling forecasts over benchmarking forecasts, but the observational sampling approach is likely to have a limiting effect on performance. During the summer growing season, phytoplankton dynamics are complex and influenced by many factors, including grazing, effects from higher trophic levels, and physical factors such as flushing (Sommer *et al.*, 2012; Stockwell *et al.*, 2020). Whilst some of these drivers may be correlated with the input features, we cannot expect a monitoring program to fully capture all the complexities of a dynamic ecosystem. Additionally, the horizontal movement of algae, particularly cyanobacteria, will also have some influence on the observed chl-a, which is only measured at a single location at Blelham Tarn (Maberly *et al.*, 2017; Feuchtmayr, M.A Clarke, De Ville, *et al.*, 2021; Xue *et al.*, 2023). Therefore, given this complexity, and potential for rapid changes, certain aspects of the summer bloom dynamics will not be sufficiently captured by the sampling efforts (Pobel, Robin and Humbert, 2011). In other words, limitations of the monitoring approach contribute uncertainty to the forecasts, regardless of the length of data used to train the model.

2.5.2 Model Selection

Appropriate model selection is critical, particularly in data-limited scenarios. With the full input parameter set, the GRU, MLP and SVM both required more than three times the number of years training data required by the RF to consistently outperform the benchmarks in a statistically significant way (Figure 2.2, Table 2.3). The model dependence of this result likely has several reasons: firstly, the MLP and GRU are higher complexity models than the RF and ridge regression models as they have a larger number of parameters (weights) to learn. This may be the reason that several studies which used

similar datasets also found decision tree models to perform better than neural networks at algae forecasting (Fornarelli *et al.*, 2013; Harris and Graham, 2017; Aláez *et al.*, 2021). The result that the more complex models performed poorly in data-limited scenarios is not surprising but highlights the importance of using naive forecasts for benchmarking, as some ML models may not be able to outperform these. Secondly, as the hyperparameters were tuned using 5-fold cross validation with twenty years training data, there was some inherent bias towards larger numbers of training years, which may have resulted in some overfitting in scenarios where a lower number of training years were available. This is a plausible explanation for the very poor performance of the SVM models trained with the full feature set and a low number of training years. This issue was alleviated by removing some of the less important input features which indicates that it was at least partly a result of overfitting to spurious correlations between chl-a and the less important input features. Often, it is argued that more complex ML models have the advantage of being able to effectively model non-linear dynamics (Cruz *et al.*, 2021), however our findings suggest that multiple linear regression with regularisation may be all that is required to model algal concentrations using fortnightly data. This finding also suggests that more complex ML architectures which would be expected to outperform linear regression given sufficient training data are likely not appropriate unless datasets are available for several decades or at higher than fortnightly resolution.

2.5.3 Importance of input features

As well as only a few years' training data being sufficient to optimise model performance, only a handful of lake measurements were necessary. We found that excluding the seven least important predictor variables generally gave similar performance to using the full parameter set (Figure 2.5). Furthermore, the SVM model's performance was actually improved by removing these parameters (Figure 2.6). Several of the more important parameters were meteorological variables taken from weather stations up to 37 km from the lake, data that are commonly available from regular meteorological stations or weather forecasts (Figure 2.4). Furthermore, it was found that the air and surface water temperature variables were highly correlated, and that therefore either one of these could be excluded with a minimal performance impact. Therefore, for fortnightly sampled lakes ecologically similar to Blelham Tarn, only a relatively small amount of in-lake monitoring would be necessary to provide a basis for using ML approaches. Nevertheless, unlike Muttill and Chao (2006) and Xiao *et al.* (2017), we found that using chl-a alone was not effective, despite this variable standing out as the most important predictor. This difference in

outcome is likely due to differences in sampling frequency and study site: Xiao *et al.* (2017) used daily monitoring data, and Muttill and Chao (2006) used data from a large estuarine harbour.

Several of the features which we found to be least important to ML forecasts are those often considered to be strongly linked with algal blooms: ammonium, SRP, and nitrate (O'Neil *et al.*, 2012; Paerl and Otten, 2013). This indicates that for Blelham Tarn, the fortnightly fluctuations of these variables are not useful for predicting chl-a. This is likely because nutrient concentrations are often at limiting values during the summer, where most of the large chl-a fluctuations occur (Figure 2.1). Page *et al.* (2018) identified that difficulties in modelling SRP fluxes in Esthwaite Water was a significant hindrance to chl-a forecast performance, and suggested that monitoring nutrients more regularly than fortnightly could ameliorate this issue. Given the geographic proximity, and similarity in trophic status between Blelham Tarn and Esthwaite Water, this could likely explain why we did not find fortnightly nutrient measurements to be particularly useful for forecasting.

Whilst there are many other studies which found that water temperature and other physical parameters were generally stronger predictors of algal concentrations than nutrient measurements (Wang *et al.*, 2019; Roussio *et al.*, 2020; Li *et al.*, 2021), there are also many studies which identify nutrients as key predictors. For example, Roussio *et al.* (2020) found that 30.5 % of the reviewed cyanobacterial bloom forecast studies identified either phosphorus or nitrogen as the most important predictor in their models. This indicates that the optimal set of features for forecasts is likely to be highly variable across different study sites and monitoring schemes. Where possible, monitoring a broad suite of water quality parameters is advantageous because this will likely cover statutory monitoring requirements; provide data useful for other scientific and management investigations; and importantly, provide a degree of flexibility in forecast models should the system change in such a way that the feature importance hierarchy is modified. For example, forecasts of cyanobacterial blooms in a temperate lake might initially be highly reliant on water temperature measurements. However, if over a number of years, the lake was to warm to the point where summer temperatures were consistently optimal for cyanobacteria growth, then nutrient availability could become a more important predictor than temperature (Reynolds, 2006; Paerl and Huisman, 2009; Carey *et al.*, 2012; Richardson *et al.*, 2018; Bonilla *et al.*, 2023). Critically, this highlights that whilst automated high frequency monitoring platforms are increasingly being adopted for water quality monitoring applications, the ability of manual sampling approaches to measure a wide range of

parameters in a consistent and reliable way is a compelling argument for their continuation (Marcé *et al.*, 2016; Park, Kim and Lee, 2020).

2.5.4 Sampling frequency

It has been previously demonstrated that increasing sampling frequency has a positive impact on the performance of phytoplankton forecasts (Derot, Yajima and Schmitt, 2020; Lin, Pierson and Mesman, 2023), and our results corroborate this. Specifically, we found that the strongest way in which sampling frequency affected performance was through limiting the minimum forecast horizon possible. However, with this effect removed, comparisons with 28 day ahead forecasts still showed that higher frequency data generally gave lower errors and, for the SVM, converged to peak performance with fewer years' training data. That this effect was either not present, or much more subtle, for the comparisons made with 14 day ahead forecasts is a reflection of the number of observations available for training in each of these scenarios. Both 14 day ahead forecast scenarios were trained with the same number of observations for a given number of training years, whereas the 28 day ahead forecasts made with 14 day samples had approximately twice as many training observations per training year than those made with 28 day samples. Therefore, alongside making considerations concerning the forecast horizon, managers designing or assessing sampling routines for forecasting should understand that the total number of observations collected annually may affect how many years' training data are required to achieve adequate performance.

2.6 Conclusion

Whilst this study was only carried out for a single lake, there are many lakes and reservoirs that suffer algal blooms that are monitored at a similar frequency to Blelham Tarn. We have shown that with four or more years' data consisting of just a few parameters sampled on a fortnightly basis, ML can outperform standard naive benchmarks, and is therefore appropriate for generating two week ahead algae forecasts. For those already working with more than ten years' fortnightly data for lakes ecologically-similar to Blelham Tarn, our results suggest that with appropriate model selection, tuning, and feature selection, the performance reached at this point is likely very close to the maximal performance that can be expected, even if more years of training data are added as the sampling continues. Ultimately, our results suggest that there are likely numerous lakes and reservoirs with sufficient existing data for ML forecasting, and that users interested in this approach may see significant improvements in forecast ability just

a few years after the initiation of a simple monitoring programme. However, it is still essential to better understand data requirements for ML algae forecasts across a diverse range of sampling approaches and lake ecologies. By adopting a similar workflow to that which has been presented in this study, managers and scientists can continue to advance our understanding of these requirements and therefore refine future algae forecasting efforts.

3 How climate change erodes short-term lake-temperature predictability: informing climate resilient lake forecasting

This chapter was published as:

Atton Beckmann, D., Werther, M., Shatwell, T., Spyrakos, E., Hunter, P., Jones, I.D., 2026. How climate change erodes short-term lake-temperature predictability: Informing climate resilient lake forecasting. Water Research X 30, 100457.

<https://doi.org/10.1016/j.wroa.2025.100457>

CRedit authorship contribution statement:

D. Atton Beckmann: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **M. Werther:** Writing – review & editing. **T. Shatwell:** Writing – review & editing. **E. Spyrakos:** Writing – review & editing, Supervision. **P. Hunter:** Writing – review & editing, Supervision. **I.D. Jones:** Writing – review & editing, Supervision, Conceptualization.

3.1 Abstract

Climate warming threatens short-term environmental forecast skill, yet its effect on water quality predictability is largely unquantified. Here, we demonstrate a new approach for assessing climate change effects on lake forecasts. Random forest (RF) and gated recurrent unit network models were trained on data from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) Local Lakes Sector (five central-European lakes, four hydrodynamic models) and then used to forecast daily lake surface temperature 14 days ahead for 2060 - 2100 under four climate scenarios. We then varied (i) sensor-sampling interval (3, 7, 14 days) and (ii) training-set length (1 - 30 years). Under the strongest forcing (SSP585), the summer mean absolute error (MAE) of worst-affected lake, Esthwaite, rose by 0.14 °C (from 1.75 to 1.89 °C), driven by higher day-to-day temperature volatility ($R^2 = 0.78$). For this lake, extending the training set from 5 to 20 years or shortening sampling from 14 to 3 days reduced summer MAE by 0.11 and 0.17 °C, effectively offsetting the volatility caused by climate change. In winter, forecast error declined for four lakes because warmer, more stratified conditions simplified surface-layer dynamics. Thus, modest investments in monitoring cadence or historical record length can preserve forecast skill, even under extreme climate change. More broadly, this highlights a largely unexplored potential use for climate scenario projections: informing the design of climate resilient lake monitoring systems.

3.2 Introduction

Climate change is altering freshwater ecosystems globally (Capon, Stewart-Koster and Bunn, 2021). Rising lake surface temperatures affect metabolic rates, species survival, and biogeography (Dokulil *et al.*, 2021), while shifts in stratification and mixing influence light, carbon, and oxygen dynamics (Woolway *et al.*, 2021). Increased frequency and severity of droughts and extreme precipitation events are also projected to increase (Dai, 2013; Tabari, 2020), heightening the risk of harmful algal blooms in many lakes (Paerl *et al.*, 2020). Simultaneously, water quality monitoring is advancing, with automated sensors complementing manual sampling and generating high-frequency, high-volume datasets (Marcé *et al.*, 2016). Machine learning (ML) is increasingly applied to manage this complexity - both for real-time monitoring and short-term forecasting of variables such as water temperature (Yousefi and Toffolon, 2022; Di Nunno *et al.*, 2023), dissolved oxygen (Ziyad Sami *et al.*, 2022), and algal abundance (Schaeffer *et al.*, 2024). By generating daily to weekly forecasts, ML can support real-time management, allowing for timely

execution of interventions such as water supply switching, aeration system activation, or issue of public health advisories (Stroom and Kardinaal, 2016).

As water quality forecasting capabilities increase, it is likely that some water managers will become increasingly reliant on such approaches to ensure consistent and safe provision of many of the ecosystem services provided by freshwater bodies, such as water supply, recreation, food, and carbon sequestration (Inácio *et al.*, 2022; Carey *et al.*, 2025). However, in the context of a rapidly changing climate, it is possible that short-term water quality forecasting will become more inaccurate for several reasons. Firstly, under climate change, water bodies should be considered non-stationary systems, and will therefore become inherently less predictable than they would be given a more stable climate (Milly *et al.*, 2008; Rollinson *et al.*, 2021). Secondly, data-driven forecasting approaches such as ML rely on representative training datasets, typically historical monitoring data, to make future forecasts. With a sufficiently large training dataset, performance improvements tend to show diminishing returns as the length of the training data is increased (Last, 2007; Atton Beckmann *et al.*, 2025). Importantly, this effect may be stronger for non-stationary problems as older training data may be less representative of the present system, and therefore less useful for informing predictions, particularly if the system is changing rapidly (Valavi *et al.*, 2022). Lastly, the forecast uncertainty associated with certain meteorological variables such as precipitation is expected to increase and this might in turn affect water quality forecast ability (Xu *et al.*, 2020).

Lake surface water temperature is a fundamental variable in aquatic systems because it strongly influences numerous physical, chemical, and biological processes. These include dissolved oxygen dynamics (Hutchings *et al.*, 2024), algal bloom development (Glibert and Burkholder, 2018), fish habitat suitability (Fang and Stefan, 2012), and the spread of pathogens (van der Wielen *et al.*, 2023). Consequently, accurate forecasting of lake surface water temperature is central to effective lake and water resource management, particularly under changing climatic conditions (Piccolroaz *et al.*, 2024).

In recent years, there has been growing interest in applying ML techniques to model and forecast lake surface water temperature across a range of spatial and temporal scales. Many studies have focused on short-term prediction or *nowcasting*, often comparing the performance of different ML algorithms (Liu and Chen, 2012; Heddam, Ptak and Zhu, 2020; Quan *et al.*, 2022; Yousefi and Toffolon, 2022). Others have developed hybrid approaches that integrate ML with process-based models to improve temperature simulations (Read *et al.*, 2019; Jia *et al.*, 2021). Additional work has explored ML-based

approaches for both short- and long-term forecasting (Saber, James and Hayes, 2020; Zhu *et al.*, 2020). However, the vast majority of these studies have assessed model performance only under present-day or historical conditions. A few have considered limited cross-condition testing - for instance, Read *et al.* (2019) examined models trained on relatively cold years and tested on warmer ones - but systematic evaluation of ML model performance under projected future climate scenarios remains largely unexplored.

This study addresses this gap by leveraging lake-specific surface water temperature projections available through the Lakes Sector of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; Golub *et al.*, 2022) to investigate how climate change may affect ML short-term forecasting skill. Understanding this relationship is crucial, as some variables and lakes are likely to become increasingly difficult to forecast under future climatic conditions (Milly *et al.*, 2008). Beyond quantifying this effect, it is also essential to assess whether modifications to the monitoring or forecasting approach can compensate for potential performance degradation. Because ML approaches tend to benefit from longer training datasets (Mahmood *et al.*, 2022; Atton Beckmann *et al.*, 2025), it is important to understand monitoring requirements for climate-resilient forecasting as soon as possible to give sufficient time for data collection.

Therefore, this study presents an approach for informing the design of water quality monitoring systems that can provide sufficient data for training data-driven short-term forecast models, even in extreme climate change scenarios. To demonstrate this, we investigate how the performance of short-term ML forecasts of surface water temperature is affected under different climate change scenarios for five moderately-sized northern temperate lakes. While the lakes in this study are geographically specific, this serves as a proof of concept for the approach and offers valuable insights relevant to many northern temperate lakes. Additionally, by comparing performance across different sampling frequencies and training dataset durations, we also test how climate-driven performance degradation can be mitigated by adjusting the monitoring strategy. Therefore, the key research questions of this study are:

1. Does the performance of ML forecasts of surface water temperature reduce in more extreme climate change scenarios?
2. Are the benefits of increased training dataset duration lessened in more extreme climate change scenarios?
3. Can management strategies such as increasing sampling frequency compensate for the effects of more extreme climate change scenarios on forecast performance?

4. Which variables can be considered drivers of any link between forecast performance and climate scenario?

Ultimately, by addressing these questions, we aim to demonstrate the value of using climate projections and modelled lake data to inform the design of water quality monitoring and forecasting systems in the context of a rapidly changing climate.

3.3 Materials and Methods

3.3.1 Dataset and study sites

The ISIMIP framework provides both historical and projected climate and socioeconomic forcing data over a range of different policy relevant climate scenarios. The ISIMIP Lake Sector has used these data to force physical lake models both for ‘general’ lakes across the globe, and ‘local’ lakes which have been calibrated to a smaller number of specific lakes (Golub *et al.*, 2022). This study uses ISIMIP Local Lakes data from five lakes located in England, Ireland, and Germany (Table 3.1).

Table 3.1: Summary of the geophysical properties of the lakes used for analysis. Note that relative depth = $50 \times Z_{\max} \times \sqrt{\pi}/\sqrt{A}$, where $Z_{\max}$ is the maximum depth, and A the area (Hutchinson, 1957).

	Lake				
	Esthwaite	Stechlin	Lough Feeagh	Müggelsee	Windermere
Country	England	Germany	Ireland	Germany	England
Latitude (°)	54.37	53.17	53.9	52.43	54.31
Longitude (°)	-2.99	13.03	-9.5	13.65	-2.95
Elevation (m)	65	59.8	15	32.3	39
Mean Depth (m)	6.9	23.2	14.5	4.9	21.3
Max Depth (m)	16	69.5	44	7.7	64
Area (km ²)	0.96	2.23	3.9	7.4	14.76
Relative Depth (%)	1.45	4.12	1.97	0.25	1.48
Mixing regime	monomictic	dimictic	monomictic	polymictic	monomictic

These lakes were chosen as they are all located in the northern temperate region (Figure 3.1), and are moderately sized (~1-15 km², Figure 3.2) and therefore can be considered representative of a large number of lakes in this part of the world given the high prevalence of lakes this size (Cael and Seekell, 2016).

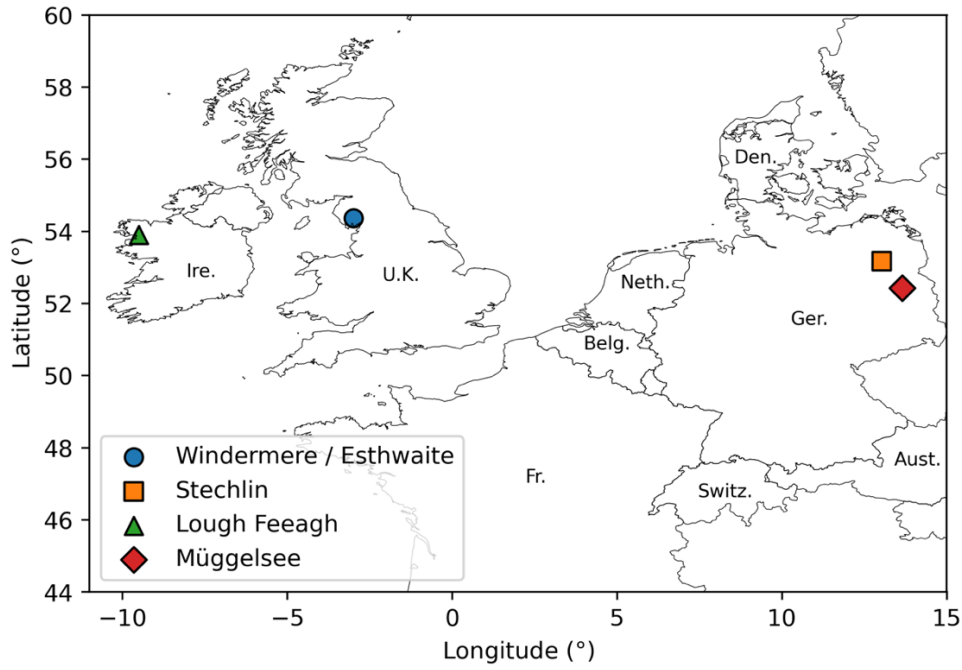


Figure 3.1: Map illustrating the geographical locations of the lakes used in this study.

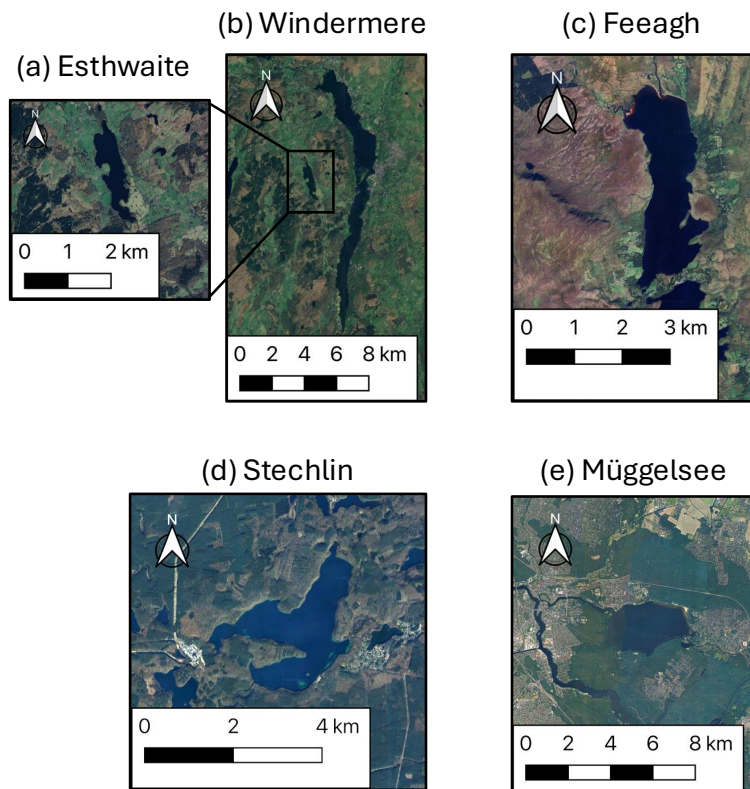


Figure 3.2: Satellite imagery showing the lakes used in this study. Imagery ©2025 provided by Google / Airbus, CNES / Airbus, Landsat / Copernicus, Maxar Technologies / Airbus, GeoBasis-DE/BKG.

Daily surface water temperature projections from the ISIMIP3b simulation round for each of these lakes from 2015-2100 were downloaded under four different scenarios, including

a pre-industrial control (PIC) and three of the shared socioeconomic pathways (SSP) described by O'Neill et al. (2016) (Table 3.2).

Table 3.2: Climate change scenarios used for experiments.

Scenario	Description	Anthropogenic Radiative Forcing in 2100 (Wm^{-2})
PIC	Model driven by invariant solar, greenhouse gases, ozone, tropospheric aerosol and land-use forcings from the year 1850.	0
SSP126	‘Sustainable and green’ pathway where substantial climate change mitigation actions are taken.	2.6
SSP370	‘Regional rivalry’ pathway where climate change action is not prioritised.	7.0
SSP585	‘Fossil-fuelled development’. Most extreme climate change scenario.	8.5

For each scenario, projections were obtained from four different physics-driven, one-dimensional lake models: Flake (Mironov *et al.*, 2010), General Lake Model (GLM) (Hipsey *et al.*, 2019), General Ocean Turbulence Model (GOTM) (Umlauf, Burchard and Bolding, 2005), and Simstrat (Gaudard *et al.*, 2019); and five different ensemble climate models: GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1) (Dunne *et al.*, 2020), IPSL-CM6A-LR (Boucher *et al.*, 2020), Max Planck Institute Earth System Model (MPI-ESM1.2) (Gutjahr *et al.*, 2019), Meteorological Research Institute Earth System Model Version 2.0 (MRI-ESM2.0) (Yukimoto *et al.*, 2019), and the UK Earth System Model (UKESM) (Sellar *et al.*, 2019).

Alongside surface water temperature data, daily meteorological data were also downloaded from the ISIMIP repository for the same five climate models. Specifically, these variables were: near-surface relative humidity, near-surface specific humidity, precipitation, snowfall flux, surface air pressure, surface downwelling longwave radiation, surface downwelling shortwave radiation, near-surface wind speed, near-surface air temperature, daily-maximum near-surface air temperature and daily minimum near-surface air temperature.

3.3.2 Forecasting model and approach

This study used entirely modelled data from ISIMIP3b lake simulations both for training and evaluation. This enabled systematic analysis of the effects of climate change on forecast skill under well-controlled conditions, independent of observational gaps or other inconsistencies. The effect of climate change was explored through a comparison of forecast performance in more extreme scenarios with the PIC scenario used as a baseline. Therefore, the 2100 anthropogenic radiative forcing associated with each scenario (Table 3.2) is used as the independent variable associated with climate change for the experimental procedure.

To ensure that the day-to-day variation of the ISIMIP-driven lake model outputs was sufficiently realistic to be a good test of forecast skill, surface water temperature measurements provided by the United Kingdom Centre for Ecology and Hydrology from Windermere (hourly data from 2016-18) and Esthwaite (fortnightly data from 2015-18), were compared with simulation outputs (Feuchtmayr, I.D. Jones, Clarke, *et al.*, 2021b; Feuchtmayr, M.A. Clarke, De Ville, *et al.*, 2021) (Figure 3.3).

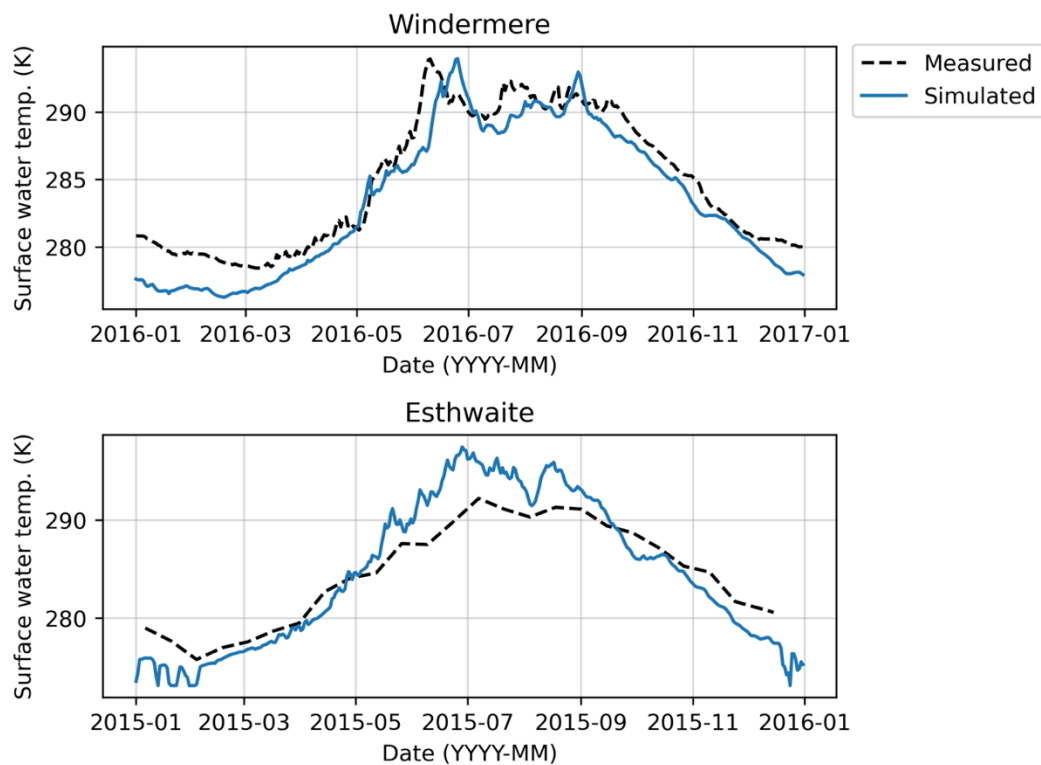


Figure 3.3: Example of measured (black dashed lines) and simulated (blue lines) surface water temperature time series from Windermere (upper plot) and Esthwaite (lower plot). In situ data from Windermere were sampled daily whereas data from Esthwaite were fortnightly. Simulated data were generated using GLM forced by climate projections from the MPI-ESM1.2 model under the SSP585 scenario.

First-order time series differencing was used to assess the daily variability for Windermere and fortnightly variability for Esthwaite. Differencing involves calculating the change in temperature between successive time steps (e.g., day-to-day or fortnight-to-fofortnight), providing a measure of short-term variability. The mean of the absolute differenced values (MADt) and the standard deviation (SDDt) of the raw differenced time series were calculated for the in situ data and the corresponding simulated data from all four lake models, providing a comparison of the magnitude and variability of short-term temperature changes. These variability metrics for the observed and simulated time series matched closely for both lakes, indicating that the simulated data were sufficiently realistic for testing forecasting skill (Table 3.3). Additionally, the mean absolute error (MAE) between the in situ and simulated time series was sufficiently low (average = 1.75 °C) to indicate that this data captured well the seasonal lake variation. However, it is important to note that a very close match between the simulated and in situ records is not expected, as the simulated time series represents a climate-scenario projection rather than a hindcast of observed conditions.

Table 3.3: Metrics comparing day-to-day and fortnightly variability of surface temperature time series: mean absolute error (MAE), mean of absolute differenced time series (MADt) and standard deviation of differenced time series (SDDt). Differences were calculated on daily data for Windermere and fortnightly data for Esthwaite.

Data Source	Lake					
	Windermere			Esthwaite		
	MAE	MADt	SDDt	MAE	MADt	SDDt
In situ	N/A	0.17	0.27	N/A	1.69	2.16
Flake	1.34	0.13	0.23	1.70	1.66	2.06
GLM	1.70	0.19	0.32	2.06	1.93	2.45
GOTM	1.99	0.18	0.29	2.11	1.77	2.28
Simstrat	1.37	0.15	0.22	1.76	1.75	2.18

Two ML forecasting models were used for the core experiments of the study. Firstly, as an example of a widely used *classical* ML approach, a Random Forest (RF) model implemented using Scikit Learn was used (Breiman, 2001; Pedregosa *et al.*, 2011). Specifically, the RF was chosen because it has been shown to be effective for similar short-term forecasting tasks even with a training dataset of relatively short duration (Giamalaki, Beaulieu and Prochaska, 2022; Bhogal *et al.*, 2023; Atton Beckmann *et al.*, 2025). In addition to the RF, a Gated Recurrent Unit (GRU) network implemented in *Tensorflow* was used (Martín Abadi *et al.*, 2015). GRUs are a type of recurrent neural

network which are generally very effective for time series forecasting tasks as they explicitly account for temporal dependencies in sequential data. Compared to Long Short-Term Memory (LSTM) networks, which are widely used for similar forecasting tasks, GRUs provide similar predictive capability whilst generally requiring smaller training datasets due to their streamlined architecture (Chung *et al.*, 2014; Waqas and Humphries, 2024).

Prior to running forecasting experiments, input data were down-sampled by keeping every T^{th} observation and discarding the others for three different sampling periods, $T = 3, 7, 14$ days. All forecasts were made 2 weeks ahead, using data from the most recent 28 days as input. Therefore, for 14-day data, the model would have access to data from the two most recent observations to make a prediction, whereas for 3-day data, the model would have access to the nine most recent observations. Each observation consisted of measurements of all eleven meteorological variables, as well as the surface water temperature estimate from that day.

The RF and GRU hyperparameters were tuned using modelled data from all lakes, lake models climate models, and scenarios for the period from 2015-2060 using 100 iterations of a randomised search with five-fold cross validation. The resulting tuned parameters are detailed in Table 3.4.

Table 3.4: Tuned hyperparameters for the Random Forest (RF) and Gated Recurrent Unit (GRU) network models. Parameters named according to the Scikit Learn and Tensorflow conventions (Pedregosa *et al.*, 2011; Martín Abadi *et al.*, 2015). GRU parameters (units and batch size) were rounded to the nearest power of two for computational efficiency.

RF		GRU	
Hyperparameter	Value	Hyperparameter	Value
max_depth	30	dropout	0.0001
max_features	0.64	units	64
min_samples_leaf	2	early stopping	True
min_samples_split	4	activation	tanh
n_estimators	415	batch size	128

Machine learning models were implemented in Python (v3.11) using the scikit learn (v1.4) and tensorflow (v2.15) packages (Pedregosa *et al.*, 2011; Martín Abadi *et al.*, 2015).

3.3.3 Statistical analysis

Forecast runs were carried out for the test years 2061-2100 at the three different sampling frequencies for all four climate scenarios and each of the lake/climate model combinations. For each test year, the model was trained using the N years directly prior to the test year for a range of different training dataset durations, $N = 1, 5, 10, 15, 20, 30$ years. As surface water temperatures were expected to be, on average, higher in the more extreme climate change scenarios, for statistical analysis it was necessary to use a relative error metric which does not scale with the magnitude of the predicted variable. Therefore, mean absolute percentage error (MAPE) was used to evaluate performance as this is a widely used relative error metric (Koutsandreas *et al.*, 2022). Alongside MAPE, MAE was also calculated as this is a widely used non-normalised average error metric (Willmott and Matsuura, 2005).

A linear mixed model (LMM) was used to test the following:

1. Is forecast performance reduced in more extreme climate change scenarios?
2. Do the benefits of training dataset duration lessen in more extreme climate change scenarios?
3. Can increasing sampling frequency or training dataset duration compensate for performance reductions in more extreme climate change scenarios?

To assess the importance of variables within the LMM, statistical significance was calculated using Satterthwaite's method to estimate degrees of freedom in order to obtain p -values (Kuznetsova, Brockhoff and Christensen, 2017). MAPE was used as the dependent variable in the LMM, with a natural logarithm transform used to approximate normality. Climate scenario was included as a fixed effect to understand how this affects forecast performance. In addition, training dataset duration and its interaction with climate scenario were included as fixed effects to understand if the benefits of long training datasets diminish in more extreme climate scenarios. Prior to analysis, the training dataset duration variable was inverse transformed to obtain an approximately linear relationship with the error. Further, sampling frequency and its interaction with training dataset duration were included as fixed effects to understand if increasing sampling frequency could compensate for any negative effects on performance due to more extreme climate change scenarios. Additionally, season and its interaction with climate scenario were included as fixed effects as this was flagged in preliminary data exploration as an important interaction. Test year was also included as a fixed effect to understand if

forecast performance could be expected to continually degrade further into the future. Finally, climate model, lake model and lake were all included as random effects to understand their contribution to the variance of forecast performance. This resulted in the following model specification:

$$\begin{aligned}
y = & \beta_0 + \beta_1 N + \beta_2 S + \beta_3 R_j + \beta_4 T + \beta_5 Y + \\
& \gamma_1 NS + \gamma_2 SR + \gamma_3 TN + \\
& u_C + u_{LM} + u_L + \varepsilon, \\
& j = 1, 2, 3, 4, \\
& u_C \sim N(0, \sigma_C^2), \\
& u_{LM} \sim N(0, \sigma_{LM}^2), \\
& u_L \sim N(0, \sigma_L^2), \\
& \varepsilon \sim N(0, \sigma_\varepsilon^2), \tag{5}
\end{aligned}$$

where y is the dependent variable, $\log(MAPE)$; N the inverse of the number of training years; S the climate scenario, which is coded as a continuous variable: Anthropogenic Radiative Forcing in 2100 (Table 3.2); R_j the season; T the sampling period; Y the test year; β_{0-5} the fixed effect coefficients; γ_{1-3} the interaction effect coefficients; u_C , u_{LM} , u_L the random effects associated with climate model, lake model and lake respectively, and ε the error, assumed to be normally distributed. Random effects were also assumed to be normally distributed. To make meaningful comparisons between the effect sizes of each of the fixed effects, all variables used in the LMM were first standardised to have a mean of zero and standard deviation of one. Following standardisation, two LMMs were fitted – one for the outputs from the GRU model, and one for the RF, allowing cross-model comparison of effect sizes. Model assumptions were assessed through visual inspection of residual plots. When plotted against fitted values, the residuals showed no substantial patterns indicative of heteroscedasticity, and Q-Q plots indicated that residuals were approximately normally distributed.

As an additional demonstration of the practical effect of climate scenario on performance, the maximum summer forecast error was compared between scenarios for each model. Maximum error was defined as the largest absolute difference between simulated and observed surface water temperature within each summer season. This metric was selected

because in summer months extreme warm events can, for example, exceed thermal limits for freshwater fish or trigger other heat-related impacts and are therefore of significant management concern (Till *et al.*, 2019). Maximum summer error therefore provides a conservative measure of model reliability under conditions critical for decision-making (Woolway, Sharma and Smol, 2022; Duan *et al.*, 2024).

Finally, additional linear regression analyses were used to explore the relationships between the suspected drivers of forecast error and climate change scenario. Firstly, to understand which variables to focus on, a feature importance analysis was carried out using a permutation importance approach using MAE as the error metric (Breiman, 2001; Ernst, 2004). The highest ranking features accounting for the vast majority (> 90 %) of the total increase in MAE found in the permutation importance analysis were selected for further analysis. These variables were then aggregated in two ways: 1) using monthly averages, and 2) using monthly averages of the absolute values of the differenced time series of each variable. This second aggregation approach was used as it is effectively a measure of volatility, which for some variables, such as temperature, may increase under climate change (Guo *et al.*, 2021). Therefore, this enabled testing if increased predictor volatility associated with more extreme climate change scenarios was also associated with increased forecast error. To do this, firstly the aggregated variables were regressed against climate change scenario. Following this, MAPE was regressed against each of the aggregations. This was investigated for a medium/high data availability scenario with a training dataset duration of 20 years, and a sampling period of seven days. All statistical analyses were carried out using R Statistical Software version 4.3.2 (R Core Team, 2021).

3.4 Results

3.4.1 Effect of climate scenario on forecast performance

In the more extreme climate change scenarios, forecast error was on average higher, particularly in the summer months (Figure 3.4a). However, for all lakes except Müggelsee, the reverse effect was observed in the winter months: forecast error was on average lowest in the most severe climate change scenarios (Figure 3.4c). These effects were statistically significant for both models ($p < 0.001$, Table 3.5).

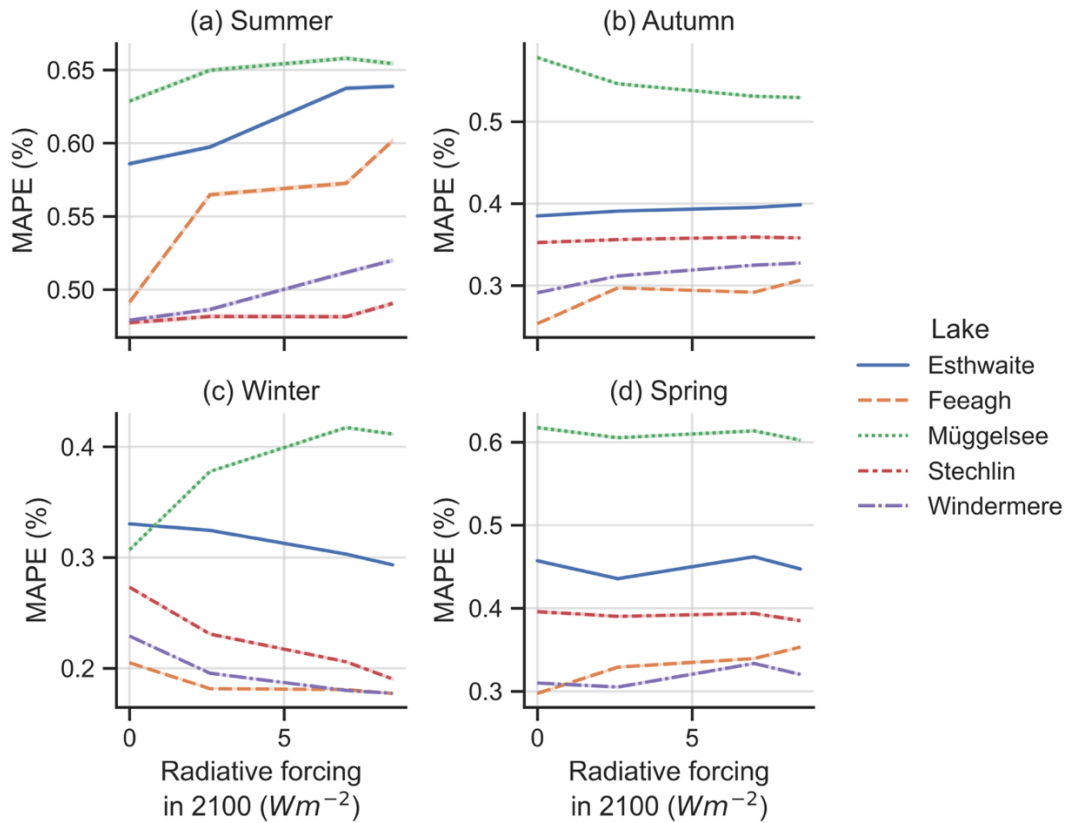


Figure 3.4: MAPE against 2100 anthropogenic radiative forcing separated by season and lake (Esthwaite – blue, Feeagh – orange, Müggelsee – green, Stechlin – red, Windermere – purple).

The standardised regression coefficient for the year variable was positive and statistically significant ($p < 0.001$) for both forecast models, but much smaller in magnitude than most other predictor coefficients. This indicates that forecast error increased further into the future, but that this effect was small in comparison to the effects of training dataset duration, sampling period, and climate change scenario.

Table 3.5: Summary of fixed effects from standardised LMM regression with $\log(\text{MAPE})$ as the dependent variable. Statistical significance is indicated using asterisk notation: *** - $p < 0.001$, ** - $p < 0.01$, * - $p < 0.05$, n.s. – not significant ($p \geq 0.05$). Values < 0.001 are denoted 0.000[^].

Fixed Effect	Estimate ($\pm$ Standard Error)		Significance	
	RF	GRU	RF	GRU
β_0 (Intercept)	-0.073 $\pm$ 0.163	0.031 $\pm$ 0.161	n.s.	n.s.
β_1 (Inverse no. of train years)	0.131 $\pm$ 0.000 [^]	0.252 $\pm$ 0.000 [^]	***	***
β_2 (Scenario)	0.023 $\pm$ 0.000 [^]	0.026 $\pm$ 0.000 [^]	***	***
β_3 (Season – spring)	0.047 $\pm$ 0.001	0.048 $\pm$ 0.001	***	***
β_3 (Season – summer)	0.614 $\pm$ 0.001	0.562 $\pm$ 0.001	***	***
β_3 (Season – winter)	-0.601 $\pm$ 0.001	-0.501 $\pm$ 0.001	***	***
β_4 (Sample period)	0.021 $\pm$ 0.000 [^]	0.013 $\pm$ 0.000 [^]	***	***
β_5 (Year)	0.003 $\pm$ 0.000 [^]	0.004 $\pm$ 0.000 [^]	***	***
γ_1 (Inverse no. of train years : Scenario)	0.003 $\pm$ 0.000 [^]	-0.002 $\pm$ 0.000 [^]	***	***
γ_2 (Scenario : Season – spring)	-0.001 $\pm$ 0.001	-0.014 $\pm$ 0.001	n.s.	***
γ_2 (Scenario : Season – summer)	0.028 $\pm$ 0.001	0.018 $\pm$ 0.001	***	***
γ_2 (Scenario : Season – winter)	-0.072 $\pm$ 0.001	-0.076 $\pm$ 0.001	***	***
γ_3 (Sample period : No. of train years)	-0.030 $\pm$ 0.000 [^]	0.002 $\pm$ 0.000 [^]	***	***

The variance associated with lake (RF = 0.12, GRU = 0.12) and the unexplained residual variance (RF = 0.68, GRU = 0.70) were of a similar order of magnitude to the largest of the fixed effect coefficients: season and training dataset duration (Table 3.5, Table 3.6). The random effects associated with climate model and lake model were much smaller suggesting approximate agreement between models.

Table 3.6: Summary of random effects from standardised LMM regression with $\log(\text{MAPE})$ as the dependent variable.

Random effect	RF		GRU	
	Variance	Std. Dev.	Variance	Std. Dev.
u_L (Lake)	0.12	0.35	0.12	0.34
u_C (Climate Model)	1.7E-3	0.041	1.6E-3	0.041
u_{LM} (Lake Model)	8.8E-3	0.094	8.6E-3	0.093
ε (Residual)	0.68	0.83	0.70	0.83

With 30 years of weekly training data, averaged across all model runs and years, the GRU model's maximum summer forecast error was 0.16 °C higher under SSP585 compared to SSP126, while the RF model's maximum summer error increased by 0.30 °C (Table 3.7).

Table 3.7: Difference in average maximum summer error (°C) between SSP585 and SSP126 for models trained with 30 years of weekly data.

Lake	RF	GRU
Esthwaite	0.28	0.16
Feeagh	0.27	0.27
Müggelsee	0.40	0.16
Stechlin	0.28	0.09
Windermere	0.27	0.12
Average	0.30	0.16

3.4.2 Interaction between training dataset duration and climate scenario

Across all climate change scenarios, training dataset duration scaled with error according to an approximate inverse power law (Figure 3.5). This has the consequence that the majority of the performance benefit was obtained with the first ten years of training data, with diminishing returns beyond this point. The effect of training dataset duration was stronger for the GRU (= 0.252) than the RF (= 0.131) model. As such, in low data-availability scenarios, the RF consistently out-performed the GRU, but with the maximum training dataset size and sampling frequency, the performance of the two models was approximately equal (Figure 3.5). The interaction between the number of training years and climate change scenario was statistically significant ($p < 0.001$) for both forecast models, but with standardised regression coefficients over 40 times smaller than that of training dataset duration this effect is effectively negligible (Table 3.5).

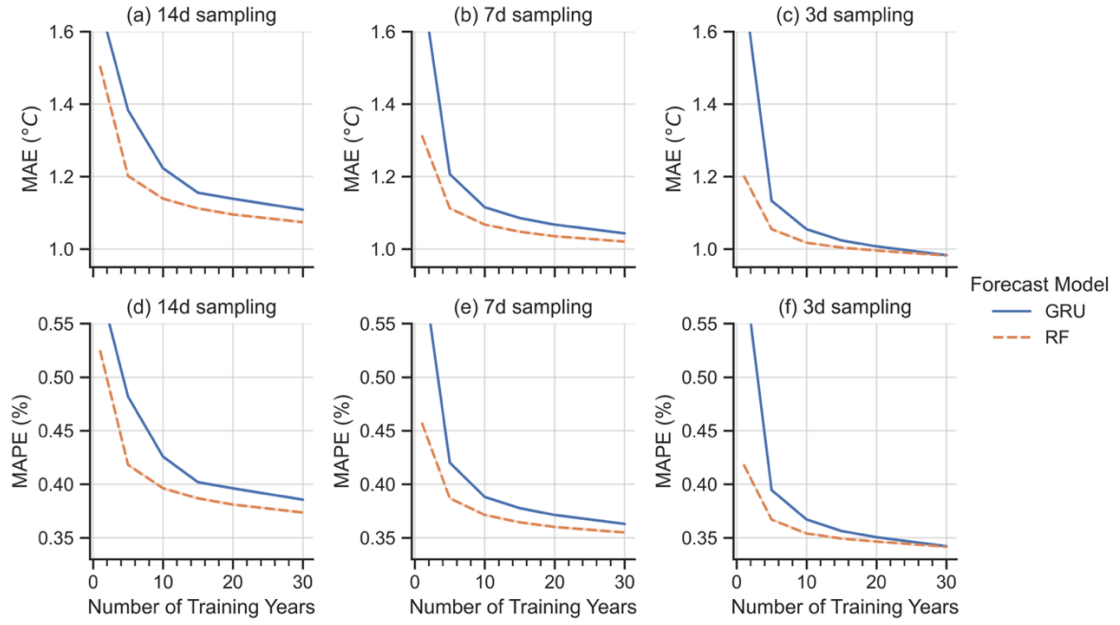


Figure 3.5: Forecast model MAE (top row) and MAPE (bottom row) plotted against number of training years for different sampling periods (14 days – left, 7 days – centre, 3 days – right). Blue solid lines show errors for the GRU model, and dashed orange lines the RF.

3.4.3 Analysis of relative effect sizes

Asides from the seasonal effects, training dataset duration had the largest effect on performance. However, this was most prevalent for low numbers of training years. Whilst the difference in performance between low and high training dataset duration scenarios was smaller at higher sampling frequencies, the overall effect of increasing sampling frequency was that this improved performance. Standardised regression coefficients for sampling frequency (RF = 0.021, GRU = 0.013) and climate change scenario (RF = 0.023, GRU = 0.026) were of similar magnitudes indicating that increasing sampling frequency has a similar effect size on performance as more extreme climate scenarios (Table 3.5, Figure 3.6). For the RF model, reducing the sampling period by 1 day would have comparable impact on average forecast performance as a 1.0 Wm^{-2} reduction in radiative forcing by 2100.

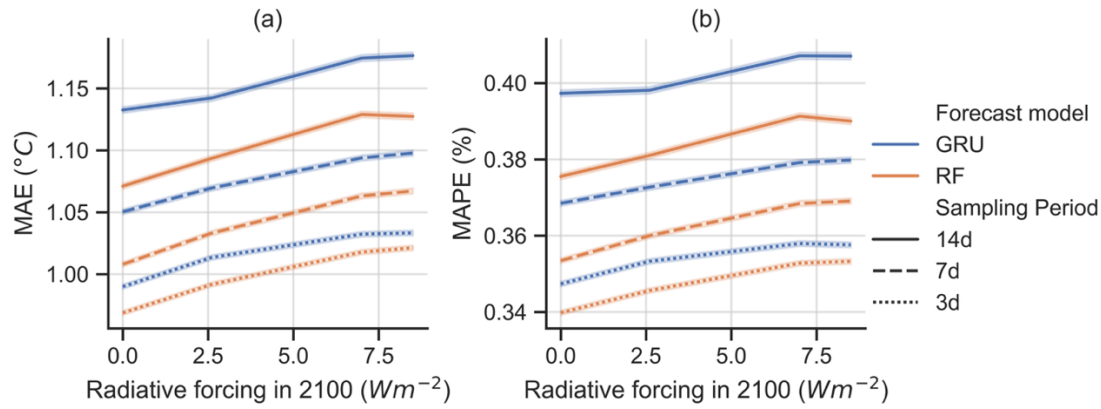


Figure 3.6: Forecast model MAE (a) and MAPE (b), plotted against the 2100 anthropogenic radiative forcing associated with each of the four climate scenarios for different sampling periods for higher data availability scenarios (number of training years ≥ 10). Sampling period is indicated by line style (14 days – solid, 7 days – dashed, 3 days – dotted), with blue lines showing errors for the GRU model and orange lines the RF.

For all lakes, various interventions such as increasing sampling frequency, or training dataset duration were sufficient to compensate for the forecast performance difference between SSP126 and SSP585, even in the summer months where this effect was strongest (Figure 3.7). Notably, the German lakes required much smaller interventions than the English and Irish lakes. For example, if using the RF model to make forecasts for Stechlin, increasing the sampling frequency from fortnightly to weekly sampling was sufficient to compensate for the effect of climate change scenario, whereas with Windermere, the equivalent compensation required a sampling frequency increase from fortnightly to 3-day sampling (Figure 3.7). Additionally, GRU forecasts generally required smaller interventions to compensate for the effects of climate change than the RF. For example, for Windermere, increasing the sampling from fortnightly to weekly was sufficient compensation for the GRU, whereas for the RF this required an increasing from fortnightly to 3-day sampling (Figure 3.7).

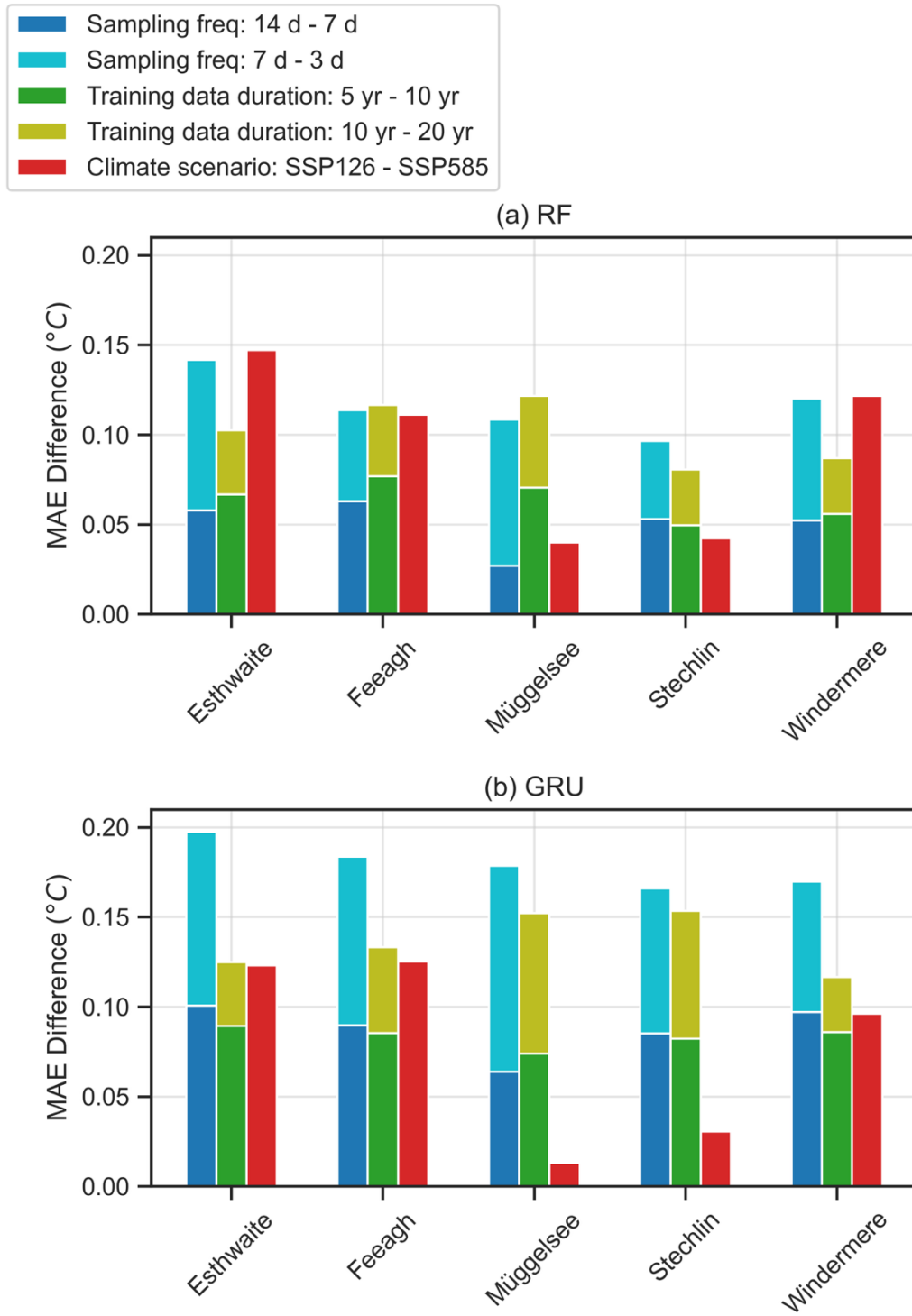


Figure 3.7: Absolute difference in summer MAE for (a) RF and (b) GRU resulting from increasing sampling frequency from 14 to 7 to 3 days (blue and light blue bars) and increasing training dataset duration from 5 to 10 to 20 years (green and olive bars) under climate scenario SSP585. Red bars show the absolute difference in summer MAE between SSP126 and SSP585.

3.4.4 Drivers of performance differences

On average across the two forecast models, the three most important variables were shortwave downwelling radiation, water surface temperature, and specific humidity. Together, these accounted for 90.2 % of the total increase in MAE found in the

permutation importance analysis (Figure 3.8), and thus were selected for further investigation in relation to climate scenario and forecast error

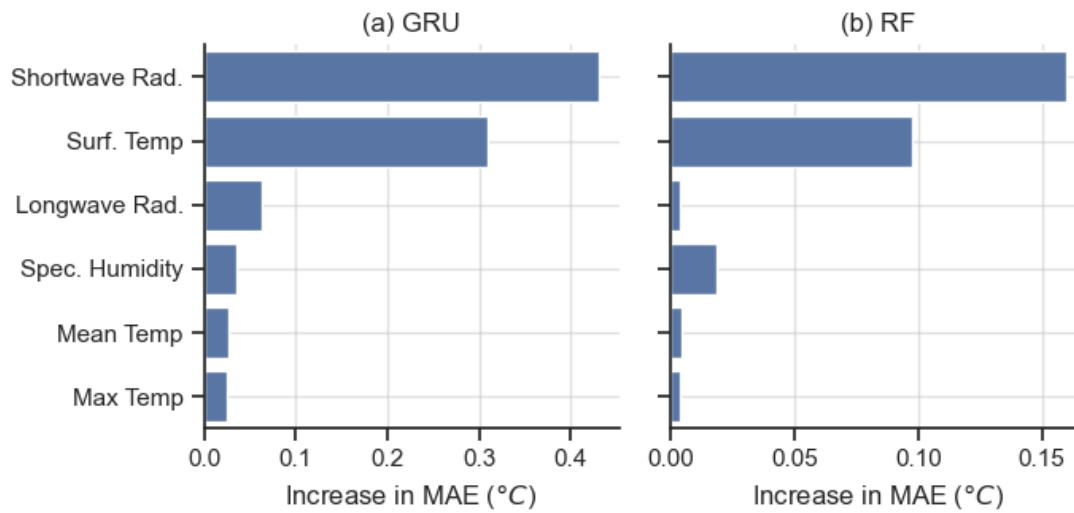


Figure 3.8: Permutation importance analysis, showing mean increase in MAE due to the selective permutation of each variable. Results are an average of all lakes, lake models, climate models, and climate scenarios, and were generated using 20 years of training data from 2040 - 2060, with 30 permutations calculated for each run. Only the six most important variables are shown.

In summer months, it was found that climate change scenario was significantly ($p < 0.05$) correlated with both surface temperature and specific humidity, as well as the absolute values of these following differencing (Figure 3.9 a, b, c, d). Conversely, shortwave downwelling radiation was not found to be associated with climate change scenario (Figure 3.9 e, f).

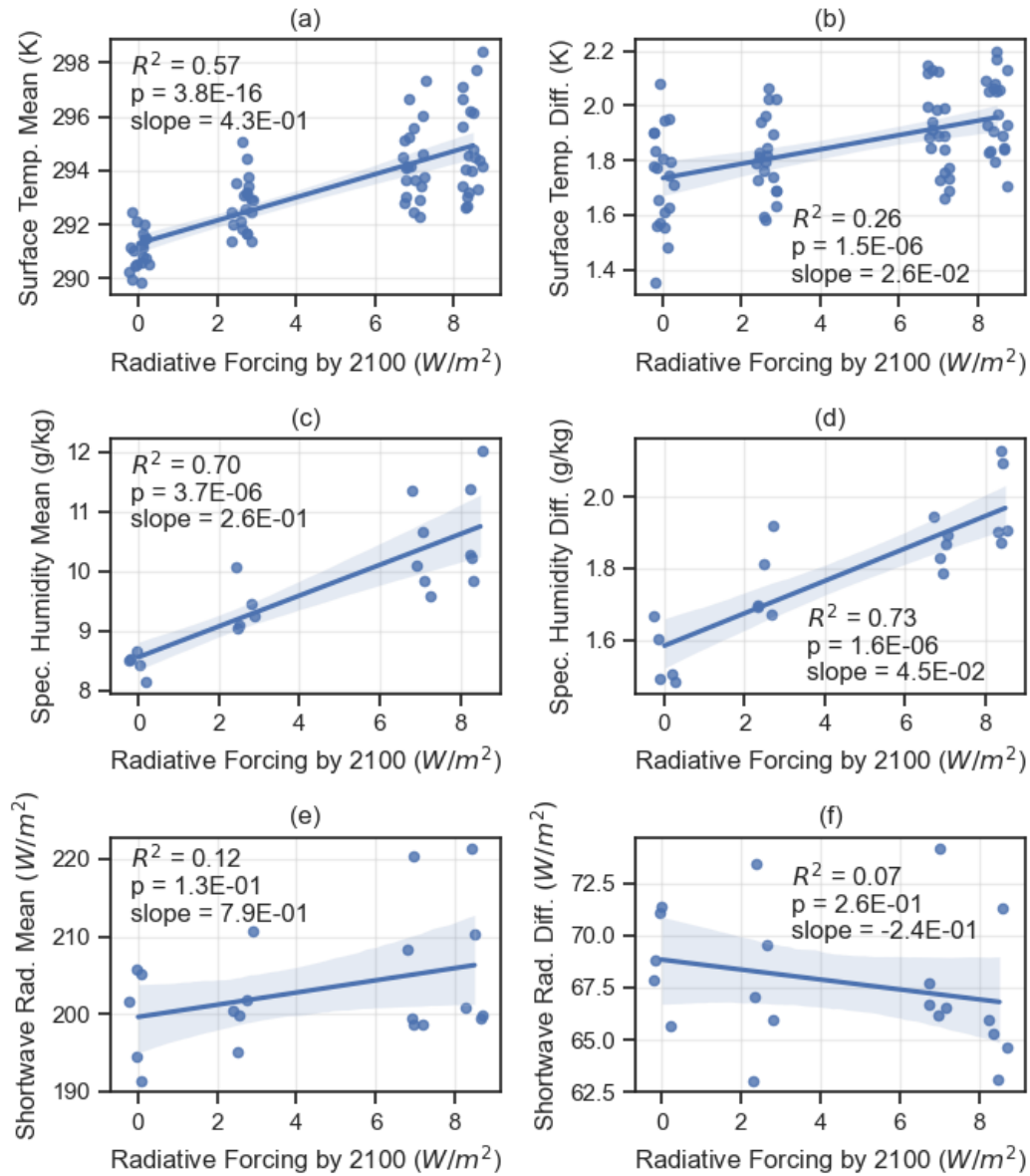


Figure 3.9: Linear regression analyses from summer months between climate change scenarios and key feature means (left column) and absolute values of the differenced time series of key features (right column). Each scatter point represents the average from a single climate/lake model combination. Note that specific humidity (c, d) and shortwave radiation (e, f) do not vary with lake model, and so have fewer points than surface temperature (a, b). 'Jitter' in the x axis has been added to improve readability. 95% confidence intervals for the regression estimates are shown by the blue translucent bars around each regression line.

Similarly, both mean and differenced versions of surface temperature and specific humidity were associated with MAPE (Figure 3.10 a, b, c, d), whereas shortwave downwelling radiation was not found to be significantly correlated with forecast error (Figure 3.10 e, f). The variable with the strongest relationship with forecast error was absolute differenced surface temperature ($R^2 = 0.78$).

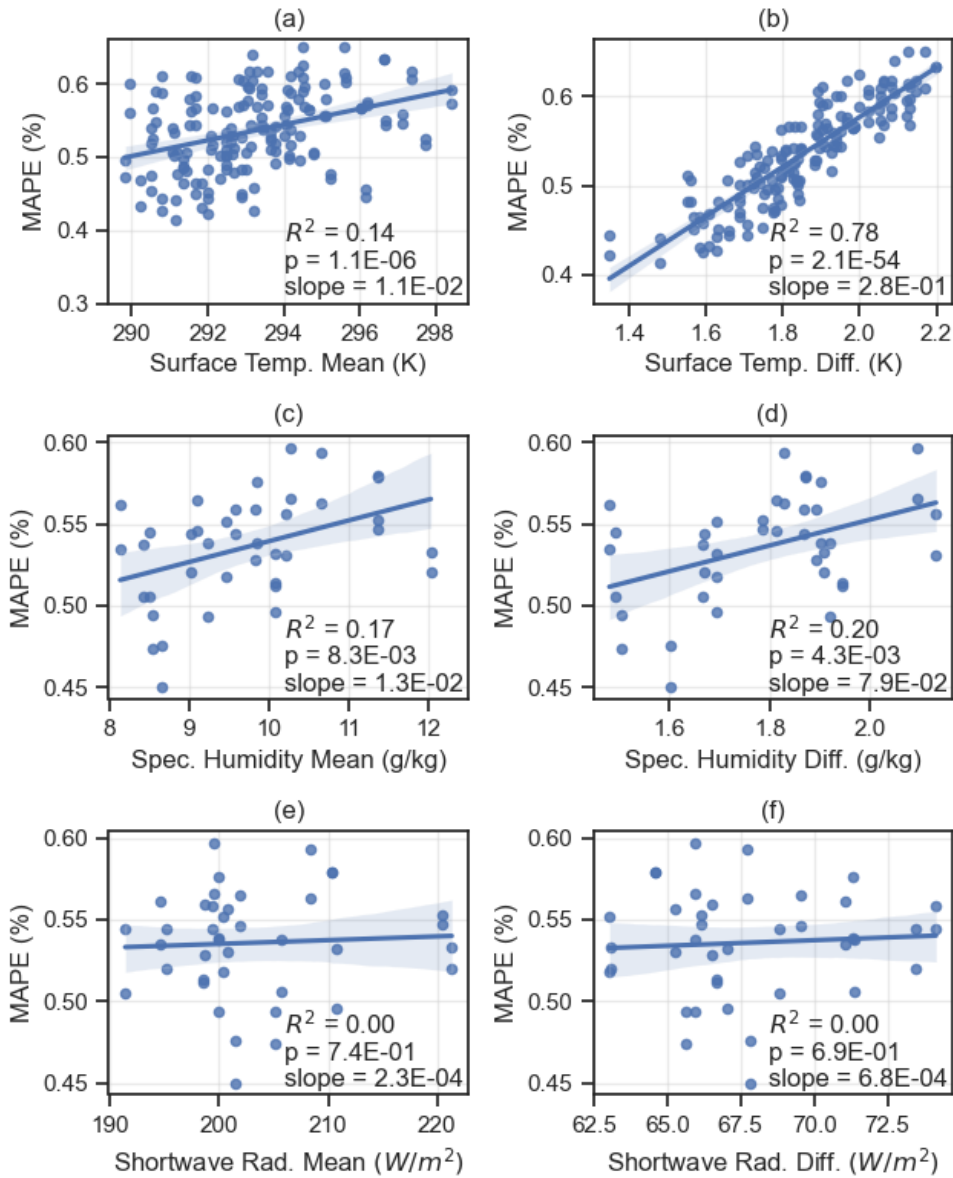


Figure 3.10: Linear regression analyses from summer months between key feature means (left column), absolute values of the differenced time series of key features (right column) and MAPE. Each scatter point represents the average from a single climate/lake model/forecast model combination. Note that specific humidity (c, d) and shortwave radiation (e, f) do not vary with lake model, and so have fewer points than surface temperature (a, b). 95% confidence intervals for the regression estimates are shown by the blue translucent bars around each regression line.

Similar, but mostly weaker patterns were observed in the other seasons. For example, absolute differenced surface temperature was significantly correlated with MAPE in all seasons (Table 3.8), but only significantly correlated with climate change scenario in autumn and summer (Table 3.9). Generally, the associations between aggregated variables and MAPE were strongest in the spring and summer months.

Table 3.8: Complete results from correlation analyses with key features and MAPE. Statistical significance is denoted as follows: *** - $p < 0.001$, ** - $p < 0.01$, * - $p < 0.05$, n.s. – not significant ($p \geq 0.05$).

Feature	Aggregation	Season	R ²	Slope	Significance
Surface temperature	Mean	Autumn	0.13	3.60E-03	***
		Spring	0.04	7.15E-03	**
		Summer	0.14	1.07E-02	***
		Winter	0.09	-4.58E-03	***
Surface temperature	Absolute differencing	Autumn	0.30	7.55E-02	***
		Spring	0.67	2.23E-01	***
		Summer	0.78	2.79E-01	***
		Winter	0.32	1.82E-01	***
Specific Humidity	Mean	Autumn	0.30	7.51E-03	***
		Spring	0.25	2.74E-02	**
		Summer	0.17	1.28E-02	**
		Winter	0.16	-1.16E-02	*
Specific Humidity	Absolute differencing	Autumn	0.26	4.98E-02	***
		Spring	0.33	1.20E-01	***
		Summer	0.20	7.94E-02	**
		Winter	0.01	1.93E-02	n.s.
Shortwave radiation	Mean	Autumn	0.05	1.11E-03	n.s.
		Spring	0.07	1.25E-03	n.s.
		Summer	0.00	2.30E-04	n.s.
		Winter	0.09	2.83E-03	n.s.
Shortwave radiation	Absolute differencing	Autumn	0.01	1.03E-03	n.s.
		Spring	0.01	7.85E-04	n.s.
		Summer	0.00	6.82E-04	n.s.
		Winter	0.09	5.32E-03	n.s.

Table 3.9: Complete results from correlation analyses with climate change scenario and key features. Statistical significance is denoted as follows: *** - $p < 0.001$, ** - $p < 0.01$, * - $p < 0.05$, n.s. – not significant ($p \geq 0.05$).

Feature	Aggregation	Season	R ²	Slope	Significance
Surface temperature	Mean	Autumn	0.62	0.431	***
		Spring	0.70	0.361	***
		Summer	0.57	0.428	***
		Winter	0.68	0.376	***
Surface temperature	Absolute differencing	Autumn	0.13	0.014	***
		Spring	0.00	-0.001	n.s.
		Summer	0.26	0.026	***
		Winter	0.00	0.000	n.s.
Specific Humidity	Mean	Autumn	0.62	0.233	***
		Spring	0.75	0.127	***
		Summer	0.70	0.258	***
		Winter	0.71	0.145	***
Specific Humidity	Absolute differencing	Autumn	0.45	0.028	**
		Spring	0.21	0.018	*
		Summer	0.73	0.045	***
		Winter	0.15	0.011	n.s.
Shortwave radiation	Mean	Autumn	0.00	-0.032	n.s.
		Spring	0.14	-0.627	n.s.
		Summer	0.12	0.787	n.s.
		Winter	0.46	-0.370	***
Shortwave radiation	Absolute differencing	Autumn	0.03	-0.065	n.s.
		Spring	0.02	0.103	n.s.
		Summer	0.07	-0.239	n.s.
		Winter	0.17	-0.116	n.s.

3.5 Discussion

3.5.1 The impact of climate change on water temperature forecasting

It was found that surface water temperature forecasts did suffer from reduced performance in more extreme climate change scenarios, although this varied between seasons, lakes and forecast models. Whilst the size of this effect may appear small, it was found to be comparable to the performance benefit gained from quite substantial monitoring improvements.

In summer, MAE in the most severely affected lake (Esthwaite) was on average 0.14 °C higher in SSP585 compared with SSP126. Additionally, across all lakes, using a 30-year

weekly training dataset, the maximum summer error for the best performing model (RF) was on average 0.30 °C higher in SSP585 than SSP126. Consequently, if these forecasts were used as an early warning system to trigger specific management interventions, then these error increases could be conceivably large enough to have negative management effects, particularly if the common approach of threshold-based decision triggers is used (Herman and Giuliani, 2018). Specifically, this could potentially determine whether a forecast system correctly triggers a response such as reservoir withdrawal for thermal regulation (Rheinheimer, Null and Lund, 2015; Weber *et al.*, 2017; Feldbauer *et al.*, 2020).

With the RF forecast model, one way to compensate for the increased error between SSP126 and SSP585 would be to increase the sampling frequency approximately fourfold. Although automated in situ sampling technology now provides a means for consistent high frequency sampling of water quality indicators, this is still costly, particularly when considering the maintenance of such systems over long time periods (Seifert-Dähnn *et al.*, 2021; Harmel *et al.*, 2023). This highlights a critical design consideration for climate-resilient monitoring systems - the trade-off between the cost of monitoring infrastructure and forecast performance degradation under climate change which, as demonstrated, can be quantified using the approach outlined in this study.

3.5.2 Variation across lakes and seasons

All lakes behaved in a similar way during the summer, that is forecast performance generally reduced in more extreme climate change scenarios (Figure 3.4a). However, all lakes except Müggelsee showed the opposite pattern in the winter – forecast performance improved in more extreme climate change scenarios (Figure 3.4c). This contrast can be explained by shifts in lake mixing regimes: climate change is expected to push some lakes from dimictic toward monomictic behaviour, reducing winter mixing events and increasing thermal stability (Shatwell, Thiery and Kirillin, 2019). Mixed conditions tend to lower surface temperature variability due to a deeper, more thermally inert surface layer, making winter conditions more predictable. Müggelsee, by contrast, is relatively much more shallow and polymictic (Table 3.1), with a more stable mixing regime that does not shift under warming (Shatwell and Köhler, 2019). It likely represents the default pattern for lakes that remain polymictic, lacking the "reset" effect that makes deeper lakes more predictable in winter. As a result, Müggelsee shows a winter forecast pattern more similar to its summer behaviour.

This underlines the need for site-specific monitoring design that is informed by climate projections, as no single rule applies uniformly across lakes. This also illustrates the importance of seasonality, suggesting that forecasts for lakes located in different climate zones, with alternative seasonal patterns, may respond differently to climate change. For example, climate change is likely increasing over-lake atmospheric instability, which varies by lake size and latitude, and may lead to a decoupling of air and surface water temperatures, potentially reducing forecast skill more strongly in some lakes than others (Woolway *et al.*, 2017). Additionally, shifts in lake thermal structure due to climate change have also been shown to be altitude dependent, which indicates that lake forecast performance degradation due to climate change may also vary across elevation gradients (Råman Vinnå *et al.*, 2021). Accordingly, it is evident that the results of this study are not globally generalisable but are still relevant for the large number of northern temperate lakes with similar latitudes, morphologies, and elevations to the five studied lakes.

3.5.3 Relevance to other water quality metrics

Under the assumption that increased difficulty in forecasting water temperature would also be associated with increased difficulty in forecasting other variables which are driven or affected by it, the results of this study may generalise to other variables. This assumption is likely most valid for variables which are largely physics-driven, such as stratification onset, or hypolimnetic anoxia (Boehrer and Schultze, 2008; Snortheim *et al.*, 2017). For variables strongly influenced by both the physics and ecology of the study site, the picture is more complex. For example, in the case of cyanobacterial bloom forecasting, increased uncertainty in surface temperature suggests higher uncertainty in growth rates (Davis *et al.*, 2009). However for waterbodies where cyanobacterial blooms transition from being rare events to more commonplace phenomena due to climate change, they may inherently become more predictable (Paerl *et al.*, 2020; Feng, Li and Xu, 2023). Given that it is possible to simulate processes such as algal growth under various climate scenarios (Zwart *et al.*, 2019), it is likely that soon a more diverse and complex set of water quality variables will become available through projects like ISIMIP (Golub *et al.*, 2022). This therefore opens the possibility for informing the design of climate-resilient monitoring and forecasting systems for more complex water quality phenomena such as algal blooms.

Additionally, whilst it was found that there was a significant interaction ($p < 0.001$) between training dataset duration and climate scenario, the relative size of this effect was small in comparison to other factors, such as climate scenario (Table 3.5). This means that older data were approximately equally useful for training models, regardless of climate

change intensity. A likely explanation for this is that surface water temperature is well-understood to be primarily governed by air temperature and surface heat exchange, which, from a process perspective, can be considered largely stationary. In other words, the underlying heat transfer equations which govern it generally do not change in structure, except during periods of ice cover (Hipsey *et al.*, 2019). In contrast, other variables such as algal biomass are influenced by a far more complex set of interacting drivers, including nutrients, light availability, and grazing, which can exhibit strong non-stationarity, particularly under ecological regime shifts (Zingone and Oksfeldt Enevoldsen, 2000). These complexities may render older data less representative of current or future conditions, underscoring that future work applying the approach presented here to additional water quality indicators would be valuable for designing climate-resilient monitoring and forecasting systems.

3.5.4 Forecast model choice implications

Whilst the structure of recurrent neural network models such as GRU is inherently suited to time series forecasting tasks, they often do not outperform other forecasting approaches, particularly in data-limited scenarios (Hewamalage, Bergmeir and Bandara, 2021; Yousefi and Toffolon, 2022; Atton Beckmann *et al.*, 2025). Accordingly, in this study, in all but the highest data availability scenario, the RF outperformed the GRU (Figure 3.5). This is likely due to the higher complexity of the GRU relative to the RF, which therefore necessitates a larger volume of training data to achieve competitive performance (Atton Beckmann *et al.*, 2025). However, with sufficient training data (30 years of 3-day samples), the GRU's performance did marginally exceed the RF. Furthermore, given that the GRU's training dataset size / performance gradient (β_1) was almost double that of the RF (Table 3.5), it can be inferred that in even higher data-availability scenarios, the GRU would be expected to outperform the RF by an increasing margin.

Whilst both models' performance was affected by climate change scenario, this effect varied according to season, and in some cases was notably different between the two forecast models (Table 3.5). In autumn, both models suffered similar performance degradations in more extreme climate scenarios. Likewise, in winter, the forecast performance improvements due to intensified climate warming were comparable between forecast models. However, in summer, climate scenario decreased the RF's performance ($\gamma_2 = 0.028$) more than the GRU's performance ($\gamma_2 = 0.018$). Similarly, in spring, the effect of climate scenario on the RF's performance was approximately equivalent to that in autumn ($\gamma_2 = \text{n.s.}$), whereas the GRU's performance was less negatively affected ($\gamma_2 = -$

0.014). Therefore, on average, whilst the RF performed better with smaller training datasets, it was also more negatively affected by climate scenario than the GRU. This suggests that where large volumes of high frequency measurement data are available, using recurrent neural network models such as GRU may be beneficial for minimising the effects of climate change on forecast performance.

3.5.5 Evaluating forecast skill with synthetic data

All the data used in this study were synthetic, and therefore there is some uncertainty concerning how the findings would translate to actual surface lake temperature forecasting with field-measured data. To some extent this was quantified by using a suite of climate and lake models, and it was found that the variances associated with the random effects of lake model ($\sigma^2 = 0.0087$) and climate model ($\sigma^2 = 0.0018$) were both small compared with the variance associated with lake ($\sigma^2 = 0.12$). This suggests good agreement between the models used in this study. Furthermore, the comparison of simulated data with in situ observations demonstrated that the magnitude and variation in daily and fortnightly changes were closely matched (Table 3.3). It was found that GLM tended to have slightly higher short-term variations than the corresponding in situ observations, whereas Flake tended to vary less. This demonstrates that it is certainly possible to systemically under- or over-estimate forecast skill when testing this with simulated data. However, by using a suite of models, any model-specific biases will likely have been reduced. Ultimately, despite the uncertainties inherent with modelled data, using these allows exploration of forecast performance across multiple climate scenarios and decades, which would be impossible with field measurements alone.

3.5.6 Designing climate-resilient water quality forecasting systems

Unlike weather forecasting, which benefits from decades of methodological refinement (Lynch, 2008), water quality forecasting is still developing, particularly in the context of ecological complexity and climate change (Dietze *et al.*, 2018). Interest in forecasting complex ecological phenomena, such as algal blooms, continues to grow (Zahir *et al.*, 2024), yet these events often have site-specific drivers that differ markedly between lakes (Ding and Qin, 2024). Climate change further complicates this challenge by simultaneously increasing the need for reliable forecasts (Bartlett and Dedekorkut-Howes, 2022) and, as this study demonstrates, potentially making the forecasting task inherently more difficult. This illustrates the importance of designing water quality monitoring infrastructure with both climate resilience and site-specific variability in mind. While

ongoing evaluation and adaptation of monitoring systems is critical, we argue that a purely iterative design approach is not efficient - particularly when it may take five or more years to determine whether a given sampling approach adequately supports the forecasting objective (Atton Beckmann *et al.*, 2025).

Instead, we propose that the design of monitoring and forecasting systems should also be informed through simulation, by testing forecasting models on data generated by physical and biological models driven by climate projections. This approach enables systematic testing of key design decisions, such as which variables to monitor, how frequently to sample, which forecast models to apply, and how much training data is needed before reliable forecasts can be generated. As demonstrated in this study, climate projections can also be used to evaluate how system performance may change under different future scenarios, supporting the development of forecasting systems capable of maintaining sufficient accuracy even under extreme climate change. Whilst this simulation-based approach is constrained by the underlying input data - here, physical, one-dimensional lake models and ISIMIP climate projections - these models are continually improving (Golub *et al.*, 2022; Frieler *et al.*, 2024). Moreover, two-way coupling of climate projections with three-dimensional lake models has already been successfully demonstrated for some regions, which is particularly important for resolving hydrodynamics in larger and more complex lake systems (Xue *et al.*, 2022). Therefore, as coupling capabilities expand and lake and climate models continue to advance, they will offer increasing potential to support the development of effective monitoring and forecasting strategies for managing freshwater resources in a changing climate.

3.6 Conclusions

This study highlights how climate change can affect the performance of short-term lake water temperature forecasts, and how simulation-based approaches can be used to design climate-resilient monitoring and forecasting systems. The key conclusions are:

- Forecast skill is sensitive to climate change, especially in summer, due to increased surface temperature variability. However, this performance degradation can be offset through targeted improvements in monitoring.
- Monitoring system design should balance cost and performance, as increasing sampling frequency or extending historical records can compensate for climate-driven forecast degradation. These trade-offs can be quantified using synthetic data.

- Responses vary by lake and season, with deeper lakes often becoming more predictable in winter under warming scenarios. This emphasizes the need for site-specific, simulation-informed monitoring strategies.
- Findings likely extend beyond temperature forecasts to other water quality indicators influenced by surface physics, such as stratification or hypoxia. For more complex ecological phenomena, responses may differ due to nonlinear regime shifts.
- Simulation-based system design offers a proactive approach to developing resilient water quality monitoring infrastructure, complementing traditional iterative methods that require years of real-world testing.
- As climate and lake models improve, this approach will become increasingly powerful in guiding the development of sustainable, future-proof forecasting systems for freshwater management.

These findings support a shift towards forward-looking, data-informed monitoring strategies that account for the challenges posed by a changing climate.

4 Widespread phytoplankton monitoring in small lakes: a case study comparing satellite imagery from Planet SuperDoves and ESA Sentinel-2

This chapter was published as:

Atton Beckmann, D., Spyrakos, E., Hunter, P., Jones, I.D., 2025. Widespread phytoplankton monitoring in small lakes: a case study comparing satellite imagery from planet SuperDoves and ESA sentinel-2. Front. Remote Sens. 6.
<https://doi.org/10.3389/frsen.2025.1549119>

CRedit authorship contribution statement:

D. Atton Beckmann: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing.

E. Spyrakos: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review and editing. **P. Hunter:**

Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review and editing. **I.D. Jones:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review and editing.

4.1 Abstract

Satellite imagery has enabled widespread monitoring of algae in larger water bodies, however until recently, the spatial resolution of available sensors has not been sufficient to apply this to smaller lakes. Therefore, this study investigated a new dataset of high-resolution metre-scale imagery for monitoring phytoplankton at spatial and temporal scales previously impossible with satellite data. Specifically, the Planet SuperDoves constellation was used to monitor a small (0.069 km^2), eutrophic lake from 2021 to 2024. Several chlorophyll-a (chl-a) algorithms were tested on both SuperDoves and Sentinel-2 data against in situ measurements. chl-a was successfully retrieved using the Ocean Colour 3 algorithm ($R^2 = 0.64$, root mean square error (RMSE) = $0.93 \mu\text{g L}^{-1}$), which outperformed the best performing Sentinel-2 chl-a algorithm ($R^2 = 0.61$, RMSE = $1.01 \mu\text{g L}^{-1}$). Furthermore, both Sentinel-2 and SuperDoves data were equally effective for algal bloom detection, each having F1-scores of 0.89 at a chl-a bloom threshold of $40 \mu\text{g L}^{-1}$. Finally, we investigated the suitability of citizen science data as a validation tool for widespread algal bloom monitoring by comparing reports of algal blooms in five small water bodies in central Scotland with corresponding SuperDoves chl-a images. This demonstrates that metre-scale satellite monitoring of algae is possible even in challenging and optically complex environments such as small, shallow water bodies. This leads towards a potential step-change in the number of remotely monitorable inland water bodies, which would be a significant advancement for global lake science, environmental management and public health protection efforts.

4.2 Introduction

Water quality in lakes is a critical issue. The ecosystem services provided by lakes are innumerable, ranging from provision of drinking water, food supply, recreation, preservation of cultural heritage, and even healthcare (Heino *et al.*, 2021; Inácio *et al.*, 2022; Alejandre *et al.*, 2024). However, water quality in many lakes across the globe is poor, and this is exacerbated by climate change and unsustainable industrialisation (Naderian *et al.*, 2024). In this context, the United Nations Sustainable Development Goal 6 is focused on ensuring availability and sustainable management of water and sanitation across the globe by 2030, but unfortunately this is still out of reach for billions of people (Evaristo *et al.*, 2023). To meet this ambitious and incredibly important goal, a massive acceleration of progress in effective and sustainable water resource management is needed, and there is therefore a need to improve the way in which we manage, monitor

and understand lake ecosystems across the world (Rajapakse, Otoo and Danso, 2023). Globally, the vast majority of lakes are very small in size ($\sim 99\% < 1 \text{ km}^2$) (Cael and Seekell, 2016). Despite this, until recently, smaller water bodies have generally been considered of lower scientific importance than their larger counterparts, and so are somewhat less-well understood (Downing, 2010; Shen *et al.*, 2025). Now, there is a growing recognition that smaller water bodies are critically important for biodiversity, water quality, human use, and as early indicators of environmental changes (Biggs, von Fumetti and Kelly-Quinn, 2017). In these water bodies the aquatic primary producers, the phytoplankton, provision a large number of ecosystem services including oxygen production, nutrient cycling, climate regulation, and food supply (Naselli-Flores and Padisák, 2023). However, large accumulations of phytoplankton known as algal blooms can present severe ecological and public health risks and are thought to be increasing globally, likely due to eutrophication and climate change (Huisman *et al.*, 2018). Therefore, there is a clear need to further our understanding of, and ability to monitor, phytoplankton dynamics, particularly in smaller lakes which have been historically less well-studied.

In recent years, satellite remote sensing has emerged as a powerful method for monitoring water quality indicators such as chlorophyll-a (chl-a), a well-established proxy for algal biomass (Zahir *et al.*, 2024). Typically, algae monitoring with remote sensing is carried out on relatively large bodies of water such as the Great Lakes in North America (Beaulne and Fotopoulos, 2024). Historically this has been due to both the spatial limitations of satellite sensors, and the significant impact these water bodies have on regional ecosystems and economies (Sterner *et al.*, 2020). However, the ability to monitor smaller water bodies with satellite remote sensing would provide many benefits. Firstly, this would provide valuable data which could enable widescale monitoring of algal blooms and therefore enable better management and provision of the critical services which are threatened by harmful algae (Liu *et al.*, 2022). Additionally, there is a potentially huge benefit to the scientific community, as such data could enable global lake water quality monitoring studies of an unprecedented scale (Kutser and Soomets, 2024).

Various satellite sensors, such as Sentinel-2 Multispectral imager (MSI) (10 - 60 m spatial resolution, 5-day repeat cycle) and the Landsat 8/9 constellation (30 m spatial resolution, 8-day repeat cycle) have already started to further our understanding of phytoplankton dynamics and algal blooms in smaller water bodies (Zabaleta *et al.*, 2023). SuperDoves satellite constellation, which were launched sequentially between 2019 and 2023 by Planet

Labs Public Benefit Company (PBC) offer daily 3 m imagery with eight spectral bands in the visible and near infra-red (NIR) portion of the light spectrum (Planet Labs PBC, 2018). This presents a considerable spatio-temporal advantage over the existing satellite sensors which have been used for studying phytoplankton.

Previously, four-band Planet imagery has been used for chl-a estimation in a large (128 km²) lake (Niroumand-Jadidi and Bovolo, 2021), and to monitor chl-a in a number of medium to large sized (approx. 5 km² - 200 km²) reservoirs (Mansaray *et al.*, 2021). More recently, Vanhellemont (2023) evaluated the capabilities of eight-band SuperDove imagery for aquatic applications, demonstrating its potential for monitoring chl-a concentrations in a coastal environment. At present, there is one other study that has compared SuperDoves with imagery from another satellite sensor for monitoring chl-a in an inland water body. Wasehun *et al.* (2025) compared Sentinel-2 and SuperDoves for monitoring chl-a in a large (~ 60 km²) reservoir, and found that SuperDoves generally outperformed Sentinel-2. Specifically, it was found that the higher spatial resolution of SuperDoves was beneficial for resolving narrow geometries where larger areas of the Sentinel-2 imagery displayed errors due to edge-effects. Evidently, the SuperDoves show promise for monitoring algae in inland waters, but at present there are no studies which have specifically evaluated eight-band SuperDove imagery for monitoring algae in smaller water bodies (< 1 km²).

Additionally, whilst the higher spatial resolution of SuperDoves may be advantageous for phytoplankton monitoring, Niroumand-Jadidi *et al.* (2022) found that down-sampling SuperDoves data to 15 m resolution improved the performance of river bathymetry retrieval by increasing the signal to noise ratio. This has not been investigated for chl-a monitoring, and so it is uncertain whether the higher spatial resolution of the SuperDoves is truly beneficial. Furthermore, as the SuperDoves generally cover a narrower spectral region than Landsat or Sentinel-2, they potentially lack bands which are useful for atmospheric correction (AC), such as those in the short wavelength infrared region (SWIR), which can be used for glint correction over water (Steinmetz and Ramon, 2018). Dogliotti *et al.* (2024) found that SuperDoves tended to have higher AC errors than Landsat and Sentinel-2 when using default “land-based” AC approaches for water reflectance retrieval. However, as there are several AC algorithms that have been developed for inland waters (e.g. Soriano-González *et al.*, 2022; Steinmetz and Ramon, 2018; Vanhellemont and Ruddick, 2018), it would be beneficial to understand to what extent the wider range of AC algorithms available for a more well-established sensor such

as Sentinel-2 MSI provide an advantage over those presently available for the SuperDoves.

Generally, very few small water bodies are routinely monitored, highlighting the need for remote sensing approaches. However, this presents a challenge, as limited ground truth data are available for validation. A potential solution is to use data collected in a regulatory monitoring context alongside citizen science data. Therefore, this study focuses on two external sources of validation data: data collected by the Scottish Environment Protection Agency (SEPA), and citizen science data recorded through the mobile phone application *Bloomin' Algae*. The use of citizen science data can foster environmental stewardship and can increase public understanding of environmental issues (Pocock *et al.*, 2023). Alongside this, it has potential to provide wide coverage, and so can be used both to validate and complement satellite data for water quality monitoring applications (Putman *et al.*, 2023).

Therefore, the focus of this study is to understand if SuperDoves imagery and the currently available associated algorithms are suitable for monitoring algae and algal blooms in small inland water bodies. Specifically, the key hypothesis being tested is: given that SuperDoves have a similar band configuration to Sentinel-2, it is expected that with appropriate algorithm selection they may have similar performance for algae monitoring in small water bodies. This is significant given that the spatial resolution of SuperDoves data potentially enables water quality monitoring in a vast number of lakes which previously would have been considered too small for satellite monitoring. To ascertain if this is possible, we first make a comparison of SuperDoves with Sentinel-2 and associated algorithms. This comparison enables answering the following questions for a small inland water body: (1) do the presently available atmospheric correction algorithms for SuperDoves perform similarly to the best performing algorithms for Sentinel-2? (2) using the best working algorithms, can chl-a be retrieved from SuperDoves imagery with similar or better performance to Sentinel-2? (3) using citizen science and environment agency data for validation, do Sentinel-2 and SuperDoves perform similarly for algal bloom detection? (4) Does the higher spatial resolution offered by SuperDoves resolve features in the chl-a distribution more clearly than it would at lower resolution? Finally, to test the wider applicability of the SuperDoves, we investigate whether citizen science algal bloom records in five other small water bodies in central Scotland match the corresponding SuperDoves imagery.

4.3 Material and methods

4.3.1 Study site

With an area of 0.069 km², Airthrey Loch is a small, eutrophic lake in central Scotland (UK) with a maximum depth of less than 4.5 m. The water body is artificial, was constructed in 1788, and is now part of the University of Stirling campus (Kelly and Smith, 1996). During the spring and summer cyanobacterial blooms regularly occur which frequently result in closures to the recreational activities which the lake supports such as fishing and water sports. Whilst there is no routine water quality monitoring of Airthrey Loch, SEPA carry out phytoplankton counts a handful of times each year to assess the safety of the water for recreation upon request of the University. Airthrey Loch is large enough that most of its area is resolvable with Sentinel-2 imagery, however, it also features narrow sections which may require the higher resolution of SuperDoves to be properly resolved (Figure 4.1).

4.3.2 In situ data collection

4.3.2.1 *Reflectance measurements*

An automated hyperspectral above water reflectance measurement platform (In situ Marine Optics Dynamic Above-water Radiance and Irradiance Collector (DALEC-SN0005)) was deployed for 19 days in 2022 from 28th July – 15th August, and 85 days in 2023 from 8th July – 30th September. Data were only collected over 96 days due to instrument maintenance and power supply outages. The DALEC was configured to continuously measure for 4 hours each day, from 10:00 – 14:00 UTC to include both the SuperDoves overpass between 10:30 – 11:30 UTC, and Sentinel-2A/B overpass between 11:30-11:50 UTC. The DALEC automatically tracks the angle of the sun with a relative azimuth angle of 135 °, and was mounted on a steel pole 1 m above the water surface at a location approximately 25 m from the shore on the western side of the lake (latitude: 56.14701 °, longitude: -3.92347 ° - see Figure 4.1).

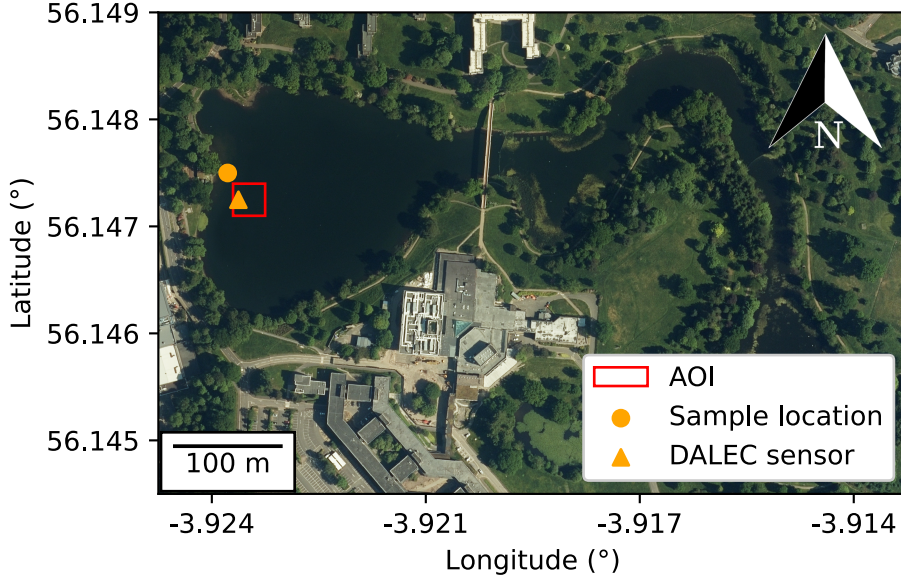


Figure 4.1: Aerial photograph of Airthrey Loch provided by Digimap®, © Getmapping Ltd. The orange triangle indicates the approximate location of the DALEC hyperspectral sensor, the orange circle the water sampling location, and the red square the 30 x 30 m area of interest (AOI) used for analysis.

It has three sensors (Figure 4.2) which independently measure L_u , the upwelling radiance, L_{sky} , the sky radiance, and E_d , the downwelling irradiance at approximately 3 nm intervals in the spectral range 380 – 1020 nm. Using these measurements, remote sensing reflectance (R_{rs}) values from the DALEC data were calculated with the following equation (Mobley, 1999):

$$R_{rs}(\lambda) = \frac{L_u(\lambda) - \rho L_{sky}(\lambda)}{E_d(\lambda)}, \quad (6)$$

where λ is the wavelength and ρ the skylight reflectance, which for all measurements, was assumed to be the standard value for low wind speeds ($< 5 \text{ m s}^{-1}$) of 0.028. Following the initial calculation of R_{rs} , the procedure established by Jiang et al. (2020) was used to remove residual skylight reflectance. Thirteen spectra with negative values, or extremely large reflectance values ($R_{rs}(580) > 0.04 \text{ sr}^{-1}$) were excluded from the data as they were assumed to be spurious measurements. The R_{rs} data from the DALEC sensor were then convolved using the spectral response functions of Planet SuperDoves and Sentinel-2A/B to allow for validation of the AC procedure.

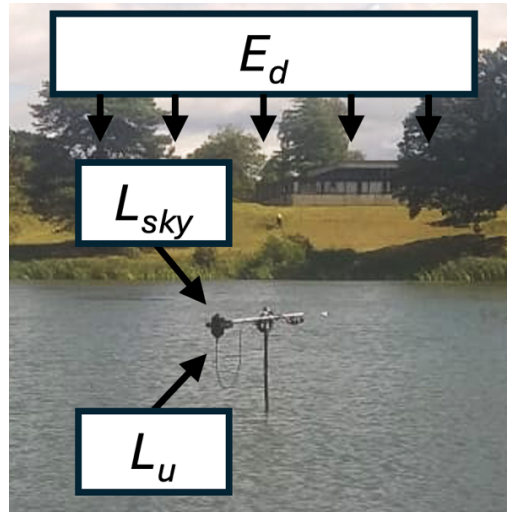


Figure 4.2: Photograph showing the DALEC sensor deployed on Airthrey Loch. The DALEC captures downwelling irradiance (E_d), indicated by multiple arrows pointing downwards; upwelling radiance (L_u), indicated by a single angled arrow (approximately 40° off nadir) pointing up from the water to the sensor; and sky radiance (L_{sky}), indicated by a single angled arrow (approximately 40° off zenith) pointing down at the sensor.

4.3.2.2 Chlorophyll-a measurements

In total, 25 water samples were taken from the study site over two periods: July – October 2022, and April – October 2023. Samples were taken at approximately fortnightly intervals to capture different stages of bloom development throughout the phytoplankton growing season. Where possible, clear, cloud free days were chosen for sampling to maximise the chances of suitable satellite match-up images. Samples were always collected between 10:00 and 13:00 UTC to match satellite overpass times. Sampling was done from the Southern edge of a jetty on the West bank of the loch which extends approximately 20 m away from the shore and is located close (~ 30 m) to the DALEC sensor (see Figure 4.1). Whilst the sampling location was approximately 30 m away from the AOI, any errors arising from this would affect both SuperDoves and Sentinel-2 match-ups, meaning that the resulting performance comparison is still meaningful. On the day of sampling, samples were filtered onto 25 mm GF/F glass microfiber filters with 0.7 mm pore size and stored at -80°C for a maximum of 4 months. Subsequently, analysis for chl-a was carried out using a High-Performance Liquid Chromatography (HPLC) method based on the approach described by Hooker et al. (2009). The generated chl-a data were checked for consistency and spurious values by examining the time series to ensure that the seasonal pattern expected for eutrophic water bodies in Scotland was present. Table 4.1 gives details on the number of chl-a match-ups obtained for each sensor.

Table 4.1: Number of match-ups obtained for reflectance measurements (DALEC), and chl-a samples for both sensors.

		SuperDoves	Sentinel-2
DALEC	Same day	24	10
	1 day difference	1	2
	Total	25 (15 separate days)	12
chl-a samples	Same day	7	5
	1 day difference	9	9
	Total	16	14

4.3.3 Satellite imagery

4.3.3.1 Collation and pre-processing

European Space Agency (ESA) Sentinel-2 (Level 1C) and Planet SuperDoves (8-band top of atmosphere reflectance) imagery covering the full spatial extent of Airthrey Loch from January 2021 through to April 2024 was downloaded from the Planet and Copernicus Dataspace repositories. A 30 x 30 m area of interest (AOI) centred 10 m east of the DALEC sensor location was selected for comparison with the DALEC measurements (Figure 4.1). This corresponds to 10 x 10 pixels for the 3 m SuperDoves data, and 3 x 3 pixels for the 10 m Sentinel-2 data. The choice to offset the centre of the AOI 10 m from the location of the DALEC was done to minimise adjacency effects, whilst still including several pixels in the direct vicinity of the DALEC.

For the SuperDoves data, the Planet useable data mask (UDM) was used to identify images with clouds over the AOI. Such images were not used for further analysis. For the Sentinel-2 data, the IdePix (Identification of pixel properties) masking tool was used (Wevers *et al.*, 2021). Sentinel-2 images with pixels in the AOI which were flagged as invalid, cloud, cloud shadow, snow/ice, bright, white, cirrus or bright white were not used for analysis.

4.3.3.2 Atmospheric correction

Several AC procedures for the generation of Level 2 surface reflectance products were tested. For SuperDoves data there are presently only two available AC procedures, and both were used for testing: the default Planet surface reflectance (SR) product, and Atmospheric Correction for Operational Land Imager (OLI) ‘lite’ (ACOLITE), which is designed specifically for aquatic applications (Vanhellemont, 2023). ACOLITE works by automatically selecting bands that are minimally influenced by surface reflectance and uses these for “dark spectrum fitting” to correct for atmospheric scattering and absorption

effects (Vanhellemont and Ruddick, 2018). For Sentinel-2, three AC approaches were tested: ACOLITE; Polymer, which uses a spectral matching approach to separate and correct atmospheric and sun glint signals from water leaving reflectance using polynomial and bio-optical modelling (Steinmetz and Ramon, 2018); and Case 2 Regional Coast Colour (C2RCC) which is a deep learning approach trained on a large database of simulated water leaving reflectances and corresponding top of atmosphere radiances (Soriano-González *et al.*, 2022). C2RCC and Polymer were selected as these have been shown to consistently outperform other Sentinel-2 atmospheric methods over inland waters (Warren *et al.*, 2019). ACOLITE was selected for comparison as it is the only procedure available for both SuperDoves and Sentinel-2. For C2RCC, three different neural network configurations were tested: case-2 waters in general (C2RCC_nets), extreme case-2 waters (C2X_nets), and complex extreme case-2 waters (C2X_COMPLEX_nets). All AC models were tested using the default settings, besides from the salinity parameters which were set to 0.3 ppt, a typical value for freshwaters in proximity to urban developments (Dugan *et al.*, 2017).

For both satellite imagery datasets, the median R_{rs} value over the AOI was calculated for each image and matched with the corresponding DALEC data. A matching window of ± 1 days was used to accommodate days with DALEC outages or those immediately before or after DALEC recording started. Matching windows of this size have been previously shown to be effective in obtaining a larger number of inland water match-ups without introducing significant error (Schröder *et al.*, 2024). Table 4.1 gives details on the number of match-ups obtained for each sensor. To evaluate the performance of the AC, the corrected satellite data were linearly regressed against the convolved DALEC reflectances. Slope, coefficient of determination (R^2), root mean square error (RMSE), median absolute percentage difference (MAPD), and mean average difference (MAD), were used as performance indicators. All algorithms were ranked according to each of these metrics and a combined total rank computed for each algorithm. The highest-ranking AC algorithms for both SuperDoves and Sentinel-2 data were identified and used for subsequent analyses.

4.3.3.3 Chlorophyll-*a* estimation

All of the algorithms used by Neil *et al.* (2019) and Vanhellemont (2023) which use bands available in both the Planet SuperDoves and Sentinel-2 data were chosen for testing. Several algorithms used by Neil *et al.* (2019) make use of the near infrared ratio in linear, quadratic and power functions. Preliminary tests found that the linear versions of these

algorithms performed best and so only these were considered for the formal analysis. This resulted in six separate algorithms which were used for formal testing:

1. Near infrared ratio (Gitelson, 1992):

$$Chla_{NIR} = a \times \frac{R_{rs}(708)}{R_{rs}(665)} + b. \quad (7)$$

2. Normalised difference chlorophyll index (NDCI) (Mishra and Mishra, 2012):

$$Chla_{NDCI} = a + b \times \left(\frac{R_{rs}(708) - R_{rs}(665)}{R_{rs}(708) + R_{rs}(665)} \right) + c \times \left(\frac{R_{rs}(708) - R_{rs}(665)}{R_{rs}(708) + R_{rs}(665)} \right)^2. \quad (8)$$

3. NASA Ocean Colour 3 band algorithm (OC3) (O'Reilly *et al.*, 1998):

$$Chla_{OC3} = 10^{(a+bX+cX^2+dX^3+eX^4)}, \quad (9)$$

where:

$$X = \log_{10} \left(\frac{R_{rs}(443) - R_{rs}(490)}{R_{rs}(560)} \right). \quad (10)$$

4. NASA Ocean Colour 2 band algorithm (OC2) (O'Reilly *et al.*, 1998):

$$Chla_{OC2} = 10^{(a+bX+cX^2+dX^3+eX^4)}, \quad (11)$$

where:

$$X = \log_{10} \left(\frac{R_{rs}(490)}{R_{rs}(560)} \right). \quad (12)$$

5. Red band difference (RBD) (Freitas and Dierssen, 2019):

$$Chla_{RBD} = a \times (R_{rs}(707) - R_{rs}(666)) + b. \quad (13)$$

6. Adapted version of the red band ratio (RBR3) (Vanhellemont, 2023):

$$Chla_{RBR3} = a \times R_{rs}(866) \times \left(\frac{1}{R_{rs}(666)} - \frac{1}{R_{rs}(708)} \right) + b, \quad (14)$$

where a , b , c , d , and e are the tuneable coefficients associated with each algorithm, and $R_{rs}(\lambda)$ is the remote sensing reflectance at the wavelength λ .

All in situ chl-a match-ups were used for tuning the coefficients. Tuning was done using least squares regression in log-space as log-transformed chl-a is generally normally distributed, and therefore more suitable for regression (Seegers *et al.*, 2018). Preliminary investigation found that after fitting, the two fourth order polynomial ocean colour algorithms often produced extremely large or negative chl-a values when applied to new imagery, suggesting overfitting (Lever, Krzywinski and Altman, 2016). Therefore, simplified linear versions ($c, d, e = 0$) of OC2 and OC3 were used to avoid this. As with the AC evaluation, the median R_{rs} values from the same 30 x 30 m grid were used. Algorithm performance was evaluated using root mean squared error (RMSE), median absolute percentage difference (MAPD), bias (mean average difference, MAD), slope and R^2 . All metrics were first calculated in log-space and then converted back to linear space as recommended by Seegers *et al.* (2018).

4.3.4 Bloom detection

To test the capability of the algorithms to detect algal blooms, records were compiled from data recorded by SEPA and the citizen science application *Bloomin' Algae*. Both sources included binary 'bloom' / 'no bloom' data. Specifically, SEPA record algal cell counts, where they define blooms to be cyanobacterial concentrations of more than 10^5 cells mL⁻¹. On request of the University of Stirling, SEPA generally do this analysis on samples taken in the aftermath of a large bloom, to check if recreational activities can be restarted. Therefore, the majority of the SEPA records were in the 'no bloom' category. *Bloomin' Algae* is a mobile phone application developed by the UK Centre for Ecology & Hydrology (UKCEH) which enables members of the public to report sightings of algal blooms. Blooms are reported by uploading photos with some brief contextual notes to an online database. All bloom reports are regularly reviewed by experts at UKCEH, and categorised as either 'confirmed', 'plausible', or 'incorrect'. For this study, only reports of blooms categorised as 'confirmed' were used. Together, these bloom reports allowed for testing of the ability of the SuperDoves and Sentinel-2 sensors to detect blooms in Airthrey Loch. Specifically, each of the bloom records was matched with the nearest satellite image, allowing ± 1 day for match-ups. The chl-a values were calculated for each of these images using the best performing chl-a algorithms previously identified for Sentinel-2 MSI and SuperDoves. The median value from the Western basin of the lake was used as all the bloom records used for this study were located at that side of the lake, as well as this being the in situ sampling location. This chl-a data was classified as either 'bloom' or 'no bloom' based on a threshold. For Airthrey Loch, there has been no work

which defines an algal bloom threshold based on chl-a levels, and so the threshold was varied from 0 to 55 $\mu\text{g L}^{-1}$. For each of these thresholds, the F1 score, a standard measure of classification performance, was calculated. Class imbalances were accounted for through weighting (Christen, Hand and Kirielle, 2023).

4.3.5 Spatial resolution analysis

To directly investigate the effect of spatial resolution on the retrieval of chl-a across the water body, a similar procedure to Li et al (2024) was followed. The 3 m resolution SuperDoves chl-a image acquired on 29 May 2023 was resampled to coarser spatial resolutions of 10 m and 20 m using average resampling. For each output pixel, the mean value of the contributing 3 m pixels was calculated, preserving the average chl-a concentration within the larger pixel footprint while reducing the spatial resolution. Histograms were then used to compare the overall distribution of chl-a values across the three spatial resolutions.

4.3.6 Bloom detection in other small water bodies

To investigate the capability of SuperDoves to detect algal blooms in other small water bodies, five small water bodies in central Scotland within a 35 km radius of Airthrey Loch were selected for additional analysis (Table 4.2). All water bodies selected were less than 1 km², and had at least one confirmed Bloomin' Algae bloom record. For each water body, the closest available SuperDoves image to a bloom report from either 2023 or 2024 was compared with an image from the end of January 2023. Algal blooms very rarely occur in Scottish water bodies at this time of year, and so these images act as a low-chl-a baseline (May *et al.*, 2022). For each image, the OC3 algorithm with the coefficients derived earlier in the study was used to visualise the chl-a distribution over each water body, and lake-median chl-a values calculated. Qualitative comparisons were made between the baseline and bloom images to examine the correspondence between Bloomin' Algae records and high chl-a concentrations measured using SuperDoves.

Table 4.2: Location, area and mean depth of water bodies used for additional analysis (Hughes et al., 2004).

Water Body	Latitude (°)	Longitude (°)	Area (km ²)	Mean depth (m)
Linlithgow Loch	55.98	-3.60	0.41	2.3
Beeccraigs Loch	55.95	-3.58	0.074	4.2
Loch Fitty	56.11	-3.42	0.62	2.3
Gartmorn Dam	56.13	-3.74	0.57	3.3
Castlehill Reservoir	56.21	-3.62	0.21	5.0

4.4 Results

4.4.1 Atmospheric correction

Most AC models performed better for the bands located between 490 and 666 nm for both MSI and SuperDoves sensors (Figure 4.3). In the majority of SuperDoves bands, ACOLITE products showed consistently higher slopes and biases than those derived from the Planet SR product, whereas the Planet SR product generally had lower RMSE and MAD (Figure 4.3, Figure 4.4). On average, these two AC algorithms ranked equally and so both were considered for subsequent analyses.

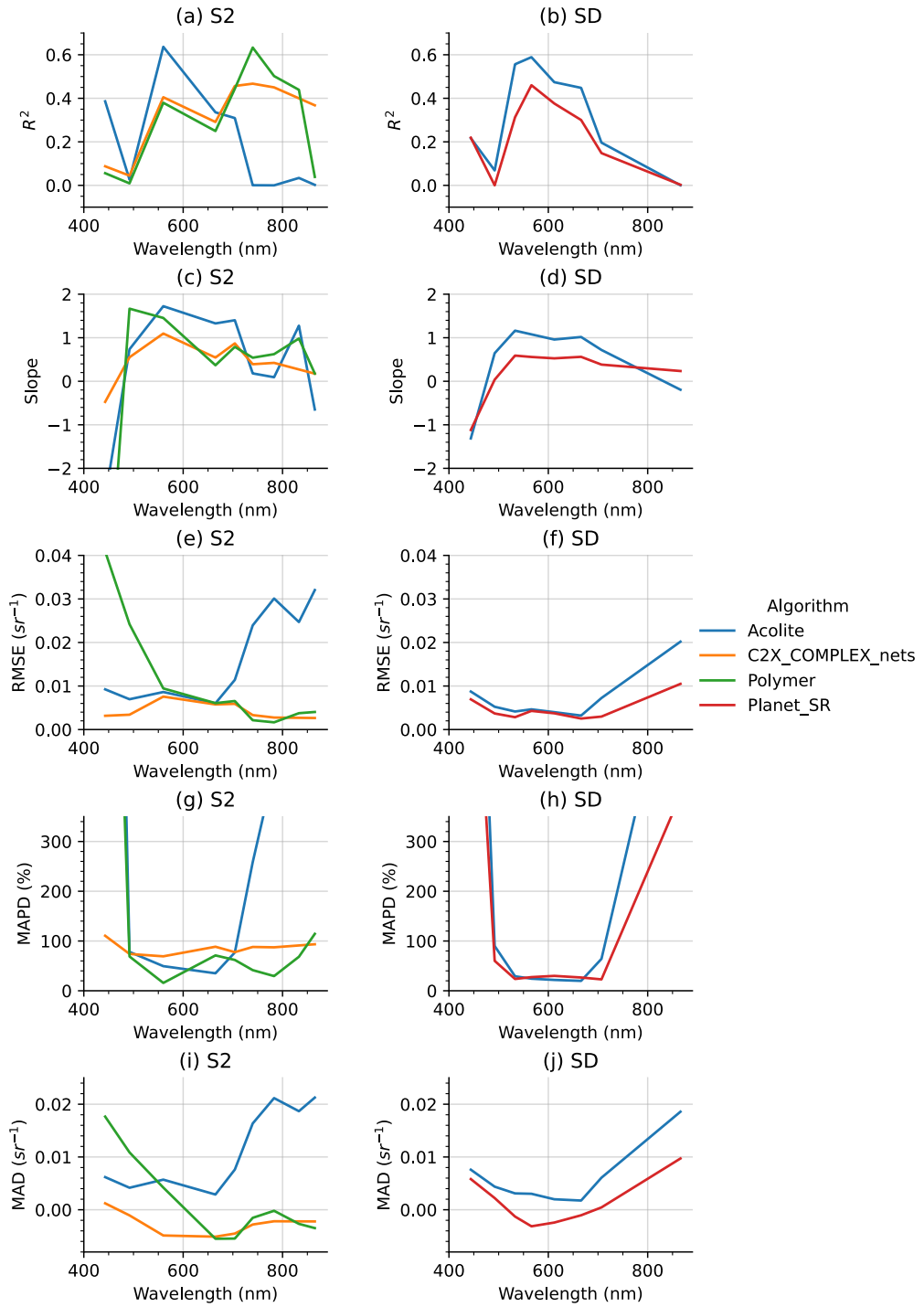


Figure 4.3: Atmospheric correction performance metrics: R^2 , slope, RMSE, MAPD, and MAD plotted against wavelength for both SuperDoves (right column) and Sentinel-2 (left column) match-up data. For clarity, only the best performing of the C2RCC algorithms (C2X Complex Nets) is shown. Blue lines indicate errors associated with ACOLITE, orange lines C2X Complex Nets, green lines Polymer, and red lines the planet surface reflectance product.

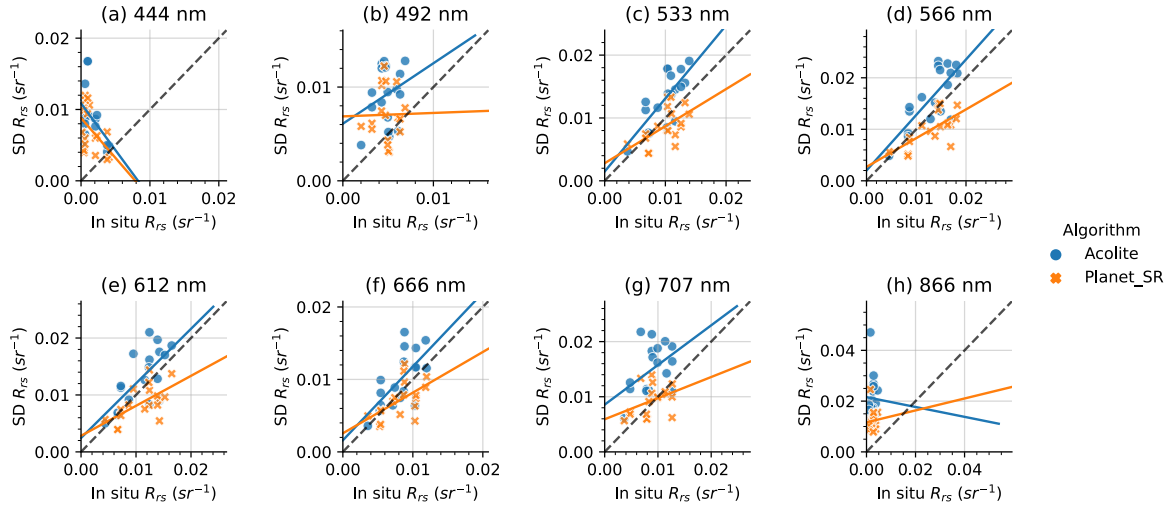


Figure 4.4: SuperDoves remote sensing reflectance plotted against in situ remote sensing reflectance measurements for the two tested atmospheric correction algorithms. Each marker represents a separate acquisition. Blue circles represent ACOLITE corrected data and orange crosses the Planet surface reflectance data. The solid coloured lines plotted were generated from ordinary least squares linear regression, and the black dashed line is the 1:1 line.

For the Sentinel-2 data, on average across all wavebands the highest-ranking algorithm was *C2X_COMPLEX_nets*, the C2RCC algorithm designed for complex waters. Whilst Polymer did outperform this algorithm at certain wavebands, *C2X_COMPLEX_nets* performed more consistently across all wavebands, and had very few outliers (Figure 4.3, Figure 4.5).

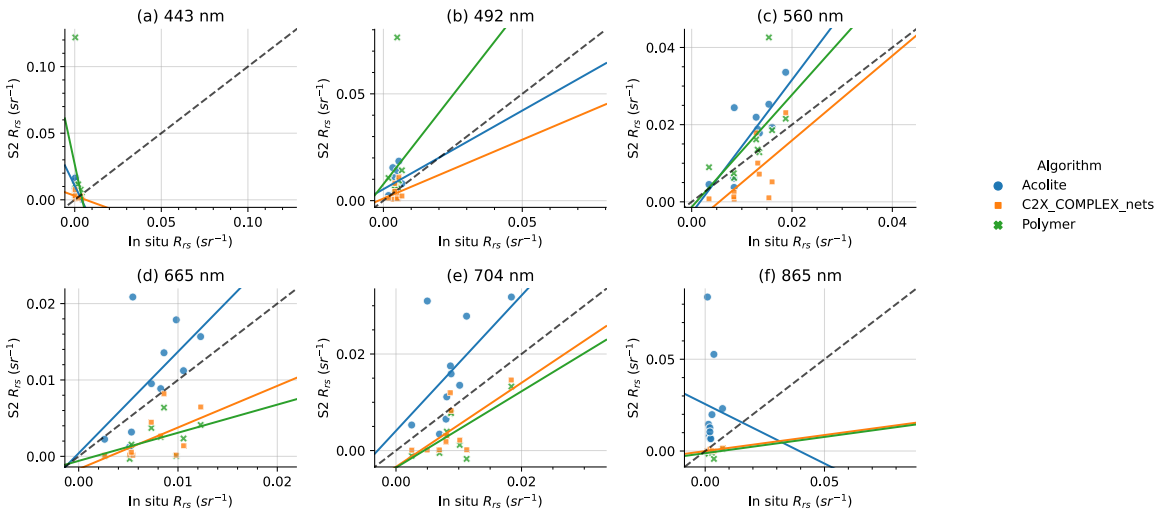


Figure 4.5: Sentinel-2 remote sensing reflectance plotted against in situ remote sensing reflectance measurements for the tested atmospheric correction algorithms. Each marker represents a separate acquisition. Blue circles represent ACOLITE corrected data, orange squares *C2X Complex Nets* corrected data, and green crosses are data corrected with Polymer. The solid coloured lines plotted were generated from ordinary least squares linear regression, and the black dashed line is the 1:1 line. For clarity, only the best performing of the C2RCC algorithms (*C2X_COMPLEX_NETS*) is shown.

4.4.2 Chlorophyll-a algorithm evaluation

For the Sentinel-2 data, all algorithms had statistically significant fits ($p \leq 0.01$) when regressed to the in situ data (Figure 4.6). In contrast, none of the chl-a algorithms evaluated with the Planet surface reflectance product had statistically significant fits ($p > 0.05$). For this reason, only the SuperDoves data corrected with ACOLITE was considered subsequently. Tuning to in situ data gave statistically significant fits ($p \leq 0.01$) and moderately high R^2 values (> 0.5) for all algorithms with Sentinel-2 data, whereas the only statistically significant tuned SuperDoves algorithms were *OC3* ($p \leq 0.001$) and *RBR3* ($p \leq 0.05$) (Figure 4.6). Out of these, only *OC3* had $R^2 > 0.5$. The best performing algorithm overall was the SuperDoves *OC3*, which had the lowest RMSE and MAPD of 0.93 mg L⁻¹ and 17.99 %, highest R^2 of 0.64, and slope of 0.64. The best performing Sentinel-2 algorithm was *NDCI*, which had the second highest R^2 of 0.61, slope of 0.61, and second lowest RMSE of 1.01 mg L⁻¹ overall.

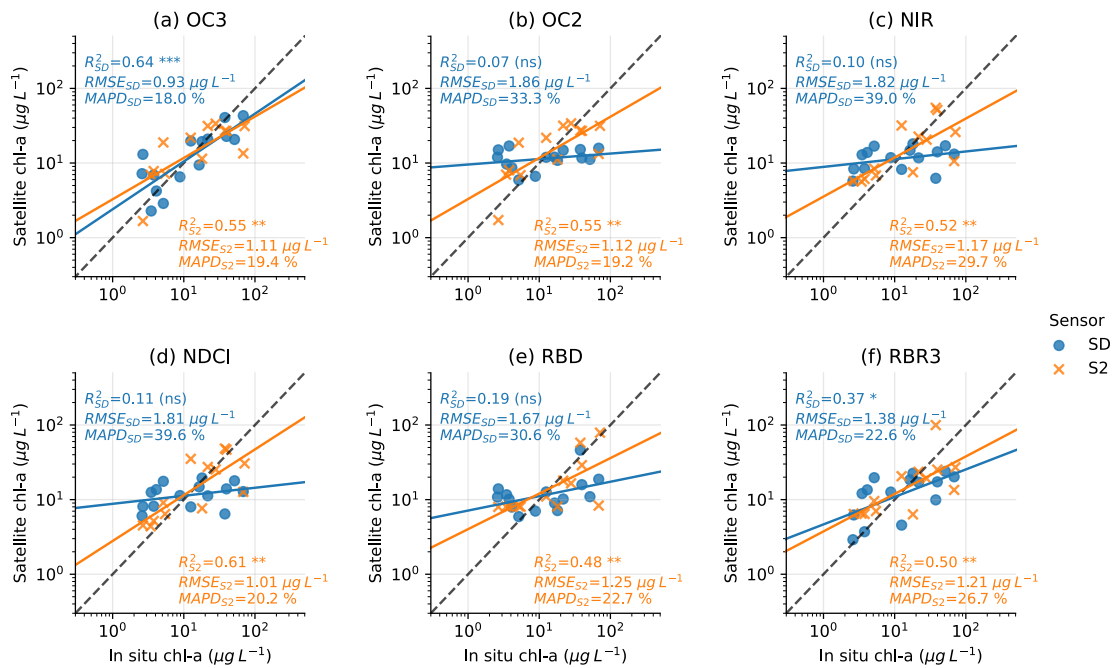


Figure 4.6: Satellite chl-a plotted against in situ measurements. Satellite chl-a was generated with the following algorithms: (a) OC3, (b) OC2, (c) NIR, (d) NDCI, (e) RBD, and (f) RBR3. Sentinel-2 data are indicated by orange crosses and SuperDoves by blue circles. Black dashed lines are the 1:1 lines, and the solid lines are the lines of best fit generated from log-log least squares regression. RMSE, MAPD and R^2 values from log-space linear regression are displayed for each sensor, alongside an indication of the statistical significance level: (ns) - $p > 0.05$; * - $p \leq 0.05$; ** - $p \leq 0.01$; *** - $p \leq 0.001$.

4.4.3 Bloom detection

The algal bloom history of Airthrey Loch since 2021 was visualised using the full timeseries of SuperDoves *OC3* and Sentinel-2 *NDCI* chl-a data (Figure 4.7). As expected

of the higher temporal imagery, SuperDoves had more cloud free images during the study period than Sentinel-2. Between January 2021 and May 2024 there were a total of 191 cloud free SuperDoves images of the loch which equates to, on average, 56 per year. For Sentinel-2 there were 87 cloud free images, equating to approximately 25 per year. In 2021 there was some discrepancy between Sentinel-2 and SuperDoves, which is particularly obvious in July. However, from 2022 onwards both sensors show similar trends. The combined *Bloomin' Algae* and SEPA bloom records consisted of 10 reports on separate dates, six of which were labelled as blooms. All six of these dates correspond to peaks in the chl-a time series generated from SuperDoves and Sentinel-2 data (Figure 4.7). Likewise, all the 'no bloom' records corresponded with lower chl-a values.

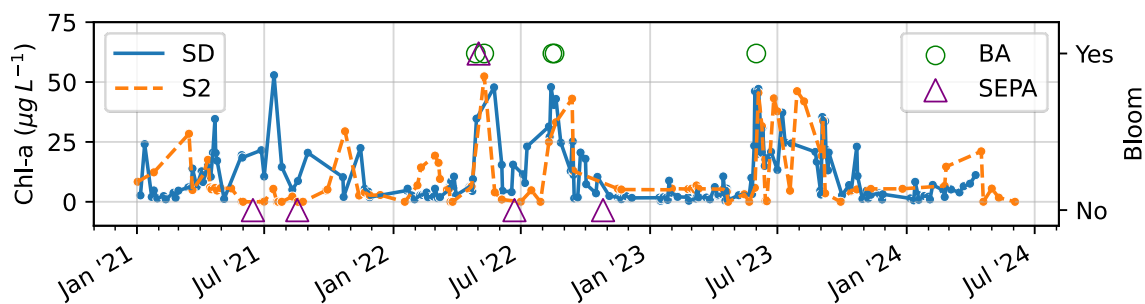


Figure 4.7: Median chl-a time series from Western basin of Airthrey loch generated with SuperDoves OC3 (solid blue line) and Sentinel-2 NDCI (dashed orange line). Purple triangles (SEPA records) and green circles (Bloomin' Algae data) indicate bloom presence if placed in the upper section of the plot, and bloom absence if placed in the lower section.

The clear separation in chl-a values between 'bloom' and 'no bloom' events is reflected in the F1 scores (Figure 4.8). For SuperDoves, a perfect score of 1.0 was obtained with the bloom threshold set between 20 – 30 $\mu\text{g L}^{-1}$, and for Sentinel-2 this was obtained in the 10 – 30 $\mu\text{g L}^{-1}$ range. Performance is still very strong with F1 score > 0.85 if the threshold is set between 10 – 45 $\mu\text{g L}^{-1}$. This indicates that both sensors, with their associated best-performing atmospheric correction and chl-a algorithms, could detect the presence and absence of algal blooms reported by SEPA and citizen scientists in the Western basin of the lake.

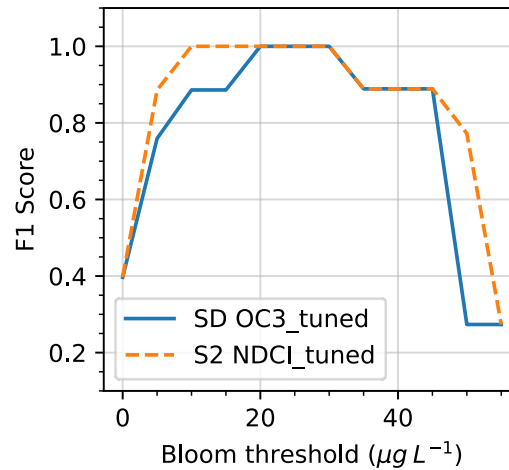


Figure 4.8: F1 Scores from bloom detection task for SuperDoves OC3 (blue solid line) and Sentinel-2 NDCI (orange dashed line) across a range of chlorophyll-a bloom thresholds.

4.4.4 Spatial analyses

The difference between low and high chl-a states is evident when comparing images from mid-May to early June (Figure 4.9). All sensors clearly depict this variation, with the early June images showing higher chl-a concentrations. There is a noticeable correlation in the spatial pattern of the bloom observed in early June, where the highest concentrations, excluding shorelines, tend to be in the northwest part of the lake. However, estimates from Sentinel-2 NDCI imagery using the C2X_COMPLEX_nets algorithm are significantly lower compared to those from the SuperDoves OC3 imagery. In the images from the 4th June, the chl-a distribution in the narrow section of the Eastern side of the lake is very different in all three of the chl-a maps. The Sentinel-2 NDCI image corrected with C2X_COMPLEX_nets shows extremely low chl-a values, whereas the Sentinel-2 OC3 ACOLITE-corrected image is dominated by very high chl-a in this part of the water body. In contrast, the corresponding SuperDoves image shows a more discernible pattern: higher chl-a concentrations largely accumulated on the Western banks of the narrow section of the loch, with lower chl-a on the opposite bank. Notably, at the narrowest points, this section of the lake is approximately only 30 m wide, which corresponds to ten SuperDoves pixels, but only three 10 m Sentinel-2 pixels. Furthermore, the Sentinel-2 443 nm and 705 nm bands have lower resolutions of 60 and 20 m respectively. This has the consequence that parts of this narrow section are likely not properly resolvable with Sentinel-2 using NDCI and OC3 as the lower resolution bands are used for chl-a retrieval. In addition to the apparent issues with the narrow section of the loch, the Sentinel-2 NDCI image corrected with C2X_COMPLEX_nets from the 4th of June has a distinctive adjacency-effect in the Western basin. The perimeter of the basin has a narrow ring of

very high chl-a values which are encircled by a ring of very low chl-a values around the perimeter of the basin.

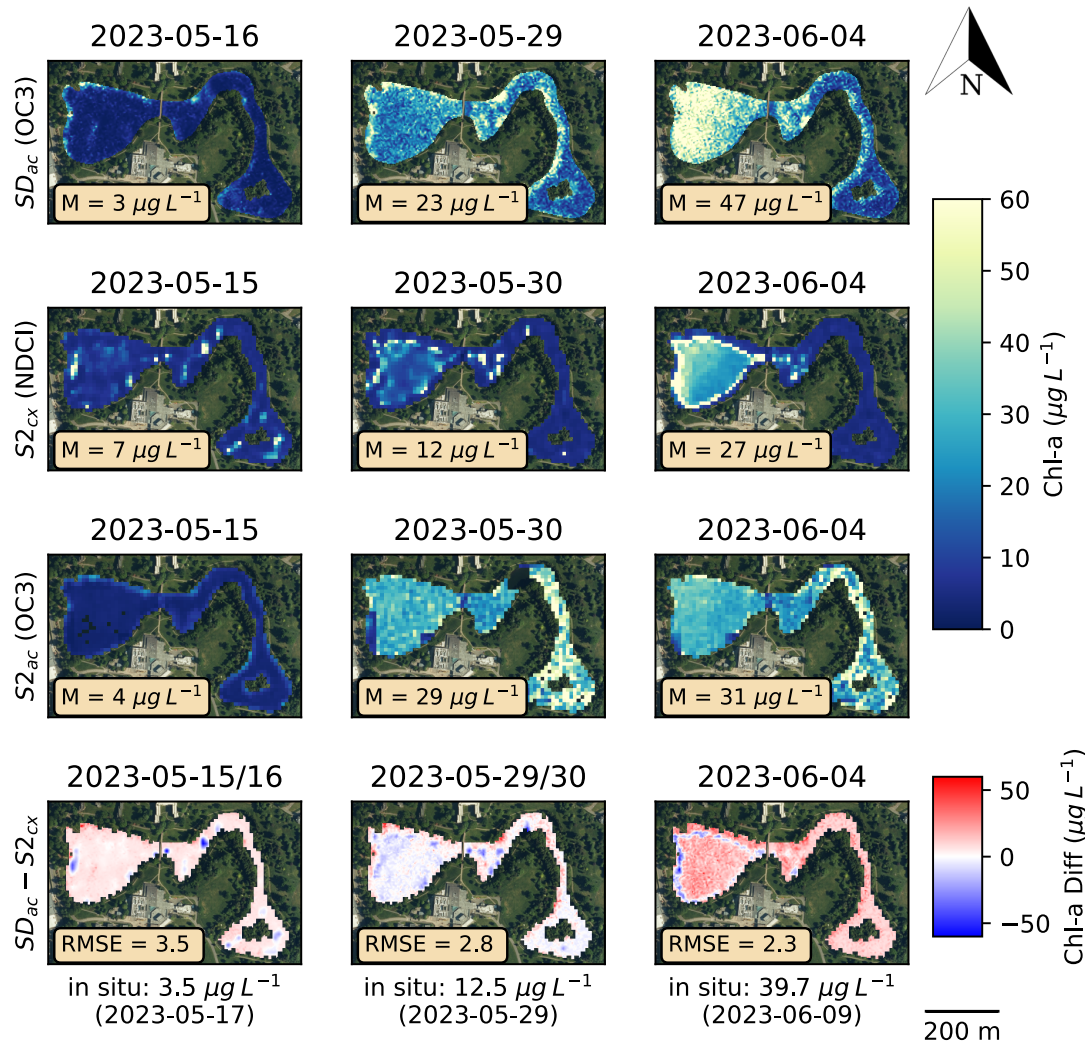


Figure 4.9: Satellite chl-a maps generated with SuperDoves ACOLITE corrected OC3 (top row), Sentinel-2 C2X_COMPLEX_nets corrected NDCI (second row from top), Sentinel-2 ACOLITE corrected OC3 (second row from bottom). The bottom row shows the difference images between the first two rows. Image median (M) values are given for each of the chl-a plots and RMSE values are given for the difference images.

Analysis of the SuperDoves image from 2023-05-29 showed clear differences between the distribution of chl-a values when the image was down-sampled from 3 m to 10 m and 20 m spatial resolutions (Figure 4.10). Notably, at 10 m resolution there are clear peaks at approximately 15 and $20 \mu g L^{-1}$, whereas in the 3 m resolution image, the distribution in this region is more even. Similarly, the high values of chl-a at 20 m resolution are exclusively located at approximately 40 and $60 \mu g L^{-1}$, whereas high values in the 3 m image are distributed throughout this region.

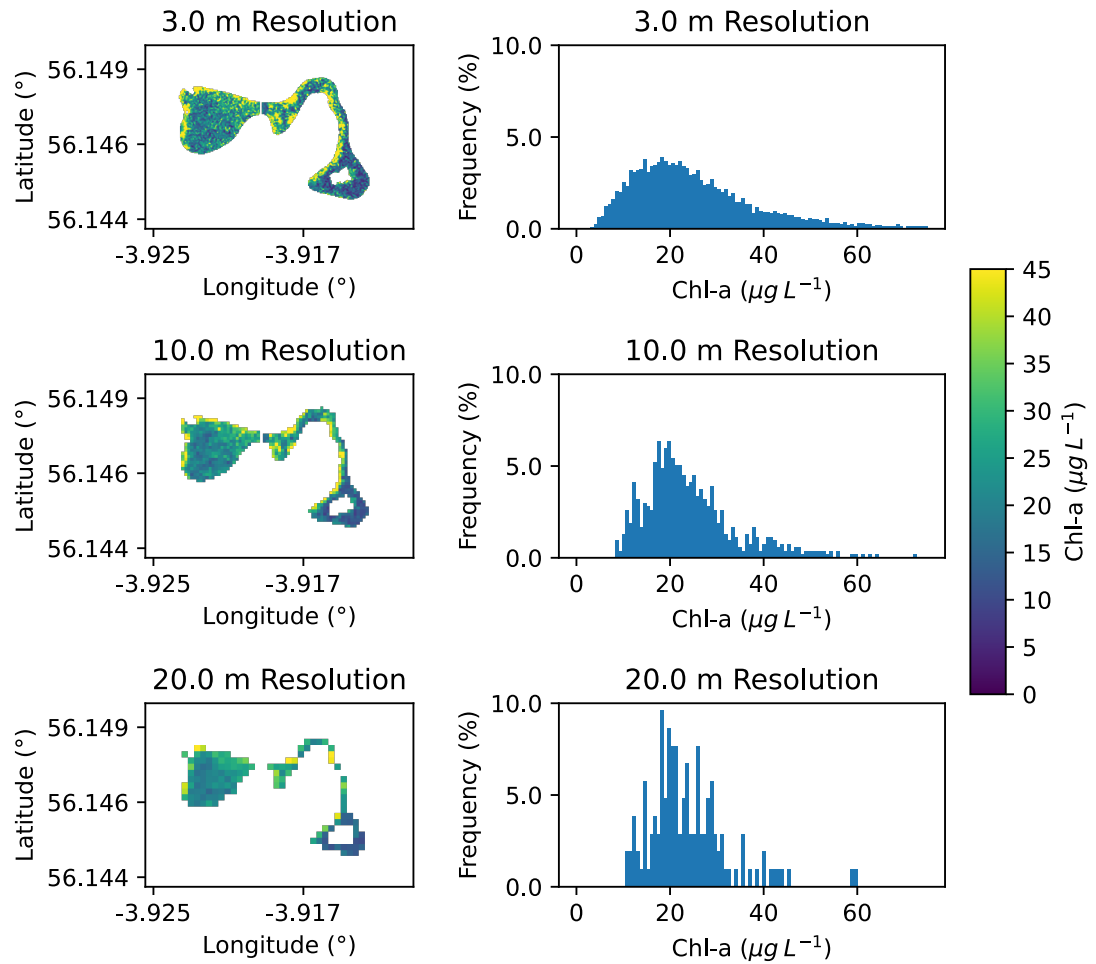


Figure 4.10: Comparison of SuperDoves chl-a image from 2023-05-29 at three different spatial resolutions (top: 3.0 m, middle: 10.0 m, bottom: 20.0 m). Histograms show the frequency of chl-a values over the entire image in each scenario.

4.4.5 Bloom detection in other small water bodies

For all five water bodies there was a clear distinction between the baseline images from January 2023 and those acquired either shortly before or after confirmed Bloomin' Algae bloom records (Figure 4.11). Castlehill Reservoir and Linlithgow Loch showed the largest median chl-a differences of 25 and 27 $\mu\text{g L}^{-1}$ between the bloom and no-bloom images. In contrast, the differences between bloom and no-bloom images in Loch Fitty and Gartmorn Dam were lowest - both 12 $\mu\text{g L}^{-1}$. Nonetheless, all bloom images show localised chl-a features of above 30 $\mu\text{g L}^{-1}$, which are not present the baseline images.

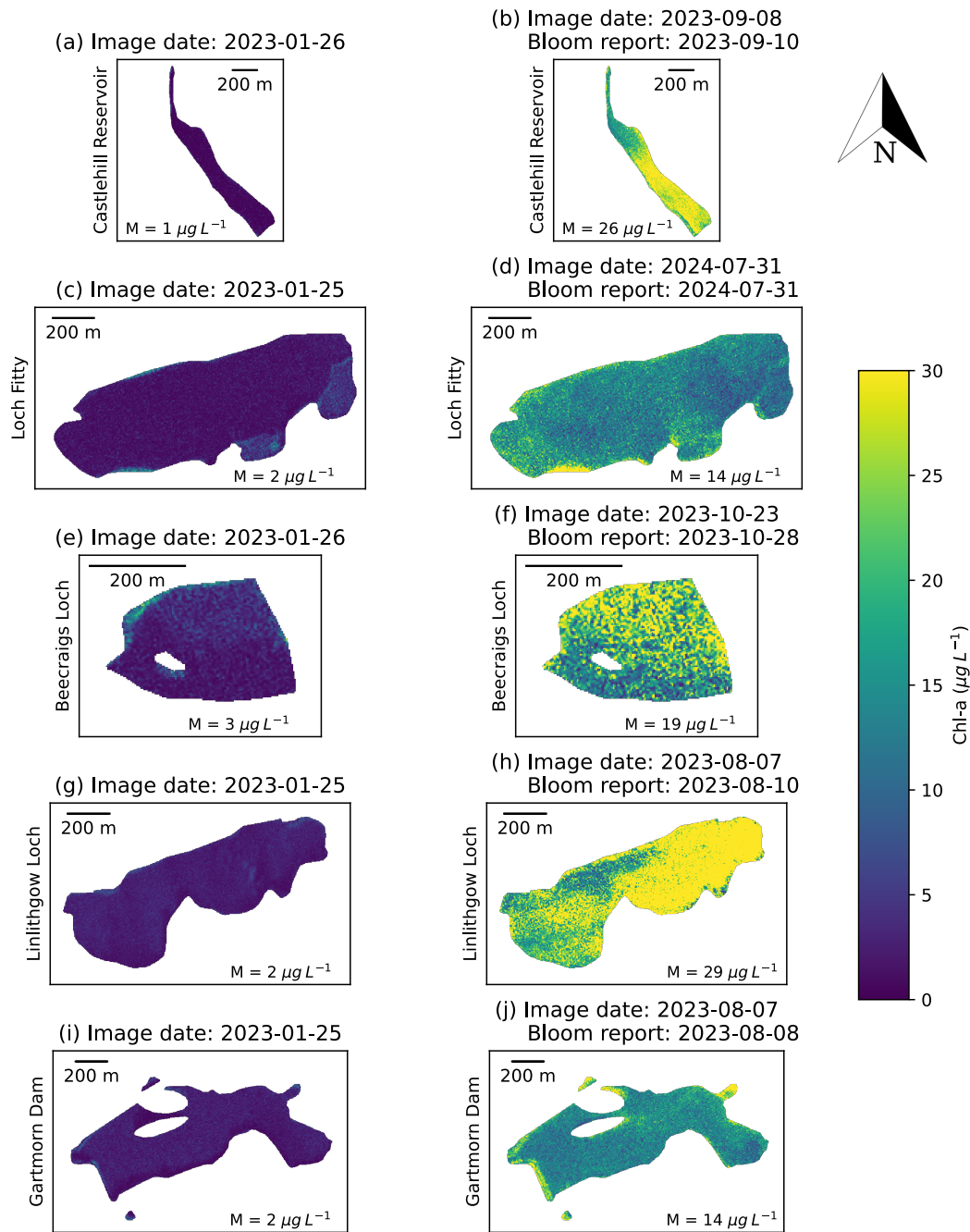


Figure 4.11: Comparison of Superdoves OC3 chl-a images from five small water bodies in central Scotland. The left column shows images from late January, and the right column shows images captured on dates close to Bloomin' Algae bloom reports. Median chl-a values for each image are also shown.

4.5 Discussion

4.5.1 Atmospheric correction

At present, algorithms for AC over water and retrieval of water quality parameters such as chl-a have not been widely tested for eight-band SuperDoves imagery. At the time of writing, this has only been done in the present study and by Vanhellemont (2023) for a coastal environment. There are currently only two available AC algorithms for SuperDoves, and of these, only ACOLITE is intended for aquatic applications. Thus, there

were notable differences in the performance of AC algorithms available for Sentinel-2 and SuperDoves. Notably, where ACOLITE suffered from elevated errors at higher wavebands above 700 nm, the C2X Complex Nets algorithm generally performed more consistently across the spectrum (Figure 4.3). This discrepancy in error across wavelengths provides a plausible explanation as to why the NIR, RBD and NDCI algorithms (that use the ratio or difference between the 666 and 707 nm bands) were not sensitive to changes in chl-a.

The finding that the ACOLITE-corrected Sentinel-2 data also showed a similar error pattern across wavebands suggests that this issue may be partly due to AC limitations. However, it is likely that this error may have other sources such as the relatively low signal-to-noise ratio of SuperDoves (Valenzuela, Reinke and Jones, 2024). This is a general issue for microsatellite constellations given their small size and limited power (Siddique, 2024). Furthermore, the correspondence between in situ and satellite R_{rs} in the 443/444 nm band was poor, with negative correlations in both SuperDoves and Sentinel-2 data. This is not unusual - it is very common for AC procedures to perform poorly over coastal and inland waters at the blue-end of the visible spectrum (Lv, Feng and Zhao, 2024). Typically this is due to difficulties in estimating the contribution of strongly absorbing aerosols at these wavelengths (Kahn *et al.*, 2016). In comparison, Vanhellemont (2023) found that the ACOLITE correction was effective across all eight wavebands. Notably, the errors in the 444 nm, 707 nm and 866 nm bands were significantly lower than those found in the present study. This discrepancy could be due to dissimilarities between the study sites (coastal compared with inland water), and possibly differences in the atmospheric conditions associated with each of these (Pahlevan *et al.*, 2021). However, for another coastal environment, Dogliotti *et al.* (2024) also found issues with AC of SuperDoves in the 444 and 866 nm bands. A likely source of this error is sun glint, which can be corrected using SWIR (900 – 1700 nm) bands which SuperDoves lacks (Lavigne *et al.*, 2023). Whilst NIR bands have been used successfully for glint correction of SuperDoves imagery over clear waters, further developments are required to improve this capability for turbid waters (Vanhellemont, 2023).

4.5.2 Chlorophyll-a retrieval and algal bloom detection

Despite challenges associated with AC, with the best algorithms, the SuperDoves chl-a retrieval error was generally lower than that of Sentinel-2, indicating the suitability of this data for phytoplankton monitoring in small water bodies such as Airthrey Loch. Retrieval of chl-a in this study was inevitably affected by errors arising from various sources.

Firstly, there was a consistent discrepancy between the location of in situ sampling and the satellite pixels that this was compared with. Furthermore, the timing of sampling did not consistently match with satellite acquisition times – many chl-a samples were taken approximately 24 hours before or after the corresponding satellite acquisition. Whilst this is an important error source, both sensors had similar numbers of same-day and one-day difference match-ups (Table 4.1). Therefore, whilst this has the consequence of a chl-a retrieval uncertainty comparable, if not greater, in magnitude to that of the average variation in ± 24 hours, the sensor comparison is nonetheless valid. Furthermore, using this approach, we were able to use both sensors to detect algal blooms recorded by SEPA and Bloomin’ algae in a reliable way across a range of chl-a thresholds. An obvious issue with defining algal blooms using a simple metric such as a chl-a threshold is that this does not necessarily correlate well with qualitative records such as the Bloomin’ Algae records which are verified using user-uploaded photos of cyanobacterial scums. Given that these scums can sometimes be relatively small in area, it is possible for reports of blooms to correspond with images with a relatively low median chl-a. Whilst this did not appear to be an issue for the Airthrey Loch data that was analysed, this was apparent for some of the images of other water bodies. Specifically, Loch Fitty and Gartmorn Dam both had relatively low median chl-a measurements close to dates corresponding to Bloomin’ Algae bloom reports (Figure 4.11).

Reliably detecting smaller scums near to water body edges using satellite imagery is challenging, because emergent vegetation are commonly found in marginal areas, and can contribute to the chl-a signal detected by a satellite sensor (Qing *et al.*, 2020; Jiang *et al.*, 2023). For cyanobacteria monitoring applications, it can be beneficial to make use of the phycocyanin absorption peak at ~ 620 nm, a spectral feature specific to cyanobacteria (Beck *et al.*, 2017). This enables distinguishing cyanobacterial scums from emergent vegetation or other chl-a sources such as green algae. Whilst Sentinel-2 does not have a band configuration which supports this, SuperDoves have a band centred at 612 nm and so could potentially be used for high-resolution cyanobacteria-specific monitoring.

4.5.3 Spatial resolution

The spatial resolution analysis of the SuperDoves chl-a data over Airthrey Loch showed that the difference in resolution between SuperDoves and Sentinel-2 is important (Figure 4.10). Specifically, down-sampling SuperDoves from 3 m to 10 m or 20 m, resulted in chl-a distributions with notable differences, including higher occurrences of lower to moderate values. Visually, this appears as a blurring of the patches of higher chl-a concentration,

which could prevent detection of smaller scums, pointing to a clear benefit of the higher resolution imagery for algal bloom monitoring applications. Whilst higher resolution imagery typically has a lower signal to noise ratio (Niroumand-Jadidi, Legleiter and Bovolo, 2022), this is likely less important for detecting high concentrations of algae, where the chl-a signal is typically strong. Therefore, for algal bloom detection and tracking, the higher spatial resolution offered by the SuperDoves is beneficial. Furthermore, the lower resolution Sentinel-2 NDCI data showed some apparent adjacency-effects in the wider basin. The very low chl-a values at the edge of the water body in the Sentinel-2 NDCI images likely included 20 m NIR pixels (705 nm) which overlap both water and land. For such pixels, high NIR reflectance would be expected as the shoreline of the lake has significant vegetation cover (Huang *et al.*, 2021). With the NDCI algorithm, high NIR reflectance should produce high chl-a values, and therefore the low chl-a values indicate an over-correction in the NIR by the C2X_COMPLEX_nets algorithm. Whilst these are only issues for a relatively small portion of Airthrey Loch, they would likely be more prevalent for smaller lakes where narrow geometries, edge-adjacent pixels, and mixed pixels tend to make up a larger portion of the total area (Jiang *et al.*, 2023). Therefore, our results corroborate and extend upon those of Wasehun *et al.* (2025), indicating that a key advantage of the higher spatial resolution SuperDoves data is that this reduces distortion in the distribution of measured chl-a values and reduces the extent of adjacency-effects.

4.5.4 Wider applicability to other small water bodies

The finding that gross changes in water quality associated with the presence and absence of algal blooms were detectable in several small water bodies indicates promise for the wider applicability of SuperDoves data for algae monitoring. Still, there are various sources of uncertainty associated with this analysis which are difficult to quantify without additional in situ data. For example, in shallow water bodies bottom effects may contribute an unwanted component to the chl-a signal (Ma *et al.*, 2011). Given that most of the water bodies considered in this study were relatively shallow, this was likely a source of error for them all (Table 4.2). With sufficient in situ chl-a data, it is possible to correct for bottom effects and other errors (Zhang *et al.*, 2018), however this poses a specific challenge when considering using high-resolution satellite imagery for widespread water quality monitoring: given the sheer number of small water bodies, gathering sufficient data for validating phytoplankton monitoring schemes and algal bloom detection systems is potentially difficult. Furthermore, routine monitoring programs

tend to be reserved for larger, more economically significant lakes (Biggs, von Fumetti and Kelly-Quinn, 2017). However, the efficacy of using irregular records from reactive environmental monitoring and citizen science to validate satellite monitoring (Figure 4.7, Figure 4.11) suggests that this approach could be used more widely than currently for comparing with remotely sensed data. Satellite imagery can suffer from errors and missing data due to atmospheric effects, cloud cover, confounding (non-algal) chl-a sources, and adjacency effects (Zahir *et al.*, 2024). Likewise, citizen science data are likely to be collected somewhat irregularly, possibly with inconsistencies, and uncertainty arising from human error and biases (Pocock *et al.*, 2023). However, as their sources of error are very much independent from each other, together these data complement each other very well. Therefore, further exploring ways to use these data alongside each other would offer a richer, more complete way of monitoring and understanding aquatic ecosystems than either would be capable of alone (Putman *et al.*, 2023).

Alongside spatial resolution, another clear advantage of the SuperDoves data is that of temporal resolution. Given that SuperDoves images are acquired daily, this imagery could be expected to give approximately a five-fold increase in the number of useable images compared with Sentinel-2. However, only slightly more than twice as many useable images were obtained with SuperDoves. This is largely a reflection of a challenge associated with using optical remote sensing data at high latitudes (e.g., central Scotland) where significant cloud cover is common throughout the year (McGrath *et al.*, 2023). On average, approximately one cloud-free SuperDoves image was obtained each week. Although a significant improvement on the Sentinel-2 data, this is far from the daily observations which Dubelaar *et al.* (2004) suggest are required to follow ecosystem dynamics of phytoplankton in detail. Therefore, for regions like Scotland with consistent and regular cloud cover, higher frequency monitoring data would likely be best obtained using automated in situ measurement platforms (Marcé *et al.*, 2016). However, even in areas where satellite temporal coverage is limited by cloud cover, there is still the potential to collect far more data than would be feasible with manual sampling efforts. Whilst there were some errors and discrepancies between the SuperDoves and Sentinel-2 chl-a estimates, we were still able to detect gross changes in water quality. Therefore, if applied on a larger scale, this capability could be of significant use to assist regulatory agencies with targeting sampling efforts to reduce operational costs.

Finally, it is worth noting that Planet SuperDoves data are not freely available, except for small-scale research projects, such as the present study (Planet Labs PBC, 2018). Whilst

Planet have various pricing structures available, including discounts for non-profit organizations, it is still possible that certain projects such as global lake studies or implementation of widespread algal bloom monitoring programs might be prohibited by cost. It is therefore very likely that freely available products such as Sentinel-2 and Landsat will continue to be widely used for water quality monitoring applications even as higher resolution imagery becomes increasingly attractive.

4.6 Conclusion

This study highlights several key findings and challenges associated with using the Planet SuperDoves satellite constellation for monitoring phytoplankton in a small, eutrophic lake. Despite some challenges, it was found that SuperDoves data are competitive with Sentinel-2 for this application and have distinct advantages arising from higher spatial and temporal resolution. We have shown that the higher spatial resolution offered by SuperDoves can enable monitoring of a vast number of smaller lakes, which has important implications for the field of limnology, environmental monitoring, and public health protection efforts. Furthermore, it has been demonstrated that alongside irregular environmental agency records, citizen science data can be a valuable complement to satellite imagery. This points to a low-cost means of validating widespread monitoring data, whilst also generating an increased sense of shared environmental stewardship in society. Ultimately, while high-resolution satellite data like SuperDoves offer great potential for monitoring phytoplankton dynamics, they should complement, rather than replace other data collection methods. Such a combined approach could enable monitoring of a huge number of unstudied lakes, and therefore provide critical data to improve the management and conservation of small, sensitive aquatic ecosystems at a scale previously impossible.

5 Discussion

The chapters presented in this thesis further knowledge in two key areas. The forecasting-focused studies (Chapters 2, 3) highlight considerations relevant to the design of monitoring systems to generate suitable training data for machine learning forecasting models. Furthermore, the insights presented in these studies provide a perspective on how we can expect data-driven models for algae and, more generally, water quality forecasting to improve as additional training data are collected over time. Additionally, Chapter 3 highlights challenges which may arise for water managers attempting forecasting in scenarios where climate warming continues at a rapid rate. In contrast, the remote sensing focused study (Chapter 4) takes a different angle as it evaluates the potential for using high-resolution cubesat imagery to monitor lakes in a more widespread and frequent way than has been previously possible. The demonstration that the Planet SuperDoves satellite constellation is potentially a powerful tool for algal monitoring in smaller water bodies is an important result, which could signal future paradigm shifts for inland water quality monitoring and forecasting practices, particularly for smaller, less well-studied lakes.

The inherent synergies of data-driven water quality forecasting and inland aquatic remote sensing are evident. Data-driven forecasting is reliant on data, and as such, developments in satellite remote sensing technologies may be key to providing data at the scale needed to make the best use of more advanced and complex machine learning architectures. Such capability is particularly important for areas where in situ data collection is infrequent or sparse. Therefore, by considering the combined relevance of all three of these studies it is possible to reflect on the possibilities which this satellite-ML synergy may unlock in the future. Similarly, these studies also point clearly towards a number of challenges and important messages relevant to the remote sensing, ecological forecasting, and water management communities. The sections of this discussion Chapter therefore reflect the central themes that emerged from the research detailed within this thesis.

5.1 Rising above the plateau – is it possible to further increase forecast performance?

5.1.1 Forecast performance plateaus

In both forecasting chapters (Chapters 2, 3), it was found that beyond a certain number of training years, the performance gains from adding additional years of training data tended to diminish, resulting in error curves which plateaued. In Chapter 2, where multiple

models were trained using up to 20 years of in situ training data, this plateau was very apparent for all models trained with more than 10 years of data. Importantly, whilst the best performing model outperformed both naïve benchmark models with only four years of training data, the presence of the plateau shortly beyond this point meant that, even in high data-availability scenarios, ML forecasting was, at best, only moderately better than the naïve persistence forecast. Importantly, this might not be the case for all lakes – many may have phytoplankton dynamics more predictable than Blelham Tarn, and so for these, ML forecasts would likely outperform naïve benchmarks by a much greater margin given a similar volume of training data. Nonetheless, being a relatively small, fertile temperate lake, Blelham Tarn is by no means unusual, and so there are likely a large number of lakes with a similar, or lower predictability. Therefore, it is important to consider potential ways to significantly improve performance beyond this plateau.

A similar trend was found in the third Chapter, which used simulated data, however the plateau was not as strong, and displayed clear, albeit minor, improvements in performance between models trained with twenty years' data versus those trained with thirty years' data. Nonetheless, if these data were extrapolated, the resultant curves would likely show effectively negligible performance improvements beyond 40-50 years' training data.

As previously discussed in Section 2.5.1, there are several possible causes for these plateaus. One important potential cause is that forecasts may be hindered by older data that are not sufficiently representative of the present system to be particularly useful for training forecasting models. This was explored in the third chapter by investigating the interaction between training dataset duration and climate scenarios. This experiment effectively investigated whether older training data would be less useful in more rapidly changing climate scenarios, and it was found that whilst there was a significant interaction, the magnitude of this effect was dwarfed by other factors. This suggests that in this case, older training data were approximately equally useful, regardless of how intensely the climate had changed over the course of the training dataset period. Still, it is unlikely that this finding would apply to algal concentrations where dramatic regime shifts may fundamentally change the ecosystem structure, and therefore the equations needed to model it (Tay *et al.*, 2021; Zhang *et al.*, 2021). Therefore, this is still a very likely cause of the algae forecast performance plateau seen in the second chapter, and could be a reason that this plateau is stronger than that seen for the surface water temperature forecasts in the third chapter. Furthermore, this difference can also be explained by the simple fact that forecasting surface water temperature outputs from one-dimensional lake physics models,

which are deterministic, is a much easier task than forecasting point measurements of chlorophyll-a in a real lake environment, which has a much higher degree of stochasticity. This is demonstrated by the difference of approximately two orders of magnitude between MAPE in the first and second chapters.

Still, despite the less-challenging nature of forecasting modelled surface water temperature, the learning curves still appeared to plateau. As explained in Section 1.5.2.6, this is typical for machine learning, where error reductions tend to be correlated with the volume of training data according to an inverse power law (Mahmood *et al.*, 2022).

5.1.2 Limited examples of critical training data

Both chapters 2 and 3 were focused on forecasting water quality variables with strong seasonal components. While water temperature typically follows quite predictable patterns through the course of a year (Hipsey *et al.*, 2019), the behaviour of algal populations in the summer, when severe blooms often occur, is markedly different from other seasons (Sommer *et al.*, 2012). Therefore, while it can be assumed that summer water temperature forecasts benefit from training data from other seasons, it is likely that the benefit of non-summer training data for algal forecasts is more limited³. Furthermore, whilst very eutrophic lakes can have blooms which consistently persist for several months each year (Duan *et al.*, 2009), there are many lakes that only experience bloom events for a short duration annually, such as those with short residence times (Harris *et al.*, 2024). In some cases, this may be fewer than one per year (Favot *et al.*, 2019; Irani Rahaghi *et al.*, 2024). In the 30 years of fortnightly monitoring data used in the second Chapter, 95 % of the measurement-to-measurement differences in chlorophyll-a were less than $17.1 \mu\text{gL}^{-1}$, demonstrating that large changes are considerably rare (Figure 5.1). Since a primary aim of short-term algae forecasting is to predict large increases in algal concentrations, this imbalance in the training data is likely important to address.

Generally, to deal with imbalances in machine learning training data, there are several potential approaches. It is common to use data augmentation, which is to create additional observations based on transformations of the rare events in order to artificially balance the data (Jiang and Ge, 2021; de la Fuente *et al.*, 2024). For example, for image recognition tasks, augmentation typically involves geometric transformations such as rotation, or

³ However, it is important to note that conditions at other times of the year can still have an impact on blooms in the summer or autumn. For example, winter water-level fluctuations due to extreme precipitation events, or reservoir drawdown have been associated with cyanobacterial blooms in summer months (Bakker and Hilt, 2016).

colour space adjustments (Xu *et al.*, 2023). The validity of such transformations when performed on images is generally testable in an intuitive way. For example, for training ML to recognise different objects in images, it is obvious that upside-down or grey-scale images of the same objects should be considered valid, and therefore likely useful training data for this task. However, in the context of forecasting ecological parameters, defining what sort of augmentation is appropriate is difficult. This is because the underlying processes driving the data are generally non-linear, and not easily understood in a fully comprehensive way – if they were, then process-based models specifying this behaviour would likely be equally as useful for short-term forecasting. An alternative approach to deal with imbalanced training data is to simply remove some of the observations exhibiting the more common behaviour, which in this context would be data outside the algal bloom season. This may effectively balance the dataset, but in the situation where the dataset is limited to 20 years, this would likely lead to too few samples for good generalisation to be learnt.

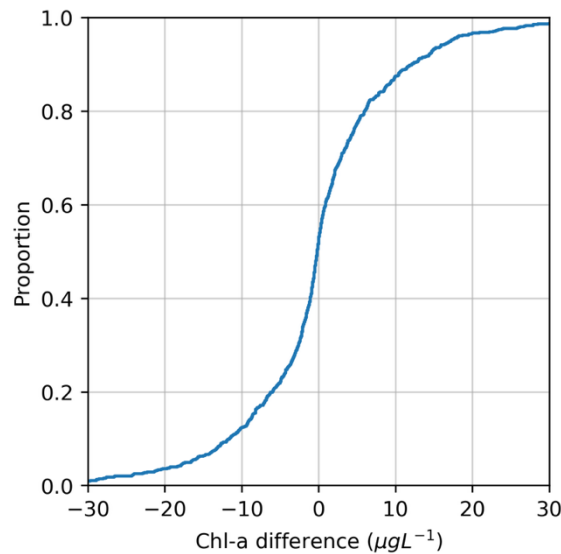


Figure 5.1: Empirical cumulative distribution function of the Blelham Tarn differenced chl-a time series.

5.1.3 Improving performance through increasing training dataset size

5.1.3.1 Forecasting lakes individually

In this thesis, it was generally found that increasing the sampling frequency had a positive effect on forecast performance, which is aligned with the findings of other studies which have investigated this (Derot, Yajima and Schmitt, 2020; Lin, Pierson and Mesman, 2023). However, improvements made through increasing sampling frequency cannot be considered a solution to the problem of data imbalance - critical events such as blooms are still generally rare on an annual basis, regardless of the sampling frequency. Therefore, to

make significant performance improvements it is essential to effectively increase the number of years of training data. However, as forecast error approximately follows an inverse power law with the number of years of training data, the dataset size increase required may be very large. Specifically, to gain significant improvements beyond the plateaus observed in this thesis may require increasing the effective number of years' training data by several orders of magnitude - a necessary condition for transitioning from a resolution-limited regime towards a variance-limited regime, in which forecast error reductions scale more efficiently with dataset size (Bahri *et al.*, 2024). If a model requires that all data be collected from a single lake, then for obvious reasons, such a requirement is infeasible in the near-term, as this could require continuous data collection for hundreds or even thousands of years.

5.1.3.2 *Forecasting multiple lakes*

A more realistic alternative is to effectively increase the number of training years by making use of data from multiple lakes. Given that there are millions of lakes across the globe, using data from even a small fraction of these to train forecasting models is a potentially feasible way to provide models with sufficient training examples to increase performance past the plateaus seen in this thesis. As such, even if data has only been collected for a relatively low number of years, if this data spans a large number of lakes, then the effective number of *lake-years* which a forecast model can learn from would be much higher than for any individual lake dataset, even those which have the world's longest running monitoring programs (Popovskaya, 2000; May and Spears, 2012; Maberly *et al.*, 2018). One caveat of this approach is that the wide variance in ecology, chemistry, morphometry, and geography across lakes means that data from some lakes would likely not be very useful for informing forecast models for lakes with very different characteristics (Downing *et al.*, 2006; Recknagel *et al.*, 2008; Maberly *et al.*, 2011; Doubek and Carey, 2017; Larsen *et al.*, 2017; McCullough *et al.*, 2019; Verspagen *et al.*, 2022). However, with a sufficiently large dataset, it becomes increasingly likely that many lakes will share similar characteristics, allowing their data to be complementary and collectively useful for training broadly applicable forecasting models.

While most algal forecasting studies to date have focused on individual lake systems, a small but growing body of work has begun to explore how data from multiple lakes can be used to improve forecast models. This shift is important, as leveraging multi-lake datasets opens possibilities for building more generalisable and scalable forecasting approaches - but it also introduces new complexities. Attempts to model algal blooms across multiple

lakes include the work of Welk et al. (2008) and Recknagel et al. (2008) who used rule-based hybrid evolutionary algorithms to forecast *Microcystis* and *Oscillatoria* concentrations. However, a key limitation of these studies was their exclusive reliance on in situ data, which restricted the number of lakes that could be feasibly included. Likely partially due to the scarcity of consistent long-term monitoring across lakes, these studies only worked with data from a total of just nine systems. More recently, Matthews (2023) advanced this field by applying univariate time series models to satellite-derived data across multiple lakes, using imagery from the OLCI sensor onboard Sentinel-3. Building on this, Schaeffer et al. (2024) demonstrated the scalability of satellite-based approaches by developing a hierarchical Bayesian spatiotemporal model alongside a suite of machine learning methods to forecast algal bloom risk in over 2,000 US lakes and reservoirs. Their study highlights the significant advantage of remote sensing in this context: namely, the ability to generate large, spatially expansive datasets that are otherwise impractical to collect via in situ sampling alone.

Using the Sentinel-3 OLCI sensor - which offers global coverage with a revisit time of approximately two days (Donlon et al., 2012) - Schaeffer et al. (2024) were able to compile ~197,000 cloud-free observations across 2,192 lakes over five years. This translates to an average of 18 observations per year per lake, or roughly one every three weeks. Algal bloom dynamics can evolve rapidly, meaning that an average sampling interval of three weeks may miss key events, illustrating the need for higher-frequency observations where possible (Dubelaar, Geerders and Jonker, 2004; Pobel, Robin and Humbert, 2011). Another critical limitation of the OLCI data is its relatively coarse spatial resolution. At 300 metres, many smaller water bodies, which are often ecologically and socially significant, and more often eutrophic, remain unresolvable (Downing, 2010; Donlon et al., 2012; Biggs, von Fumetti and Kelly-Quinn, 2017). As the vast majority of lakes are relatively small, this spatial constraint significantly narrows the range of lakes that can be included in forecasting efforts and limits the overall applicability of the approach (Downing et al., 2006; Cael and Seekell, 2016).

This is where the fourth chapter of this thesis becomes particularly relevant. By evaluating the use of Planet SuperDoves imagery, which offers much higher spatial resolution, for monitoring small lakes, this work opens a pathway towards expanding algal forecasting models to a vastly greater number of systems. Not only does this present an opportunity for forecasting at a more widespread scale, it also has the potential to improve forecast skill by increasing the number of effective lake-years available for training and validation.

While there are undoubtedly challenges associated with handling such a large and complex dataset, the benefits in terms of improved forecast performance and broader coverage are likely to be substantial, with direct value for scientists, water managers, and recreational users alike.

5.2 Achieving widespread monitoring in smaller lakes

A key proposition of the fourth chapter in this thesis is that high spatial resolution imagery such as the Planet SuperDoves presents a means to systematically monitor a much larger and more diverse number of water bodies than lower resolution imagery. Alongside this, as explained in Section 5.1.3.2, such capability may also be critical to providing larger datasets for training forecast algorithms. However, there are several key challenges associated with the realisation of reliable widespread monitoring of small water bodies using satellite imagery.

5.2.1 The challenges of small water bodies

One of the key advantages of high-resolution imagery such as SuperDoves is that monitoring relatively small water bodies is possible. In the fourth chapter, it was shown that the higher resolution of the SuperDoves imagery enabled detection of features on a fine spatial scale such as shoreline algal scums. Importantly, these are often of most concern for human and animal health, and therefore the ability to resolve these in smaller water bodies is of considerable value. However, monitoring smaller water bodies presents several challenges which are particularly relevant given that the SuperDoves imagery is not specifically designed for aquatic applications, and is newer and therefore less-well established for inland water monitoring than other sensors such as Sentinel-2 MSI or Sentinel-3 OLCI.

Firstly, in the same way that the vast majority of lakes are small in terms of area, this majority also tend to be relatively shallow (Wetzel, 1990). Accordingly, most of the lakes which become newly monitorable as the resolution of satellite imagery is increased are shallow. Monitoring shallow lakes using remote sensing is challenging due to bottom reflectance effects which can be influenced by submerged vegetation (Ma *et al.*, 2011), and also the tendency for other optically active constituents such as suspended sediments to be more variable and heterogenous (Tyler *et al.*, 2006). The primary study site in the fourth chapter of this thesis, Airthrey Loch, is very shallow, but as it tends to have high phytoplankton biomass in the summer, bottom reflectance effects in the deeper parts of the lake are largely negligible. Secchi depth water transparency measurements may be used to

distinguish water which can be considered *optically deep*, that is without significant bottom reflectance effects, and the converse, known as optically shallow waters. If the water depth is greater than three times the Secchi depth, then it is usually considered optically deep, and otherwise it is likely that some of the remote sensing signal is influenced by bottom reflectance (Højerslev, 1974; Ma *et al.*, 2011). Both lake depth (bathymetry) and Secchi depth can be estimated using satellite imagery (Arsen *et al.*, 2014; Armon *et al.*, 2020; Kabir *et al.*, 2025).

In addition to the challenges associated with depth, small lakes, by virtue of their geometry, tend to have a larger portion of surface area adjacent to the shoreline than larger lakes. The consequence of this is that, generally, a higher percentage of satellite pixels over small lakes will be influenced by land adjacency effects. Additionally, often being shallow, small lakes have a relatively large littoral zone compared to larger, deeper lakes. (Seekell *et al.*, 2021). A potential consequence of this is that satellite pixels over small lakes are more likely to contain signals from multiple biological reflectance sources such as submerged, floating or emergent aquatic plants as these are typically found in the littoral zone. It follows that areas with both high cyanobacteria and macrophyte populations are more common in smaller lakes. As such, whilst algorithms have been developed for distinguishing cyanobacteria from floating or submerged macrophytes (Oyama, Matsushita and Fukushima, 2015; Liang *et al.*, 2017), it is possible that masking out all pixels affected by macrophytes in smaller water bodies might result in an underestimation of cyanobacterial bloom extent, or simply an insufficient number of valid pixels to make confident estimates from. Therefore, it may be important to develop approaches which can accurately flag joint cyanobacteria and macrophyte presence, as the presence of one does not exclude the other (Bilous *et al.*, 2023).

During the data collection carried out for the fourth chapter of this thesis, a number of Secchi depth measurements and additional field observations, such as photographs, were taken during the measurement campaign. This was generally done in an ad-hoc manner, focusing primarily on visible changes to the water transparency, macrophyte presence and water level. As such, this was an efficient way of generating an informative record of which parts of the water body were likely to have possible influence from bottom effects, macrophyte cover and other potentially confounding reflectance sources. Whilst this was an effective approach for studying a single water body, such an approach would likely be impractical for studying large numbers of water bodies due to the time-consuming nature of such work. Therefore, for studying large numbers of lakes, it is likely more practical to

use automated pixel-identification routines, which flag pixels potentially affected by bottom effects (Arabi *et al.*, 2020), macrophyte cover (Liang *et al.*, 2017), algal scums (Matthews, Bernard and Robertson, 2012), or land-adjacency effects (Jiang *et al.*, 2023). Critically, alongside chlorophyll-a retrieval, algorithms for estimating variables such as Secchi-disk depth and macrophyte cover have now been successfully implemented using SuperDoves imagery (Grendaité *et al.*, 2025; Kabir *et al.*, 2025). This demonstrates that the development of automated pixel flagging routines for this imagery is feasible, suggesting additional promise for the use of this imagery for widespread operational water-quality monitoring applications.

5.2.2 The costs of high-resolution imagery

Whilst freely available for small-scale non-commercial research projects, such as the investigation carried out in the fourth chapter, using the high-resolution Planet SuperDoves imagery for any large-scale operational application would be costly. At the time of writing, the largest advertised Planet monitoring subscription covers an area of 300 km² and costs between 15,000 and 55,900 USD annually, depending on the region of interest (Planet Labs PBC, 2025a). For reference, the eight largest inland water bodies in Scotland have a combined area of approximately 290 km², and the largest inland water body in the British Isles, Lough Neagh, has an area of approximately 382 km² (Hughes *et al.*, 2004).

Alongside the purchase cost of these data, another major consideration is that of data storage and processing costs. The number of pixels covering a given area scales in proportion to the inverse of the square of the pixel size, or resolution. Accordingly, even though Planet SuperDoves has fewer bands than Sentinel-2 MSI or Sentinel-3 OLCI imagery, the storage requirements and therefore processing overheads associated with this data would be much larger for working with equivalent areas. In fact, when also considering the higher effective revisit frequency of the SuperDoves data, the memory requirements for storing these data would be 3-4 orders of magnitude larger than those for Sentinel-2/3 data (Table 5.1). Whilst this is not necessarily an insurmountable barrier to the realisation of high-resolution widespread lake monitoring, it is a very practical constraint which could likely lead potential funders and managers of such projects to ask: *is the added value of high-resolution imagery worth the additional costs?*

Table 5.1: Approximate uncompressed storage requirements for an entire year of imagery for every lake in the world for three different satellite sensors. Calculated under the assumption that the total global area of all lakes is $5 \times 10^6 \text{ km}^2$ (Verpoorter et al., 2014), and that each band has a storage requirement of 16 bits per pixel. Images with cloud cover are assumed to have the same storage requirements as cloud-free imagery.

Satellite	Resolution (m)	Pixel Area	Bands	N Pixels	Revisit period (d)	Memory / year
Sentinel-3 OLCI	300	0.09 km^2	21	5.56×10^7	~ 2	0.43 TB
Sentinel-2 MSI	10, 20, 60	100 m^2	13	5.00×10^{10}	~ 5	27.9 TB
SuperDoves	3	9 m^2	8	5.56×10^{11}	1	3.24 PB

Imagery with high temporal and spatial resolution may enable monitoring smaller water bodies and better capturing phytoplankton dynamics, thereby improving algal bloom forecasting capabilities. However, whilst the economic impact of these benefits may be possible to approximately estimate, it is also most likely highly variable between lakes. This is because the economic impacts of harmful algal blooms are far-reaching and diverse – they affect the recreational value of water bodies, fishery income, water utility operations, and property prices (Pretty *et al.*, 2003; Hoagland and Scatasta, 2006; Dodds *et al.*, 2009; Carias *et al.*, 2024). Therefore, there would certainly be large variations in the benefits offered by high-resolution satellite monitoring and associated forecasting capabilities depending on which lake is studied. Whilst there may be obvious justifications for funding monitoring and forecasting programs for individual lakes such as Lake Erie (Wynne *et al.*, 2013), Windermere (Maberly and Elliott, 2012), Lough Neagh (Elliott, McElarney and Allen, 2016), Loch Leven (May and Spears, 2012), or Müggelsee (Recknagel *et al.*, 2025), which are all economically significant, high profile, and relatively large, the vast majority of smaller, lesser-known lakes are likely not seen as important or critical to monitor (Biggs, von Fumetti and Kelly-Quinn, 2017). At present, this is likely an important barrier to the adoption of widespread high-resolution lake monitoring.

Nonetheless, regardless of whether it is true that lakes perceived as less important are genuinely less valuable to monitor, the obvious benefit of monitoring these lakes is the benefit that these data may offer to forecasting models. As previously explained in Section 5.1.3.2, by monitoring lakes in a widespread manner, many more useful training examples of critical events such as blooms can be captured for training forecast models, as well as furthering scientific understanding more generally. As such, it is very likely that the benefit of monitoring many lakes collectively, rather than individually, would be in effect,

greater than the sum of its parts (Schaeffer *et al.*, 2024). It would therefore be beneficial to focus some future research on further evidencing this which, if successful, would provide strong justification for the funding of large-scale, operational lacustrine monitoring and forecasting programs driven by high-resolution satellite imagery.

5.2.3 Sustaining operational funding

Building and maintaining reliable operational monitoring and forecasting services requires continual, secure funding. For example, the United States National Oceanic and Atmospheric Administration (NOAA) Lake Erie algal bloom forecast bulletin service, which uses satellite imagery, has been running since 2008, and as such, has a well-developed operational pipeline for timely processing of new imagery for forecast generation (Wynne *et al.*, 2013; Stumpf *et al.*, 2016; Manning *et al.*, 2019). The continued funding which this project has received to-date is critical to reliably producing operational forecasts for Lake Erie year after year. Importantly, this is quite unique – there are very few operational lacustrine satellite monitoring and forecasting systems in the world, and that Lake Erie has such a system which receives continued funding is a testament to its socio-economic importance, and history of very severe, high profile cyanobacterial blooms (Jetoo, Grover and Krantzberg, 2015; Reutter, 2019).

In the UK, operational satellite monitoring of chlorophyll-a, and associated products has been demonstrated for hundreds of lakes in England and Scotland as part of the University of Stirling UK Lakes Observatory (UKLO) project⁴. However, unlike the NOAA Lake Erie monitoring and forecasting systems, the UKLO was initially funded as a short-term proof of concept project and therefore is not presently active. Whilst the UKLO has now received further funding for future work, its operational gap still illustrates what is perhaps a key challenge to overcome for widespread satellite monitoring and forecasting of algae in lakes to be realised – such systems can almost certainly be implemented, however their longevity and therefore success is contingent on sufficient funding which is both continual and reliable. Whilst both the NOAA Lake Erie forecast system and UKLO have been effectively governmentally funded, one alternative is to run such a system as a commercial business, charging users such as water utility companies for the service. This model has been pioneered by the company CyanoLakes, which primarily uses Sentinel-3 OLCI data

⁴ The UK Lakes Observatory project was funded by NERC from February 2019 - August 2020, under the award: *Delivering resilience to climate impacts on UK freshwater quality: towards national-scale cyanobacterial bloom monitoring and forecasting* (NE/S016856/1)

to provide cyanobacterial health advisory data to water utility companies, government agencies and the general public (Matthews, 2022).

5.2.4 Accelerating sensor technologies

At the time of writing, it has been more than six years since the first launches of SuperDoves satellites in 2019. Whilst significant advances in the application of SuperDoves to aquatic monitoring have been made during this period, this could be considered relatively slow progress relative to the pace at which commercial satellite sensor technology is being developed. Critically, there are still several important challenges to overcome before the SuperDoves can be effectively used for operational inland water quality monitoring. In a recent study on the feasibility of using SuperDoves for inland and coastal water quality monitoring, Kabir et al. (2025) highlighted that further development of atmospheric correction routines and inter-sensor harmonisation is required to resolve issues related to calibration, bias, and consistency between the cubesats. This corroborates the findings in Chapter 4, where it was found that whilst broad trends in the SuperDoves and Sentinel-2 MSI chlorophyll-a time series were comparable, closer comparison revealed atmospheric correction issues and significant biases.

Planet have recently announced that in 2026 they will begin testing their next generation monitoring fleet: *Owl*, which will eventually supersede the SuperDoves (Marshall, 2025). Therefore, it is possible that by the time inland aquatic monitoring capabilities for the SuperDoves have matured to a stage where they are operationally useful, they may have already begun to be replaced by new sensors. However, in this case, Planet intends for the Owl imagery to be backwards-compatible with SuperDoves data – both sensors will share the same eight spectral bands, and so algorithms which have been developed for SuperDoves should be mostly transferrable to Owl. Nonetheless, this is indicative of a core challenge for the field of inland aquatic remote sensing, which generally receives considerably less funding than defence and intelligence satellite applications which comprise the majority of Planet's revenue (Planet Labs PBC, 2025c). As such, for fields such as inland water quality monitoring, which are likely inherently slower to develop due to limited funding, it is critical that commercial providers of satellite data continue to ensure that backwards-compatible products are available from new sensors. Furthermore, it is key that large, diverse in situ datasets are continually readily available so that algorithm development, calibration, and validation can happen quickly and effectively shortly after data from new sensors becomes available.

5.3 Synergies across diverse data sources

5.3.1 Citizen science and satellite imagery

The fourth chapter of this thesis includes some discussion on how remote sensing products and citizen science records are well-suited to application in tandem. One reason for this is because both these data sources have the potential to enable relatively low-cost, and widespread monitoring. Satellite constellation-based approaches such as the Planet SuperDoves demonstrate that daily, high-resolution imaging of the entire Earth is possible, and citizen science projects such as *Bloomin' Algae* have generated thousands of verified algal bloom records across several different countries in recent years⁵. Alongside the potential for widespread monitoring capabilities, both citizen science and remote sensing datasets are inherently subject to errors and patchiness or missing data. In the case of satellite imagery, these issues may arise from atmospheric effects (Warren *et al.*, 2019); cloud cover; spurious reflectance sources such as boats; bottom effects (Ma *et al.*, 2011), adjacency effects (Paulino *et al.*, 2022; Jiang *et al.*, 2023); sensor noise (Kabir *et al.*, 2025); or other interference. In contrast, the errors and patchiness associated with citizen-collected records arise from non-expert identification skills, and non-random, potentially biased sampling (Crall *et al.*, 2011; Bird *et al.*, 2014). Importantly, the sources of errors and missing data associated with each of these data collection approaches are independent, meaning that there is potential for gap filling, and increased certainty in the validity of observations which are corroborated by both.

Previous efforts to integrate citizen science and remote sensing data have included the development of an online metadata hub for collation of satellite imagery and in situ bloom reports (Beaulne and Fotopoulos, 2024); an examination of the correlation between offshore satellite-measured Sargassum bloom risk reports and citizen-collected observations of *Sargassum spp.* on shorelines (Putman *et al.*, 2023); and integrations of in situ reflectance data, satellite imagery and citizen measurements made using the HydroColor and CiteClops/EyeOnWater mobile phone applications (Busch *et al.*, 2016; Baltodano, Agramont and van Griensven, 2024). However, given the relatively small number of studies in this area, there is likely potential for further work which explores this synergy in other ways.

⁵ On the 27th of November 2025 the Bloomin' Algae online map showed that since January 2017, a total of 4233 records were logged. Of these, 2183 were marked as 'correct', 638 'plausible', and 1412 'incorrect' (UKCEH, 2025).

For example, it may be useful to design algal bloom reporting applications with satellite imagery in mind. Typically, applications such as Bloomin' Algae focus on reporting the locations of blooms at a single point. However, algal scums can cover relatively large areas of a lake surface, and in smaller-sized water bodies, it would be possible for an observer to view a large portion of the affected area. Therefore, it is conceivable that a citizen scientist could not only report the presence of an algal scum, but also provide a mapping illustrating how much of the water body was affected. The addition of this step would be a minor change to the workflow of an application like Bloomin' Algae, but could potentially provide very useful information for validating satellite products for algal scum detection (Figure 5.2). This could be particularly valuable for smaller water bodies where algal scums are often largely located in relatively narrow strips along shorelines, which can be challenging to remotely sense due to adjacency effects (Paulino *et al.*, 2022). Therefore, these data could potentially be used for validation or calibration of shoreline scum detection algorithms. Additionally, including a means for experienced samplers such as environment protection agency workers to mark their records as 'expert' or similar is a powerful way to filter observations where this is required for more robust validation purposes (Beaulne and Fotopoulos, 2024).



Figure 5.2: Mock design for a spatial mapping component of a mobile phone application for citizen science algal bloom monitoring.

5.3.2 Combining ML and process-based modelling data streams

Whilst the popularity of ML for tackling applied ecological and environmental problems is increasing, there continues to be a strong scientific interest in process-based models, and how the strengths of these two approaches can be pooled, or possibly *co-evolved* (Razavi *et al.*, 2022). Synergies between ML and process-based models have been identified in several areas, including ML emulation of process-based models, and process-, or physics-guided ML.

ML emulation aims to alleviate a potential disadvantage of complex process-based models, such as three-dimensional hydrodynamic models, which is that these can be very computationally intensive. Consequently, for real-time applications, such models can be either prohibitively slow, or too expensive to feasibly use operationally. Therefore, some studies have investigated the potential for emulating complex process-models using more *lightweight* ML models. Whilst the initial training process, where the inputs and outputs of the process-based model is used to train a ML emulator, can be very resource intensive, the resulting model is typically more suited to real-time applications, generating comparable outputs faster, with lower computational overhead (Lim and Wang, 2022; Mohammed, Michel Tornyeviadzi and Seidu, 2022).

Process-guided ML however, aims to use knowledge of physics, biology or chemistry to attempt to ensure that the outputs generated by a ML model are realistic (Karpatne *et al.*, 2017). For example, known physical equations, such as the temperature-density relationship for water, or conservation of energy can be included as part of the loss-function, so that models are penalised for violating these (Read *et al.*, 2019; Daw *et al.*, 2022). Furthermore, both approaches can be combined by using outputs from process-based models as pre-training data to improve process-guided ML performance and generalisability (Jia *et al.*, 2019; Read *et al.*, 2019).

In addition to these applications, the third chapter of this thesis proposes a new way in which process-based and ML models may be used in tandem. This approach takes advantage of the principal that process-based models are generally more reliable for making predictions for scenarios which have not yet been observed (Franks, 2018). Then alongside this, the method of the third chapter also acknowledges that ML approaches are generally considered more accurate for short-term forecasting (Rousso *et al.*, 2020; Cruz *et al.*, 2021; Zahir *et al.*, 2024). As such, this work leverages each of these

strengths to examine how water quality forecasting efforts may be affected under different climate scenarios.

Importantly, this approach need not only be applied to climate scenarios. For example, algal bloom forecasting performance could be tested under different ecosystem assemblage scenarios to predict the wider impacts of shifts due to the proliferation of invasive species such as *dreissenid* mussels (Nalepa and Schloesser, 2013; Tang *et al.*, 2014; Reynolds and Aldridge, 2021). Additionally, this approach could be used to ask fundamental questions about monitoring system design for algal bloom forecasting. Given that in a process-simulation it is possible to effectively “measure” every influencing variable, it is therefore possible to evaluate many more variations and approaches to monitoring than would be possible with real, manually collected data. Whilst this was explored to some extent in the third chapter by varying sampling frequency and training dataset duration, this could be extended considerably.

For example, using data from a three-dimensional process-based model would enable the forecast performance benefits of sampling at multiple depths and locations to be evaluated. Accordingly, such models could be used to simulate data outputs from automated monitoring platforms, potentially allowing for the placement of these to be optimised for forecasting applications. Whilst the development of three-dimensional lake models typically requires a water body which is already well- monitored and instrumented (Kim, Cerco and Johnson, 2006; Liu and Chen, 2012; Missaghi, Hondzo and Melching, 2013), this kind of investigation could potentially reveal whether a current monitoring configuration was efficient for generating forecast training data. Furthermore, by studying in detail the effects of lake monitoring practices on forecasting ability, it is possible that a widely applicable set of lake-monitoring design principles could be developed. Such knowledge would be of significant benefit to future lake monitoring and forecasting efforts.

Finally, it is important to note that process-based models can be tuned using satellite imagery as calibration data (Siebers *et al.*, 2025). As such, this additional synergy potentially enables some of the suggested joint applications of ML and process-based models to be implemented for smaller, less-well studied lakes where in situ observations are often sparse.

5.3.3 Manual sampling alongside automated data collection

Technologies such as satellite imaging and high frequency autonomous measurement platforms provide a means to monitor lakes at frequencies and geographic scales which would be impossible using traditional manual approaches. In several ways, this thesis demonstrates that the benefits of this capability are significant: Chapters 2 and 3 show that the ability to monitor consistently for several years is essential to the success of data-driven forecasting approaches, and that, additionally, collecting data at higher sampling frequencies can further improve performance. Alongside this, the fourth chapter shows that new satellite imagery may further our ability to monitor lakes in widespread manner, both in terms of increased spatial resolution, but also with higher revisit rates. Therefore, collectively, these studies demonstrate the value and significant potential for furthering monitoring and forecasting capabilities using data collected by automatic measurement systems.

However, manual sampling efforts were also fundamental to each of the studies presented in this thesis (Figure 5.3). The fourth chapter made extensive use of manually collected water samples alongside bloom records collected by citizen scientists, and environment agency samplers. This enabled thorough calibration and validation of chlorophyll-a retrieval algorithms, and evaluation of algal bloom detection capabilities. Similarly, in the third chapter, the underlying process-based lake models were all calibrated and validated using a combination of manually- and automatically- collected in situ water samples. Lastly, aside from weather station data, all data used in the second chapter was entirely collected manually, through a regular, long-term monitoring programme which has spanned decades. Whilst this is a clear demonstration of the value of these data, it was found in both forecasting-focused chapters that increased sampling frequency would improve performance.

In many cases, automated approaches are considered the only economically viable solution for high frequency monitoring as manual collection of near-daily data over long time periods is typically considered prohibitively expensive (Marcé *et al.*, 2016).

However, the initial capital expenditure and continued operational costs associated with installing and maintaining automated in situ sensor platforms can be significant, and therefore these are not always a cost-effective monitoring solution (Seifert-Dähnn *et al.*, 2021). Therefore, satellite imagery is an obvious data solution for widespread lake monitoring and forecasting, particularly because imagery collected by space agencies such as ESA and NASA is, ignoring compute costs, freely available. However, as is

demonstrated by the fourth chapter of this thesis, in situ measurements are still critical to the calibration and validation of satellite data products. Similarly, conventional sampling and lab-analyses of water samples remains critical for regular validation of automated in situ sensor data streams as they can suffer from sensor drift, bio-fouling, and other data quality issues that, without correction, can reduce the utility of the data (Rozemeijer *et al.*, 2025). Therefore, while data from automated sensor platforms is likely to be increasingly widely used, conventional manual sampling should remain critical.

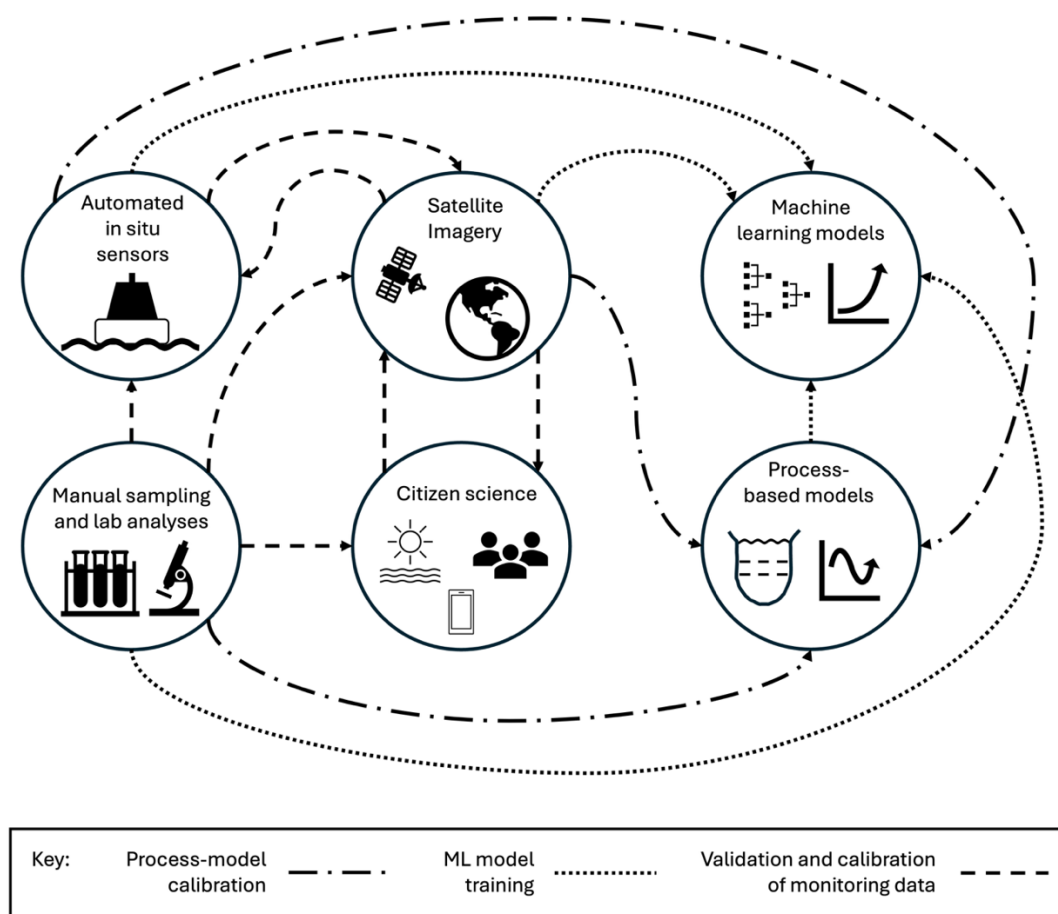


Figure 5.3: Diagram illustrating synergies between various data sources and modelling approaches used in this thesis.

In some cases, it may be beneficial to streamline sampling efforts by selecting only a small number of key variables to measure at higher frequencies. As demonstrated by Chapter 2, a moderate reduction in the number of variables used to generate algal forecasts will not necessarily reduce forecast performance significantly. For example, if one was primarily interested in optimising the monitoring of Blelham Tarn for algal forecasting, then the finding that fortnightly nutrient sampling data did not significantly improve forecast performance would be critical. In this instance, it would therefore make sense to reduce the frequency of manual nutrient sampling and potentially install a high frequency monitoring platform focused on measuring the most important variables, which were

chlorophyll-a, water temperature, air temperature, and dissolved oxygen. However, the Blelham Tarn monitoring data is used for a large number of scientific studies, and so reducing nutrient sampling frequency would likely be detrimental to other work. Nonetheless, both manual sampling and automated monitoring of a large number variables can be expensive, and so studying which of these are most useful for the intended applications is a valuable way to potentially reduce costs and therefore streamline monitoring efforts.

Additionally, while our technical capacity to monitor and understand water systems is greater than ever before, public understanding of water quality issues remains limited (Seelen *et al.*, 2019). More broadly, many people are becoming increasingly disconnected from the natural world (Soga and Gaston, 2016; Kesebir and Kesebir, 2017). Therefore, if not thoughtfully implemented, emerging monitoring technologies may unintentionally deepen this disconnection. Specifically, there is the potential for tools that supposedly reduce the need for people to directly observe, understand, and care for water to weaken the value that society places on freshwater ecosystems - not only in terms of biodiversity but also in terms of their cultural, emotional, and social significance. This points clearly to need for continued, or likely increased investment in monitoring efforts which directly involve and engage members of the public with water bodies and ecosystems.

One obvious way in which this can be done is through citizen science projects - involving the public with scientific monitoring efforts should increase societal awareness and engagement with important environmental issues (Dickinson *et al.*, 2012). However, it is important to realise that this issue, which is in effect a decline in direct interactions with nature, extends beyond members of the public into scientific communities, and even those who study natural ecosystems (Soga and Gaston, 2025). Referring to this phenomenon as the “extinction of experience”, Soga and Gaston (2025) highlight that in the field of ecology this is driven, in part, by increased availability of automated sampling technologies which reduce the necessity for fieldwork-based research and education. Therefore, as we continue to embrace the benefits of automated water monitoring technologies, it is critical that we also continue to directly observe, measure, and interact with water bodies so that we do not forget fundamental skills or lose sight of why inland waters are so important.

5.4 “Artificial Intelligence” trajectories – Self-Supervised Learning and Foundation models

5.4.1 The recent successes of large language models

Since 2020, the development of powerful large language models (LLM) with the ability to generate text responses in a coherent and seemingly intelligent way has created a huge interest in what is now commonly referred to as “Artificial Intelligence” (AI) across many industries (Rashid and Kausik, 2024). This rapid acceleration in interest and application of AI has been driven by the development of the Transformer, an artificial neural network architecture based on the multi-head attention mechanism (Vaswani *et al.*, 2017). By training on massive datasets, generative pre-trained transformers (GPT) with billions of parameters have been developed (Brown *et al.*, 2020; Yenduri *et al.*, 2024). These now form the backbone of the LLMs and AI chat interfaces which have been developed by companies such as OpenAI in the last five years. Fundamental to the success of this approach is the pre-training step, which consists of a self-supervised learning process. During this process, GPT LLMs learn to generate text in a realistic way according to the patterns they learn from massive training datasets of unlabelled text. Following this, pre-trained models are fine-tuned using supervised learning to perform specific tasks, such as taking on the role of a general-purpose *chat-bot*.

5.4.2 Large pre-trained cross-domain time series forecasters

Ultimately, time series forecasting is a very different challenge from generating useful text outputs. Specifically, whilst certain structures in texts, such as sentences, can often be rearranged without a loss of meaning, the sequential dependencies in time series data are fundamental. This means that simply re-using exactly the same self-supervised pre-training approaches which have been effective for LLMs and computer vision tasks is unlikely to be most effective for time series forecasting (Zhao *et al.*, 2024). However, it is important to realise that for tasks such as algal bloom forecasting, where relatively little domain-specific training data exists, it may be extremely beneficial to pre-train models on larger datasets of cross-domain time series. Following this, in a similar manner to LLMs, pre-trained time series models can be fine-tuned using domain-specific data, such as the chlorophyll-a time series which was used to train the forecasting models in the second chapter of this thesis (Ma *et al.*, 2024).

The advantage of this approach is that before being optimised for a specific forecasting task, these models, known as *foundation models*, first learn to recognise common patterns which occur in time series across multiple domains, such as an increase in one variable preceding the increase of another. Critically, this pre-training step would be able to take advantage of much larger datasets than those which were available for the site-specific approaches used in this thesis. As such, forecasting models developed in this way can potentially outperform those trained exclusively on domain-specific data (Jin *et al.*, 2023; Kamarthi and Prakash, 2024). Notably, there have been recent successes in applying foundation models to weather and climate forecasting (Allen *et al.*, 2025; Bodnar *et al.*, 2025). However, Luo *et al.* (2025) argue that pre-training on generic time series may not be suitable for environmental modelling applications, as generic time series data, which could include data from diverse sources such as road traffic monitoring networks, electrical supply infrastructure, or economic tracking, is unlikely to be sufficiently comparable to environmental data for this to be useful training data. Instead, they propose a self-supervised learning approach which dynamically switches between various specialised models using a “scenario-encoder” built upon a process-based lake model. Whilst this approach is very different from pre-training on generic time series, it is still fundamentally focused on leveraging self-supervised learning alongside data from additional sources in order to train a model which better generalises to unseen test data.

Importantly, the success of large, pre-trained foundation models for natural language processing and computer vision is inspiring efforts following similar approaches for time series forecasting tasks. Whilst these approaches are currently less well-developed, it is likely that if they are shown to be superior to exclusively domain-specific approaches for environmental and ecological forecasting, then this may result in a transformation of how lake forecasting is carried out. Accordingly, it may become increasingly beneficial to produce large, multi-lake time series datasets in consistent formats so that this can be easily used for training large forecasting models. For this, satellite imagery is an obvious data source for providing widespread observations, but as discussed in Section 5.3, pairing this with in situ and citizen science records would be advantageous as foundation models can work well with multi-modal data (Allen *et al.*, 2025; Wu *et al.*, 2025). Furthermore, making lake datasets publicly accessible could facilitate wider use of these for generating better generic forecasting models. However, making such data completely publicly available comes with the risk that the collectors and curators of these data may not always receive sufficient acknowledgements for their efforts, or miss potential publishing opportunities (Haeussler *et al.*, 2014; Gewin, 2016).

Nonetheless, with sufficient data from lakes and lake-adjacent fields, such as other ecosystems, it may be possible to train models which are generic for a time series sub-category, for example ecological and environmental applications. Such models would benefit from a larger pre-training dataset than a lake-specific model, with the potential advantage that time series which are unlikely to be useful for this domain, such as economic data, are excluded. Given that there is growing interest in the field of ecological forecasting, it is likely that there would be many groups keen to contribute data to such an effort (Dietze *et al.*, 2018, 2024; Tulloch, Hagger and Greenville, 2020; Thomas *et al.*, 2023).

5.4.3 Foundation models for remote sensing

In the same way that progress in natural language processing and computer vision are inspiring new developments for time series tasks, there is also significant interest in applying these principles to remote sensing data. Numerous foundation models for remote sensing data are currently in active development, all aimed at performing over a broad range of tasks. Typically these models are, to some extent, inspired by computer vision approaches. This is because analysis of remote sensing imagery often shares parallels with typical computer vision tasks such as object recognition and image segmentation, which is the portioning of an image into distinct regions with separate characteristics (Brar *et al.*, 2025). However, computer vision foundation models are typically trained on datasets which largely consist of ground-level images, without any temporal dimension. It is therefore beneficial to develop remote sensing specific foundation models which can learn from temporal patterns in series of images of the earth taken from space (Tseng *et al.*, 2023).

Recently, remote sensing foundation models which can make use of data from multiple satellite sensors, alongside other data such as meteorological or elevation data, have been developed. Examples of these include IBM's *TerraMind* (Jakubik *et al.*, 2025); NASA Harvest's *Galileo* (Tseng *et al.*, 2025); and Wuhan University's *SkySense++* (Wu *et al.*, 2025). Additionally, some models such as *Presto* and *Galileo* explicitly focus on extracting temporal information from satellite time series (Tseng *et al.*, 2023, 2025). At present, the applications of these models have included land-use mapping, crop mapping, oil spill segmentation, flood monitoring, wildfire monitoring, ocean colour retrieval, and tree species classification (Jain *et al.*, 2024; Butsko *et al.*, 2025; Wu *et al.*, 2025; Yang, Wang and Li, 2025). Furthermore, there is also interest in combining these efforts with LLM capabilities, for example developing an LLM agent to select an appropriate remote sensing

foundation model for a specified task (Chen *et al.*, 2025). Moreover, Planet Labs recently announced that they will partner with Anthropic to use their AI platform, *Claude*, to “help users get value from our data faster” (Planet Labs PBC, 2025b).

Whilst significant progress has been made in developing foundation models for remote sensing applications, the vast majority of this work has focused on land-based applications. Only a handful of studies have developed foundation models for ocean colour applications, and these have exclusively focused on marine environments (Dawson *et al.*, 2025; Yang, Wang and Li, 2025). Importantly, Yang *et al.* (2025) demonstrated that their ocean colour foundation model was able to learn, without labelled data, to make some differentiation between case-I waters, where all optically active constituents are correlated with chlorophyll-a, and the more optically complex case-II waters. This is particularly relevant to inland waters, as applying different algorithms according to the optical water type of the water in question has been shown to be powerful for improving retrieval performance over multiple water bodies (Spyrakos *et al.*, 2018; Neil *et al.*, 2019; Xiong *et al.*, 2025). As such, this suggests that, if trained on widespread inland water satellite imagery, it would be possible for a foundation model to self-learn important distinctions and similarities between the wide variety of optical conditions found in these waters.

Therefore, whilst there is still considerable progress to be made before such approaches could be considered for operational inland water quality monitoring, there is certainly potential. Furthermore, given the pace at which AI applications have developed in the last decade, it is very likely that foundation models for inland water quality remote may also develop rapidly in coming years. Ultimately, however, for such models to be competitive with the presently available alternatives, it is critical that there is sufficient labelled training data for fine tuning to specific applications. Therefore, regardless of how ubiquitous foundation models become for inland water quality remote sensing applications, it will still be paramount to continue to make regular in situ measurements of key parameters so that these can ultimately be used to train and validate retrieval algorithms. Therefore, whilst new AI approaches such as foundation models certainly have the potential to improve remote sensing parameter retrieval performance, they are unlikely to dramatically change the way in which the science of remote sensing is carried out.

5.5 Climate change – understanding impacts and preparing for future risks

Throughout this thesis, climate change has stood as a key justification for why this work is important and will likely become increasingly so into the future. Firstly, the second and fourth chapters are explicitly concerned with forecasting and monitoring algal blooms. Whilst algal blooms have been an environmental concern for several decades, it is now understood that, for many water bodies, blooms are becoming more frequent, long-lasting, and severe due to anthropogenic climate change (Hou *et al.*, 2022; Duan *et al.*, 2024; D. Wang *et al.*, 2024; Feng *et al.*, 2024; Zhang *et al.*, 2025). Therefore, as climate change continues to affect inland waters, the importance of furthering bloom monitoring and forecasting capabilities is continually increasing. This is important because both the second and fourth chapters of this thesis indicate that achieving strong monitoring and forecasting capabilities is not trivial.

The second chapter shows that for data-driven forecasting to be effective, it requires consistent, long term monitoring data, which must be collected for several years before useful forecasts can be made. Alongside this, whilst the fourth chapter demonstrates the promise of new high-resolution imagery for extending remote monitoring capabilities to smaller water bodies, it also highlights the costs associated with using commercial imagery, and the potential need for more widespread in situ monitoring to complement this. As such, both these chapters suggest that developing the tools and systems required to track and anticipate algal blooms requires significant investment. Nonetheless, if the effects of climate change on algal bloom frequency and severity continue as predicted, then it is certain that increasing monitoring and forecasting capability is wise.

If climate change is used as a key justification for investment in algal bloom monitoring and forecasting infrastructure, then this must be backed up by long-term forecasts of future bloom conditions. Furthermore, it is also important to continually evaluate if the predicted climate effects on algal blooms are playing out as forecasted. For such evaluation, long-term monitoring efforts are paramount. For example, various recent remote sensing studies have found that algal blooms have increased in many lakes in recent decades, and this serves to demonstrate that, to some extent, increased investment in monitoring and forecasting blooms has likely been justified (Fang *et al.*, 2022; Hou *et al.*, 2022; Wang *et al.*, 2025). However, these studies have typically focused on larger water bodies due to the limited spatial resolution of available satellite sensors. As such, whilst this provides strong

evidence for continued investment in monitoring many larger lakes, it gives no clear information on smaller-sized lakes which are typically less-well studied and therefore often overlooked (Downing, 2010; Biggs, von Fumetti and Kelly-Quinn, 2017).

However, it is likely that in approximately 10-15 years, there will be a sufficient time series of Planet SuperDoves-like high-resolution satellite imagery to carry out a much larger scale study of recent trends in algal bloom occurrence in all but the very smallest lakes. Such an effort would certainly benefit from future developments in inland water atmospheric correction and parameter retrieval algorithms. For example, it may be possible to use the SuperDoves 612 nm band to estimate phycocyanin concentrations, as this parameter has an absorption maximum near 620 nm (Simis, Peters and Gons, 2005). This would enable distinguishing between cyanobacteria-dominated blooms and other events with high chlorophyll-a signals. As such, it is possible that future studies will be able to use this imagery to study the effect of climate change on cyanobacterial bloom occurrence in a very large number of lakes and therefore provide critical information on where to focus additional monitoring, forecasting or restoration efforts. Furthermore, as previously discussed, such studies may be able to utilise large-scale AI approaches such as inland water specific foundation models, which may increase algorithm performance and generalisation across lakes. There are obviously challenges to such work, such as the lower signal-to-noise ratio of SuperDoves compared with sensors such as Sentinel-2 MSI, and noise due to differences between SuperDoves sensors (Kabir *et al.*, 2025). Together, these sources of noise may make it challenging to discern subtle trends over long time periods, especially given that the satellites are regularly decommissioned and replaced (M. Kim *et al.*, 2022). Therefore, it will likely be critical to attempt to harmonise measurements across SuperDoves sensors so that these are sufficiently comparable for long-term trend detection.

The relevance of the third chapter of this thesis in the context of a changing climate is clear – it evaluates the effect of climate change on short-term ML forecasts of water temperature. In this chapter, the concept of climate-resilient monitoring and forecasting system design is introduced. In effect, the proposal is that by using modelling to understand how climate change may affect our ability to make critical short-term forecasts, it is possible to design monitoring infrastructure which is sufficient to compensate for potential performance reductions driven by climate change. In the field of water quality monitoring and forecasting, this proposition is novel, and importantly, it may be valuable to apply the underlying principle further. For instance, by using modelling to

identify particular water bodies where algal blooms are likely to become a serious problem in future years, monitoring infrastructure could be pre-emptively installed at these critical sites so that data collection, and therefore potential mitigation strategies, can be initiated before water quality is severely compromised.

For example, May et al. (2022) used climate change projections to forecast which standing water bodies in Scotland will exceed an average monthly water temperature of 17 °C for one or more months per year in future decades. This was used as an indication of the potential future prevalence of algal blooms due to climate change, and therefore, with further refinements, such a modelling effort could be used to selectively target water bodies for increased surveillance. One way in which targeted monitoring of this kind could be further optimised would be by incorporating site-specific modelling of potential bloom impacts, such as water supply closures, economic impacts to recreational industries, or ecological damage, into the analysis (Hoagland and Scatasta, 2006; Amorim and Moura, 2021). Accordingly, this would enable implementation of increased monitoring for sites with both a high future risk of algal blooms, but also where the consequences of this would be most severe. Extending this further, it would be possible to perform a cost-benefit analysis for various monitoring and forecasting infrastructure investment options. For example, using process-based models to project future algal dynamics, the benefits of installing and maintaining a high frequency monitoring system could be compared with alternatives such as satellite monitoring, or lower frequency manual sampling. By simulating each of these data sources, this would enable estimation of the performance of data-driven short-term forecasts under different monitoring scenarios. Therefore, the potential economic benefits of various algal bloom monitoring and forecasting infrastructure investment options could be evaluated in advance, providing valuable information to decision makers. Ultimately, therefore, there could be great value in extending the principles laid out in the third chapter of this thesis so that climate-resilient monitoring and forecasting systems for algal blooms may be implemented in a cost-effective way for the water bodies where such infrastructure is, and will be, most needed.

6 Conclusions

This thesis presents three studies which all address important problems related to monitoring and forecasting water quality in lakes, particularly algal blooms. The key findings are as follows:

Experiments on long-term fortnightly monitoring data from Blelham Tarn showed that data-driven short-term algal forecasting models required five or more years of training data to be more effective than simple naïve models. In addition to increasing the training dataset size, forecast performance can also be enhanced through increasing sampling frequency and selection of an optimal forecast model and input features. Explicitly evaluating data requirements, particularly training dataset length, for algal bloom forecasting is a novel application of long-term monitoring data, one which it would be beneficial to repeat for other lakes, and sampling setups. Nonetheless, building algal bloom forecast models exclusively using data from a single lake is ultimately limited as even several decades of data only contains a relatively small number of bloom events to learn from. Furthermore, it was found that forecast performance improvements tended to show diminishing returns as more years of training data were added, suggesting that without increasing the training dataset size by several orders of magnitude, forecast skill effectively plateaus. However, with the recent development of self-supervised AI pre-training approaches, there is now an opportunity to use this to further algal bloom forecasting capabilities, potentially using high-resolution satellite imagery from many lakes simultaneously.

The comparison of new satellite imagery from Planet Labs SuperDoves constellation with data from ESA's Sentinel-2 MSI showed that the newer, 3 m resolution imagery has potential for monitoring algal blooms in small water bodies. This opens the possibility of monitoring a large number of very small water bodies which previously would not have been possible to monitor. The benefits of this potential capability include provision of training data for algal bloom forecast models, and enabling long-term studies of climate effects on algal blooms in small-sized lakes at regional or even global scales. This is particularly important because, despite being numerous, smaller lakes are often less-well studied than larger water bodies and are therefore often overlooked. However, for these applications, using high-resolution imagery from cubesats such as the SuperDoves still presents a number of challenges.

Firstly, widespread monitoring of vast numbers of lakes requires collection of large calibration and validation datasets. Accordingly, further exploring the opportunities offered by non-conventional satellite match-up data such as citizen science or agency monitoring records may be key to validating retrieval algorithms at scale. Secondly, other than for limited small-scale research projects, the SuperDoves imagery is not freely available. Furthermore, the computational and storage costs associated with working with daily, high-resolution imagery are considerably higher than those associated with freely available, lower resolution alternatives. Finally, cubesat constellations tend to have lower signal to noise ratios than alternatives, and calibration varies to some extent between each sensor. Therefore, further efforts are required to improve atmospheric correction procedures and harmonisation across sensors.

Ultimately, climate change poses a number of critical challenges to inland waters, and this is an important part of the justification for why the work in this thesis is important and timely. Experiments using forecasted water temperature time series data for five northern temperate lakes showed that, in summer, short-term forecast skill was reduced in more extreme climate change scenarios. Furthermore, it was demonstrated that adjusting sampling frequency, or training dataset duration could compensate for this skill reduction. Critically, this highlights a new potential principle for the design of climate-resilient water quality monitoring and forecasting systems. Specifically, using process-based lake models driven by climate projections, it is possible to evaluate whether a particular monitoring and forecasting system design will remain fit-for-purpose under potential future changes to the climate. With additional modelling efforts, there would be the possibility to test, under future climate conditions, the performance of various options for algal bloom monitoring and forecasting systems, including satellite-based monitoring or high frequency in situ measurement platforms.

Collectively, this thesis demonstrates that there is great value in considering a wide diversity of data sources and modelling approaches for furthering our ability to monitor, understand, and anticipate algal blooms. Generally, each of the wide variety of data sources used in this thesis has their own advantages and disadvantages, and so recognising and harnessing synergies between these is likely critical to effectively responding to the risk which algal blooms pose to water quality in lakes.

7 Bibliography

- 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.
- Ahn, C.-Y., Oh, H.-M. and Park, Y.-S. (2011) ‘Evaluation of Environmental Factors on Cyanobacterial Bloom in Eutrophic Reservoir Using Artificial Neural Networks¹’, *Journal of Phycology*, 47(3), pp. 495–504. Available at: <https://doi.org/10.1111/j.1529-8817.2011.00990.x>.
- Ahn, J.M., Kim, J. and Kim, K. (2023) ‘Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting’, *Toxins*, 15(10), p. 608. Available at: <https://doi.org/10.3390/toxins15100608>.
- Ai, H. *et al.* (2023) ‘Short-term Lake Erie algal bloom prediction by classification and regression models’, *Water Research*, 232, p. 119710. Available at: <https://doi.org/10.1016/j.watres.2023.119710>.
- Akinnawo, S.O. (2023) ‘Eutrophication: Causes, consequences, physical, chemical and biological techniques for mitigation strategies’, *Environmental Challenges*, 12, p. 100733. Available at: <https://doi.org/10.1016/j.envc.2023.100733>.
- Aláez, F.M.B. *et al.* (2021) ‘Machine Learning Methods Applied to the Prediction of Pseudo-nitzschia spp. Blooms in the Galician Rias Baixas (NW Spain)’, *ISPRS International Journal of Geo-Information*, 10(4), p. 199. Available at: <https://doi.org/10.3390/ijgi10040199>.
- Alawadi, F. (2010) ‘Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI)’, *Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010. Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010*, SPIE, pp. 45–58. Available at: <https://doi.org/10.1117/12.862096>.
- Alejandre, J.C.P. *et al.* (2024) ‘Scotland’s model of a blue-green prescribing programme for primary mental healthcare: participatory co-creation research with Q-Methodology’, *The Lancet*, 404, p. S95. Available at: [https://doi.org/10.1016/S0140-6736\(24\)02040-3](https://doi.org/10.1016/S0140-6736(24)02040-3).
- Allen, A. *et al.* (2025) ‘End-to-end data-driven weather prediction’, *Nature*, 641(8065), pp. 1172–1179. Available at: <https://doi.org/10.1038/s41586-025-08897-0>.
- Almuqati, M.T. *et al.* (2024) ‘Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview’, *International Journal on Advanced Science, Engineering and Information Technology*, 14(4), pp. 1449–1455. Available at: <https://doi.org/10.18517/ijaseit.14.4.20191>.
- Amara, N.D. *et al.* (2025) ‘GAN-Enhanced Deep Learning Approach for Forecasting the Potentially Toxic Cyanobacteria in Dams’, in A. Bennour *et al.* (eds) *Intelligent Systems and Pattern Recognition*. Cham: Springer Nature Switzerland, pp. 42–52. Available at: https://doi.org/10.1007/978-3-031-82153-0_4.
- Amorim, C.A. and Moura, A. do N. (2021) ‘Ecological impacts of freshwater algal blooms on water quality, plankton biodiversity, structure, and ecosystem functioning’, *Science of The Total Environment*, 758, p. 143605. Available at: <https://doi.org/10.1016/j.scitotenv.2020.143605>.
- de Amorim, L.B.V., Cavalcanti, G.D.C. and Cruz, R.M.O. (2023) ‘The choice of scaling technique matters for classification performance’, *Applied Soft Computing*, 133, p. 109924. Available at: <https://doi.org/10.1016/j.asoc.2022.109924>.

- Anderson, D.M., Glibert, P.M. and Burkholder, J.M. (2002) 'Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences', *Estuaries*, 25(4), pp. 704–726. Available at: <https://doi.org/10.1007/BF02804901>.
- Arabi, B. *et al.* (2020) 'The impact of sea bottom effects on the retrieval of water constituent concentrations from MERIS and OLCI images in shallow tidal waters supported by radiative transfer modeling', *Remote Sensing of Environment*, 237, p. 111596. Available at: <https://doi.org/10.1016/j.rse.2019.111596>.
- Arias, P.A. *et al.* (2021) 'Technical summary', in V. Masson-Delmotte *et al.* (eds) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
- Armon, M. *et al.* (2020) 'Determining Bathymetry of Shallow and Ephemeral Desert Lakes Using Satellite Imagery and Altimetry', *Geophysical Research Letters*, 47(7), p. e2020GL087367. Available at: <https://doi.org/10.1029/2020GL087367>.
- Arsen, A. *et al.* (2014) 'Remote Sensing-Derived Bathymetry of Lake Poopó', *Remote Sensing*, 6(1), pp. 407–420. Available at: <https://doi.org/10.3390/rs6010407>.
- Asai, R. *et al.* (2000) 'Development of a fluorometric sensor for the measurement of phycobilin pigment and application to freshwater phytoplankton', *Field Analytical Chemistry & Technology*, 4(1), pp. 53–61. Available at: [https://doi.org/10.1002/\(SICI\)1520-6521\(2000\)4:1%3C53::AID-FACT6%3E3.0.CO;2-C](https://doi.org/10.1002/(SICI)1520-6521(2000)4:1%3C53::AID-FACT6%3E3.0.CO;2-C).
- Atkinson, K.M. (1999) *Some English Lakes as Diverse and Active Ecosystems: A Factual Summary and Source Book*. Freshwater Biological Association.
- Atton Beckmann, D. *et al.* (2025) 'Are more data always better? – Machine learning forecasting of algae based on long-term observations', *Journal of Environmental Management*, 373, p. 123478. Available at: <https://doi.org/10.1016/j.jenvman.2024.123478>.
- Azevedo, S.M.F.O. *et al.* (2002) 'Human intoxication by microcystins during renal dialysis treatment in Caruaru—Brazil', *Toxicology*, 181–182, pp. 441–446. Available at: [https://doi.org/10.1016/S0300-483X\(02\)00491-2](https://doi.org/10.1016/S0300-483X(02)00491-2).
- Bahri, Y. *et al.* (2024) 'Explaining neural scaling laws', *Proceedings of the National Academy of Sciences*, 121(27), p. e2311878121. Available at: <https://doi.org/10.1073/pnas.2311878121>.
- Bai, X. *et al.* (2017) 'The Adaptive-Clustering and Error-Correction Method for Forecasting Cyanobacteria Blooms in Lakes and Reservoirs', *Advances in Mathematical Physics*, 2017, p. e9037358. Available at: <https://doi.org/10.1155/2017/9037358>.
- Bailey-Watts, A.E. (1981) 'Loch Leven phytoplankton succession'. NERC/ITE, pp. 88–92. Available at: <https://nora.nerc.ac.uk/id/eprint/10398/> (Accessed: 9 May 2024).
- Bakker, E.S. and Hilt, S. (2016) 'Impact of water-level fluctuations on cyanobacterial blooms: options for management', *Aquatic Ecology*, 50(3), pp. 485–498. Available at: <https://doi.org/10.1007/s10452-015-9556-x>.
- Baltodano, A., Agramont, A. and van Griensven, A. (2024) 'Combining Remote Sensing and Citizen Science to Bridge the Water Quality Data Gap: Lake Titicaca Case Study', *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3621–3625. Available at: <https://doi.org/10.1109/IGARSS53475.2024.10640780>.

- Bartlett, J.A. and Dedekorkut-Howes, A. (2022) 'Adaptation strategies for climate change impacts on water quality: a systematic review of the literature', *Journal of Water and Climate Change*, 14(3), pp. 651–675. Available at: <https://doi.org/10.2166/wcc.2022.279>.
- Beaulne, D. and Fotopoulos, G. (2024) 'Development of an algal bloom satellite and *in situ* metadata hub with case studies in Canada', *Ecological Informatics*, 79, p. 102447. Available at: <https://doi.org/10.1016/j.ecoinf.2023.102447>.
- Beck, R. *et al.* (2017) 'Comparison of Satellite Reflectance Algorithms for Estimating Phycocyanin Values and Cyanobacterial Total Biovolume in a Temperate Reservoir Using Coincident Hyperspectral Aircraft Imagery and Dense Coincident Surface Observations', *Remote Sensing*, 9(6), p. 538. Available at: <https://doi.org/10.3390/rs9060538>.
- Bengis, R. *et al.* (2016) 'Eco-epidemiological and pathological features of wildlife mortality events related to cyanobacterial bio-intoxication in the Kruger National Park, South Africa', *Journal of the South African Veterinary Association*, 87(1), p. 9. Available at: <https://doi.org/10.4102/jsava.v87i1.1391>.
- Bertani, I. *et al.* (2017) 'Tracking cyanobacteria blooms: Do different monitoring approaches tell the same story?', *Science of The Total Environment*, 575, pp. 294–308. Available at: <https://doi.org/10.1016/j.scitotenv.2016.10.023>.
- Beutel, M.W. (2006) 'Inhibition of ammonia release from anoxic profundal sediments in lakes using hypolimnetic oxygenation', *Ecological Engineering*, 28(3), pp. 271–279. Available at: <https://doi.org/10.1016/j.ecoleng.2006.05.009>.
- Bevacqua, E., Schleussner, C.-F. and Zscheischler, J. (2025) 'A year above 1.5 °C signals that Earth is most probably within the 20-year period that will reach the Paris Agreement limit', *Nature Climate Change*, 15(3), pp. 262–265. Available at: <https://doi.org/10.1038/s41558-025-02246-9>.
- Bezy, J.-L. *et al.* (1999) 'The ENVISAT Medium Resolution Imaging Spectrometer (MERIS)', *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293). IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)*, pp. 1432–1434 vol.2. Available at: <https://doi.org/10.1109/IGARSS.1999.774655>.
- Bhugal, D. *et al.* (2023) 'A Comparative Study of Water Temperature Forecasting Methods', *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 325–329. Available at: <https://doi.org/10.1109/WI-IAT59888.2023.00052>.
- Bialevich, V., Zachleder, V. and Bišová, K. (2022) 'The Effect of Variable Light Source and Light Intensity on the Growth of Three Algal Species', *Cells*, 11(8), p. 1293. Available at: <https://doi.org/10.3390/cells11081293>.
- Biggs, J., von Fumetti, S. and Kelly-Quinn, M. (2017) 'The importance of small waterbodies for biodiversity and ecosystem services: implications for policy makers', *Hydrobiologia*, 793(1), pp. 3–39. Available at: <https://doi.org/10.1007/s10750-016-3007-0>.
- Bilello, M.A. (1968) 'Water Temperatures in a Shallow Lake During Ice Formation, Growth, and Decay', *Water Resources Research*, 4(4), pp. 749–760. Available at: <https://doi.org/10.1029/WR004i004p00749>.
- Bilous, O.P. *et al.* (2023) 'Interactions between Aquatic Plants and Cyanobacterial Blooms in Freshwater Reservoir Ecosystems', *Water*, 15(4), p. 672. Available at: <https://doi.org/10.3390/w15040672>.

- Bird, T.J. *et al.* (2014) ‘Statistical solutions for error and bias in global citizen science datasets’, *Biological Conservation*, 173, pp. 144–154. Available at: <https://doi.org/10.1016/j.biocon.2013.07.037>.
- Bishop, I. *et al.* (2025) ‘FreshWater Watch: Investigating the Health of Freshwater Ecosystems, from the Bottom Up’, *Citizen Science: Theory and Practice*, 10(1). Available at: <https://doi.org/10.5334/cstp.754>.
- Bláha, L., Babica, P. and Maršálek, B. (2009) ‘Toxins produced in cyanobacterial water blooms – toxicity and risks’, *Interdisciplinary Toxicology*, 2(2), pp. 36–41. Available at: <https://doi.org/10.2478/v10102-009-0006-2>.
- Bodnar, C. *et al.* (2025) ‘A foundation model for the Earth system’, *Nature*, 641(8065), pp. 1180–1187. Available at: <https://doi.org/10.1038/s41586-025-09005-y>.
- Boehrer, B. and Schultze, M. (2008) ‘Stratification of lakes’, *Reviews of Geophysics*, 46(2). Available at: <https://doi.org/10.1029/2006RG000210>.
- Bojer, C.S. (2022) ‘Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities’, *International Journal of Forecasting*, 38(4), pp. 1555–1561. Available at: <https://doi.org/10.1016/j.ijforecast.2021.11.003>.
- Bonilla, S. *et al.* (2023) ‘Nutrients and not temperature are the key drivers for cyanobacterial biomass in the Americas’, *Harmful Algae*, 121, p. 102367. Available at: <https://doi.org/10.1016/j.hal.2022.102367>.
- Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) ‘A training algorithm for optimal margin classifiers’, *Proceedings of the fifth annual workshop on Computational learning theory*. New York, NY, USA: Association for Computing Machinery (COLT ’92), pp. 144–152. Available at: <https://doi.org/10.1145/130385.130401>.
- Boucher, O. *et al.* (2020) ‘Presentation and Evaluation of the IPSL-CM6A-LR Climate Model’, *Journal of Advances in Modeling Earth Systems*, 12(7), p. e2019MS002010. Available at: <https://doi.org/10.1029/2019MS002010>.
- Bowerman, B.L., O’Connell, R.T. and Koehler, A.B. (2005) *Forecasting, Time Series, and Regression: An Applied Approach*. Thomson Brooks/Cole.
- Brar, K.K. *et al.* (2025) ‘Image segmentation review: Theoretical background and recent advances’, *Information Fusion*, 114, p. 102608. Available at: <https://doi.org/10.1016/j.inffus.2024.102608>.
- Breiman, L. (1996) ‘Bagging predictors’, *Machine Learning*, 24(2), pp. 123–140. Available at: <https://doi.org/10.1007/BF00058655>.
- Breiman, L. (2001) ‘Random Forests’, *Machine Learning*, 45(1), pp. 5–32. Available at: <https://doi.org/10.1023/A:1010933404324>.
- Brenttrup, J.A. *et al.* (2016) ‘The potential of high-frequency profiling to assess vertical and seasonal patterns of phytoplankton dynamics in lakes: an extension of the Plankton Ecology Group (PEG) model’, *Inland Waters*, 6(4), pp. 565–580. Available at: <https://doi.org/10.5268/IW-6.4.890>.
- Brierley, B. *et al.* (2007) ‘Guidance on the quantitative analysis of phytoplankton in Freshwater Samples’, p. 24.
- Brockmann, C. *et al.* (2016) ‘Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters’, 740, p. 54.

Brooks, A.S. and Torke, B.G. (1977) 'Vertical and Seasonal Distribution of Chlorophyll a in Lake Michigan', *Journal of the Fisheries Research Board of Canada*, 34(12), pp. 2280–2287. Available at: <https://doi.org/10.1139/f77-306>.

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>.

Brown, T.B. *et al.* (2020) 'Language Models are Few-Shot Learners'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2005.14165>.

Brugger, J. *et al.* (2021) 'A Pronounced Spike in Ocean Productivity Triggered by the Chicxulub Impact', *Geophysical Research Letters*, 48(12), p. e2020GL092260. Available at: <https://doi.org/10.1029/2020GL092260>.

Buratti, F. *et al.* (2017) 'Cyanotoxins: producing organisms, occurrence, toxicity, mechanism of action and human health toxicological risk evaluation', *ARCHIVES OF TOXICOLOGY*, 91(3), pp. 1049–1130. Available at: <https://doi.org/10.1007/s00204-016-1913-6>.

Busari, I. *et al.* (2025) 'Advancing harmful algal bloom predictions using chlorophyll-a as an Indicator: Combining deep learning and EnKF data assimilation method', *Journal of Environmental Management*, 382, p. 125441. Available at: <https://doi.org/10.1016/j.jenvman.2025.125441>.

Busch, J.A. *et al.* (2016) 'Citizens and satellites: Assessment of phytoplankton dynamics in a NW Mediterranean aquaculture zone', *International Journal of Applied Earth Observation and Geoinformation*, 47, pp. 40–49. Available at: <https://doi.org/10.1016/j.jag.2015.11.017>.

Butsko, C. *et al.* (2025) *Deploying Geospatial Foundation Models in the Real World: Lessons from WorldCereal*, arXiv.org. Available at: <https://arxiv.org/abs/2508.00858v1> (Accessed: 27 November 2025).

Caballero, C.B. *et al.* (2025) 'The need for advancing algal bloom forecasting using remote sensing and modeling: Progress and future directions', *Ecological Indicators*, 172, p. 113244. Available at: <https://doi.org/10.1016/j.ecolind.2025.113244>.

Caballero, E. *et al.* (2023) 'Broken Neural Scaling Laws'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2210.14891>.

Cael, B.B. and Seekell, D.A. (2016) 'The size-distribution of Earth's lakes', *Scientific Reports*, 6(1), p. 29633. Available at: <https://doi.org/10.1038/srep29633>.

Cao, Z. *et al.* (2020) 'A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes', *Remote Sensing of Environment*, 248, p. 111974. Available at: <https://doi.org/10.1016/j.rse.2020.111974>.

Capon, S.J., Stewart-Koster, B. and Bunn, S.E. (2021) 'Future of Freshwater Ecosystems in a 1.5°C Warmer World', *Frontiers in Environmental Science*, 9. Available at: <https://doi.org/10.3389/fenvs.2021.784642>.

Carey, C.C. *et al.* (2012) 'Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate', *Water Research*, 46(5), pp. 1394–1407. Available at: <https://doi.org/10.1016/j.watres.2011.12.016>.

Carey, C.C. *et al.* (2025) 'A framework for developing a real-time lake phytoplankton forecasting system to support water quality management in the face of global change', *Ambio*, 54(3), pp. 475–487. Available at: <https://doi.org/10.1007/s13280-024-02076-7>.

- Carias, J. *et al.* (2024) 'Economic valuation of Harmful Algal Blooms (HAB): Methodological challenges, policy implications, and an empirical application', *Journal of Environmental Management*, 365, p. 121566. Available at: <https://doi.org/10.1016/j.jenvman.2024.121566>.
- Carmichael, W.W. (2001) 'Health Effects of Toxin-Producing Cyanobacteria: "The CyanoHABs"', *Human and Ecological Risk Assessment: An International Journal*, 7(5), pp. 1393–1407. Available at: <https://doi.org/10.1080/20018091095087>.
- Carpenter, S.R., Stanley, E.H. and Vander Zanden, M.J. (2011) 'State of the World's Freshwater Ecosystems: Physical, Chemical, and Biological Changes', *Annual Review of Environment and Resources*, 36(1), pp. 75–99. Available at: <https://doi.org/10.1146/annurev-environ-021810-094524>.
- Carvalho, L. *et al.* (2013) 'Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control', *Journal of Applied Ecology*, 50(2), pp. 315–323. Available at: <https://doi.org/10.1111/1365-2664.12059>.
- Casterlin, M.E. and Reynolds, W.W. (1977) 'Seasonal algal succession and cultural eutrophication in a north temperate lake', *Hydrobiologia*, 54(2), pp. 99–108. Available at: <https://doi.org/10.1007/BF00034983>.
- Chai, T. and Draxler, R.R. (2014) 'Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature', *Geoscientific Model Development*, 7(3), pp. 1247–1250. Available at: <https://doi.org/10.5194/gmd-7-1247-2014>.
- Chen, B. *et al.* (2025) 'REMSA: An LLM Agent for Foundation Model Selection in Remote Sensing'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2511.17442>.
- Chen, C. *et al.* (2024) 'Combining physical-based model and machine learning to forecast chlorophyll-a concentration in freshwater lakes', *Science of The Total Environment*, 907, p. 168097. Available at: <https://doi.org/10.1016/j.scitotenv.2023.168097>.
- Cho, K. *et al.* (2014) 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation'. arXiv. Available at: <https://doi.org/10.48550/arXiv.1406.1078>.
- Chorus, I. and Welker, M. (eds) (2021) *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*. 2nd edn. London: CRC Press. Available at: <https://doi.org/10.1201/9781003081449>.
- Christen, P., Hand, D.J. and Kirielle, N. (2023) 'A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives', *ACM Comput. Surv.*, 56(3), p. 73:1-73:24. Available at: <https://doi.org/10.1145/3606367>.
- Chung, J. *et al.* (2014) 'Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling'. arXiv. Available at: <https://doi.org/10.48550/arXiv.1412.3555>.
- Cleary, E.J. (1962) 'Robot Monitor System for the Ohio Valley', *Journal AWWA*, 54(11), pp. 1347–1352. Available at: <https://doi.org/10.1002/j.1551-8833.1962.tb00973.x>.
- Codd, G.A. *et al.* (2005) 'Harmful Cyanobacteria', in J. Huisman, H.C.P. Matthijs, and P.M. Visser (eds) *Harmful Cyanobacteria*. Dordrecht: Springer Netherlands (Aquatic Ecology Series), pp. 1–23. Available at: https://doi.org/10.1007/1-4020-3022-3_1.
- Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(3), pp. 273–297. Available at: <https://doi.org/10.1007/BF00994018>.

- Coulson, S., Woods, M. and Making Sense EU (2021) 'Citizen Sensing: An Action-Orientated Framework for Citizen Science', *Frontiers in Communication*, 6. Available at: <https://doi.org/10.3389/fcomm.2021.629700>.
- Crall, A.W. *et al.* (2011) 'Assessing citizen science data quality: an invasive species case study', *Conservation Letters*, 4(6), pp. 433–442. Available at: <https://doi.org/10.1111/j.1755-263X.2011.00196.x>.
- Cruz, R.C. *et al.* (2021) 'A Review of Recent Machine Learning Advances for Forecasting Harmful Algal Blooms and Shellfish Contamination', *Journal of Marine Science and Engineering*, 9(3), p. 283. Available at: <https://doi.org/10.3390/jmse9030283>.
- Dai, A. (2013) 'Increasing drought under global warming in observations and models', *Nature Climate Change*, 3(1), pp. 52–58. Available at: <https://doi.org/10.1038/nclimate1633>.
- D'Anglada, L. *et al.* (eds) (2016) *Harmful Algal Blooms (HABs) and Public Health: Progress and Current Challenges*. MDPI - Multidisciplinary Digital Publishing Institute. Available at: <https://doi.org/10.3390/books978-3-03842-156-6>.
- Davis, T.W. *et al.* (2009) 'The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms', *Harmful Algae*, 8(5), pp. 715–725. Available at: <https://doi.org/10.1016/j.hal.2009.02.004>.
- Davoli, F. *et al.* (2019) 'Small satellites and CubeSats: Survey of structures, architectures, and protocols', *International Journal of Satellite Communications and Networking*, 37(4), pp. 343–359. Available at: <https://doi.org/10.1002/sat.1277>.
- Daw, A. *et al.* (2022) 'Physics-Guided Neural Networks (PGNN): An Application in Lake Temperature Modeling', *Knowledge Guided Machine Learning*. Chapman and Hall/CRC.
- Dawson, G. *et al.* (2025) 'A Sentinel-3 foundation model for ocean colour'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2509.21273>.
- Demoulin, C.F. *et al.* (2019) 'Cyanobacteria evolution: Insight from the fossil record', *Free Radical Biology and Medicine*, 140, pp. 206–223. Available at: <https://doi.org/10.1016/j.freeradbiomed.2019.05.007>.
- Demšar, J. (2006) 'Statistical Comparisons of Classifiers over Multiple Data Sets', *Journal of Machine Learning Research*, 7(1), pp. 1–30.
- Derot, J., Yajima, H. and Schmitt, F.G. (2020) 'Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas', *Ecological Informatics*, 60, p. 101174. Available at: <https://doi.org/10.1016/j.ecoinf.2020.101174>.
- Derrac, J. *et al.* (2011) 'A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms', *Swarm and Evolutionary Computation*, 1(1), pp. 3–18. Available at: <https://doi.org/10.1016/j.swevo.2011.02.002>.
- Di Nunno, F. *et al.* (2023) 'A stacked machine learning model for multi-step ahead prediction of lake surface water temperature', *Science of The Total Environment*, 890, p. 164323. Available at: <https://doi.org/10.1016/j.scitotenv.2023.164323>.
- Dickinson, J.L. *et al.* (2012) 'The current state of citizen science as a tool for ecological research and public engagement', *Frontiers in Ecology and the Environment*, 10(6), pp. 291–297. Available at: <https://doi.org/10.1890/110236>.

- Dietze, M. *et al.* (2024) ‘Near-term ecological forecasting for climate change action’, *Nature Climate Change*, 14(12), pp. 1236–1244. Available at: <https://doi.org/10.1038/s41558-024-02182-0>.
- Dietze, M.C. *et al.* (2018) ‘Iterative near-term ecological forecasting: Needs, opportunities, and challenges’, *Proceedings of the National Academy of Sciences*, 115(7), pp. 1424–1432. Available at: <https://doi.org/10.1073/pnas.1710231115>.
- Ding, K. and Qin, B. (2024) ‘Global Drivers of Algal Bloom Occurrence: The Role of Climate, Hydrology, and Morphometry’, *Water*, 16(24), p. 3624. Available at: <https://doi.org/10.3390/w16243624>.
- Dodds, W.K. *et al.* (2009) ‘Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages’, *Environmental Science & Technology*, 43(1), pp. 12–19. Available at: <https://doi.org/10.1021/es801217q>.
- Dogliotti, A.I. *et al.* (2024) ‘Using the automated HYPERNETS hyperspectral system for multi-mission satellite ocean colour validation in the Río de la Plata, accounting for different spatial resolutions’, *Frontiers in Remote Sensing*, 5. Available at: <https://doi.org/10.3389/frsen.2024.1354662>.
- Dokulil, M.T. *et al.* (2021) ‘Increasing maximum lake surface temperature under climate change’, *Climatic Change*, 165(3), p. 56. Available at: <https://doi.org/10.1007/s10584-021-03085-1>.
- Dolah, F.M.V., Roelke, D. and Greene, R.M. (2001) ‘Health and Ecological Impacts of Harmful Algal Blooms: Risk Assessment Needs’, *Human and Ecological Risk Assessment: An International Journal*, 7(5), pp. 1329–1345. Available at: <https://doi.org/10.1080/20018091095032>.
- Donlon, C. *et al.* (2012) ‘The Sentinel-3 Mission: Overview and status’, *2012 IEEE International Geoscience and Remote Sensing Symposium. 2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1711–1714. Available at: <https://doi.org/10.1109/IGARSS.2012.6351194>.
- Doubek, J.P. and Carey, C.C. (2017) ‘Catchment, morphometric, and water quality characteristics differ between reservoirs and naturally formed lakes on a latitudinal gradient in the conterminous United States’, *Inland Waters*, 7(2), pp. 171–180. Available at: <https://doi.org/10.1080/20442041.2017.1293317>.
- Downing, J.A. *et al.* (2006) ‘The global abundance and size distribution of lakes, ponds, and impoundments’, *Limnology and Oceanography*, 51(5), pp. 2388–2397. Available at: <https://doi.org/10.4319/lo.2006.51.5.2388>.
- Downing, J.A. (2010) ‘Emerging global role of small lakes and ponds: little things mean a lot’, *Limnetica*, 29(1), pp. 9–24. Available at: <https://doi.org/10.23818/limn.29.02>.
- Drucker, H. *et al.* (1996) ‘Support Vector Regression Machines’, *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://papers.nips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html (Accessed: 25 June 2025).
- Drusch, M. *et al.* (2012) ‘Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services’, *Remote Sensing of Environment*, 120, pp. 25–36. Available at: <https://doi.org/10.1016/j.rse.2011.11.026>.
- Duan, H. *et al.* (2009) ‘Two-Decade Reconstruction of Algal Blooms in China’s Lake Taihu’, *Environmental Science & Technology*, 43(10), pp. 3522–3528. Available at: <https://doi.org/10.1021/es8031852>.

- Duan, Z. *et al.* (2024) 'Warming surface and Lake heatwaves as key drivers to harmful algal Blooms: A case study of Lake Dianchi, China', *Journal of Hydrology*, 632, p. 130971. Available at: <https://doi.org/10.1016/j.jhydrol.2024.130971>.
- Dubelaar, G.B.J., Geerders, P.J.F. and Jonker, R.R. (2004) 'High frequency monitoring reveals phytoplankton dynamics', *Journal of Environmental Monitoring*, 6(12), pp. 946–952. Available at: <https://doi.org/10.1039/B409350J>.
- Dugan, H.A. *et al.* (2017) 'Salting our freshwater lakes', *Proceedings of the National Academy of Sciences*, 114(17), pp. 4453–4458. Available at: <https://doi.org/10.1073/pnas.1620211114>.
- Dunne, J.P. *et al.* (2020) 'The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics', *Journal of Advances in Modeling Earth Systems*, 12(11), p. e2019MS002015. Available at: <https://doi.org/10.1029/2019MS002015>.
- Elliott, J.A., McElarney, Y.R. and Allen, M. (2016) 'The past and future of phytoplankton in the UK's largest lake, Lough Neagh', *Ecological Indicators*, 68, pp. 142–149. Available at: <https://doi.org/10.1016/j.ecolind.2015.07.015>.
- Elliott, J.A. and Thackeray, S.J. (2004) 'The simulation of phytoplankton in shallow and deep lakes using PROTECH', *Ecological Modelling*, 178(3), pp. 357–369. Available at: <https://doi.org/10.1016/j.ecolmodel.2004.02.012>.
- Emslie, S.D. *et al.* (1996) 'Integrated taphonomy of an avian death assemblage in marine sediments from the late Pliocene of Florida', *Palaeogeography, Palaeoclimatology, Palaeoecology*, 124(1), pp. 107–136. Available at: [https://doi.org/10.1016/0031-0182\(96\)00005-3](https://doi.org/10.1016/0031-0182(96)00005-3).
- Ernst, M.D. (2004) 'Permutation Methods: A Basis for Exact Inference', *Statistical Science*, 19(4), pp. 676–685.
- Evaristo, J. *et al.* (2023) 'Water woes: the institutional challenges in achieving SDG 6', *Sustainable Earth Reviews*, 6(1), p. 13. Available at: <https://doi.org/10.1186/s42055-023-00067-2>.
- Fang, C. *et al.* (2022) 'Global divergent trends of algal blooms detected by satellite during 1982–2018', *Global Change Biology*, 28(7), pp. 2327–2340. Available at: <https://doi.org/10.1111/gcb.16077>.
- Fang, X. and Stefan, H.G. (2012) 'Impacts of Climatic Changes on Water Quality and Fish Habitat in Aquatic Systems', in W.-Y. Chen *et al.* (eds) *Handbook of Climate Change Mitigation*. New York, NY: Springer US, pp. 531–569. Available at: https://doi.org/10.1007/978-1-4419-7991-9_16.
- Favot, E.J. *et al.* (2019) 'Climate variability promotes unprecedented cyanobacterial blooms in a remote, oligotrophic Ontario lake: evidence from paleolimnology', *Journal of Paleolimnology*, 62(1), pp. 31–52. Available at: <https://doi.org/10.1007/s10933-019-00074-4>.
- Fee, E.J. (1976) 'The vertical and seasonal distribution of chlorophyll in lakes of the Experimental Lakes Area, northwestern Ontario: Implications for primary production estimates', *Limnology and Oceanography*, 21(6), pp. 767–783. Available at: <https://doi.org/10.4319/lo.1976.21.6.0767>.
- Feldbauer, J. *et al.* (2020) 'Managing climate change in drinking water reservoirs: potentials and limitations of dynamic withdrawal strategies', *Environmental Sciences Europe*, 32(1), p. 48. Available at: <https://doi.org/10.1186/s12302-020-00324-7>.
- Feng, C., Li, L. and Xu, C. (2023) 'Advancements in predicting and modeling rare event outcomes for enhanced decision-making', *BMC Medical Research Methodology*, 23(1), p. 243. Available at: <https://doi.org/10.1186/s12874-023-02060-x>.

- Feng, L. *et al.* (2024) 'Harmful algal blooms in inland waters', *Nature Reviews Earth & Environment*, 5(9), pp. 631–644. Available at: <https://doi.org/10.1038/s43017-024-00578-2>.
- Feuchtmayr, H., Jones, I.D., Clarke, M., *et al.* (2021a) 'Hourly data from automatic water monitoring buoy from Blelham Tarn, 2016 to 2018'. NERC EDS Environmental Information Data Centre. Available at: <https://doi.org/10.5285/603629d9-618c-4a26-8a1d-235a4c8f4791>.
- Feuchtmayr, H., Jones, I.D., Clarke, M., *et al.* (2021b) 'Hourly data from automatic water monitoring buoy from Windermere South Basin, 2016 to 2018'. NERC EDS Environmental Information Data Centre. Available at: <https://doi.org/10.5285/e1b6e387-1bd6-444a-8de9-035df0b8baec>.
- Feuchtmayr, H., Clarke, M.A., De Ville, M.M., *et al.* (2021) *Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Blelham Tarn, 2014 to 2018 (Dataset)*, EIDC. Available at: <https://catalogue.ceh.ac.uk/id/ae8c850d-211e-4560-8b37-437b6e0e2a16> (Accessed: 12 September 2023).
- Feuchtmayr, H., Clarke, M.A., De Ville, M.M., *et al.* (2021) 'Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Esthwaite Water, 2014 to 2018'. NERC Environmental Information Data Centre. Available at: <https://doi.org/10.5285/89ff964d-4aed-4b66-8669-71d42c5add52>.
- Fink, G. *et al.* (2018) 'Phosphorus Loadings to the World's Largest Lakes: Sources and Trends', *Global Biogeochemical Cycles*, 32(4), pp. 617–634. Available at: <https://doi.org/10.1002/2017GB005858>.
- Fleming, L.E. *et al.* (2002) 'Blue green algal (cyanobacterial) toxins, surface drinking water, and liver cancer in Florida', *Harmful Algae*, 1(2), pp. 157–168. Available at: [https://doi.org/10.1016/S1568-9883\(02\)00026-4](https://doi.org/10.1016/S1568-9883(02)00026-4).
- Foley, B. *et al.* (2012) 'Long-term changes in oxygen depletion in a small temperate lake: effects of climate change and eutrophication', *Freshwater Biology*, 57(2), pp. 278–289. Available at: <https://doi.org/10.1111/j.1365-2427.2011.02662.x>.
- Fornarelli, R. *et al.* (2013) 'An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers', *Water Resources Research*, 49(6), pp. 3626–3641. Available at: <https://doi.org/10.1002/wrcr.20268>.
- Fournier, C. *et al.* (2024) 'LSTM networks provide efficient cyanobacterial blooms forecasting even with incomplete spatio-temporal data', *Water Research*, 267, p. 122553. Available at: <https://doi.org/10.1016/j.watres.2024.122553>.
- Franks, P.J.S. (2018) 'Recent Advances in Modelling of Harmful Algal Blooms', in P.M. Glibert *et al.* (eds) *Global Ecology and Oceanography of Harmful Algal Blooms*. Cham: Springer International Publishing (Ecological Studies), pp. 359–377. Available at: https://doi.org/10.1007/978-3-319-70069-4_19.
- Freitas, F.H. and Dierssen, H.M. (2019) 'Evaluating the seasonal and decadal performance of red band difference algorithms for chlorophyll in an optically complex estuary with winter and summer blooms', *Remote Sensing of Environment*, 231, p. 111228. Available at: <https://doi.org/10.1016/j.rse.2019.111228>.
- Frey, L.J. and Fisher, D.H. (1999) 'Modeling decision tree performance with the power law', *Seventh International Workshop on Artificial Intelligence and Statistics. Seventh International Workshop on Artificial Intelligence and Statistics*, PMLR. Available at: <https://proceedings.mlr.press/r2/frey99a.html> (Accessed: 15 December 2023).

- Frieler, K. *et al.* (2024) ‘Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP3a)’, *Geoscientific Model Development*, 17(1), pp. 1–51. Available at: <https://doi.org/10.5194/gmd-17-1-2024>.
- de la Fuente, N. *et al.* (2024) ‘Enhancing Image Classification in Small and Unbalanced Datasets Through Synthetic Data Augmentation’, in K. Drechsler *et al.* (eds) *Clinical Image-Based Procedures*. Cham: Springer Nature Switzerland, pp. 11–21. Available at: https://doi.org/10.1007/978-3-031-73083-2_2.
- Ganf, G.G. and Oliver, R.L. (1982) ‘Vertical Separation of Light and Available Nutrients as a Factor Causing Replacement of Green Algae by Blue-Green Algae in the Plankton of a Stratified Lake’, *Journal of Ecology*, 70(3), pp. 829–844. Available at: <https://doi.org/10.2307/2260107>.
- Gao, B.-C. *et al.* (2000) ‘Atmospheric correction algorithm for hyperspectral remote sensing of ocean color from space’, *Applied Optics*, 39(6), pp. 887–896. Available at: <https://doi.org/10.1364/AO.39.000887>.
- Gaudard, A. *et al.* (2019) ‘Toward an open access to high-frequency lake modeling and statistics data for scientists and practitioners – the case of Swiss lakes using Simstrat v2.1’, *Geoscientific Model Development*, 12(9), pp. 3955–3974. Available at: <https://doi.org/10.5194/gmd-12-3955-2019>.
- Gewin, V. (2016) ‘Data sharing: An open mind on open data’, *Nature*, 529(7584), pp. 117–119. Available at: <https://doi.org/10.1038/nj7584-117a>.
- Giamalaki, K., Beaulieu, C. and Prochaska, J.X. (2022) ‘Assessing Predictability of Marine Heatwaves With Random Forests’, *Geophysical Research Letters*, 49(23), p. e2022GL099069. Available at: <https://doi.org/10.1029/2022GL099069>.
- Giannuzzi, L. *et al.* (2011) ‘An Acute Case of Intoxication with Cyanobacteria and Cyanotoxins in Recreational Water in Salto Grande Dam, Argentina’, *Marine Drugs*, 9(11), pp. 2164–2175. Available at: <https://doi.org/10.3390/md9112164>.
- Gibbons, K.J. and Bridgeman, T.B. (2020) ‘Effect of temperature on phosphorus flux from anoxic western Lake Erie sediments’, *Water Research*, 182, p. 116022. Available at: <https://doi.org/10.1016/j.watres.2020.116022>.
- Gilerson, A.A. *et al.* (2010) ‘Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands’, *Optics Express*, 18(23), pp. 24109–24125. Available at: <https://doi.org/10.1364/OE.18.024109>.
- Gitelson, A. (1992) ‘The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration’, *International Journal of Remote Sensing*, 13(17), pp. 3367–3373. Available at: <https://doi.org/10.1080/01431169208904125>.
- Gleick, P.H. and Srinivasan, V. (2011) ‘Water Resources’, *Encyclopedia of Climate and Weather*. Edited by S. H. Schneider, T. L. Root, and M. D. Mastrandrea. Oxford University Press. Available at: <https://www.oxfordreference.com/view/10.1093/acref/9780199765324.001.0001/acref-9780199765324-e-0511> (Accessed: 8 June 2022).
- Glibert, P.M. and Burkholder, J.M. (2018) ‘Causes of Harmful Algal Blooms’, *Harmful Algal Blooms*. John Wiley & Sons, Ltd, pp. 1–38. Available at: <https://doi.org/10.1002/9781118994672.ch1>.
- Golub, M. *et al.* (2022) ‘A framework for ensemble modelling of climate change impacts on lakes worldwide: the ISIMIP Lake Sector’, *Geoscientific Model Development*, 15(11), pp. 4597–4623. Available at: <https://doi.org/10.5194/gmd-15-4597-2022>.

- Gons, H.J., Auer, M.T. and Effler, S.W. (2008) 'MERIS satellite chlorophyll mapping of oligotrophic and eutrophic waters in the Laurentian Great Lakes', *Remote Sensing of Environment*, 112(11), pp. 4098–4106. Available at: <https://doi.org/10.1016/j.rse.2007.06.029>.
- Gons, H.J., Rijkeboer, M. and Ruddick, K.G. (2002) 'A chlorophyll-retrieval algorithm for satellite imagery (Medium Resolution Imaging Spectrometer) of inland and coastal waters', *Journal of Plankton Research*, 24(9), pp. 947–951. Available at: <https://doi.org/10.1093/plankt/24.9.947>.
- González Vilas, L. *et al.* (2014) 'Support Vector Machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain)', *Progress in Oceanography*, 124, pp. 66–77. Available at: <https://doi.org/10.1016/j.pocean.2014.03.003>.
- Gordon, H. (2019) *Physical Principles of Ocean Color Remote Sensing*. Available at: <https://doi.org/10.33596/ppocrs-19>.
- Granger, C.W.J. and Newbold, P. (1974) 'Spurious regressions in econometrics', *Journal of Econometrics*, 2(2), pp. 111–120. Available at: [https://doi.org/10.1016/0304-4076\(74\)90034-7](https://doi.org/10.1016/0304-4076(74)90034-7).
- Grendaité, D. *et al.* (2025) 'Mapping aquatic vegetation genera in temperate lakes with remote sensing data', *International Journal of Remote Sensing*, 46(17), pp. 6428–6450. Available at: <https://doi.org/10.1080/01431161.2025.2538828>.
- Grover, J.P. and Chrzanowski, T.H. (2006) 'Seasonal dynamics of phytoplankton in two warm temperate reservoirs: association of taxonomic composition with temperature', *Journal of Plankton Research*, 28(1), pp. 1–17. Available at: <https://doi.org/10.1093/plankt/fbi095>.
- Guo, F. *et al.* (2021) 'Trends of temperature variability: Which variability and what health implications?', *Science of The Total Environment*, 768, p. 144487. Available at: <https://doi.org/10.1016/j.scitotenv.2020.144487>.
- Gupta, A., Hantush, M.M. and Govindaraju, R.S. (2023) 'Sub-monthly time scale forecasting of harmful algal blooms intensity in Lake Erie using remote sensing and machine learning', *Science of The Total Environment*, 900, p. 165781. Available at: <https://doi.org/10.1016/j.scitotenv.2023.165781>.
- Gutjahr, O. *et al.* (2019) 'Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP)', *Geoscientific Model Development*, 12(7), pp. 3241–3281. Available at: <https://doi.org/10.5194/gmd-12-3241-2019>.
- Haeussler, C. *et al.* (2014) 'Specific and general information sharing among competing academic researchers', *Research Policy*, 43(3), pp. 465–475. Available at: <https://doi.org/10.1016/j.respol.2013.08.017>.
- Halabaku, E. and Bytyçi, E. (2024) 'Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests', *Intelligent Automation & Soft Computing*, 39(6), pp. 987–1006. Available at: <https://doi.org/10.32604/iasc.2024.059429>.
- Hallegraeff, G.M. *et al.* (2021) 'Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts', *Communications Earth & Environment*, 2(1), pp. 1–10. Available at: <https://doi.org/10.1038/s43247-021-00178-8>.
- Hamilton, D.P. *et al.* (2014) 'Costs of harmful blooms of freshwater cyanobacteria', *Cyanobacteria*. John Wiley & Sons, Ltd, pp. 245–256. Available at: <https://doi.org/10.1002/9781118402238.ch15>.

- Harmel, R.D. *et al.* (2023) 'A Review of Data Quality and Cost Considerations for Water Quality Monitoring at the Field Scale and in Small Watersheds', *Water*, 15(17), p. 3110. Available at: <https://doi.org/10.3390/w15173110>.
- Harris, T.D. *et al.* (2024) 'What makes a cyanobacterial bloom disappear? A review of the abiotic and biotic cyanobacterial bloom loss factors', *Harmful Algae*, 133, p. 102599. Available at: <https://doi.org/10.1016/j.hal.2024.102599>.
- Harris, T.D. and Graham, J.L. (2017) 'Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset', *Lake and Reservoir Management*, 33(1), pp. 32–48. Available at: <https://doi.org/10.1080/10402381.2016.1263694>.
- He, Y., Wang, X. and Xu, F. (2022) 'How reliable is chlorophyll-*a* as algae proxy in lake environments? New insights from the perspective of *n*-alkanes.', *Science of The Total Environment*, 836, p. 155700. Available at: <https://doi.org/10.1016/j.scitotenv.2022.155700>.
- Heddiam, S., Ptak, M. and Zhu, S. (2020) 'Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN', *Journal of Hydrology*, 588, p. 125130. Available at: <https://doi.org/10.1016/j.jhydrol.2020.125130>.
- Hei, P. *et al.* (2025) 'Unraveling eutrophication controversies: Innovative strategies and holistic perspectives', *Critical Reviews in Environmental Science and Technology*, 55(3), pp. 147–168. Available at: <https://doi.org/10.1080/10643389.2024.2392987>.
- Heino, J. *et al.* (2021) 'Lakes in the era of global change: moving beyond single-lake thinking in maintaining biodiversity and ecosystem services', *Biological Reviews of the Cambridge Philosophical Society*, 96(1), pp. 89–106. Available at: <https://doi.org/10.1111/brv.12647>.
- Herman, J.D. and Giuliani, M. (2018) 'Policy tree optimization for threshold-based water resources management over multiple timescales', *Environmental Modelling & Software*, 99, pp. 39–51. Available at: <https://doi.org/10.1016/j.envsoft.2017.09.016>.
- Herrero, A., Muro-Pastor, A.M. and Flores, E. (2001) 'Nitrogen Control in Cyanobacteria', *Journal of Bacteriology*, 183(2), pp. 411–425. Available at: <https://doi.org/10.1128/jb.183.2.411-425.2001>.
- Hewamalage, H., Bergmeir, C. and Bandara, K. (2021) 'Recurrent Neural Networks for Time Series Forecasting: Current status and future directions', *International Journal of Forecasting*, 37(1), pp. 388–427. Available at: <https://doi.org/10.1016/j.ijforecast.2020.06.008>.
- Hill, P.R. *et al.* (2020) 'HABNet: Machine Learning, Remote Sensing-Based Detection of Harmful Algal Blooms', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, pp. 3229–3239. Available at: <https://doi.org/10.1109/JSTARS.2020.3001445>.
- Hipsey, M.R. *et al.* (2019) 'A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON)', *Geoscientific Model Development*, 12(1), pp. 473–523. Available at: <https://doi.org/10.5194/gmd-12-473-2019>.
- Hipsey, M.R. (2022) 'Modelling Aquatic Eco-Dynamics: Overview of the AED modular simulation platform'. Available at: <https://doi.org/10.5281/zenodo.6516222>.
- Ho, J.C., Michalak, A.M. and Pahlevan, N. (2019) 'Widespread global increase in intense lake phytoplankton blooms since the 1980s', *Nature*, 574(7780), pp. 667–670. Available at: <https://doi.org/10.1038/s41586-019-1648-7>.

- Hoagland, P. and Scatista, S. (2006) 'The Economic Effects of Harmful Algal Blooms', in E. Granéli and J.T. Turner (eds) *Ecology of Harmful Algae*. Springer Berlin Heidelberg (Ecological Studies), pp. 391–402. Available at: https://doi.org/10.1007/978-3-540-32210-8_30.
- Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8), pp. 1735–1780. Available at: <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Højerslev, N.K. (1974) *Inherent and apparent optical properties of the Baltic*. Copenhagen: Københavns universitet, Institut for fysisk oceanografi.
- Holm, S. (1979) 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, 6(2), pp. 65–70.
- Hooker, S. *et al.* (2009) *The Third SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-3)*. Greenbelt, Maryland (United States of America). Available at: <https://publications.jrc.ec.europa.eu/repository/handle/JRC57305> (Accessed: 30 April 2024).
- Hooker, S.B. and Esaias, W.E. (1993) 'An overview of the SeaWiFS Project', *Eos, Transactions American Geophysical Union*, 74(21), pp. 241–246. Available at: <https://doi.org/10.1029/93EO00945>.
- Hou, X. *et al.* (2022) 'Global mapping reveals increase in lacustrine algal blooms over the past decade', *Nature Geoscience*, 15(2), pp. 130–134. Available at: <https://doi.org/10.1038/s41561-021-00887-x>.
- Hovis, W.A. *et al.* (1980) 'Nimbus-7 Coastal Zone Color Scanner: System Description and Initial Imagery', *Science*, 210(4465), pp. 60–63.
- Hu, C. *et al.* (2004) 'Assessment of estuarine water-quality indicators using MODIS medium-resolution bands: Initial results from Tampa Bay, FL', *Remote Sensing of Environment*, 93(3), pp. 423–441. Available at: <https://doi.org/10.1016/j.rse.2004.08.007>.
- Hu, C. (2009) 'A novel ocean color index to detect floating algae in the global oceans', *Remote Sensing of Environment*, 113(10), pp. 2118–2129. Available at: <https://doi.org/10.1016/j.rse.2009.05.012>.
- Hu, W. (2016) 'A review of the models for Lake Taihu and their application in lake environmental management', *Ecological Modelling*, 319, pp. 9–20. Available at: <https://doi.org/10.1016/j.ecolmodel.2015.07.028>.
- Hughes, M. *et al.* (2004) 'The Development of a GIS-based Inventory of Standing Waters in Great Britain together with a Risk-based Prioritisation Protocol', *Water, Air and Soil Pollution: Focus*, 4(2), pp. 73–84. Available at: <https://doi.org/10.1023/B:WAFO.0000028346.27904.83>.
- Huisman, J. *et al.* (2018) 'Cyanobacterial blooms', *Nature Reviews Microbiology*, 16(8), pp. 471–483. Available at: <https://doi.org/10.1038/s41579-018-0040-1>.
- Hunter, P.D. *et al.* (2010) 'Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes', *Remote Sensing of Environment*, 114(11), pp. 2705–2718. Available at: <https://doi.org/10.1016/j.rse.2010.06.006>.
- Hutchings, A.M. *et al.* (2024) 'Temperature and dissolved oxygen trends in English estuaries over the past 30 years', *Estuarine, Coastal and Shelf Science*, 306, p. 108892. Available at: <https://doi.org/10.1016/j.ecss.2024.108892>.
- Hutchinson, G.E. (1957) *A Treatise on Limnology. Vol. 1, Geography, Physics, and Chemistry*. Available at: <https://www.cambridge.org/core/journals/geological-magazine/article/abs/treatise-on-limnology-vol-1-geography-physics-and-chemistry-by-g-e-hutchinson-john-wiley-chapman>

and-hall-london-pp-xiv-1015-with-228-figs-and-132-tables-1957-price-7-16s/40F202509FC47603082079A4C5A0D92E (Accessed: 23 May 2025).

Hyndman, R.J. and Athanasopoulos, G. (2021) *Forecasting: Principles and Practice (3rd ed)*. Available at: <https://otexts.com/fpp3/> (Accessed: 13 June 2022).

Ibelings, B.W. *et al.* (2007) 'Resilience of Alternative Stable States during the Recovery of Shallow Lakes from Eutrophication: Lake Veluwe as a Case Study', *Ecosystems*, 10(1), pp. 4–16. Available at: <https://doi.org/10.1007/s10021-006-9009-4>.

Ibelings, B.W. *et al.* (2016) 'Cyanobacterial blooms. Ecology, prevention, mitigation and control: Editorial to a CYANOCOST Special Issue', *Aquatic Ecology*, 50(3), pp. 327–331. Available at: <https://doi.org/10.1007/s10452-016-9595-y>.

Igwaran, A. *et al.* (2024) 'Cyanobacteria Harmful Algae Blooms: Causes, Impacts, and Risk Management', *Water, Air, & Soil Pollution*, 235(1), p. 71. Available at: <https://doi.org/10.1007/s11270-023-06782-y>.

Inácio, M. *et al.* (2022) 'Mapping lake ecosystem services: A systematic review', *Science of The Total Environment*, 847, p. 157561. Available at: <https://doi.org/10.1016/j.scitotenv.2022.157561>.

Ioannou, I. *et al.* (2013) 'Deriving ocean color products using neural networks', *Remote Sensing of Environment*, 134, pp. 78–91. Available at: <https://doi.org/10.1016/j.rse.2013.02.015>.

IOCCG (2000) 'Remote Sensing of Ocean Colour in Coastal and Other Optically Complex Waters'. No 3, IOCCG Dartmouth, Canada.

IPCC (2023) 'Summary for Policymakers', *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Edited by H. Lee and J. Romero. Geneva, Switzerland: IPCC. Available at: <https://doi.org/10.59327/IPCC/AR6-9789291691647.001>.

Irani Rahaghi, A. *et al.* (2024) 'Combined Earth observations reveal the sequence of conditions leading to a large algal bloom in Lake Geneva', *Communications Earth & Environment*, 5(1), p. 229. Available at: <https://doi.org/10.1038/s43247-024-01351-5>.

Izadi, M. *et al.* (2021) 'A Remote Sensing and Machine Learning-Based Approach to Forecast the Onset of Harmful Algal Bloom', *Remote Sensing*, 13(19), p. 3863. Available at: <https://doi.org/10.3390/rs13193863>.

Jackson-Blake, L.A. *et al.* (2022) 'Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network', *Hydrology and Earth System Sciences*, 26(12), pp. 3103–3124. Available at: <https://doi.org/10.5194/hess-26-3103-2022>.

Jain, P. *et al.* (2024) *SenCLIP: Enhancing zero-shot land-use mapping for Sentinel-2 with ground-level prompting*, *arXiv.org*. Available at: <https://doi.org/10.1109/WACV61041.2025.00552>.

Jakubik, J. *et al.* (2025) 'TerraMind: Large-Scale Generative Multimodality for Earth Observation'. *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2504.11171>.

Janssen, A.B.G. *et al.* (2014) 'Alternative stable states in large shallow lakes?', *Journal of Great Lakes Research*, 40(4), pp. 813–826. Available at: <https://doi.org/10.1016/j.jglr.2014.09.019>.

Jetoo, S., Grover, V.I. and Krantzberg, G. (2015) 'The Toledo Drinking Water Advisory: Suggested Application of the Water Safety Planning Approach', *Sustainability*, 7(8), pp. 9787–9808. Available at: <https://doi.org/10.3390/su7089787>.

- Jia, X. *et al.* (2019) ‘Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles’, *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*. Society for Industrial and Applied Mathematics (Proceedings), pp. 558–566. Available at: <https://doi.org/10.1137/1.9781611975673.63>.
- Jia, X. *et al.* (2021) ‘Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles’, *ACM/IMS Trans. Data Sci.*, 2(3), p. 20:1-20:26. Available at: <https://doi.org/10.1145/3447814>.
- Jiang, D. *et al.* (2023) ‘A data-driven approach to flag land-affected signals in satellite derived water quality from small lakes’, *International Journal of Applied Earth Observation and Geoinformation*, 117, p. 103188. Available at: <https://doi.org/10.1016/j.jag.2023.103188>.
- Jiang, D., Matsushita, B. and Yang, W. (2020) ‘A simple and effective method for removing residual reflected skylight in above-water remote sensing reflectance measurements’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 165, pp. 16–27. Available at: <https://doi.org/10.1016/j.isprsjprs.2020.05.003>.
- Jiang, X. and Ge, Z. (2021) ‘Data Augmentation Classifier for Imbalanced Fault Classification’, *IEEE Transactions on Automation Science and Engineering*, 18(3), pp. 1206–1217. Available at: <https://doi.org/10.1109/TASE.2020.2998467>.
- Jin, M. *et al.* (2023) ‘Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.10196>.
- Kabir, S. *et al.* (2025) ‘Feasibility of PlanetScope SuperDove constellation for water quality monitoring of inland and coastal waters’, *Frontiers in Remote Sensing*, 6. Available at: <https://doi.org/10.3389/frsen.2025.1624783>.
- Kahn, R.A. *et al.* (2016) ‘The Sensitivity of SeaWiFS Ocean Color Retrievals to Aerosol Amount and Type’, *Journal of Atmospheric and Oceanic Technology*, 33(6), pp. 1185–1209. Available at: <https://doi.org/10.1175/JTECH-D-15-0121.1>.
- Kamarthi, H. and Prakash, B.A. (2024) ‘Large Pre-trained time series models for cross-domain Time series analysis tasks’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2311.11413>.
- Karimi, B., Hashemi, S.H. and Aghighi, H. (2024) ‘Application of Landsat-8 and Sentinel-2 for retrieval of chlorophyll-*a* in a shallow freshwater lake’, *Advances in Space Research*, 74(1), pp. 117–129. Available at: <https://doi.org/10.1016/j.asr.2024.03.056>.
- Karpatne, A. *et al.* (2017) ‘Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data’, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. Available at: <https://doi.org/10.1109/TKDE.2017.2720168>.
- Kelly, L. a. and Smith, S. (1996) ‘The nutrient budget of a small eutrophic loch and the effectiveness of straw bales in controlling algal blooms’, *Freshwater Biology*, 36(2), pp. 411–418. Available at: <https://doi.org/10.1046/j.1365-2427.1996.00059.x>.
- Kesebir, S. and Kesebir, P. (2017) ‘A Growing Disconnection From Nature Is Evident in Cultural Products’, *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(2), pp. 258–269. Available at: <https://doi.org/10.1177/1745691616662473>.
- Kim, M. *et al.* (2022) *System characterization report on Planet’s SuperDove, Open-File Report*. 2021-1030-F. U.S. Geological Survey. Available at: <https://doi.org/10.3133/ofr20211030F>.
- Kim, S.-C., Cerco, C.F. and Johnson, B.H. (2006) ‘Three-dimensional Management Model for Lake Washington, Part I: Introduction and Hydrodynamic Modeling’, *Lake and Reservoir Management*, 22(2), pp. 103–114. Available at: <https://doi.org/10.1080/07438140609353888>.

- Kim, T. *et al.* (2022) ‘Simultaneous feature engineering and interpretation: Forecasting harmful algal blooms using a deep learning approach’, *Water Research*, 215, p. 118289. Available at: <https://doi.org/10.1016/j.watres.2022.118289>.
- Koenigswald, W. v., Braun, A. and Pfeiffer, T. (2004) ‘Cyanobacteria and seasonal death: A new taphonomic model for the Eocene Messel lake’, *Paläontologische Zeitschrift*, 78(2), pp. 417–424. Available at: <https://doi.org/10.1007/BF03009232>.
- Koutroumbas, K. and Theodoridis, S. (2008) *Pattern Recognition*. San Diego, UNITED STATES: Elsevier Science & Technology. Available at: <http://ebookcentral.proquest.com/lib/stir/detail.action?docID=534973> (Accessed: 12 September 2023).
- Koutsandreas, D. *et al.* (2022) ‘On the selection of forecasting accuracy measures’, *Journal of the Operational Research Society*, 73(5), pp. 937–954. Available at: <https://doi.org/10.1080/01605682.2021.1892464>.
- Kozik, N.P. *et al.* (2022) ‘Rapid marine oxygen variability: Driver of the Late Ordovician mass extinction’, *Science Advances*, 8(46), p. eabn8345. Available at: <https://doi.org/10.1126/sciadv.abn8345>.
- Kutser, T. and Soomets, T. (2024) ‘Satellite data is revealing long time changes in the world largest lakes’, *Scientific Reports*, 14(1), p. 14391. Available at: <https://doi.org/10.1038/s41598-024-65250-7>.
- Kuznetsova, A., Brockhoff, P.B. and Christensen, R.H.B. (2017) ‘lmerTest Package: Tests in Linear Mixed Effects Models’, *Journal of Statistical Software*, 82, pp. 1–26. Available at: <https://doi.org/10.18637/jss.v082.i13>.
- Larsen, A.S. *et al.* (2017) ‘Physical and chemical characteristics of lakes across heterogeneous landscapes in arctic and subarctic Alaska’, *Journal of Geophysical Research: Biogeosciences*, 122(4), pp. 989–1008. Available at: <https://doi.org/10.1002/2016JG003729>.
- Last, M. (2007) ‘Predicting and Optimizing Classifier Utility with the Power Law’, *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pp. 219–224. Available at: <https://doi.org/10.1109/ICDMW.2007.31>.
- Lavanderos, L. *et al.* (1990) ‘Landsat image and sample design for water reservoirs (Rapel dam Central Chile)’, *Environmental Monitoring and Assessment*, 14(1), pp. 9–22. Available at: <https://doi.org/10.1007/BF00394354>.
- Lavigne, H. *et al.* (2023) ‘Turbid water sun glint removal for high resolution sensors without SWIR’, *Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2023. Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2023*, SPIE, pp. 30–45. Available at: <https://doi.org/10.1117/12.2683912>.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep learning’, *Nature*, 521(7553), pp. 436–444. Available at: <https://doi.org/10.1038/nature14539>.
- Lever, J., Krzywinski, M. and Altman, N. (2016) ‘Model selection and overfitting’, *Nature Methods*, 13(9), pp. 703–704. Available at: <https://doi.org/10.1038/nmeth.3968>.
- Li, Jianzhong *et al.* (2024) ‘Evaluating the capabilities of China’s new satellite HJ-2 for monitoring chlorophyll *a* concentration in eutrophic lakes’, *International Journal of Applied Earth Observation and Geoinformation*, 126, p. 103618. Available at: <https://doi.org/10.1016/j.jag.2023.103618>.

- Li, K. *et al.* (2021) 'Multi-factor analysis of algal blooms in gate-controlled urban water bodies by data mining', *Science of The Total Environment*, 753, p. 141821. Available at: <https://doi.org/10.1016/j.scitotenv.2020.141821>.
- Liang, Q. *et al.* (2017) 'A MODIS-Based Novel Method to Distinguish Surface Cyanobacterial Scums and Aquatic Macrophytes in Lake Taihu', *Remote Sensing*, 9(2), p. 133. Available at: <https://doi.org/10.3390/rs9020133>.
- Lim, T. and Wang, K. (2022) 'Comparison of machine learning algorithms for emulation of a gridded hydrological model given spatially explicit inputs', *Computers & Geosciences*, 159, p. 105025. Available at: <https://doi.org/10.1016/j.cageo.2021.105025>.
- Lin, S., Pierson, D.C. and Mesman, J.P. (2023) 'Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake', *Geoscientific Model Development*, 16(1), pp. 35–46. Available at: <https://doi.org/10.5194/gmd-16-35-2023>.
- Lin, T.-F. *et al.* (2019) *Taste and Odour in Source and Drinking Water: Causes, Controls, and Consequences*. London, UNITED KINGDOM: IWA Publishing. Available at: <http://ebookcentral.proquest.com/lib/stir/detail.action?docID=5888218> (Accessed: 8 December 2021).
- Lindenmayer, D.B. and Likens, G.E. (2010) 'The science and application of ecological monitoring', *Biological Conservation*, 143(6), pp. 1317–1328. Available at: <https://doi.org/10.1016/j.biocon.2010.02.013>.
- Liu, S. *et al.* (2022) 'Remote sensing to detect harmful algal blooms in inland waterbodies', *Science of The Total Environment*, 851, p. 158096. Available at: <https://doi.org/10.1016/j.scitotenv.2022.158096>.
- Liu, W.-C. and Chen, W.-B. (2012) 'Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models', *Computers & Geosciences*, 45, pp. 13–25. Available at: <https://doi.org/10.1016/j.cageo.2012.03.010>.
- Liu, X., Lu, X. and Chen, Y. (2011) 'The effects of temperature and nutrient ratios on *Microcystis* blooms in Lake Taihu, China: An 11-year investigation', *Harmful Algae*, 10(3), pp. 337–343. Available at: <https://doi.org/10.1016/j.hal.2010.12.002>.
- Lomeo, D. *et al.* (2024) 'Remote sensing and spatial analysis reveal unprecedented cyanobacteria bloom dynamics associated with elephant mass mortality', *Science of The Total Environment*, 957, p. 177525. Available at: <https://doi.org/10.1016/j.scitotenv.2024.177525>.
- Lones, M.A. (2024) 'How to avoid machine learning pitfalls: a guide for academic researchers'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2108.02497>.
- Lorenzen, C.J. (1966) 'A method for the continuous measurement of *in vivo* chlorophyll concentration', *Deep Sea Research and Oceanographic Abstracts*, 13(2), pp. 223–227. Available at: [https://doi.org/10.1016/0011-7471\(66\)91102-8](https://doi.org/10.1016/0011-7471(66)91102-8).
- Loussaief, A., Ktari, R. and Hadj Kacem, Y. (2025) 'HGS-At-LSTM: attention-based long short-term memory model combined with halving grid search optimizer for harmful algal bloom forecasting', *International Journal of Data Science and Analytics* [Preprint]. Available at: <https://doi.org/10.1007/s41060-025-00783-2>.
- Luo, S. *et al.* (2025) 'Learning to Retrieve for Environmental Knowledge Discovery: An Augmentation-Adaptive Self-Supervised Learning Framework'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2509.14563>.

- Luo, Y. *et al.* (2017) ‘Dynamic monitoring and prediction of Dianchi Lake cyanobacteria outbreaks in the context of rapid urbanization’, *Environmental Science and Pollution Research*, 24(6), pp. 5335–5348. Available at: <https://doi.org/10.1007/s11356-016-8155-2>.
- Luong, H.A. *et al.* (2024) ‘Long-term study of phytoplankton dynamics in a supply reservoir reveals signs of trophic state shift linked to changes in hydrodynamics associated with flow management and extreme events’, *Water Research*, 256, p. 121547. Available at: <https://doi.org/10.1016/j.watres.2024.121547>.
- Lürling, M. *et al.* (2013) ‘Comparison of cyanobacterial and green algal growth rates at different temperatures’, *Freshwater Biology*, 58(3), pp. 552–559. Available at: <https://doi.org/10.1111/j.1365-2427.2012.02866.x>.
- Lv, J., Feng, L. and Zhao, D. (2024) ‘Validation of Global Gridded Aerosol Models in Inland/Coastal Water Atmospheric Correction for MODIS, VIIRS, and Landsat’, *IEEE Transactions on Geoscience and Remote Sensing*, 62, pp. 1–10. Available at: <https://doi.org/10.1109/TGRS.2024.3427836>.
- Lynch, P. (2008) ‘The origins of computer weather prediction and climate modeling’, *Journal of Computational Physics*, 227(7), pp. 3431–3444. Available at: <https://doi.org/10.1016/j.jcp.2007.02.034>.
- Ma, Q. *et al.* (2024) ‘A Survey on Time-Series Pre-Trained Models’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2305.10716>.
- Ma, R. *et al.* (2011) ‘Approximate bottom contribution to remote sensing reflectance in Taihu Lake, China’, *Journal of Great Lakes Research*, 37(1), pp. 18–25. Available at: <https://doi.org/10.1016/j.jglr.2010.12.002>.
- Maberly, S.C. *et al.* (2011) ‘A survey of the lakes of the English Lake District: the Lakes Tour 2010’. Available at: <https://nora.nerc.ac.uk/id/eprint/14563/> (Accessed: 1 October 2025).
- Maberly, S.C. *et al.* (2017) *Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Blelham Tarn, 1945 to 2013 (Dataset)*, EIDC. Available at: <https://catalogue.ceh.ac.uk/id/393a5946-8a22-4350-80f3-a60d753beb00> (Accessed: 12 September 2023).
- Maberly, S.C. *et al.* (2018) ‘From Ecological Informatics to the Generation of Ecological Knowledge: Long-Term Research in the English Lake District’, in F. Recknagel and W.K. Michener (eds) *Ecological Informatics: Data Management and Knowledge Discovery*. Cham: Springer International Publishing, pp. 455–482. Available at: https://doi.org/10.1007/978-3-319-59928-1_21.
- Maberly, S.C. *et al.* (2020) ‘Global lake thermal regions shift under climate change’, *Nature Communications*, 11(1), p. 1232. Available at: <https://doi.org/10.1038/s41467-020-15108-z>.
- Maberly, S.C. and Elliott, J.A. (2012) ‘Insights from long-term studies in the Windermere catchment: external stressors, internal interactions and the structure and function of lake ecosystems’, *Freshwater Biology*, 57(2), pp. 233–243. Available at: <https://doi.org/10.1111/j.1365-2427.2011.02718.x>.
- Mahmood, R. *et al.* (2022) ‘How Much More Data Do I Need? Estimating Requirements for Downstream Tasks’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2207.01725>.
- Makridakis, S. (1993) ‘Accuracy measures: theoretical and practical concerns’, *International Journal of Forecasting*, 9(4), pp. 527–529. Available at: [https://doi.org/10.1016/0169-2070\(93\)90079-3](https://doi.org/10.1016/0169-2070(93)90079-3).

- Malthus, T.J., Ohmsen, R. and Woerd, H.J. van der (2020) 'An Evaluation of Citizen Science Smartphone Apps for Inland Water Quality Assessment', *Remote Sensing*, 12(10), p. 1578. Available at: <https://doi.org/10.3390/rs12101578>.
- Manning, N.F. *et al.* (2019) 'Extending the forecast model: Predicting Western Lake Erie harmful algal blooms at multiple spatial scales', *Journal of Great Lakes Research*, 45(3), pp. 587–595. Available at: <https://doi.org/10.1016/j.jglr.2019.03.004>.
- Mansaray, A.S. *et al.* (2021) 'Comparing PlanetScope to Landsat-8 and Sentinel-2 for Sensing Water Quality in Reservoirs in Agricultural Watersheds', *Remote Sensing*, 13(9), p. 1847. Available at: <https://doi.org/10.3390/rs13091847>.
- Marcé, R. *et al.* (2016) 'Automatic High Frequency Monitoring for Improved Lake and Reservoir Management', *Environmental Science & Technology*, 50(20), pp. 10780–10794. Available at: <https://doi.org/10.1021/acs.est.6b01604>.
- Marshall, W. (2025) *Introducing Owl: Planet's Most Advanced Satellite Mission Yet*. Available at: <https://www.planet.com/pulse/introducing-owl-planet-s-most-advanced-satellite-mission-yet/> (Accessed: 25 November 2025).
- Martín Abadi *et al.* (2015) 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems'. Available at: <https://www.tensorflow.org/>.
- Matthews, M.W. (2022) 'The CyanoLakes mobile app: Weather-like information for lakes from satellite imagery', *Water Wheel*, 21(4), pp. 32–34. Available at: <https://doi.org/10.10520/ejc-waterb-v21-n4-a7>.
- Matthews, M.W. (2023) 'Near-term forecasting of cyanobacteria and harmful algal blooms in lakes using simple univariate methods with satellite remote sensing data', *Inland Waters*, 13(1), pp. 62–73. Available at: <https://doi.org/10.1080/20442041.2022.2145839>.
- Matthews, M.W., Bernard, S. and Robertson, L. (2012) 'An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters', *Remote Sensing of Environment*, 124, pp. 637–652. Available at: <https://doi.org/10.1016/j.rse.2012.05.032>.
- May, L. *et al.* (2022) *Assessing climate change impacts on the water quality of Scottish standing waters. CRW2020_01. Scotland's Centre of Expertise for Waters (CREW)*. Available at: <https://www.crew.ac.uk/publications> (Accessed: 2 May 2022).
- May, L. and Spears, B.M. (2012) 'A history of scientific research at Loch Leven, Kinross, Scotland', in Linda May and Bryan M. Spears (eds) *Loch Leven: 40 years of scientific research: Understanding the links between pollution, climate change and ecological response*. Dordrecht: Springer Netherlands, pp. 3–9. Available at: https://doi.org/10.1007/978-94-007-4333-5_2.
- McCrackin, M.L. *et al.* (2017) 'Recovery of lakes and coastal marine ecosystems from eutrophication: A global meta-analysis', *Limnology and Oceanography*, 62(2), pp. 507–518. Available at: <https://doi.org/10.1002/lno.10441>.
- McCullough, I.M. *et al.* (2019) 'Geographic patterns of the climate sensitivity of lakes', *Ecological Applications: A Publication of the Ecological Society of America*, 29(2), p. e01836. Available at: <https://doi.org/10.1002/eap.1836>.
- McGrath, C.N. *et al.* (2023) 'An assessment of high temporal frequency satellite data for historic environment applications. A case study from Scotland', *Archaeological Prospection*, 30(3), pp. 267–282. Available at: <https://doi.org/10.1002/arp.1890>.

- McLaughlin, R.L. (1983) 'Forecasting Models: Sophisticated or Naive?', *Journal of forecasting*, 2(3).
- Medina, M. *et al.* (2024) 'An early-warning forecast model for red tide (*Karenia brevis*) blooms on the southwest coast of Florida', *Harmful Algae*, 139, p. 102729. Available at: <https://doi.org/10.1016/j.hal.2024.102729>.
- Meis, S., Thackeray, S.J. and Jones, I.D. (2009) 'Effects of recent climate change on phytoplankton phenology in a temperate lake', *Freshwater Biology*, 54(9), pp. 1888–1898. Available at: <https://doi.org/10.1111/j.1365-2427.2009.02240.x>.
- Mellios, N., Moe, S.J. and Laspidou, C. (2020) 'Machine Learning Approaches for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes', *Water*, 12(4), p. 1191. Available at: <https://doi.org/10.3390/w12041191>.
- Met Office (2019) *Dataset Collection Record: Met Office MIDAS Open: UK Land Surface Stations Data (1853-current: Dataset)*. Available at: <https://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1> (Accessed: 12 September 2023).
- Mienye, I.D. and Jere, N. (2024) 'A Survey of Decision Trees: Concepts, Algorithms, and Applications', *IEEE Access*, 12, pp. 86716–86727. Available at: <https://doi.org/10.1109/ACCESS.2024.3416838>.
- Milly, P.C.D. *et al.* (2008) 'Stationarity Is Dead: Whither Water Management?', *Science*, 319(5863), pp. 573–574. Available at: <https://doi.org/10.1126/science.1151915>.
- Milly, P.C.D. *et al.* (2015) 'On Critiques of “Stationarity is Dead: Whither Water Management?”', *Water Resources Research*, 51(9), pp. 7785–7789. Available at: <https://doi.org/10.1002/2015WR017408>.
- Mironov, D. *et al.* (2010) 'Implementation of the lake parameterisation scheme FLake into the numerical weather prediction model COSMO.', *Boreal environment research*, 15(2).
- Mishra, S. *et al.* (2013) 'Quantifying cyanobacterial phycocyanin concentration in turbid productive waters: A quasi-analytical approach', *Remote Sensing of Environment*, 133, pp. 141–151. Available at: <https://doi.org/10.1016/j.rse.2013.02.004>.
- Mishra, S. and Mishra, D.R. (2012) 'Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-*a* concentration in turbid productive waters', *Remote Sensing of Environment*, 117, pp. 394–406. Available at: <https://doi.org/10.1016/j.rse.2011.10.016>.
- Missaghi, S., Hondzo, M. and Melching, C. (2013) 'Three-Dimensional Lake Water Quality Modeling: Sensitivity and Uncertainty Analyses', *Journal of Environmental Quality*, 42(6), pp. 1684–1698. Available at: <https://doi.org/10.2134/jeq2013.04.0120>.
- Mobley, C.D. (1999) 'Estimation of the remote-sensing reflectance from above-surface measurements', *Applied Optics*, 38(36), pp. 7442–7455. Available at: <https://doi.org/10.1364/AO.38.007442>.
- Mobley, C.D. *et al.* (2015) 'Optical Modeling of Ocean Waters: Is the Case 1 - Case 2 Classification Still Useful?', *Oceanography*, 17(2), pp. 60–67. Available at: <https://doi.org/10.5670/oceanog.2004.48>.
- Mohammed, H., Michel Tornyeviadzi, H. and Seidu, R. (2022) 'Emulating process-based water quality modelling in water source reservoirs using machine learning', *Journal of Hydrology*, 609, p. 127675. Available at: <https://doi.org/10.1016/j.jhydrol.2022.127675>.

- Mooij, W.M. *et al.* (2010) 'Challenges and opportunities for integrating lake ecosystem modelling approaches', *Aquatic Ecology*, 44(3), pp. 633–667. Available at: <https://doi.org/10.1007/s10452-010-9339-3>.
- Morrison, A.T., Billings, J.D. and Doherty, K.W. (2000) 'The McLane moored profiler: an autonomous platform for oceanographic measurements', *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No.00CH37158)*. *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference (Cat. No.00CH37158)*, pp. 353–358 vol.1. Available at: <https://doi.org/10.1109/OCEANS.2000.881284>.
- Mortimer, C.H. (1941) 'The Exchange of Dissolved Substances Between Mud and Water in Lakes', *Journal of Ecology*, 29(2), pp. 280–329. Available at: <https://doi.org/10.2307/2256395>.
- Mortimer, C.H. (1942) 'The Exchange of Dissolved Substances between Mud and Water in Lakes', *Journal of Ecology*, 30(1), pp. 147–201. Available at: <https://doi.org/10.2307/2256691>.
- Moses, W.J. *et al.* (2009) 'Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—successes and challenges', *Environmental Research Letters*, 4(4), p. 045005. Available at: <https://doi.org/10.1088/1748-9326/4/4/045005>.
- Moss, B. (2007) 'Water pollution by agriculture', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491), pp. 659–666. Available at: <https://doi.org/10.1098/rstb.2007.2176>.
- Muhetaer, G. *et al.* (2020) 'Effects of Light Intensity and Exposure Period on the Growth and Stress Responses of Two Cyanobacteria Species: *Pseudanabaena galeata* and *Microcystis aeruginosa*', *Water*, 12(2), p. 407. Available at: <https://doi.org/10.3390/w12020407>.
- Murtagh, F. (1991) 'Multilayer perceptrons for classification and regression', *Neurocomputing*, 2(5), pp. 183–197. Available at: [https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/10.1016/0925-2312(91)90023-5).
- Muttil, N. and Chau, K.-W. (2006) 'Neural network and genetic programming for modelling coastal algal blooms', *International Journal of Environment and Pollution*, 28, pp. 223–238.
- Naderian, D. *et al.* (2024) 'A water quality database for global lakes', *Resources, Conservation and Recycling*, 202, p. 107401. Available at: <https://doi.org/10.1016/j.resconrec.2023.107401>.
- Nalepa, T.F. and Schloesser, D.W. (eds) (2013) 'Role of Selective Grazing by Dreissenid Mussels in Promoting Toxic *Microcystis* Blooms and Other Changes in Phytoplankton Composition in the Great Lakes', *Quagga and Zebra Mussels*. 2nd edn. CRC Press.
- Naselli-Flores, L. and Padisák, J. (2023) 'Ecosystem services provided by marine and freshwater phytoplankton', *Hydrobiologia*, 850(12–13), pp. 2691–2706. Available at: <https://doi.org/10.1007/s10750-022-04795-y>.
- Neil, C. *et al.* (2019) 'A global approach for chlorophyll-a retrieval across optically complex inland waters based on optical water types', *Remote Sensing of Environment*, 229, pp. 159–178. Available at: <https://doi.org/10.1016/j.rse.2019.04.027>.
- Nguyen, H.Q. *et al.* (2025) 'A comparative analysis of data-driven modeling approaches to forecast cyanobacteria algal blooms in eutrophic lake discharge canals', *Journal of Environmental Management*, 380, p. 124834. Available at: <https://doi.org/10.1016/j.jenvman.2025.124834>.
- Niroumand-Jadidi, M. and Bovolo, F. (2021) 'WATER QUALITY RETRIEVAL AND ALGAL BLOOM DETECTION USING HIGH-RESOLUTION CUBESAT IMAGERY', *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3–2021, pp. 191–195. Available at: <https://doi.org/10.5194/isprs-annals-V-3-2021-191-2021>.

- Niroumand-Jadidi, M., Legleiter, C.J. and Bovolo, F. (2022) 'River Bathymetry Retrieval From Landsat-9 Images Based on Neural Networks and Comparison to SuperDove and Sentinel-2', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, pp. 5250–5260. Available at: <https://doi.org/10.1109/JSTARS.2022.3187179>.
- Nixdorf, B. (1994) 'Polymixis of a shallow lake (Großer Müggelsee, Berlin) and its influence on seasonal phytoplankton dynamics', *Hydrobiologia*, 275(1), pp. 173–186. Available at: <https://doi.org/10.1007/BF00026709>.
- Nogueira, S.F. *et al.* (2025) 'Design and development of a cost-effective buoy using 3D printing for coastal monitoring', *HardwareX*, 22, p. e00657. Available at: <https://doi.org/10.1016/j.ohx.2025.e00657>.
- NRFA (2023a) *NRFA Station Data for 73013 - Rothay at Miller Bridge House*. Available at: <https://nrfa.ceh.ac.uk/data/station/info/73013> (Accessed: 12 September 2023).
- NRFA (2023b) *NRFA Station Data for 73014 - Brathay at Jeffy Knotts, National River Flow Archive*. Available at: <https://nrfa.ceh.ac.uk/data/station/info/73014> (Accessed: 12 September 2023).
- Nürnberg, G.K. (2009) 'Assessing internal phosphorus load – Problems to be solved', *Lake and Reservoir Management*, 25(4), pp. 419–432. Available at: <https://doi.org/10.1080/00357520903458848>.
- Olejarz, J. *et al.* (2021) 'The Great Oxygenation Event as a consequence of ecological dynamics modulated by planetary change', *Nature Communications*, 12(1), p. 3985. Available at: <https://doi.org/10.1038/s41467-021-23286-7>.
- Oliver, R.L. and Ganf, G.G. (2002) 'Freshwater Blooms', in B.A. Whitton and M. Potts (eds) *The Ecology of Cyanobacteria: Their Diversity in Time and Space*. Dordrecht: Springer Netherlands, pp. 149–194. Available at: https://doi.org/10.1007/0-306-46855-7_6.
- Olson, R.J. and Sosik, H.M. (2007) 'A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot', *Limnology and Oceanography: Methods*, 5(6), pp. 195–203. Available at: <https://doi.org/10.4319/lom.2007.5.195>.
- Onderka, M. (2007) 'Correlations between several environmental factors affecting the bloom events of cyanobacteria in Liptovska Mara reservoir (Slovakia)—A simple regression model', *Ecological Modelling*, 209(2), pp. 412–416. Available at: <https://doi.org/10.1016/j.ecolmodel.2007.07.028>.
- O'Neil, J. *et al.* (2012) 'The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change', *HARMFUL ALGAE*, 14, pp. 313–334. Available at: <https://doi.org/10.1016/j.hal.2011.10.027>.
- O'Neill, B.C. *et al.* (2016) 'The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6', *Geoscientific Model Development*, 9(9), pp. 3461–3482. Available at: <https://doi.org/10.5194/gmd-9-3461-2016>.
- O'Reilly, C.M. *et al.* (2015) 'Rapid and highly variable warming of lake surface waters around the globe', *Geophysical Research Letters*, 42(24), p. 10,773–10,781. Available at: <https://doi.org/10.1002/2015GL066235>.
- O'Reilly, J. (2000) 'Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4', *SeaWiFS Postlaunch Calibration and Validation Analyses*, 11, pp. 9–23.

- O'Reilly, J.E. *et al.* (1998) 'Ocean color chlorophyll algorithms for SeaWiFS', *Journal of Geophysical Research: Oceans*, 103(C11), pp. 24937–24953. Available at: <https://doi.org/10.1029/98JC02160>.
- O'Shea, R.E. *et al.* (2021) 'Advancing cyanobacteria biomass estimation from hyperspectral observations: Demonstrations with HICO and PRISMA imagery', *Remote Sensing of Environment*, 266, p. 112693. Available at: <https://doi.org/10.1016/j.rse.2021.112693>.
- Oyama, Y. *et al.* (2015) 'Monitoring levels of cyanobacterial blooms using the visual cyanobacteria index (VCI) and floating algae index (FAI)', *International Journal of Applied Earth Observation and Geoinformation*, 38, pp. 335–348. Available at: <https://doi.org/10.1016/j.jag.2015.02.002>.
- Oyama, Y., Matsushita, B. and Fukushima, T. (2015) 'Distinguishing surface cyanobacterial blooms and aquatic macrophytes using Landsat/TM and ETM + shortwave infrared bands', *Remote Sensing of Environment*, 157, pp. 35–47. Available at: <https://doi.org/10.1016/j.rse.2014.04.031>.
- Paerl, H. *et al.* (2020) 'Mitigating a global expansion of toxic cyanobacterial blooms: confounding effects and challenges posed by climate change', *MARINE AND FRESHWATER RESEARCH*, 71(5), pp. 579–592. Available at: <https://doi.org/10.1071/MF18392>.
- Paerl, H. and Huisman, J. (2009) 'Climate change: a catalyst for global expansion of harmful cyanobacterial blooms', *Environmental Microbiology Reports*, 1(1), pp. 27–37. Available at: <https://doi.org/10.1111/j.1758-2229.2008.00004.x>.
- Paerl, H. and Otten, T. (2013) 'Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls', *MICROBIAL ECOLOGY*, 65(4), pp. 995–1010. Available at: <https://doi.org/10.1007/s00248-012-0159-y>.
- Paerl, H. 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>.
- Pagano, T.S. and Durham, R.M. (1993) 'Moderate Resolution Imaging Spectroradiometer (MODIS)', *Sensor Systems for the Early Earth Observing System Platforms. Sensor Systems for the Early Earth Observing System Platforms*, SPIE, pp. 2–17. Available at: <https://doi.org/10.1117/12.152835>.
- Page, T. *et al.* (2017) 'Constraining uncertainty and process-representation in an algal community lake model using high frequency in-lake observations', *Ecological Modelling*, 357, pp. 1–13. Available at: <https://doi.org/10.1016/j.ecolmodel.2017.04.011>.
- Page, T. *et al.* (2018) 'Adaptive forecasting of phytoplankton communities', *Water Research*, 134, pp. 74–85. Available at: <https://doi.org/10.1016/j.watres.2018.01.046>.
- Pahlevan, N. *et al.* (2020) 'Seamless retrievals of chlorophyll-*a* from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach', *Remote Sensing of Environment*, 240, p. 111604. Available at: <https://doi.org/10.1016/j.rse.2019.111604>.
- Pahlevan, N. *et al.* (2021) 'ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters', *Remote Sensing of Environment*, 258, p. 112366. Available at: <https://doi.org/10.1016/j.rse.2021.112366>.
- Palmer, S.C.J., Kutser, T. and Hunter, P.D. (2015) 'Remote sensing of inland waters: Challenges, progress and future directions', *Remote Sensing of Environment*, 157, pp. 1–8. Available at: <https://doi.org/10.1016/j.rse.2014.09.021>.

- Park, J., Kim, K.T. and Lee, W.H. (2020) ‘Recent Advances in Information and Communications Technology (ICT) and Sensor Technology for Monitoring Water Quality’, *Water*, 12(2), p. 510. Available at: <https://doi.org/10.3390/w12020510>.
- Park, J., Patel, K. and Lee, W.H. (2024) ‘Recent advances in algal bloom detection and prediction technology using machine learning’, *Science of The Total Environment*, 938, p. 173546. Available at: <https://doi.org/10.1016/j.scitotenv.2024.173546>.
- Park, Y. *et al.* (2015) ‘Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea’, *Science of The Total Environment*, 502, pp. 31–41. Available at: <https://doi.org/10.1016/j.scitotenv.2014.09.005>.
- Pascanu, R., Mikolov, T. and Bengio, Y. (2013) ‘On the difficulty of training Recurrent Neural Networks’. arXiv. Available at: <https://doi.org/10.48550/arXiv.1211.5063>.
- Pasztaleniec, A. and Poniewozik, M. (2010) ‘Phytoplankton based assessment of the ecological status of four shallow lakes (Eastern Poland) according to Water Framework Directive – a comparison of approaches’, *Limnologica - Ecology and Management of Inland Waters*, 40(3), pp. 251–259. Available at: <https://doi.org/10.1016/j.limno.2009.07.001>.
- Pattinson, N.B. *et al.* (2023) *Digital innovation in citizen science to enhance water quality monitoring in developing countries*. International Water Management Institute. Available at: <https://doi.org/10.5337/2024.201>.
- Paulino, R.S. *et al.* (2022) ‘Assessment of Adjacency Correction over Inland Waters Using Sentinel-2 MSI Images’, *Remote Sensing*, 14(8), p. 1829. Available at: <https://doi.org/10.3390/rs14081829>.
- Pedregosa, F. *et al.* (2011) ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Peretyatko, A. *et al.* (2010) ‘Assessment of the risk of cyanobacterial bloom occurrence in urban ponds: probabilistic approach’, *Annales de Limnologie - International Journal of Limnology*, 46(2), pp. 121–133. Available at: <https://doi.org/10.1051/limn/2010009>.
- Piccolroaz, S. *et al.* (2024) ‘Lake Water Temperature Modeling in an Era of Climate Change: Data Sources, Models, and Future Prospects’, *Reviews of Geophysics*, 62(1), p. e2023RG000816. Available at: <https://doi.org/10.1029/2023RG000816>.
- Pitois, S.G. *et al.* (2025) ‘RAPID: real-time automated plankton identification dashboard using Edge AI at sea’, *Frontiers in Marine Science*, 11. Available at: <https://doi.org/10.3389/fmars.2024.1513463>.
- Planet Labs PBC (2018) ‘Planet Application Program Interface: In Space for Life on Earth’. Planet. Available at: <https://api.planet.com>.
- Planet Labs PBC (2025a) *Flexible Pricing for Satellite Imagery & Data | Planet*. Available at: <https://www.planet.com/pricing/?tab=monitoring> (Accessed: 21 October 2025).
- Planet Labs PBC (2025b) *Planet and Anthropic Partner to Use Claude’s Advanced AI Capabilities to Turn Geospatial Satellite Imagery into Actionable Insights*. Available at: <https://investors.planet.com/news/news-details/2025/Planet-and-Anthropic-Partner-to-Use-Claudes-Advanced-AI-Capabilities-to-Turn-Geospatial-Satellite-Imagery-into-Actionable-Insights/default.aspx> (Accessed: 27 November 2025).

Planet Labs PBC (2025c) *Planet Fiscal 2Q'26 Update Presentation*. Available at: https://s29.q4cdn.com/903184914/files/doc_earnings/2026/q2/presentation/Planet-Fiscal-2Q-26-Update-Presentation.pdf (Accessed: 25 November 2025).

Planet Team (2017) 'Planet application program interface: In space for life on Earth'. Planet. Available at: <https://www.planet.com/markets/education-and-research/>.

Pobel, D., Robin, J. and Humbert, J.-F. (2011) 'Influence of sampling strategies on the monitoring of cyanobacteria in shallow lakes: Lessons from a case study in France', *Water Research*, 45(3), pp. 1005–1014. Available at: <https://doi.org/10.1016/j.watres.2010.10.011>.

Pocock, M.J.O. *et al.* (2023) 'The benefits of citizen science and nature-noticing activities for well-being, nature connectedness and pro-nature conservation behaviours', *People and Nature*, 5(2), pp. 591–606. Available at: <https://doi.org/10.1002/pan3.10432>.

Poole, M.A. and O'Farrell, P.N. (1971) 'The Assumptions of the Linear Regression Model', *Transactions of the Institute of British Geographers*, (52), pp. 145–158. Available at: <https://doi.org/10.2307/621706>.

Popovskaya, G.I. (2000) 'Ecological monitoring of phytoplankton in Lake Baikal', *Aquatic Ecosystem Health & Management*, 3(2), pp. 215–225. Available at: <https://doi.org/10.1080/14634980008657017>.

Pretty, J.N. *et al.* (2003) 'Environmental Costs of Freshwater Eutrophication in England and Wales', *Environmental Science & Technology*, 37(2), pp. 201–208. Available at: <https://doi.org/10.1021/es020793k>.

Putman, N.F. *et al.* (2023) 'Improving satellite monitoring of coastal inundations of pelagic Sargassum algae with wind and citizen science data', *Aquatic Botany*, 188, p. 103672. Available at: <https://doi.org/10.1016/j.aquabot.2023.103672>.

Qing, S. *et al.* (2020) 'Distinguishing and mapping of aquatic vegetations and yellow algae bloom with Landsat satellite data in a complex shallow Lake, China during 1986–2018', *Ecological Indicators*, 112, p. 106073. Available at: <https://doi.org/10.1016/j.ecolind.2020.106073>.

Qu, K. (2024) 'Research on linear regression algorithm', *MATEC Web of Conferences*, 395, p. 01046. Available at: <https://doi.org/10.1051/matecconf/202439501046>.

Quan, Q. *et al.* (2022) 'Research on water temperature prediction based on improved support vector regression', *Neural Computing and Applications*, 34(11), pp. 8501–8510. Available at: <https://doi.org/10.1007/s00521-020-04836-4>.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Rajapakse, J., Otoo, M. and Danso, G. (2023) 'Progress in delivering SDG6: Safe water and sanitation', *Cambridge Prisms: Water*, 1, p. e6. Available at: <https://doi.org/10.1017/wat.2023.5>.

Ralston, D.K. and Moore, S.K. (2020) 'Modeling harmful algal blooms in a changing climate', *Harmful Algae*, 91, p. 101729. Available at: <https://doi.org/10.1016/j.hal.2019.101729>.

Råman Vinnå, L. *et al.* (2021) 'The vulnerability of lakes to climate change along an altitudinal gradient', *Communications Earth & Environment*, 2(1), p. 35. Available at: <https://doi.org/10.1038/s43247-021-00106-w>.

Ramírez, S.B., van Meerveld, I. and Seibert, J. (2023) 'Citizen science approaches for water quality measurements', *Science of The Total Environment*, 897, p. 165436. Available at: <https://doi.org/10.1016/j.scitotenv.2023.165436>.

- Ramsbottom, A.E. (1976) *Depth Charts of the Cumbrian Lakes*. Freshwater Biological Association.
- Rashid, A.B. and Kausik, M.A.K. (2024) 'AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications', *Hybrid Advances*, 7, p. 100277. Available at: <https://doi.org/10.1016/j.hybadv.2024.100277>.
- Razavi, S. *et al.* (2022) 'Coevolution of machine learning and process-based modelling to revolutionize Earth and environmental sciences: A perspective', *Hydrological Processes*, 36(6), p. e14596. Available at: <https://doi.org/10.1002/hyp.14596>.
- Read, J.S. *et al.* (2019) 'Process-Guided Deep Learning Predictions of Lake Water Temperature', *Water Resources Research*, 55(11), pp. 9173–9190. Available at: <https://doi.org/10.1029/2019WR024922>.
- Recknagel, F. *et al.* (1997) 'Artificial neural network approach for modelling and prediction of algal blooms', *Ecological Modelling*, 96(1), pp. 11–28. Available at: [https://doi.org/10.1016/S0304-3800\(96\)00049-X](https://doi.org/10.1016/S0304-3800(96)00049-X).
- Recknagel, F. *et al.* (2008) 'Adaptive agents for forecasting seasonal outbreaks of blue-green algal populations in lakes categorised by circulation type and trophic state', *SIL Proceedings, 1922-2010*, 30(2), pp. 191–197. Available at: <https://doi.org/10.1080/03680770.2008.11902108>.
- Recknagel, F. *et al.* (2025) 'Early warning of harmful cyanobacteria blooms based on high frequency *in situ* monitoring and intelligible machine learning modelling: The case study of Lake Müggelsee (Germany)', *Water Research*, 287, p. 124514. Available at: <https://doi.org/10.1016/j.watres.2025.124514>.
- Reid, A.J. *et al.* (2019) 'Emerging threats and persistent conservation challenges for freshwater biodiversity', *Biological Reviews*, 94(3), pp. 849–873. Available at: <https://doi.org/10.1111/brv.12480>.
- Reutter, J.M. (2019) 'Lake Erie: Past, Present, and Future', *Encyclopedia of Water*. John Wiley & Sons, Ltd, pp. 1–15. Available at: <https://doi.org/10.1002/9781119300762.wsts0085>.
- Reynolds, C.S. (2006) *The Ecology of Phytoplankton*. Cambridge: Cambridge University Press (Ecology, Biodiversity and Conservation). Available at: <https://doi.org/10.1017/CBO9780511542145>.
- Reynolds, S.A. and Aldridge, D.C. (2021) 'Impacts of invasive quagga mussels (*Dreissena rostriformis bugensis*) on reservoir water quality, as revealed by progressive-change BACIPS analysis', *Water Research*, 197, p. 117105. Available at: <https://doi.org/10.1016/j.watres.2021.117105>.
- Rheinheimer, D.E., Null, S.E. and Lund, J.R. (2015) 'Optimizing Selective Withdrawal from Reservoirs to Manage Downstream Temperatures with Climate Warming', *Journal of Water Resources Planning and Management*, 141(4), p. 04014063. Available at: [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000447](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000447).
- 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>.
- Rivas-Villar, D. *et al.* (2021) 'Fully automatic detection and classification of phytoplankton specimens in digital microscopy images', *Computer Methods and Programs in Biomedicine*, 200, p. 105923. Available at: <https://doi.org/10.1016/j.cmpb.2020.105923>.

- Rollinson, C.R. *et al.* (2021) ‘Working across space and time: nonstationarity in ecological research and application’, *Frontiers in Ecology and the Environment*, 19(1), pp. 66–72. Available at: <https://doi.org/10.1002/fee.2298>.
- Rose, V. *et al.* (2021) ‘Effects of Grazing and Nutrients on Phytoplankton Blooms and Microplankton Assemblage Structure in Four Temperate Lakes Spanning a Eutrophication Gradient’, *Water*, 13(8), p. 1085. Available at: <https://doi.org/10.3390/w13081085>.
- Rosenblatt, F. (1962) *Principles Of Neurodynamics*. Available at: <http://archive.org/details/principles-of-neurodynamics> (Accessed: 25 June 2025).
- Rousso, B.Z. *et al.* (2020) ‘A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes’, *Water Research*, 182. Available at: <https://doi.org/10.1016/j.watres.2020.115959>.
- Rozemeijer, J. *et al.* (2025) ‘Best practice in high-frequency water quality monitoring for improved management and assessment; a novel decision workflow’, *Environmental Monitoring and Assessment*, 197(4), p. 353. Available at: <https://doi.org/10.1007/s10661-025-13795-z>.
- Ruddick, K.G. *et al.* (2024) ‘HYPERNETS: a network of automated hyperspectral radiometers to validate water and land surface reflectance (380–1680 nm) from all satellite missions’, *Frontiers in Remote Sensing*, 5. Available at: <https://doi.org/10.3389/frsen.2024.1372085>.
- Ruddick, K.G., Ovidio, F. and Rijkeboer, M. (2000) ‘Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters’, *Applied Optics*, 39(6), pp. 897–912. Available at: <https://doi.org/10.1364/AO.39.000897>.
- Rumelhart, D.E. and McClelland, J.L. (1987) ‘Learning Internal Representations by Error Propagation’, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, pp. 318–362. Available at: <https://ieeexplore.ieee.org/document/6302929> (Accessed: 25 June 2025).
- Saadi, O.E. *et al.* (1995) ‘Murray River water, raised cyanobacterial cell counts, and gastrointestinal and dermatological symptoms’, *Medical Journal of Australia*, 162(3), pp. 122–125. Available at: <https://doi.org/10.5694/j.1326-5377.1995.tb138473.x>.
- Saber, A., James, D.E. and Hayes, D.F. (2020) ‘Long-term forecast of water temperature and dissolved oxygen profiles in deep lakes using artificial neural networks conjugated with wavelet transform’, *Limnology and Oceanography*, 65(6), pp. 1297–1317. Available at: <https://doi.org/10.1002/lno.11390>.
- Sala, O.E. *et al.* (2000) ‘Global Biodiversity Scenarios for the Year 2100’, *Science*, 287(5459), pp. 1770–1774. Available at: <https://doi.org/10.1126/science.287.5459.1770>.
- Sarker, I.H. (2021a) ‘Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions’, *SN Computer Science*, 2(6), p. 420. Available at: <https://doi.org/10.1007/s42979-021-00815-1>.
- Sarker, I.H. (2021b) ‘Machine Learning: Algorithms, Real-World Applications and Research Directions’, *SN Computer Science*, 2(3), p. 160. Available at: <https://doi.org/10.1007/s42979-021-00592-x>.
- Sayers, M. *et al.* (2016) ‘Cyanobacteria blooms in three eutrophic basins of the Great Lakes: a comparative analysis using satellite remote sensing’, *International Journal of Remote Sensing*, 37(17), pp. 4148–4171. Available at: <https://doi.org/10.1080/01431161.2016.1207265>.

- Schaeffer, B.A. *et al.* (2024) 'Forecasting freshwater cyanobacterial harmful algal blooms for Sentinel-3 satellite resolved U.S. lakes and reservoirs', *Journal of Environmental Management*, 349, p. 119518. Available at: <https://doi.org/10.1016/j.jenvman.2023.119518>.
- Schallenberg *et al.* (2013) 'Ecosystem Services of Lakes', pp. 203–225.
- Scheffer, M. and van Nes, E.H. (2007) 'Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size', *Hydrobiologia*, 584(1), pp. 455–466. Available at: <https://doi.org/10.1007/s10750-007-0616-7>.
- Schindler, D.W. (1974) 'Eutrophication and Recovery in Experimental Lakes: Implications for Lake Management', *Science*, 184(4139), pp. 897–899. Available at: <https://doi.org/10.1126/science.184.4139.897>.
- Schindler, D.W. (2009) 'Lakes as sentinels and integrators for the effects of climate change on watersheds, airsheds, and landscapes', *Limnology and Oceanography*, 54(6part2), pp. 2349–2358. Available at: https://doi.org/10.4319/lo.2009.54.6_part_2.2349.
- Schmidt, R.M. (2019) 'Recurrent Neural Networks (RNNs): A gentle Introduction and Overview'. arXiv. Available at: <https://doi.org/10.48550/arXiv.1912.05911>.
- Schröder, T. *et al.* (2024) 'Exploring Spatial Aggregations and Temporal Windows for Water Quality Match-Up Analysis Using Sentinel-2 MSI and Sentinel-3 OLCI Data', *Remote Sensing*, 16(15), p. 2798. Available at: <https://doi.org/10.3390/rs16152798>.
- Schroeder, T. *et al.* (2022) 'An ensemble neural network atmospheric correction for Sentinel-3 OLCI over coastal waters providing inherent model uncertainty estimation and sensor noise propagation', *Remote Sensing of Environment*, 270, p. 112848. Available at: <https://doi.org/10.1016/j.rse.2021.112848>.
- Schulze, K. *et al.* (2013) 'PlanktoVision - an automated analysis system for the identification of phytoplankton', *BMC Bioinformatics*, 14(1), p. 115. Available at: <https://doi.org/10.1186/1471-2105-14-115>.
- Scott, M.D., Ramaswamy, L. and Lawson, V. (2016) 'CyanoTRACKER: A Citizen Science Project for Reporting Harmful Algal Blooms', *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), pp. 391–397. Available at: <https://doi.org/10.1109/CIC.2016.058>.
- Seegers, B.N. *et al.* (2018) 'Performance metrics for the assessment of satellite data products: an ocean color case study', *Optics Express*, 26(6), pp. 7404–7422. Available at: <https://doi.org/10.1364/OE.26.007404>.
- Seekell, D. *et al.* (2021) 'Patterns and Variation of Littoral Habitat Size Among Lakes', *Geophysical Research Letters*, 48(20), p. e2021GL095046. Available at: <https://doi.org/10.1029/2021GL095046>.
- Seelen, L.M.S. *et al.* (2019) 'Saving water for the future: Public awareness of water usage and water quality', *Journal of Environmental Management*, 242, pp. 246–257. Available at: <https://doi.org/10.1016/j.jenvman.2019.04.047>.
- Seifert-Dähnn, I. *et al.* (2021) 'Costs and benefits of automated high-frequency environmental monitoring – The case of lake water management', *Journal of Environmental Management*, 285, p. 112108. Available at: <https://doi.org/10.1016/j.jenvman.2021.112108>.

- Sellar, A.A. *et al.* (2019) 'UKESM1: Description and Evaluation of the U.K. Earth System Model', *Journal of Advances in Modeling Earth Systems*, 11(12), pp. 4513–4558. Available at: <https://doi.org/10.1029/2019MS001739>.
- Sellner, K.G., Doucette, G.J. and Kirkpatrick, G.J. (2003) 'Harmful algal blooms: causes, impacts and detection', *Journal of Industrial Microbiology and Biotechnology*, 30(7), pp. 383–406. Available at: <https://doi.org/10.1007/s10295-003-0074-9>.
- Selva, D. and Krejci, D. (2012) 'A survey and assessment of the capabilities of Cubesats for Earth observation', *Acta Astronautica*, 74, pp. 50–68. Available at: <https://doi.org/10.1016/j.actaastro.2011.12.014>.
- Shatwell, T. and Köhler, J. (2019) 'Decreased nitrogen loading controls summer cyanobacterial blooms without promoting nitrogen-fixing taxa: Long-term response of a shallow lake', *Limnology and Oceanography*, 64(S1), pp. S166–S178. Available at: <https://doi.org/10.1002/lno.11002>.
- Shatwell, T., Thiery, W. and Kirillin, G. (2019) 'Future projections of temperature and mixing regime of European temperate lakes', *Hydrology and Earth System Sciences*, 23(3), pp. 1533–1551. Available at: <https://doi.org/10.5194/hess-23-1533-2019>.
- Shen, W. *et al.* (2025) 'Restoring small water bodies to improve lake and river water quality in China', *Nature Communications*, 16(1), p. 294. Available at: <https://doi.org/10.1038/s41467-024-55714-9>.
- Sherstinsky, A. (2020) 'Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network', *Physica D: Nonlinear Phenomena*, 404, p. 132306. Available at: <https://doi.org/10.1016/j.physd.2019.132306>.
- Shukla, A., Matharu, P.S. and Bhattacharya, B. (2023) 'Design and development of a continuous water quality monitoring buoy for health monitoring of river Ganga', *Engineering Research Express*, 5(4), p. 045073. Available at: <https://doi.org/10.1088/2631-8695/ad0d40>.
- Shyalika, C., Wickramarachchi, R. and Sheth, A. (2023) 'A Comprehensive Survey on Rare Event Prediction'. arXiv. Available at: <https://doi.org/10.48550/arXiv.2309.11356>.
- Siddique, I. (2024) 'Small Satellites: Revolutionizing Space Exploration and Earth Observation'. Rochester, NY: Social Science Research Network. Available at: <https://papers.ssrn.com/abstract=4908526> (Accessed: 11 December 2024).
- Siebers, M.A.C. *et al.* (2025) 'Improving algal bloom modelling in eutrophic lakes by calibrating the General Lake Model with satellite remote sensing products', *Water Research X*, 28, p. 100386. Available at: <https://doi.org/10.1016/j.wroa.2025.100386>.
- Siegel, D.A. *et al.* (2013) 'Regional to global assessments of phytoplankton dynamics from the SeaWiFS mission', *Remote Sensing of Environment*, 135, pp. 77–91. Available at: <https://doi.org/10.1016/j.rse.2013.03.025>.
- Simis, S.G.H., Peters, S.W.M. and Gons, H.J. (2005) 'Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water', *Limnology and Oceanography*, 50(1), pp. 237–245. Available at: <https://doi.org/10.4319/lo.2005.50.1.0237>.
- Singleton, J., De Young, B. and Bachmayer, R. (2019) 'Design and Development of a Novel Autonomous Moored Profiler', *OCEANS 2019 MTS/IEEE SEATTLE*, pp. 1–10. Available at: <https://doi.org/10.23919/OCEANS40490.2019.8962852>.
- Smits, A.P. *et al.* (2025) 'Too much and not enough data: Challenges and solutions for generating information in freshwater research and monitoring', *Ecosphere*, 16(3), p. e70205. Available at: <https://doi.org/10.1002/ecs2.70205>.

- Snorheim, C.A. *et al.* (2017) 'Meteorological drivers of hypolimnetic anoxia in a eutrophic, north temperate lake', *Ecological Modelling*, 343, pp. 39–53. Available at: <https://doi.org/10.1016/j.ecolmodel.2016.10.014>.
- Soga, M. and Gaston, K.J. (2016) 'Extinction of experience: the loss of human–nature interactions', *Frontiers in Ecology and the Environment*, 14(2), pp. 94–101. Available at: <https://doi.org/10.1002/fee.1225>.
- Soga, M. and Gaston, K.J. (2025) 'Extinction of experience among ecologists', *Trends in Ecology & Evolution*, 40(3), pp. 212–215. Available at: <https://doi.org/10.1016/j.tree.2024.12.010>.
- 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>.
- Song, Z. *et al.* (2023) 'Atmospheric correction of absorbing aerosols for satellite ocean color remote sensing over coastal waters', *Remote Sensing of Environment*, 290, p. 113552. Available at: <https://doi.org/10.1016/j.rse.2023.113552>.
- Soranno, P.A. (1997) 'Factors affecting the timing of surface scums and epilimnetic blooms of blue-green algae in a eutrophic lake', *Canadian Journal of Fisheries and Aquatic Sciences*, 54(9), pp. 1965–1975. Available at: <https://doi.org/10.1139/f97-104>.
- Soriano-González, J. *et al.* (2022) 'Towards the Combination of C2RCC Processors for Improving Water Quality Retrieval in Inland and Coastal Areas', *Remote Sensing*, 14(5), p. 1124. Available at: <https://doi.org/10.3390/rs14051124>.
- Sosik, H.M., Olson, R.J. and Armbrust, E.V. (2010) 'Flow Cytometry in Phytoplankton Research', in D.J. Suggett, O. Prášil, and M.A. Borowitzka (eds) *Chlorophyll a Fluorescence in Aquatic Sciences: Methods and Applications*. Dordrecht: Springer Netherlands, pp. 171–185. Available at: https://doi.org/10.1007/978-90-481-9268-7_8.
- Spaulding, S.A. *et al.* (2024) 'Chlorophyll a in lakes and streams of the United States (2005–2022)', *Scientific Data*, 11(1), p. 611. Available at: <https://doi.org/10.1038/s41597-024-03453-3>.
- Spyrakos, E. *et al.* (2018) 'Optical types of inland and coastal waters', *Limnology and Oceanography*, 63(2), pp. 846–870. Available at: <https://doi.org/10.1002/lno.10674>.
- Steinmetz, F. and Ramon, D. (2018) 'Sentinel-2 MSI and Sentinel-3 OLCI consistent ocean colour products using POLYMER', *Remote Sensing of the Open and Coastal Ocean and Inland Waters. Remote Sensing of the Open and Coastal Ocean and Inland Waters*, SPIE, pp. 46–55. Available at: <https://doi.org/10.1117/12.2500232>.
- Sterner, R.W. (2008) 'On the Phosphorus Limitation Paradigm for Lakes', *International Review of Hydrobiology*, 93(4–5), pp. 433–445. Available at: <https://doi.org/10.1002/iroh.200811068>.
- Sterner, R.W. *et al.* (2020) 'Ecosystem services of Earth's largest freshwater lakes', *Ecosystem Services*, 41, p. 101046. Available at: <https://doi.org/10.1016/j.ecoser.2019.101046>.
- Stockwell, J.D. *et al.* (2020) 'Storm impacts on phytoplankton community dynamics in lakes', *Global Change Biology*, 26(5), pp. 2756–2784. Available at: <https://doi.org/10.1111/gcb.15033>.
- Strokal, M. *et al.* (2021) 'Urbanization: an increasing source of multiple pollutants to rivers in the 21st century', *npj Urban Sustainability*, 1(1), p. 24. Available at: <https://doi.org/10.1038/s42949-021-00026-w>.

- Stroom, J.M. and Kardinaal, W.E.A. (2016) 'How to combat cyanobacterial blooms: strategy toward preventive lake restoration and reactive control measures', *Aquatic Ecology*, 50(3), pp. 541–576. Available at: <https://doi.org/10.1007/s10452-016-9593-0>.
- Stumpf, R.P. *et al.* (2016) 'Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie', *Journal of Great Lakes Research*, 42(6), pp. 1174–1183. Available at: <https://doi.org/10.1016/j.jglr.2016.08.006>.
- Sylvester, M. *et al.* (1972) 'Application of automatic monitors for state water quality surveillance'. Available at: <https://hdl.handle.net/10217/233865> (Accessed: 8 December 2025).
- Tabari, H. (2020) 'Climate change impact on flood and extreme precipitation increases with water availability', *Scientific Reports*, 10(1), p. 13768. Available at: <https://doi.org/10.1038/s41598-020-70816-2>.
- Talib, A. *et al.* (2008) 'Forecasting and explanation of algal dynamics in two shallow lakes by recurrent artificial neural network and hybrid evolutionary algorithm', *Mathematics and Computers in Simulation*, 78(2), pp. 424–434. Available at: <https://doi.org/10.1016/j.matcom.2008.01.037>.
- Tammeorg, O. *et al.* (2020) 'Internal phosphorus loading due to sediment anoxia in shallow areas: implications for lake aeration treatments', *Aquatic Sciences*, 82(3), p. 54. Available at: <https://doi.org/10.1007/s00027-020-00724-0>.
- Tang, H. *et al.* (2014) 'Quagga mussel (*Dreissena rostriformis bugensis*) selective feeding of phytoplankton in Saginaw Bay', *Journal of Great Lakes Research*, 40, pp. 83–94. Available at: <https://doi.org/10.1016/j.jglr.2013.11.011>.
- Tanguy, M. *et al.* (2021) *Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019) [CEH-GEAR] (Dataset)*, EIDC. Available at: <https://catalogue.ceh.ac.uk/id/dbf13dd5-90cd-457a-a986-f2f9dd97e93c> (Accessed: 12 September 2023).
- Tapics, T., Gregory-Eaves, I. and Huot, Y. (2025) 'High-frequency diatom dynamics seen in an ice- and snow-covered temperate lake using an imaging-in-flow cytometer', *Hydrobiologia*, 852(11), pp. 2887–2905. Available at: <https://doi.org/10.1007/s10750-025-05802-8>.
- Tay, C.J. *et al.* (2021) 'Managing Regime Shift in Lake Systems by Modelling and Simulation', in M.H. Mohd *et al.* (eds) *Modelling, Simulation and Applications of Complex Systems*. Singapore: Springer, pp. 215–232. Available at: https://doi.org/10.1007/978-981-16-2629-6_11.
- Tebbs, E.J., Remedios, J.J. and Harper, D.M. (2013) 'Remote sensing of chlorophyll-*a* as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline–alkaline, flamingo lake, using Landsat ETM +', *Remote Sensing of Environment*, 135, pp. 92–106. Available at: <https://doi.org/10.1016/j.rse.2013.03.024>.
- Teixeira, M.G. *et al.* (1993) 'Gastroenteritis epidemic in the area of the Itaparica Dam, Bahia, Brazil.', *Bulletin of the Pan American Health Organization*.
- Teles, L.O. *et al.* (2006) 'Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks', *Environmental Management*, 38(2), pp. 227–237. Available at: <https://doi.org/10.1007/s00267-005-0074-9>.
- Thackeray, S.J., Jones, I.D. and Maberly, S.C. (2008) 'Long-term change in the phenology of spring phytoplankton: species-specific responses to nutrient enrichment and climatic change', *Journal of Ecology*, 96(3), pp. 523–535. Available at: <https://doi.org/10.1111/j.1365-2745.2008.01355.x>.

- Thomas, M.K. *et al.* (2018) ‘The predictability of a lake phytoplankton community, over time-scales of hours to years’, *Ecology Letters*, 21(5), pp. 619–628. Available at: <https://doi.org/10.1111/ele.12927>.
- Thomas, R.Q. *et al.* (2020) ‘A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time’, *Water Resources Research*, 56(11), p. e2019WR026138. Available at: <https://doi.org/10.1029/2019WR026138>.
- Thomas, R.Q. *et al.* (2023) ‘The NEON Ecological Forecasting Challenge’, *Frontiers in Ecology and the Environment*, 21(3), pp. 112–113. Available at: <https://doi.org/10.1002/fee.2616>.
- Tigli, M. *et al.* (2025) ‘The future of algal blooms in lakes globally is in our hands’, *Water Research*, 268, p. 122533. Available at: <https://doi.org/10.1016/j.watres.2024.122533>.
- Till, A. *et al.* (2019) ‘Fish die-offs are concurrent with thermal extremes in north temperate lakes’, *Nature Climate Change*, 9(8), pp. 637–641. Available at: <https://doi.org/10.1038/s41558-019-0520-y>.
- Torres, R. *et al.* (2011) ‘Forecasting of cyanobacterial density in Torrão reservoir using artificial neural networks’, *Journal of Environmental Monitoring*, 13(6), pp. 1761–1767. Available at: <https://doi.org/10.1039/C1EM10127G>.
- Tseng, G. *et al.* (2023) *Lightweight, Pre-trained Transformers for Remote Sensing Timeseries*, *arXiv.org*. Available at: <https://arxiv.org/abs/2304.14065v4> (Accessed: 27 November 2025).
- Tseng, G. *et al.* (2025) *Galileo: Learning Global & Local Features of Many Remote Sensing Modalities*, *arXiv.org*. Available at: <https://arxiv.org/abs/2502.09356v3> (Accessed: 27 November 2025).
- Tulloch, A.I.T., Hagger, V. and Greenville, A.C. (2020) ‘Ecological forecasts to inform near-term management of threats to biodiversity’, *Global Change Biology*, 26(10), pp. 5816–5828. Available at: <https://doi.org/10.1111/gcb.15272>.
- Tyler, A.N. *et al.* (2006) ‘Remote sensing of the water quality of shallow lakes: A mixture modelling approach to quantifying phytoplankton in water characterized by high-suspended sediment’, *International Journal of Remote Sensing*, 27(8), pp. 1521–1537. Available at: <https://doi.org/10.1080/01431160500419311>.
- Uehlinger, U. (1985) ‘An in situ pulse light fluorometer for chlorophyll determination as a monitor for vertical and horizontal phytoplankton distribution in lakes*’, *Journal of Plankton Research*, 7(5), pp. 605–615. Available at: <https://doi.org/10.1093/plankt/7.5.605>.
- UK Interagency Freshwater Group (2015) ‘Common Standards Monitoring Guidance for Freshwater Lakes’. Peterborough.
- UKCEH (2025) *Bloomin’ Algae maps*, *Bloomin’ Algae maps*. Available at: <https://bloominalgae.ceh.ac.uk/> (Accessed: 27 November 2025).
- Umlauf, L., Burchard, H. and Bolding, K. (2005) ‘GOTM sourcecode and test case documentation’, *Devel version-pre*, 4.
- Vadadi-Fülöp, C. and Hufnagel, L. (2014) ‘Climate change and plankton phenology in freshwater: current trends and future commitments’, *Journal of Limnology*, 73(1). Available at: <https://doi.org/10.4081/jlimnol.2014.770>.
- Valavi, E. *et al.* (2022) ‘Time and the Value of Data’. *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2203.09118>.

- Valenzuela, A., Reinke, K. and Jones, S. (2024) 'Assessing the spatial resolution distance of satellite images: SuperDove versus Landsat 8', *International Journal of Remote Sensing*, 45(12), pp. 4120–4159. Available at: <https://doi.org/10.1080/01431161.2024.2357839>.
- Van Nguyen, M. *et al.* (2025) 'Landsat 8 OLI atmospheric correction neural network for inland waters in tropical regions', *International Journal of Environmental Science and Technology*, 22(8), pp. 6769–6788. Available at: <https://doi.org/10.1007/s13762-024-06080-y>.
- Vanhellemont, Q. (2023) 'Evaluation of eight band SuperDove imagery for aquatic applications', *Optics Express*, 31(9), pp. 13851–13874. Available at: <https://doi.org/10.1364/OE.483418>.
- Vanhellemont, Q. and Ruddick, K. (2018) 'Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications', *Remote Sensing of Environment*, 216, pp. 586–597. Available at: <https://doi.org/10.1016/j.rse.2018.07.015>.
- Vapnik, V., Golowich, S. and Smola, A. (1996) 'Support Vector Method for Function Approximation, Regression Estimation and Signal Processing', *Advances in Neural Information Processing Systems*. MIT Press. Available at: https://papers.nips.cc/paper_files/paper/1996/hash/4f284803bd0966cc24fa8683a34afc6e-Abstract.html (Accessed: 25 June 2025).
- Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*. New York, NY: Springer. Available at: <https://doi.org/10.1007/978-1-4757-2440-0>.
- Varga, E. *et al.* (2024) 'Environmental drivers of phytoplankton community dynamics in an agriculturally-influenced tributary in the lower Great Lakes', *Science of The Total Environment*, 939, p. 173411. Available at: <https://doi.org/10.1016/j.scitotenv.2024.173411>.
- Vaswani, A. *et al.* (2017) 'Attention Is All You Need'. arXiv. Available at: <https://doi.org/10.48550/arXiv.1706.03762>.
- Velo-Suárez, L. and Gutiérrez-Estrada, J.C. (2007) 'Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucía, Spain)', *Harmful Algae*, 6(3), pp. 361–371. Available at: <https://doi.org/10.1016/j.hal.2006.11.002>.
- Vermote, E.F. *et al.* (1997) 'Second Simulation of the Satellite Signal in the Solar Spectrum, 6S: an overview', *IEEE Transactions on Geoscience and Remote Sensing*, 35(3), pp. 675–686. Available at: <https://doi.org/10.1109/36.581987>.
- Verpoorter, C. *et al.* (2014) 'A global inventory of lakes based on high-resolution satellite imagery', *Geophysical Research Letters*, 41(18), pp. 6396–6402. Available at: <https://doi.org/10.1002/2014GL060641>.
- Verspagen, J.M.H. *et al.* (2022) 'Large-scale variation in phytoplankton community composition of >1000 lakes across the USA', *Environmental Research: Ecology*, 1(1), p. 015001. Available at: <https://doi.org/10.1088/2752-664X/ac788c>.
- Vincent, A.N. *et al.* (2023) 'Machine Learning and Deep Learning Techniques for Predictive Modeling of Marine Ecosystem – A case of Flic en Flac Region, Mauritius', *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*. 2023 *Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pp. 1329–1335. Available at: <https://doi.org/10.1109/SmartTechCon57526.2023.10391505>.
- Vörös, L. and Padisák, J. (1991) 'Phytoplankton biomass and chlorophyll-a in some shallow lakes in central Europe', *Hydrobiologia*, 215(2), pp. 111–119. Available at: <https://doi.org/10.1007/BF00014715>.

- Vrede, T. *et al.* (2009) 'Effects of N : P loading ratios on phytoplankton community composition, primary production and N fixation in a eutrophic lake', *Freshwater Biology*, 54(2), pp. 331–344. Available at: <https://doi.org/10.1111/j.1365-2427.2008.02118.x>.
- Wang, D. *et al.* (2024) 'Satellite Tracking Reveals the Speed Up of the Lacustrine Algal Bloom Drift in Response to Climate Change', *Environmental Science & Technology*, 58(26), pp. 11727–11736. Available at: <https://doi.org/10.1021/acs.est.4c03391>.
- Wang, J.-H. *et al.* (2019) 'Meteorological factors and water quality changes of Plateau Lake Dianchi in China (1990–2015) and their joint influences on cyanobacterial blooms', *Science of The Total Environment*, 665, pp. 406–418. Available at: <https://doi.org/10.1016/j.scitotenv.2019.02.010>.
- Wang, M. and Shi, W. (2005) 'Estimation of ocean contribution at the MODIS near-infrared wavelengths along the east coast of the U.S.: Two case studies', *Geophysical Research Letters*, 32(13). Available at: <https://doi.org/10.1029/2005GL022917>.
- Wang, X. *et al.* (2024) 'Colonial Microcystis' biomass affects its shift to diatom aggregates under aeration mixing', *Scientific Reports*, 14(1), p. 4058. Available at: <https://doi.org/10.1038/s41598-024-53920-5>.
- 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>.
- Waqas, M. and Humphries, U.W. (2024) 'A critical review of RNN and LSTM variants in hydrological time series predictions', *MethodsX*, 13, p. 102946. Available at: <https://doi.org/10.1016/j.mex.2024.102946>.
- Warren, M.A. *et al.* (2019) 'Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters', *Remote Sensing of Environment*, 225, pp. 267–289. Available at: <https://doi.org/10.1016/j.rse.2019.03.018>.
- Wasehun, E.T. *et al.* (2025) 'Comparative analysis of Sentinel-2 and PlanetScope imagery for chlorophyll-a prediction using machine learning models', *Ecological Informatics*, 85, p. 102988. Available at: <https://doi.org/10.1016/j.ecoinf.2024.102988>.
- Weber, M. *et al.* (2017) 'Optimizing withdrawal from drinking water reservoirs to reduce downstream temperature pollution and reservoir hypoxia', *Journal of Environmental Management*, 197, pp. 96–105. Available at: <https://doi.org/10.1016/j.jenvman.2017.03.020>.
- Wei, J. *et al.* (2020) 'Seasonal succession of phytoplankton in two temperate artificial lakes with different water sources', *Environmental Science and Pollution Research*, 27(34), pp. 42324–42334. Available at: <https://doi.org/10.1007/s11356-020-10387-x>.
- Weirich, C.A. and Miller, T.R. (2014) 'Freshwater Harmful Algal Blooms: Toxins and Children's Health', *Current Problems in Pediatric and Adolescent Health Care*, 44(1), pp. 2–24. Available at: <https://doi.org/10.1016/j.cppeds.2013.10.007>.
- Welk, A. *et al.* (2008) 'Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms', *Ecological Informatics*, 3(1), pp. 46–54. Available at: <https://doi.org/10.1016/j.ecoinf.2007.12.002>.
- Werther, M., Odermatt, D., Simis, S.G.H., Gurlin, D., Lehmann, M.K., *et al.* (2022) 'A Bayesian approach for remote sensing of chlorophyll-*a* and associated retrieval uncertainty in oligotrophic and mesotrophic lakes', *Remote Sensing of Environment*, 283, p. 113295. Available at: <https://doi.org/10.1016/j.rse.2022.113295>.

- Werther, M., Odermatt, D., Simis, S.G.H., Gurlin, D., Jorge, D.S.F., *et al.* (2022) ‘Characterising retrieval uncertainty of chlorophyll-*a* algorithms in oligotrophic and mesotrophic lakes and reservoirs’, *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, pp. 279–300. Available at: <https://doi.org/10.1016/j.isprsjprs.2022.06.015>.
- Wetzel, R.G. (1990) ‘Land-water interfaces: Metabolic and limnological regulators’, *SIL Proceedings, 1922-2010*, 24(1), pp. 6–24. Available at: <https://doi.org/10.1080/03680770.1989.11898687>.
- Wevers, J. *et al.* (2021) ‘IdePix for Sentinel-2 MSI Algorithm Theoretical Basis Document’. Available at: <https://doi.org/10.5281/zenodo.5788067>.
- Whitton, B.A. and Potts, M. (2012) ‘Introduction to the Cyanobacteria’, in B.A. Whitton (ed.) *Ecology of Cyanobacteria II: Their Diversity in Space and Time*. Dordrecht: Springer Netherlands, pp. 1–13. Available at: https://doi.org/10.1007/978-94-007-3855-3_1.
- van der Wielen, P.W.J.J. *et al.* (2023) ‘Influence of Temperature on Growth of Four Different Opportunistic Pathogens in Drinking Water Biofilms’, *Microorganisms*, 11(6), p. 1574. Available at: <https://doi.org/10.3390/microorganisms11061574>.
- Willmott, C.J. and Matsuura, K. (2005) ‘Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance’, *Climate Research*, 30, pp. 79–82. Available at: <https://doi.org/10.3354/cr030079>.
- Wolf, D., Georgic, W. and Klaiber, H.A. (2017) ‘Reeling in the damages: Harmful algal blooms’ impact on Lake Erie’s recreational fishing industry’, *Journal of Environmental Management*, 199, pp. 148–157. Available at: <https://doi.org/10.1016/j.jenvman.2017.05.031>.
- Woo Kim, Y. *et al.* (2022) ‘Validity evaluation of a machine-learning model for chlorophyll *a* retrieval using Sentinel-2 from inland and coastal waters’, *Ecological Indicators*, 137, p. 108737. Available at: <https://doi.org/10.1016/j.ecolind.2022.108737>.
- Wood, S.A. *et al.* (2010) ‘Low dissolved inorganic nitrogen and increased heterocyte frequency: precursors to *Anabaena planktonica* blooms in a temperate, eutrophic reservoir’, *Journal of Plankton Research*, 32(9), pp. 1315–1325. Available at: <https://doi.org/10.1093/plankt/fbq048>.
- Woolway, R.I. *et al.* (2017) ‘Latitude and lake size are important predictors of over-lake atmospheric stability’, *Geophysical Research Letters*, 44(17), pp. 8875–8883. Available at: <https://doi.org/10.1002/2017GL073941>.
- Woolway, R.I. *et al.* (2021) ‘Phenological shifts in lake stratification under climate change’, *Nature Communications*, 12(1), p. 2318. Available at: <https://doi.org/10.1038/s41467-021-22657-4>.
- Woolway, R.I. and Merchant, C.J. (2017) ‘Amplified surface temperature response of cold, deep lakes to inter-annual air temperature variability’, *Scientific Reports*, 7(1), p. 4130. Available at: <https://doi.org/10.1038/s41598-017-04058-0>.
- Woolway, R.I. and Merchant, C.J. (2019) ‘Worldwide alteration of lake mixing regimes in response to climate change’, *Nature Geoscience*, 12(4), pp. 271–276. Available at: <https://doi.org/10.1038/s41561-019-0322-x>.
- Woolway, R.I., Sharma, S. and Smol, J.P. (2022) ‘Lakes in Hot Water: The Impacts of a Changing Climate on Aquatic Ecosystems’, *BioScience*, 72(11), pp. 1050–1061. Available at: <https://doi.org/10.1093/biosci/biac052>.

World Health Organization (2017) *Guidelines for drinking-water quality: fourth edition incorporating first addendum*. 4th ed + 1st add. Geneva: World Health Organization. Available at: <https://apps.who.int/iris/handle/10665/254637> (Accessed: 8 December 2021).

Wu, K. *et al.* (2025) 'A semantic-enhanced multi-modal remote sensing foundation model for Earth observation', *Nature Machine Intelligence*, 7(8), pp. 1235–1249. Available at: <https://doi.org/10.1038/s42256-025-01078-8>.

Wu, S.S. *et al.* (2023) 'Time-Series Forecasting of Chlorophyll-a in Coastal Areas Using LSTM, GRU and Attention-Based RNN Models', *JOURNAL OF ENVIRONMENTAL INFORMATICS*, 41(2), pp. 104–117.

Wu, Yunhai *et al.* (2014) 'Phosphorus release from lake sediments: Effects of pH, temperature and dissolved oxygen', *KSCE Journal of Civil Engineering*, 18(1), pp. 323–329. Available at: <https://doi.org/10.1007/s12205-014-0192-0>.

WWF (2024) *Living Planet Report 2024 - A system in Peril*. WWF, Gland, Switzerland. Available at: <https://www.wwf.org.uk/our-reports/living-planet-report-2024> (Accessed: 5 December 2025).

Wynne, T.T. *et al.* (2013) 'Evolution of a cyanobacterial bloom forecast system in western Lake Erie: Development and initial evaluation', *Journal of Great Lakes Research*, 39, pp. 90–99. Available at: <https://doi.org/10.1016/j.jglr.2012.10.003>.

Xiao, X. *et al.* (2017) 'A novel single-parameter approach for forecasting algal blooms', *Water Research*, 108, pp. 222–231. Available at: <https://doi.org/10.1016/j.watres.2016.10.076>.

Xiao, X. *et al.* (2024) 'Current status and prospects of algal bloom early warning technologies: A Review', *Journal of Environmental Management*, 349, p. 119510. Available at: <https://doi.org/10.1016/j.jenvman.2023.119510>.

Xiong, K. *et al.* (2025) 'Advanced classification of optical water types and ensemble learning models for Chl-a inversion in Dongting and Poyang lakes using Sentinel-2 remote sensing: assessing the impact of extreme drought events', *Ecological Indicators*, 177, p. 113738. Available at: <https://doi.org/10.1016/j.ecolind.2025.113738>.

Xu, L. *et al.* (2020) 'Potential Precipitation Predictability Decreases Under Future Warming', *Geophysical Research Letters*, 47(22), p. e2020GL090798. Available at: <https://doi.org/10.1029/2020GL090798>.

Xu, M. *et al.* (2023) 'A Comprehensive Survey of Image Augmentation Techniques for Deep Learning', *Pattern Recognition*, 137, p. 109347. Available at: <https://doi.org/10.1016/j.patcog.2023.109347>.

Xue, K. *et al.* (2023) 'Horizontal and vertical migration of cyanobacterial blooms in two eutrophic lakes observed from the GOCI satellite', *Water Research*, 240, p. 120099. Available at: <https://doi.org/10.1016/j.watres.2023.120099>.

Xue, P. *et al.* (2017) 'Improving the Simulation of Large Lakes in Regional Climate Modeling: Two-Way Lake–Atmosphere Coupling with a 3D Hydrodynamic Model of the Great Lakes', *Journal of Climate*, 30(5), pp. 1605–1627. Available at: <https://doi.org/10.1175/JCLI-D-16-0225.1>.

Xue, P. *et al.* (2022) 'Climate projections over the Great Lakes Region: using two-way coupling of a regional climate model with a 3-D lake model', *Geoscientific Model Development*, 15(11), pp. 4425–4446. Available at: <https://doi.org/10.5194/gmd-15-4425-2022>.

Yang, W. *et al.* (2010) 'An Enhanced Three-Band Index for Estimating Chlorophyll-a in Turbid Case-II Waters: Case Studies of Lake Kasumigaura, Japan, and Lake Dianchi, China', *IEEE*

Geoscience and Remote Sensing Letters, 7(4), pp. 655–659. Available at: <https://doi.org/10.1109/LGRS.2010.2044364>.

Yang, Y., Wang, H. and Li, X. (2025) ‘Deep Learning for Enhanced Ocean Color Remote Sensing: A Foundation Model Approach’, *IEEE Transactions on Geoscience and Remote Sensing*, 63, pp. 1–18. Available at: <https://doi.org/10.1109/TGRS.2025.3600411>.

Yarger, H. *et al.* (1973) ‘Water turbidity detection using ERTS-1 imagery’. Available at: <https://ntrs.nasa.gov/citations/19730019544> (Accessed: 5 December 2025).

Yenduri, G. *et al.* (2024) ‘GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions’, *IEEE Access*, 12, pp. 54608–54649. Available at: <https://doi.org/10.1109/ACCESS.2024.3389497>.

Yousefi, A. and Toffolon, M. (2022) ‘Critical factors for the use of machine learning to predict lake surface water temperature’, *Journal of Hydrology*, 606, p. 127418. Available at: <https://doi.org/10.1016/j.jhydrol.2021.127418>.

Yu, T. and Zhu, H. (2020) ‘Hyper-Parameter Optimization: A Review of Algorithms and Applications’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2003.05689>.

Yukimoto, S. *et al.* (2019) ‘The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component’, *Journal of the Meteorological Society of Japan. Ser. II*, 97(5), pp. 931–965. Available at: <https://doi.org/10.2151/jmsj.2019-051>.

Zabaleta, B. *et al.* (2023) ‘Satellite assessment of eutrophication hot spots and algal blooms in small and medium-sized productive reservoirs in Uruguay’s main drinking water basin’, *Environmental Science and Pollution Research*, 30(15), pp. 43604–43618. Available at: <https://doi.org/10.1007/s11356-023-25334-9>.

Zahir, M. *et al.* (2024) ‘A review on monitoring, forecasting, and early warning of harmful algal bloom’, *Aquaculture*, 593, p. 741351. Available at: <https://doi.org/10.1016/j.aquaculture.2024.741351>.

Zamani, M.G. *et al.* (2023) ‘Forecasting water quality variable using deep learning and weighted averaging ensemble models’, *Environmental Science and Pollution Research*, 30(59), pp. 124316–124340. Available at: <https://doi.org/10.1007/s11356-023-30774-4>.

Zhang, C., Ding, W. and Zhang, L. (2024) ‘Impacts of Missing Buoy Data on LSTM-Based Coastal Chlorophyll-a Forecasting’, *Water*, 16(21), p. 3046. Available at: <https://doi.org/10.3390/w16213046>.

Zhang, H. *et al.* (2021) ‘Hydrologic and nutrient-driven regime shifts of cyanobacterial and eukaryotic algal communities in a large shallow lake: Evidence from empirical state indicator and ecological network analyses’, *Science of The Total Environment*, 783, p. 147059. Available at: <https://doi.org/10.1016/j.scitotenv.2021.147059>.

Zhang, S. *et al.* (2025) ‘Climate change promotes harmful algal blooms in China’s lakes and reservoirs despite significant nutrient control efforts’, *Water Research*, 277, p. 123307. Available at: <https://doi.org/10.1016/j.watres.2025.123307>.

Zhang, Y. *et al.* (2018) ‘Evaluation of the Influence of Aquatic Plants and Lake Bottom on the Remote-Sensing Reflectance of Optically Shallow Waters’, *Atmosphere-Ocean*, 56(4), pp. 277–288. Available at: <https://doi.org/10.1080/07055900.2018.1454295>.

- Zhang, Y. *et al.* (2023) 'Harmful cyanobacteria-diatom/dinoflagellate blooms and their cyanotoxins in freshwaters: A nonnegligible chronic health and ecological hazard', *Water Research*, 233, p. 119807. Available at: <https://doi.org/10.1016/j.watres.2023.119807>.
- Zhao, S. *et al.* (2024) 'Rethinking self-supervised learning for time series forecasting: A temporal perspective', *Knowledge-Based Systems*, 305, p. 112652. Available at: <https://doi.org/10.1016/j.knosys.2024.112652>.
- Zhou, T. *et al.* (2022) 'Assessing Spatial and Temporal Distribution of Algal Blooms Using Gini Coefficient and Lorenz Asymmetry Coefficient', *Frontiers in Environmental Science*, 10. Available at: <https://doi.org/10.3389/fenvs.2022.810902>.
- Zhu, S. *et al.* (2020) 'Forecasting surface water temperature in lakes: A comparison of approaches', *Journal of Hydrology*, 585, p. 124809. Available at: <https://doi.org/10.1016/j.jhydrol.2020.124809>.
- Zhu, X. *et al.* (2023) 'A hybrid decomposition and Machine learning model for forecasting Chlorophyll-a and total nitrogen concentration in coastal waters', *Journal of Hydrology*, 619, p. 129207. Available at: <https://doi.org/10.1016/j.jhydrol.2023.129207>.
- Zimmermann, G. *et al.* (1993) 'MOS/PRIRODA: an imaging VIS/NIR spectrometer for ocean remote sensing', *Imaging Spectrometry of the Terrestrial Environment. Imaging Spectrometry of the Terrestrial Environment*, SPIE, pp. 201–206. Available at: <https://doi.org/10.1117/12.157057>.
- Zingone, A. and Oksfeldt Enevoldsen, H. (2000) 'The diversity of harmful algal blooms: a challenge for science and management', *Ocean & Coastal Management*, 43(8), pp. 725–748. Available at: [https://doi.org/10.1016/S0964-5691\(00\)00056-9](https://doi.org/10.1016/S0964-5691(00)00056-9).
- Ziyad Sami, B.F. *et al.* (2022) 'Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan', *Scientific Reports*, 12(1), p. 3649. Available at: <https://doi.org/10.1038/s41598-022-06969-z>.
- Zwart, J.A. *et al.* (2019) 'Cross-Scale Interactions Dictate Regional Lake Carbon Flux and Productivity Response to Future Climate', *Geophysical Research Letters*, 46(15), pp. 8840–8851. Available at: <https://doi.org/10.1029/2019GL083478>.