Low-cost Optical Detection of Pollen Bioaerosols with Machine Learning for Human Health

By

Sophie A. Mills

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Geography, Earth and Environmental Sciences

College of Life and Environmental Sciences

University of Birmingham

July 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

“道行之而成。”

[A path is made by walking on it.]

- Zhuangzi (Qiwulun 11)

“知，不知，上矣；不知，知，病也。”

[To know yet to think that one does not know is best; Not to know yet to think that one knows leads to difficulty.]

- Laozi (Dao De Jing, Chapter 71)

Abstract

Pollen and associated subpollen particles are responsible for up to 40% of populations suffering from allergies and the situation in many countries is becoming more severe with changing lifestyles, environment, and climate. Meanwhile, pollen can interact with cloud processes, affecting cloud albedo, lifetime, and precipitation patterns. However, our means to characterise airborne pollen, and bioaerosols in general, are severely limited, making it difficult to answer many important questions and assess risk to public health. Conventional pollen monitoring instruments are generally manual samplers with crucial limitations of labour, time, and cost. There are recent advancements towards developing automated pollen monitoring instruments, however, these are expensive and not economically viable to populate large monitoring networks with high spatial resolution.

The objective of this work is to address the limitations of conventional methods by investigating alternative methods that can provide useful information on pollen bioaerosols for fundamental understanding and public health. This thesis presents and evaluates novel, low-cost methods for monitoring airborne pollen, using optical particle counters (OPCs) and machine learning, and investigating physical properties of pollen under varying atmospheric conditions, using an acoustic levitator, macroscope and computer vision techniques. The superior ability of supervised machine learning models to distinguish pollen trends from OPC data is demonstrated, as well as their potential to provide useful, high spatiotemporal resolution data in unique locations and for public health. This work provides comprehensive detail on how to train, interpret and employ such models for purpose, including scrutinising how the models learn to distinguish between different pollen types. Collectively, these studies demonstrate the potential for these low-lost techniques to provide novel information on pollen bioaerosol characteristics that was previously inaccessible. This novel information could be vital for our comprehension of the bioaerosol component of atmospheric aerosols, climate models, pollen forecasts, and public health advice and warnings.

Acknowledgements

Firstly, I would like to sincerely thank my supervisors, Francis Pope and Rob MacKenzie, for the supervision and support they have given me these last few years. It has been a long journey through my PhD, not to mention COVID as well, and I am sincerely grateful for their supportive and encouraging presence during this time. It has provided a healthy balance to my own self-doubts and I'm sure they will have also observed a great deal of positive change in me through the course of my PhD. Their patience and openness have allowed me to follow my own interests, make my own discoveries and grow professionally and personally.

All this work, and the journey I have been through, would not have been possible without funding from the Natural Environment Research Council (NERC) and the CENTA2 NE/S007350/1 grant provided by them. Thank you also Mica Jones, Stephanie Smith and Jennifer Thomson who helped with all things admin and finance related for our cohort. I also express my gratitude to the UK Aerosol Society and Ecological Continuity Trust for providing funding to travel, present my work abroad and extend my experiences.

I am grateful to my various colleagues, within Francis' and Rob's groups, within my cohort, the department and the BIFoR community - all who have contributed to the unique experience that is a PhD. Aileen and Adam, who helped me find my footing at the start. Andy, from whom during our long discussions I learned many things and gained experience in general electronics and software areas, not just related to my PhD. Nick, for all vital help you have provided for practical elements on campus. The BIFoR Team - Peter, Kris, Nick, Thomas and Gael - who went above and beyond to make various practical elements of my research in the field be realised in practice.

Thank you to all my friends who supported and helped me through the good and difficult times, including Donghwi, Michelle, Marissa, Aistė, Alistair, Luke, Yuqing, Adam and Echo, and various housemates who have been there for me, in particular, Alistair, Michael, and Kat. A huge thanks to my counsellor, Lindsay, for helping me process and get through the tough times, cultivate confidence, and personally develop in a positive direction over time.

Thank you to the University of Birmingham Meditation Society and the kind community at the Birmingham Buddhist Centre for contributing to my wellbeing and personal growth. I am also heartily grateful to all those who contributed to my musical experience at the university - playing in orchestras and other opportunities - it has been wonderful to be a part of and a joy to play and perform together.

Lastly, I thank my family, the most consistent presence of all. Dad, thank you for taking an excitement and interest in my work and the time you've spent working stuff out with me, I may not have got where I am without this influence. Fortunately, I think I inherited your enthusiasm and excitability for nerdy stuff which has brought me far. Mum, thanks for always caring the most and being the most understanding, and keeping home always like home to come back to. Charlotte, thanks for understanding me deeply and appreciating my humour. Thanks for invariably being on my side and by my side as a holiday partner and chauffee. Thank you to Ozzie and Pippa for providing me something soft and furry to stroke and hug when I needed it, making me laugh a lot, and reminding me to enjoy the simple pleasures of life too.

List of abbreviations

AI – Artificial Intelligence

AOD – Aerosol Optical Depth

BIFoR – Birmingham Institute of Forest Research

CCN – Cloud Condensation Nuclei

CNN - Convolutional Neural Network

CO_2 – Carbon dioxide

EDB – Electrodynamic Balance

FACE – Free Air Carbon dioxide Enrichment

GSM - Global System for Mobile Communications

HPHE – 'High pollen high error'

HPLE – 'High pollen low error'

IN – Ice Nuclei

LiDAR – Light Detection and Ranging

LPHE – 'Low pollen high error'

LPLE – 'Low pollen low error'

MAE – Mean Absolute Error

ML – Machine Learning

MLP – Multilayer Perceptron

MSE – Mean Squared Error

NN – Neural Network

NO_2 – Nitrogen Dioxide

OPC – Optical Particle Counter

PBAP – Primary Biological Aerosol Particles

PCA – Principle Component Analysis

PM – Particulate Matter

PMF – Positive Matrix Factorisation

PNC – Particle Number Concentration

R^2 – coefficient of determination

ReLU – Rectified Linear Unit

RF – Random Forest

RH – Relative Humidity

RMSE – Root Mean Squared Error

RNN – Recurrent Neural Network

RRMSE – Relative Root Mean Squared Error

SHAP – SHapley Additive exPlanation

SI – Supporting Information

SOA – Secondary Organic Aerosols

SPP – Subpollen Particles

UFP – Ultra Fine Particles

VOC – Volatile Organic Compounds

WIBS - Wideband-Integrated Bioaerosol Sensor

XAI – Explainable Artificial Intelligence

List of Figures

Figure 1.1: *Summary of different aerosol sources. Taken from Myhre, Myhre et al. (2013).* ---------- ***[5]***

Figure 1.2: *Diagram of the direct effect of aerosols on radiative forcing (a) and indirect effect via cloud interactions (b). Taken from Myhre, Myhre et al. (2013).* ---------- ***[7]***

Figure 1.3: *Various bioaerosols on a size scale. Taken from Fröhlich-Nowoisky et al. (2016).* ---------- ***[10]***

Figure 1.4: *Simulated annual average near-surface bioaerosol concentrations for bacteria, fungal spores and pollen. Taken from Hoose et al. (2010).* ---------- ***[11]***

Figure 1.5: *Picture of a Burkard (Hirst-type) trap belonging to the University of Worcester, UK.* ---------- ***[20]***

Figure 1.6: *UK pollen monitoring sites. Taken from https://www.worcester.ac.uk/about/academic-schools/school-of-science-and-the-environment/science-and-the-environment-research/national-pollen-and-aerobiology-research-unit/pollen-calendar.aspx (Accessed June 2023).* ---------- ***[22]***

Figure 1.7: *Generalisation error and bias-variance trade-off relevant to machine learning. Taken from Pichler and Hartig (2023).* ---------- ***[34]***

Figure 1.8: *Diagram of implementation of dropout in neural networks. Taken from Srivastava et al. (2014).* ---------- ***[39]***

Figure 2.0: *Graphical abstract for Chapter 2.* ---------- ***[53]***

Figure 2.1: *Time series for Simple Pollen Proxy.* ---------- ***[72]***

Figure 2.2: *Time series for PMF Pollen Proxy.* ---------- ***[74]***

Figure 2.3: *Bar plots of correlations between OPC input features and Hirst target variables and Random Forest feature importances.* ---------- ***[76]***

Figure 2.4*: Time series for NN and RF pollen proxies.* ---------- ***[79]***

Figure 2.5: *Polar plots of detected pollen source concentrations as a function of wind direction and speed.* ---------- ***[86]***

Figure 2.6: *Monthly and diurnal trends for model-constructed pollen proxies compared to Hirst baseline.* ---------- ***[88]***

Figure 3.0: *Graphical abstract for Chapter 3.* ---------- ***[103]***

Figure 3.1: *Methodology by which various hyperparameter configurations were tested.* ---------- ***[117]***

Figure 3.2: *Effect of varying hyperparameters on model performance, quantified by the R^2 score.* ---------- ***[126]***

Figure 3.3: *Summary of R^2 scores achieved for final hourly and daily resolution models.* ---------- ***[130]***

Figure 3.4: *SHAP 'bee swarm' plots for all pollen types bins-only models.* ---------- ***[135]***

Figure 3.5: *Matrix plot summarising observable nonlinear correlations between input features and target pollen variables based on SHAP bee swarm plots in Figure 3.4.* ---------- ***[136]***

Figure 3.6: *Bar plots showing feature value for bins 0-11 for HPLE (high pollen low error), LPLE (low pollen high error), HPHE (high pollen high error) and LPHE (low pollen high error) events for each target.* ---------- ***[138]***

Figure 3.7: *Summary of particle sizes from OPC bins that demonstrate some positive correlation with each target variable according to further SHAP analysis.* ---------- ***[143]***

Figure 4.2: *Map and location of the Birmingham Institute of Forest Research (BIFoR) Free Air CO_2 Enrichment (FACE) facility, taken from Hart et al. (2019).* ---------- ***[165]***

Figure 4.3: *Picture and diagram of Alphasense OPC-N3 sensors on the flux tower at BIFoR FACE.* ---------- ***[166]***

Figure 4.3: *Time series of PM_{10},* Quercus *pollen number concentration, temperature, and RH, uncorrected and corrected results.* ---------- ***[171]***

Figure 4.4: *Zoomed in time series from Fig. 4.3 with only corrected data and from the start of the 'tower' period (5th April) up to 16th April 2022.* ---------- ***[172]***

Figure 4.5: *Pearson correlation coefficients visualised as matrix plots for each sensor pairing for corrected PM_{10} and* Quercus *pollen concentration for each of colocation and tower periods.* ---------- ***[173]***

Figure 4.6: *Bar plots of mean values for each variable – PM_{10},* Quercus *pollen number concentration, temperature and RH - with 95% confidence intervals for each sensor and period.* ---------- ***[174]***

Figure 4.7: *Diurnal trend of* Quercus *pollen.* ---------- ***[175]***

Figure 4.8: *Polar plots showing corrected PM_{10} mass concentration and* Quercus *pollen proxy number concentration as a function of wind speed and direction for each sensor height during the tower period.* ---------- ***[176]***

Figure 5.4: *Diagram illustrating setup of acoustic levitation chamber, humidification system and macroscope for experiments investigating the hygroscopic behaviour of pollen.* ---------- ***[207]***

Figure 5.5: *Example of multiple Lilium orientalis pollen grains being levitated at once but experiencing instability and fluctuation in structure and orientation.* ---------- ***[212]***

Figure 5.6: *Visualised results for surface-fixed Lilium orientalis pollen grains undergoing incremental RH changes with error bars.* ---------- ***[216]***

Figure 5.7: *Visualised results for surface-fixed Populus deltoides pollen grains undergoing incremental RH changes with error bars.* ---------- ***[218]***

Figure 5.8: *Visualised results for levitated Lilium orientalis pollen grains undergoing incremental RH changes with error bars.* ---------- ***[220]***

Figure 5.9: *Box-and-whisker plots summarising the change in area and aspect ratios for the RH interval 60-90% for each experiment group with overlaying swarm plot showing individual data points.* ---------- ***[222]***

List of Tables

Table 1.1: *Estimates of global emissions, number and mass concentrations in the biosphere for different types of Primary Biological Aerosol Particles (PBAP), adapted from Després et al. (2012) and Fröhlich-Nowoisky et al. (2016) and references therein.* ---------- ***[12]***

Table 1.2: *Summarised statistics of pollen grain numbers for various tree features for a range of 10 different anemophilous tree species as reported by Molina et al. (1996).* ---------- ***[15]***

Table 1.3: *Confusion matrix.* ---------- ***[30]***

Table 2.1: *Summary of Neural Network (NN) and Random Forest (RF) models and evaluation metrics.* ---------- ***[80]***

Table 2.2: *Mean-averaged metrics for models by category.* ---------- ***[82]***

Table 3.1*. Summary table of hyperparameter configurations trialled for this study.* ---------- ***[117]***

Table 3.2: *Summary of R^2 scores achieved for all final models, both hourly and daily time resolution.* ---------- ***[131]***

Table 5.1: *Previous studies investigating the hygroscopic behaviour of pollen.* ---------- ***[203]***

Table 5.2: *Representative snapshots taken for each pollen type and experiment setup showing the image quality achieved as well as the variation across different instances.* ---------- ***[214]***

Table 5.3: *Mean changes in area and aspect ratios within in experiment group and the results of paired t-tests performed between area ratios at 60 and 90% RH respectively.* ---------- ***[222]***

Table 5.4: *Comparison of the method in this study and vapour sorption analyser method reported in literature, including calculated mass increase values for increasing RH between 60-90% for surface-fixed pollen experiments in this study with reported literature values.* ---------- ***[226]***

TABLE OF CONTENTS

Abstract ii
Acknowledgements iii
List of abbreviations v
List of Figures vii
List of Tables x
TABLE OF CONTENTS xi
CHAPTER 1: Introduction 1
1.1 Atmospheric Aerosols 2
1.1.1 The atmosphere 2
1.1.2 Aerosols overview 3
1.1.3 Interaction between aerosols and climate 5
1.2 Bioaerosols 9
1.2.1 Overview 9
1.2.2 Pollen 13
1.2.3 Subpollen particles 16
1.2.4 Plants, pollen and climate change 17
1.3 Instruments and monitoring techniques 19
1.3.1 Monitoring airborne pollen – current standard 19
1.3.2 Monitoring pollen – other techniques 22
1.3.3 Low-cost automated sensor technology 24
1.4 Artificial Intelligence & Machine Learning 27
1.4.1 Introduction to Artificial Intelligence and Machine Learning 27
1.4.2 Types of learning 28
1.4.3 Metrics and evaluation of machine learning models 30
1.4.4 Generalisation error and bias-variance trade-off 33
1.4.5 Ensemble Methods and Random Forests 35
1.4.6 Neural Networks and Deep Learning 37
1.5 Research gaps and thesis objectives 41
1.6 Thesis overview 43
1.6 References 45
2. CHAPTER 2: Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests 56
2.1 Abstract 57
2.2 Introduction 59

2.3 Materials and methods 64
2.3.1 Instrumentation 64
2.3.2 Context 66
2.3.3 Data pre-processing 66
2.3.4 Simple (large particle) pollen proxy 67
2.3.5 PMF pollen proxy 68
2.3.6 Supervised machine learning methods 69
2.3.7 Neural Network method 70
2.3.8 Random Forest method 71
2.3.9 Evaluating the spatial footprint of pollen vs pollen proxy 73
2.3.10 Evaluating model ability to distinguish high pollen events 73
2.4 Results and discussion 75
2.4.1 Simple pollen proxy 75
2.4.2 PMF pollen proxy 77
2.4.3 Neural Network and Random Forest methods 79
2.5 Conclusions 94
2.6 Contributors 96
2.7 Acknowledgements 96
2.8 Data availability 97
2.9 Comment 98
2.10 References 99
3. CHAPTER 3: Machine learning methods for low-cost pollen monitoring – model optimisation and interpretability 106
3.1 Abstract 107
3.2 Introduction 109
3.3 Methods 116
3.3.1 Context and Instrumentation 116
3.3.2 Neural network hyperparameter tuning 117
3.3.3 Final models 121
3.3.4 Model interpretation using SHAP values 123
3.4 Results and Discussion 125
3.4.1 Hyperparameter tuning 125
3.4.2 Improved models from chosen hyperparameter values 130
3.4.3 Model interpretation and input-output relationships 134
3.4.4 Total Pollen 140
3.4.5 *Poaceae* 141
3.4.6 *Quercus* 141

3.4.7 *Betula* ... 142
3.4.8 *Pinus* ... 142
3.4.9 Summary ... 143
3.4.10 Importance of input features and context ... 144
3.5 Conclusions ... 147
3.6 Contributors ... 149
3.7 Acknowledgements ... 149
3.8 Data availability ... 150
4.9 Comment ... 151
4.10 References ... 152
4. CHAPTER 4: Can low-cost sensors be used to observe airborne pollen at a forest site at varying heights through the canopy? ... 157
4.1 Abstract ... 158
4.2 Introduction ... 160
4.3 Methods ... 165
4.3.1 Context ... 165
4.3.2 Instrument deployment ... 166
4.3.3 Model training ... 168
4.3.4 General analyses ... 169
4.4 Results and discussion ... 172
4.4.1 Results summary ... 172
4.4.2 Sensor congruency ... 179
4.4.3 Observed PM_{10} characteristics ... 180
4.4.4 Observed *Quercus* pollen characteristics ... 182
4.4.5 Observed *Quercus* pollen characteristics across different heights ... 183
4.4.7 Comments on *Poaceae* and total pollen for comparison ... 187
4.5 Conclusions ... 189
4.6 Contributors ... 191
4.7 Acknowledgements ... 191
4.8 Comment ... 193
4.9 References ... 194
5. CHAPTER 5: Acoustic levitation of pollen and visualisation of hygroscopic behaviour ... 197
5.1 Abstract ... 198
5.2 Introduction ... 199
5.3 Methods ... 205
5.3.1 Pollen samples ... 205

5.3.2 Acoustic levitation chamber apparatus 206
5.3.3 Hygroscopicity experiments 208
5.3.4 Image and data processing 209
5.4 Results and discussion 212
5.4.1 Acoustic levitation and visual imaging of pollen grains 212
5.4.2 Surface-fixed pollen hygroscopicity experiments 216
5.4.3 Levitated hygroscopicity experiments 220
5.4.4 Summarised results 223
5.4.5 Comparison with previous methods 224
5.5 Conclusions 229
5.6 Contributors 231
5.7 Acknowledgements 231
5.8 Data and code availability 231
5.9 Comment 232
5.10 References 233
6. CHAPTER 6: Summary and future directions 236
6.1 Thesis summary 237
6.2 Limitations and future directions 239
6.3 Potential value 241
Appendices 242
7. Supporting information for Chapter 2 243
8. Supporting information for Chapter 3 250
Section 1: General OPC bin information 250
Section 2: Extended methods section 251
Section 3: Various dataset statistics relevant to the XAI/SHAP analysis on high and low pollen and error categories 257
Section 4: Extended SHAP analysis figures 259
Section 5: Effect of removing certain bin input features from models 266
9. Supporting information for Chapter 4 267
10. Supporting information for Chapter 5 278

CHAPTER 1: Introduction

1.1 Atmospheric Aerosols

1.1.1 The atmosphere

The Earth's atmosphere comprises a series of layers which each have their own specific characteristics. The troposphere is the layer that starts from the Earth's surface and extends up to about 10 km above sea level. Most clouds and weather systems occur within the troposphere, largely because 99% of atmospheric water vapour is contained within this layer. With increasing height, both pressure and temperature generally drop, but conditions within this layer are susceptible to great variance and turbulence. The stratosphere sits above the troposphere up to around 50 km above the Earth's surface. This is where the ozone layer resides, which absorbs ultraviolet (UV) radiation from the sun. Temperature generally increases with height, resulting in relatively minimal turbulence within this layer. Above the stratosphere is the mesosphere, which extends up to 85 km and continues to drop greatly in pressure and temperature with increasing height. The thermosphere beyond this (up to 500-1000 km) has very low pressure and air density but absorbs much of the high energy (X-ray and UV) radiation from the sun, making the temperatures very high but also largely dictated by variations in solar radiation. (This description is adapted from a website resource supplied by the University Corporation for Atmospheric Research (UCAR) Centre for Science Education: https://scied.ucar.edu/learning-zone/atmosphere/layers-earths-atmosphere, accessed June 2023).

For the purposes of this thesis, we are generally interested in the troposphere, as this is where most meteorology and variation occurs and has the most direct effect on our daily lives. When the term 'atmosphere' is used henceforth in this thesis, it generally refers to the troposphere or biosphere. The biosphere more specifically refers to the

portion of the atmosphere near the Earth's surface where there is interface with ecosystems, with natural emissions and anthropogenic activity.

One more important term to establish is the atmospheric boundary layer. This is defined as the part of the atmosphere which directly interacts with the Earth's surface and where turbulence is generated (https://www.metoffice.gov.uk/research/foundation/parametrizations/boundary-layer, accessed June 2023). The depth of the boundary layer can range variably from just metres to kilometres above the surface and exhibits seasonal and diurnal variations dictated by meteorology (Rodrigues et al., 2021). Turbulence is generated by the transfer of energy between the sun and earth's surface and redistributes heat, moisture, wind speeds (or kinetic energy) and gaseous and particulate constituents of the atmosphere. Thus, it plays an important role in modulating weather and air quality, including aerosols (see, e.g., Q. Li et al., 2020; W. Li et al., 2020).

1.1.2 Aerosols overview

Aerosols are defined as suspensions of solid, liquid or mixed-phase particles present in a gaseous medium and are present throughout the troposphere with highly variable size distributions and chemical compositions (Myhre, Myhre et al., 2013, Mushtaq et al., 2022). This high variability is due to the wide range of aerosol types, sources and atmospheric processing they can undergo, which can affect particle compositions over functions of time and location (McMurry et al., 1999). Among this diversity, aerosols can be classified into different categories based on origin, chemical composition, or size.

Primary atmospheric aerosols originate from sources from the Earth's surface and includes natural mineral dust and sea salt as well as black carbon and organic carbon, often from anthropogenic sources such as combustion (see Figure 1.1). Secondary

aerosols originate from chemical processes in the atmosphere, by precursor gases condensing onto pre-existing particles or nucleating into new particles, and include sulphates, nitrates and organic carbon (Myhre, Myhre et al., 2013). These formation processes often occur within clouds, which are closely linked with aerosol particles (Ervens et al., 2011). Aerosols can further be classified into 'coarse', 'fine' and 'ultrafine' particulate matter which refers to those of aerodynamic diameters within 10, 2.5 and 0.1 μm respectively (Mushtaq et al., 2022).

Aerosol distribution across the globe can be estimated from Aerosol Optical Depth (AOD), measured by remote-sensing satellites (Mushtaq et al., 2022). AOD measurements, as well as estimated contributions of different aerosol types, show great spatial and temporal variability across the globe. Though AOD measurements tend to be highest around Asia and tropical regions of Africa, with sulphates and biogenic carbon being significant components in each case respectively (Myhre, Myhre et al., 2013). Atmospheric aerosols do not remain static but evolve through processes of coagulation, condensation and chemical reactions which depend on meteorological conditions. In other words, aerosol processes and transport are largely influenced by the atmospheric boundary layer. Thus, it can be difficult to accurately characterise and quantify the aerosol components of the atmosphere and large uncertainties remain.

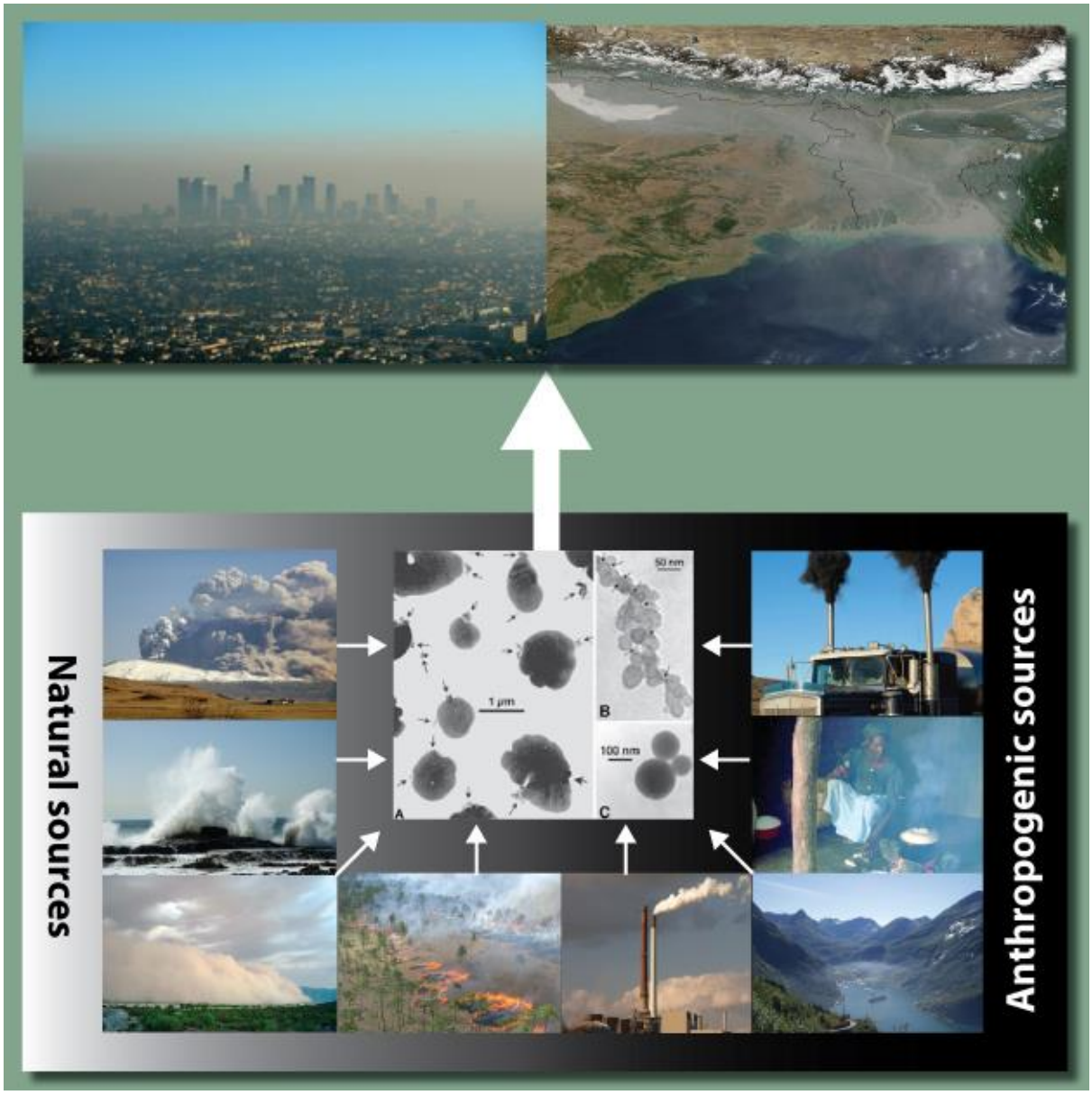


Figure 1.1: *Summary of different aerosol sources. Taken from Myhre, Myhre et al. (2013).*

1.1.3 Interaction between aerosols and climate

Aerosols are important components of the atmosphere to consider because they can interact with various processes and influence feedbacks within the Earth system and impact climate (Carslaw et al., 2010). For example, atmospheric aerosols can have a direct effect by either scattering or absorbing incoming solar radiation depending on

chemical composition. Scattering solar radiation results in a cooling effect on the Earth, while absorbing it results in a warming effect. There are also indirect effects that occur via aerosol-cloud interactions, whereby cloud albedo and lifetime is affected by aerosol concentrations which in turn affects radiative forcing and climate (see Figure 1.2). The IPCC reports that aerosols, especially the aerosol-cloud interactions component, are still a major source of uncertainty, impacting the reliability of future climate predictions (Myhre, Shindell et al., 2013).

Cloud formation is fundamentally influenced by aerosol particles since many of them can act as cloud condensation nuclei (CCN) and also ice nuclei (IN). The concentration of aerosols, and in turn CNN concentrations, can influence the size and number concentration of cloud droplets. This affects the properties of the cloud, in particular albedo which dictates the extent to which the clouds can reflect solar radiation, and is an important factor for climate models. For example, an increased concentration of CNN will result in more numerous cloud droplets, but each of smaller size for a certain water content. This increases the cloud albedo and thus the cooling effect of reflecting solar radiation (Twomey, 1977). Meanwhile, it may also mean that the cloud droplets require more time to reach sizes that will result in precipitation. Thus, the lifetime and precipitation patterns of clouds will also be affected (Albrecht, 1989).

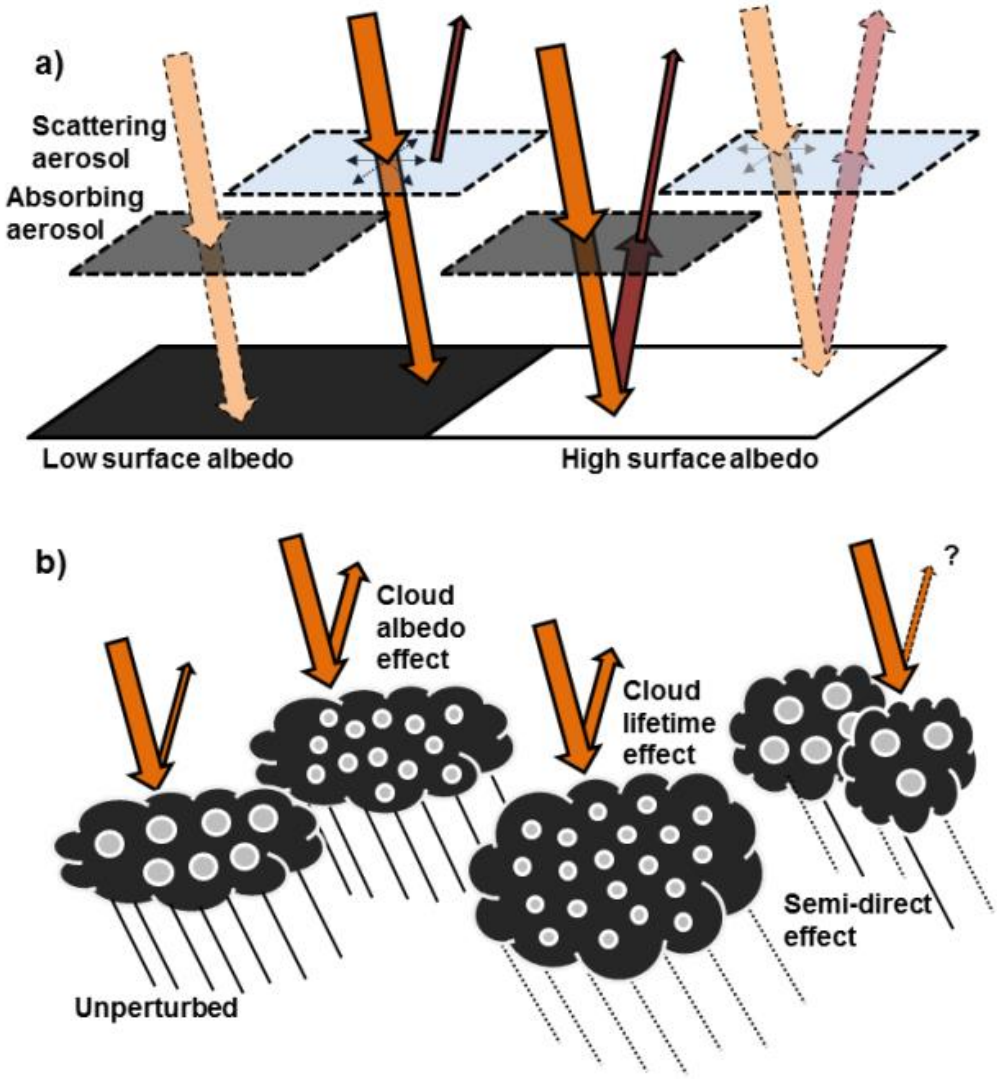


Figure 1.2: *Diagram of the direct effect of aerosols on radiative forcing (a) and indirect effect via cloud interactions (b). Taken from Myhre, Myhre et al. (2013).*

The potential impact of different aerosols on climate can be estimated by calculations of radiative forcing, which quantifies the change in energy balance of the Earth/atmosphere system due to a perturbation of natural or anthropogenic origin (Boucher et al., 2001; Myhre, Myhre et al., 2013). However, to calculate this effect requires knowledge of aerosol particle density distributions across the atmosphere and the characteristics of each aerosol type present (see, e.g., Andrews et al., 2011). Current

limitations in obtaining this information contributes to large uncertainties due to aerosol effects when predicting climate sensitivity and future states of the atmosphere (Andreae et al., 2005; Bellouin et al., 2019).

Aerosols warrant concern as they directly contribute to poor air quality, the world's leading environmental health risk and a major cause of mortality globally (Shaffer et al., 2019; McDuffie et al., 2021). In this context, aerosols are generally referred to as particulate matter (PM) which has been correlated with adverse health outcomes including cardiovascular and respiratory diseases (Shiraiwa et al., 2017; Contini et al., 2021). In particular, ultrafine particles (UFP) (< 100 nm) are known to penetrate more readily into the respiratory system due to their small size, resulting in a stronger inflammatory effect than coarser particles and increased adverse biological effects. Large numbers of toxic free radicals can be deposited from the high surface areas of UFPs within the lungs and diffuse into any organ system (Groma et al., 2022).

A range of different instrument types exist to characterise the diverse range of atmospheric aerosols, measuring properties including size distribution, number concentration, density, chemical composition, CNN concentration, optical coefficients and refractive index. A summary of these various instruments is illustrated by McMurry (2000) which classifies them according to their capacity to resolve size, time and composition.

1.2 Bioaerosols

1.2.1 Overview

Primary Biological Aerosol Particles (PBAP) are described as "solid airborne particles derived from biological organisms, including microorganisms and fragments of biological materials such as plant debris and animal dander" by Després et al (2012) according to the IGAP (1992). This includes many different types of particles, of diverse shapes and sizes, from airborne viruses, cell fragments, archaea, algae and cyanobacteria to fungal spores, hyphae, pollen and other plant debris. Their aerodynamic diameter can range over four orders of magnitude, generally between 10 nm and 100 μm (see Figure 1.3; Fröhlich-Nowoisky et al., 2016; Huffman et al., 2019).

The term 'bioaerosol' is generally used synonymously with PBAP, but can be less distinct as to what the meaning encompasses. For example, there are also biogenic secondary organic aerosols (SOA), formed by atmospheric oxidation and gas-to-particle conversion of volatile organic compounds (VOCs) released from biological organisms, however the term 'PBAPs' excludes such particles. Here, we use the term bioaerosols to mean PBAPs, excluding SOAs formed from gaseous precursors, but including fragments derived from PBAPs that may be released by atmospheric processes, since these can present similar concerns for climate and public health as the primary particles they derive from.

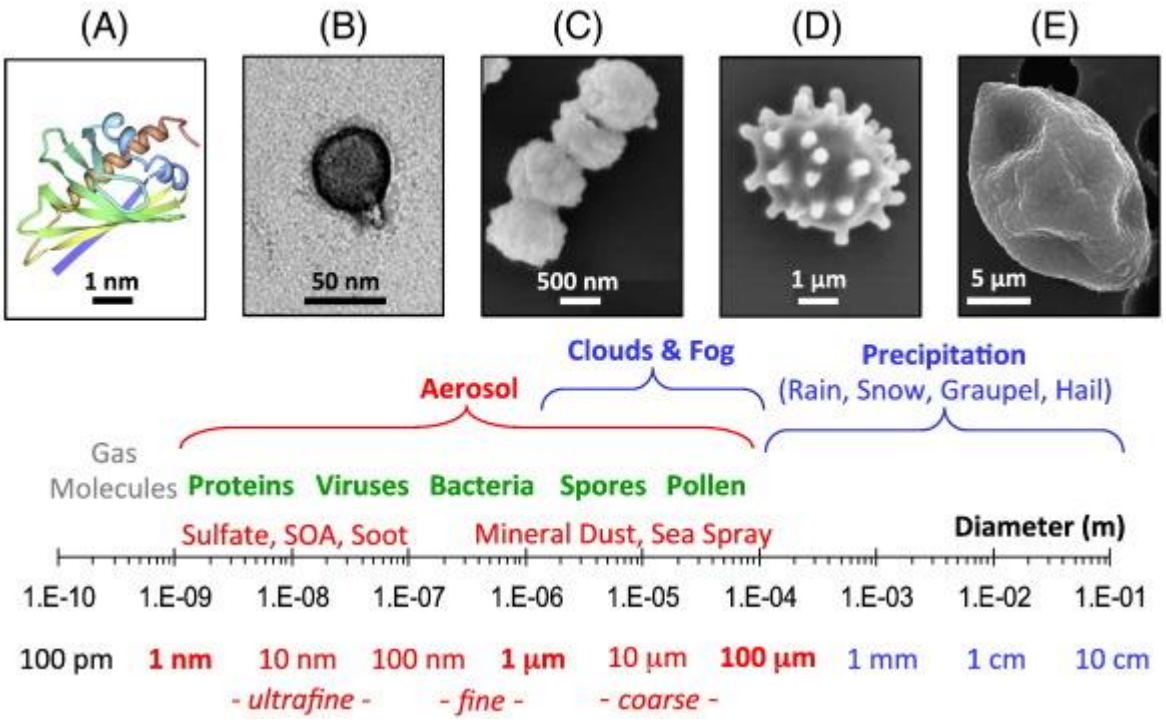


***Figure 1.3**: Various bioaerosols on a size scale. Taken from Fröhlich-Nowoisky et al. (2016).*

With such a wide variety of sources present in every biome of the biosphere, land and ocean, bioaerosols can be found unanimously across the globe. Mechanisms for their release can be passive, for example via meteorology and decay mechanisms, or active, by reproduction mechanisms or insects. It is estimated that the global emission source of PBAP may total as much as 1000 Tg a^{-1} (Carslaw et al., 2010), while particle fluxes vary greatly depending on vegetation cover of the Earth's surface (Lindemann et al., 1982). A summary of estimated global emissions, number and mass concentrations for the various types of bioaerosols is displayed in Table 1.1, adapted from comprehensive reviews by Després et al. (2012) and Fröhlich-Nowoisky et al. (2016). Figure 1.4, taken from Hoose et al. (2010), broadly visualises the simulated global near-surface distributions of bacteria, fungal spores and pollen. Bacteria bioaerosol concentrations

appear to be highest across continental regions, excluding polar regions, up to ~10^5 m^{-3} and above in some areas. The distribution of pollen bioaerosols similarly is highest for continental regions, generally but to ~10^2 m^{-3}, but excluding desert regions as well as polar regions. Meanwhile, fungal spore bioaerosols appear to be particularly concentrated (up to number concentrations of ~10^5 m^{-3}) around marsh regions of North America and tropical regions of south America, central Africa and southern Asia, i.e. continental regions where relative humidity levels may be particularly high.

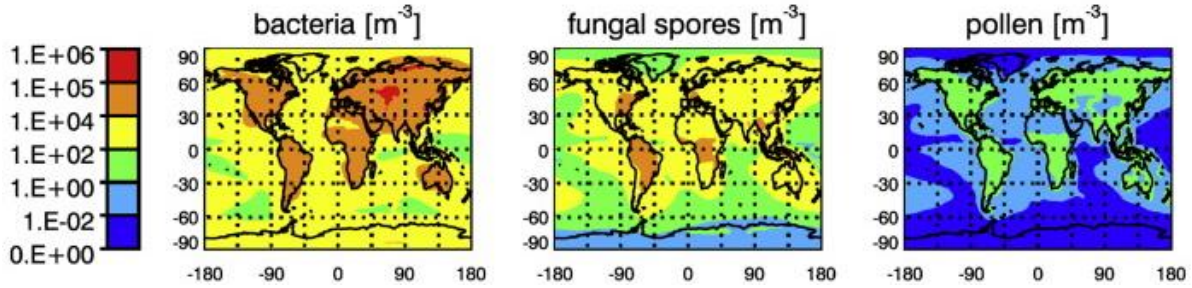


Figure 1.4: *Simulated annual average near-surface bioaerosol concentrations for bacteria (left), fungal spores (middle) and pollen (right). Taken from Hoose et al. (2010).*

Though many bioaerosols arise from natural processes, human activity can also increase the number and alter the composition of airborne microorganisms, particularly in urban and agricultural areas. For example, bioaerosol concentrations can be elevated around composting sites or crop fields, and behaviour and allergenicity can be heavily influenced by anthropogenic air pollution and PM concentrations (Shammi et al., 2021). Matthias et al. (2000) also reported the bioaerosol fraction of total aerosol particles to be higher in populated urban/rural areas (24% by number concentration) than remote continental (20%) and marine (17%) regions. They are not only confined to the

atmosphere outdoors but are also present in indoor environments. For example, fungal spores arising from mould in buildings, and can cause significant concerns for human health (Verhoeff et al., 1997; Ghosh et al., 2015).

Table 1.1: *Estimates of global emissions, number and mass concentrations in the biosphere for different types of Primary Biological Aerosol Particles (PBAP), adapted from Després et al. (2012) and Fröhlich-Nowoisky et al. (2016) and references therein.*

	Global emissions (Tg a^{-1})	**Number concentration (m^{-3})**	**Mass concentration (µg m^{-3})**
Viral particles	-	~10^4	~10^{-3}
Bacteria	0.4-28	~10^4	~0.1
Algae	-	~10^2-10^3	~10^{-3}
Fungal spores	8-190	~10^3-10^4	~0.1-1
Fern spores	-	~10-10^3	~1
Pollen	47-84	~10-10^3	~1
Total PBAP (dominated by plant debris & fungal spores)	<10		
Total PBAP (including cellular fragments)	~10^3		

Bioaerosols, as airborne pathogens, can cause infectious respiratory diseases, including tuberculosis and influenza (Peccia et al., 2011), can have acute toxic effects, and sensitise allergies and atopic diseases, such as allergic rhinitis and asthma (Verhoeff et al., 1997; Burge et al., 2000, Douwes et al. 2003; Humbal et al., 2018). Studies also suggest they can increase risk of cancer and affect the immune system, though the full

extent of possible effects of PBAPs is not fully understood (Peccia et al., 2011; Humbal et al., 2018). Bioaerosols can be used harmfully as agents of bioterrorism and biowarfare, which has brought them to the concern of national defence (Lim et al., 2005). Meanwhile, as with non-biological aerosols, bioaerosols can act as CCN and IN and so are also of relevance to climate and Earth system feedbacks (Ariya et al., 2004; Christner et al., 2008; Prenni et al., 2009).

Reviews evaluating the different methods available for monitoring bioaerosols can be found by Griffiths et al. (1994) and Huffman et al. (2019). This includes modern real-time techniques based on methods such as fluorescence spectroscopy, elastic scattering, microscopy, holography, mass spectrometry, Raman spectroscopy and remote sensing techniques (e.g. LiDAR).

1.2.2 Pollen

Pollen grains occur naturally in the process of plant reproductive cycles for the purpose of transporting the male gametes (reproductive cells) to other individuals for fertilisation of the female gamete. For this purpose, different species of pollen have evolved into different shapes and sizes to facilitate this process of dissemination. While some are transported by animals, such as insects or mammals, a large proportion have evolved to be disseminated by wind - such plants are called anemophilous species. Thus, pollen can be present in the atmosphere across the biosphere with great variance, depending on vegetation cover and season, and travelling significant distances at substantial concentrations depending on meteorology (Skjøth et al., 2007; Alarcón et al., 2022).

Pollen grains are among the largest particulates present in the atmosphere, ranging between 10 and 100 μm (Després et al., 2012; Reponen, 2011, pp723; Bradley, 2015, pp.

408-409). Pollen types that are not evolved to travel by wind tend to be larger, and can exceed 100 μm in diameter. Anemophilous types tend to be on the smaller end of this diameter range (typically 17-58 μm (Stanley and Linskins, 1974)), since a smaller size facilitates suspension in the atmosphere for longer periods and distances before being deposited. Thus it is these types that are most atmospherically relevant.

Pollen, among other bioaerosols, can be heavily influenced by meteorological conditions and tend to exhibit diurnal emissions behaviour, influenced by the atmospheric boundary layer. For example, Jones and Harrison (2004) concluded in a review of numerous studies that mechanisms of bioaerosol (including pollen) release tended to correlate negatively with RH, while many pollen taxa had a positive correlation of emissions with temperature. A diurnal pattern can be observed (see, e.g., Juprasong et al., 2022; Ríos et al., 2016) with concentrations rising in the morning as surfaces dry out and wind speed increases, which can be delayed by overcast or humid conditions. Some studies observed a bimodal concentration profile with a second peak in the afternoon, which could be attributed to changing source strength, increased afternoon wind speeds, or downward mixing from higher concentrations above. Higher concentrations can also be stimulated by heavy rainfall or other events of physical disturbance. Meanwhile, concentrations and emission profiles may be limited by depletion of the source (e.g. catkins, or fruiting bodies).

Pollen grains are released specifically from anthers found within individual flowers which generally are clustered into groups, or 'inflorescences' across the branches in the tree. For many anemophilous trees these inflorescences are produced in the form of catkins, which may contain in the range of 10^1-10^2 (e.g. ~20 for oak) flowers per inflorescence and 2-10^2 (around 7 for oak) anthers per flower (Molina et al., 1996).

Table 1.2 summarises the general scale of number of pollen grains found per anther, flower, inflorescence, tree and metre of crown diameter for 10 different anemophilous tree species investigated by Molina et al. (1996). The specific ranges for oak (*Quercus* taxon) are added beside this for context, since this taxon will be particularly relevant in the context of this thesis. Ultimately, these figures can vary significantly among and across different taxa/species and will be dependent on the age and size of the tree. (Evidently, from their size, the trees studied by Molina et al. (1996) were likely not fully mature.)

Table 1.2: *Summarised statistics of pollen grain numbers for various tree features (e.g. per flower, per inflorescence, per tree) for a range of 10 different anemophilous tree species as reported by Molina et al. (1996).*

	General range over 10 species	**Specific range for oak**
Pollen grains per anther	10^3-10^5	10^3
Pollen grains per flower	10^4-10^5	10^4
Pollen grains per inflorescence	10^5-10^7	10^5-10^6
Pollen grains per tree	10^6-10^8	10^8
Pollen grains per metre of crown diameter	10^8-10^{10}	10^{10}
Tree height	3-9 m	4-5 m
Tree (crown) diameter	4-12 m	11-12 m

1.2.3 Subpollen particles

Pollen grains are known to release numerous subpollen particles (SPP) on rupturing in the atmosphere, often, but not exclusively, stimulated by meteorological conditions such as high relative humidity (Miguel et al., 2005; Matthews et al., 2023). There are currently limited studies characterising SPP but it is understood that they can encompass a range of particles including starch granules and other cytoplasmic debris that also contain allergenic material and more easily penetrate the respiratory system (Miguel et al., 2005; Taylor et al., 2004). SPP have generally been reported between the range of sub-μm to several μm (i.e. <10 μm) (see, e.g., Miguel et al., 2005) but naturally vary for different pollen species. For example, diameter ranges of 20 nm-6.5 μm and 0.25-3.2 μm have been measured for ragweed (Stone et al., 2021) and birch (Burkart et al., 2021) respectively, and characterisation of airborne concentrations are inevitably limited to the size range of the instrument/method used for measurement. SPPs may be more numerous than individual pollen grains in the atmosphere and pose a greater risk to public health, since they are of respirable size (<5 μm) and retain key components that induce airway inflammation (Pazmandi et al., 2012).

Hendrickson et al. (2023) and Matthews et al. (2023) measured SPP emission factors, stimulated by wind, to be between 10^3-10^5 particles/pollen grain and 10^{13}-10^{15} particles/m^2 for oak (*Quercus*), ryegrass (*Lolium*) and giant ragweed (*Ambrosia*) species. This was for particles between 0.010-1.0 μm measured by a condensation particle counter. Using a portable aerosol spectrometer measuring up to 20 μm, they observed that the majority of SPP particles were submicron (<1 μm) but meanwhile there were also supermicron-sized particles counted. Matthews et al. (2023) also conclude that while pollen grains may play a more important role in IN activity, SPP may have greater

impact on clouds as CCN. Bioaerosol number concentrations relative to background (sulfate) aerosols dictate how significant the CCN effect is (Griffiths et al., 2012), and subpollen particles will generally be more numerous than pollen grains.

1.2.4 Plants, pollen and climate change

Plants are very sensitive to climate change. They can be affected by increased CO_2 levels which can lead to greater CO_2 uptake and a fertilisation effect, with potentially increased growth (i.e. carbon storage) or production (i.e. turnover) of certain features - e.g. trunk/stem, roots, branches, leaves or flowers. Meanwhile, changes in meteorological conditions, such as temperature or rainfall, or nutrient availability and balance can also affect the processes of growth and production occurring within plants or trees, i.e., their phenology. Naturally, the production of pollen from flowers and catkins may also be susceptible to the changing climate.

Firstly, studies demonstrate that increasing temperatures are changing plant phenology patterns and causing pollen seasons to lengthen, since the temperature threshold that stimulates pollen release from plants is being hit earlier in the year (Anderegg et al., 2021; Zhang and Steiner, 2022). Pollen seasons could lengthen by as much as 20 days and increase airborne pollen concentrations by 20-40%, which alone increases the period and severity of exposure for individuals suffering from allergies. Additionally, it has been shown that increasing CO_2 levels can be responsible for increased pollen production for ragweed (Ziska and Caulfield, 2000; Wayne et al., 2002; Choi et al., 2018), grass (Albertine et al., 2014) and oak (Kim et al., 2018) species. Furthermore, there is evidence that air pollution may interact with aeroallergens (Sedghy et al., 2018; Grewling et al., 2019), modulate allergenicity of pollen and promote airway sensitisation (Choi et al., 2021). For example, the effect of NO_2, a major traffic pollutant, significantly

affects pollen chemistry and increases allergenicity (Zhao et al., 2015). Another study demonstrated correlation between high pollutant concentrations and patients associated with allergies (Leru et al., 2022).

This is a serious issue for public health. Higher pollen levels are associated with increased visits to hospital emergency departments and medication sales (Ito et al., 2015; Sheffield et al., 2008) and the situation is projected to worsen with some particularly allergenic species, such as ragweed, expanding to new areas in the future (Lake et al., 2016). Ziska (2021) highlights the effects we need to consider and where unknowns still exist, particularly due to the complexity of botanical ecosystems. There are still many aspects of pollen and SPPs that need to be better understood, including detailed information on spatiotemporal emission and transport characteristics and the impact of varied atmospheric conditions on residency lifetime, cloud interactions and human health.

1.3 Instruments and monitoring techniques

1.3.1 Monitoring airborne pollen – current standard

The current 'gold' standard for measuring airborne pollen (and fungal spore) concentrations is the Hirst-type (or Burkard, https://burkard.co.uk/) volumetric trap, shown in Figure 1.5. This technology is now 70 years old (Hirst, 1952) yet is still used at the majority of active national monitoring stations (Buters et al., 2018). This is a manual sampler where samples of particulates from ambient air are deposited on sticky tape, which is transferred onto slides and taken back to the lab. The number of pollen grains of each taxon must then be counted under the microscope by a trained professional. This is a laborious process, more so for higher time resolutions, and there is an inevitable time delay for obtaining the final pollen concentrations data (Tummon et al., 2021). While they are relatively robust and cost-efficient (in terms of initial capital expenditure, CAPEX, but not subsequent operating expenditure, OPEX), this method still gives rise to significant error and discrepancy between instruments, depending on the time resolution compared.

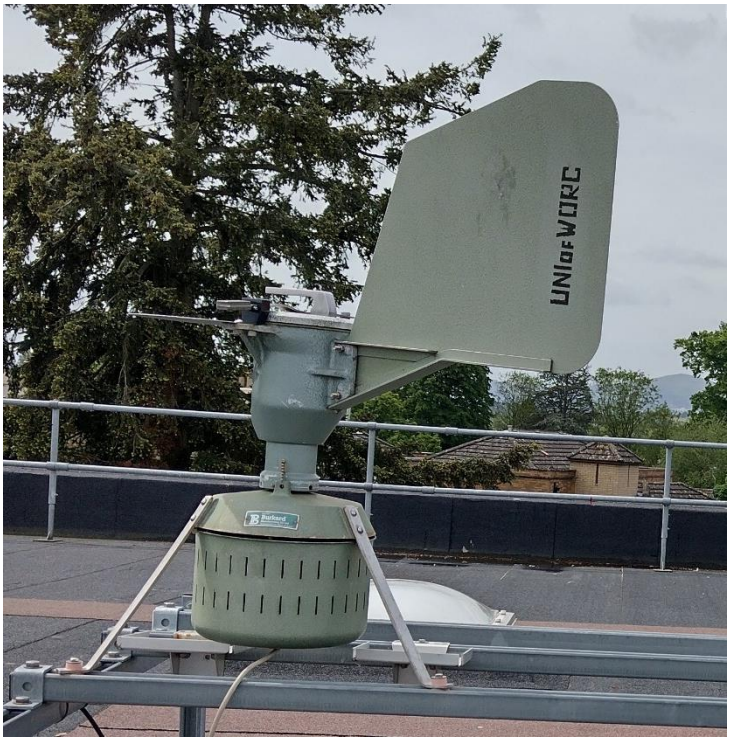


Figure 1.5: *Picture of a Burkard (Hirst-type) trap belonging to the University of Worcester, UK.*

Adamov et al. (2021) investigated the error between 3 Hirst parallel at various time resolutions. They reported that, for time resolutions of 24 and 12 hours, the instruments measured total pollen in agreement at least 99 and 98% of the time, however this lowered to 93 and 85% respectively for 6 and 2 hour resolutions (i.e. ~15% chance that no pollen will be observed even is if present in the air for 2 hourly resolution). These percentages were lower for individual pollen taxa, ranging from 50% (Betula) to 84% (Poaceae) for daily and 41-60% for 2 hourly resolution. They observed that pollen taxa with shorter pollen seasons and lower concentrations (e.g. Betula, Fraxinus) had less consistency while taxa that were easily recognisable during times of lower pollen diversity (e.g. Poaceae) had higher consistency. This is likely due to sub-sampling procedures or human error during counting. Such methods are sensitive to sampling

flow rate as pollen is collected and methodology specifics observed by the person undertaking the counting (e.g. surface area of slide studied).

Currently, there are only 11 regulatory pollen monitoring stations across the UK for a population of well over 60 million people (see Figure 1.6). This equates to approximately one monitoring site for every ~22,000 km^2. This demonstrates the current difficulty we have in accurately measuring airborne pollen concentrations and highlights the limited quality of information we can provide at present to the public, as well as the wider research community who could benefit, including atmospheric scientists and climate modellers. More recently, there have been developments in the field of automated instruments for measuring pollen (see section 1.3.2). However, when comparing pollen concentrations obtained from other instruments it is important to keep in mind the significant error of the current Hirst standard, particularly at higher time resolutions.

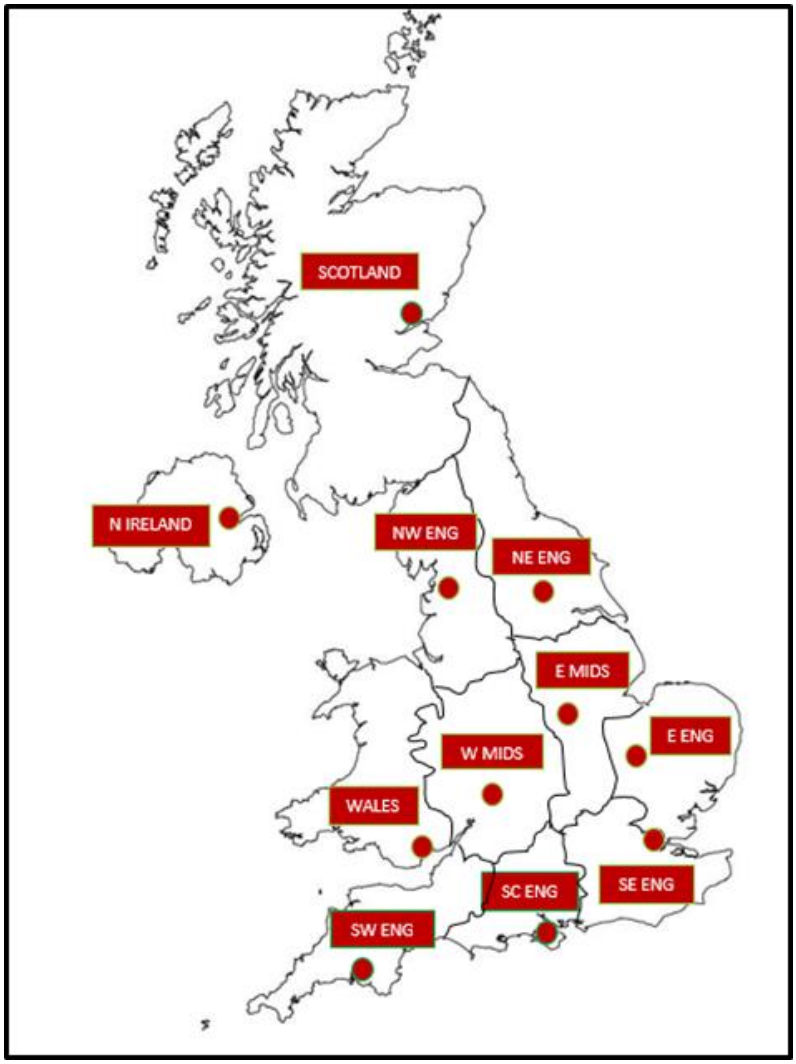


Figure 1.6: *UK pollen monitoring sites. Taken from https://www.worcester.ac.uk/about/academic-schools/school-of-science-and-the-environment/science-and-the-environment-research/national-pollen-and-aerobiology-research-unit/pollen-calendar.aspx (Accessed June 2023).*

1.3.2 Monitoring pollen – other techniques

New advancements in technology have given rise to a number of different instruments to address the limitations of the Hirst and automate pollen monitoring. These include instruments such as the APS-300 from PollenSense (Jiang et al., 2022), the BAA500 from Hund-Wetzlar (Oteros et al., 2015; Oteros et al., 2020; Plaza et al., 2022), the Plair Rapid-

E (Šaulienė et al., 2019; Smith et al., 2022) and the Swisens Poleno (Sauvegeat et al., 2020; Sofiev et al., 2022), which generally all employ machine learning techniques. The APS-300 and the BAA500 work similarly to the Hirst in that particles are impacted on tape, but then the optical microscopy imaging and classification are performed by (computer vision) machine learning algorithms. Meanwhile, the Rapid-E and Poleno are cytometers which collect data from in-flight particles and use machine learning algorithms trained on specific data for various pollen types to distinguish them. The Rapid-E uses laser scattering signal, fluorescence spectrum and fluorescence lifetime information while the Swisens Poleno uses laser scattering, fluorescence and holographic imaging techniques (Sauvageat et al., 2020).

These instruments have so far demonstrated very promising results and better precision between automated instruments than the manual Hirsts (Oteros et al., 2020). They may also be more effective at fully capturing airborne pollen seasons and diversity (Plaza et al., 2022). Valuable advantages of increased time-resolution (Smith et al., 2022; Chappuis et al., 2019) and real-time data (Crouzy et al., 2016; Sauvageat et al., 2019) also come with these methods. Importantly, machine learning - in particular, deep learning - methods have enabled these devices to become operational (Schaefer et al., 2021) due to their ability to learn patterns from data after sufficient experience.

Another instrument that can be used to measure bioaerosols more generally is the Wideband-Integrated Bioaerosol Sensor (WIBS), or the Spectral Intensity Bioaerosol Sensor (SIBS) which expands upon the WIBS with greater spectral resolution (Nasir et al., 2019). This is instrument measures in-flight single particle fluorescence across three channels of different wavelength ranges but also uses light scattering calculations to evaluate particle shape (sphericity) (O'Connor et al., 2014; Savage et al., 2017), and

unsupervised (see section 1.4.2) machine learning techniques have been employed to discriminate different bioaerosol particles (Ruske et al., 2018). It has also been demonstrated to characterise pollen fragments (i.e. SPP) (Hughes et al., 2020). Although, this method has not yet achieved the accuracy to measure pollen concentrations nor the ability to distinguish taxa that the other techniques (designed specifically for pollen and using supervised machine learning) have demonstrated. These specialised instruments however can be very expensive, costing in tens to hundreds of thousands of pounds.

Further details on recent developments and available instruments for automatic pollen monitoring can be found by Maya-Manzano et al. (2020) and Buters et al. (2022). However, there is no official standard yet in this area and Tummon et al. (2022) highlights the relevant considerations that are needed to provide high-quality data from automatic monitoring networks. This includes considerations of surrounding land-use and vegetation, proximity to local particle sources and representativity.

1.3.3 Low-cost automated sensor technology

Low-cost sensors can provide a trade-off of high accuracy for many other benefits including lower costs, high spatiotemporal resolution, real-time data, flexibility, portable convenience and the possibility to collect data from unique places. While regulatory grade instruments are necessary to retain experimental rigour and accurate measurements, low-cost sensors clearly have an important role to play beside them. By low-cost sensors, we mean sensors that are significantly more affordable in terms of acquisition and running costs than their regulatory counterparts and range from under £100 to within £1000.

There are various low-cost sensors that have become commercially available for monitoring ambient air pollution and specifically PM. A general review of such PM

sensors from various companies - including AlphaSense, Plantower and Shinyei – can be found in Alfano et al. (2020). Generally they use light scattering and Mie Theory to determine the count particles passing through a laser and determine approximate size to varying resolutions. Thus, they are often called Optical Particle Counters (OPCs). When using low-cost sensors, it is good practice to be aware of data quality and implications for the context being studied. Giordano et al. (2021) summarises the challenges and limitations concerned with low-cost sensors, as well as best practices for obtaining high quality data when using them.

The particle measurements from these sensors are sensitive to the optics (diode, transistor and lens) and geometry of the measurement cell, and rely on assumptions of particle refractive index and sphericity. Thus, complex composition and morphology, for example of bioaerosols, may cause difficulties. Low-cost sensors can also struggle with lower limits of detection compared to reference instruments which becomes relevant in cleaner environments (e.g. in remote or polar regions). Sometimes sensors have been observed to drift over many months of use or report erroneously after exposure to abnormally high concentrations of dust (Tryner et al., 2020). It is possible for the inlet pathway to sometimes become blocked by dust or insects. Regular calibration with an appropriate reference instrument is generally necessary for reliable data and corrections may be required – among low-cost sensors when looking at relative concentrations and against a reference instrument where absolute concentrations are important.

Sensors that do not dry the sampled air before measurement can also record PM mass concentrations that are significantly skewed by high RH conditions (>85%). This is due to the mass change that hygroscopic particles undergo with increasing humidity, and

would also be relevant to pollen (Pope, 2010; explored further in Chapter 5). In such cases it is often best practice to apply a correction factor based on κ-Köhler theory (see, e.g., Crilley et al., 2017; 2020). (This may, however, be less relevant for measured number concentrations, depending on the particle size range bin widths observed by the sensor.)

1.4 Artificial Intelligence & Machine Learning

1.4.1 Introduction to Artificial Intelligence and Machine Learning

The term 'Artificial Intelligence' (AI) has increasingly draw attention in recent years, yet it remains somewhat difficult to define accurately. It can refer to algorithms that are 'capable of achieving similar to human-like performance in specific decision or recognition tasks' (Pichler and Hartig, 2023). However, experienced AI practitioners, such as Nello Christianini (2023), argue that we must reconsider our concept of 'intelligence' beyond our narrow standard of 'human intelligence'. For example, there are tasks that we are specifically good at, relating to logic or reasoning, however there are also tasks that other organisms or machines can perform better than us, thus using our ability to solve a given task as a benchmark for 'intelligence' is not appropriate. This begs the question: how do we then define intelligence? Wang (2008) addresses precisely this problem and concludes that while various AI definitions differ, there can be 'multiple valid working definitions that serve as ultimate research goals'. In this case, we use the term AI synonymously with 'machine learning' and is fundamentally concerned with creating computer systems or models that automatically improve with experience (Jordan and Mitchell, 2015).

Machine learning (ML) encompasses the multitude of algorithms which serve as tools in AI to learn statistical patterns from data which prove useful for a given task and generally improve on predictive performance with experience (Goodfellow et al., 2016, pp 96). In contrast to classical methods, reliance on laws and assumptions based on theoretical understanding are discarded in favour of simply learning statistical patterns from the data. Though, this comes with the trade-off of requiring large datasets before

useful predictions can be made and an increased difficulty for 'human intelligence' to interpret (Christianini, 2023).

To learn, machine learning algorithms are tasked with a 'loss function' that provides a penalty or score for how well the algorithm is achieving the given task and it must optimise this. The algorithm is a fixed structure, chosen by the practitioner from the various generic machine learning algorithms available and tested to date, but within the structure are parameters (or 'weights') which are updated as the algorithm iterates through the data searching for a configuration of parameter values that optimise the given loss function. While the algorithmic structures come from generic families, there are often various hyperparameters which control the specific structure or the way the algorithm learns, iterates through data or updates its parameters (Goodfellow et al., 2016, pp 416—420).

Aside from simply learning patterns of the data it is given to adjust its parameters to, the algorithm must also be able to take new 'unseen' data and make predictions or decisions from it which satisfy the given task. Therefore, in machine learning you need a training and test dataset – to learn to map the relationship between input and output and evaluate the performance of the fitted model, respectively. In some cases, a validation dataset is also used to tune hyperparameters (or inform the algorithm when to stop training) before the test dataset is used for final evaluation.

1.4.2 Types of learning

Machine learning algorithms can be classified into different types of learning, which depend on the type of task at hand and related data requirements. Supervised learning is used when 'labels' or examples of correct execution of the task can be presented to the algorithm (Goodfellow et al., 2016, pp 102). In such cases, the loss function is tasked to

minimise the error calculated between the model output and the 'correct' target value on each iteration. This type of learning includes decision tree-based and neural network methods and is very useful and widely used today to solve regression (i.e. continuous output) and classification (i.e. discrete output) tasks. However, it can be demanding and reach limitations in terms of needing sufficient labelled data.

Unsupervised learning does not require labelled examples to learn from but instead is generally tasked with finding patterns among the data while optimising a general loss function for the given algorithm. The three kinds of task for unsupervised learning, outlined by Pichler and Hartig (2023), are dimension reduction, clustering and anomaly detection. This involves algorithms such as Principal Component Analysis (PCA), *k*-means, density based or hierarchical clustering or neural network-based autoencoders. Generally these algorithms require the practitioner to make some decision beforehand, either arbitrarily or based on previous knowledge of the data, such as the number of dimensions or clusters to result in for meaningful interpretation of the data.

Reinforcement learning is another type, similar in process to supervised learning, but rather than using labelled data these algorithms learn from feedback provided by interacting with a (virtual or real) environment. This type of learning can be thought of somewhat like playing a game, where the algorithm learns the rules and how to 'do well' by evaluating the consequences of actions taken, whether these cause a gain or loss in points, money, or 'health', for example.

The type that we will primarily focus on for the purposes of this work is supervised learning.

1.4.3 Metrics and evaluation of machine learning models

Machine learning has abandoned classical methods that rely on observed distributions, p-values and confidence intervals and has prioritised the task of optimising predictive performance by empirical inference as opposed to using logical reasoning and causal inference. This presents the problem of how to evaluate model performance. Generally, the performance of trained machine learning models is evaluated by how well it performs on a hold-out 'test' dataset that has not been used for training, measured by a metric which depends on the type of task.

For classification problems, the simplest metric that can be used is accuracy which is described by Eq. 1. However, accuracy can give a false sense of achieving good performance when classes are unbalanced.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad \text{Eq. 1}$$

Another approach can be to construct a confusion matrix which summarises the complete model performance by four categories described in Table 1.3.

Table 1.3: *Confusion matrix.*

	Positive prediction	**Negative prediction**
Positive actual observation	**True positive:** both prediction and actual observation are positive.	**False negative**: the predictions was negative while the actual observation was positive.
Negative actual observation	**False positive**: the prediction was positive while the actual observation was negative.	**True negative**: both prediction and actual observation are negative.

From these categories, precision (Eq. 2) and recall (Eq. 3) can be calculated which can be combined to calculate the F1 Score (Eq. 4; the harmonic mean of precision and recall) (Goodfellow et al., 2016, pp 411-412). This is a more rigorous metric for accuracy, the higher the score the better.

$$Precision = \frac{True\ positives}{True\ positive + False\ positive} \quad \text{Eq. 2}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad \text{Eq. 3}$$

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall} \quad \text{Eq. 4}$$

A metric that works well for multiclass classification and gives high accuracy is logarithmic loss. For this, the classifying algorithm must assign probabilities for each possible class for all samples, then the logarithmic loss metric penalises false classifications. Its calculation is described by Eq. 5, where N is the number of samples, M is the number of classes, y_{ij} indicates whether sample i belongs to class j or not, and p_{ij} indicates the probability of sample i belonging to class j. As the loss diminishes closer to 0, it indicates higher accuracy.

$$Logarithmic\ loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}\right) \quad \text{Eq. 5}$$

The equivalent metric generally used for evaluating regression models during training is Mean Squared Error (MSE; Eq. 6) (Goodfellow et al., 2016, pp.105), which takes the average of the square of the difference between actual and predicted values. This has the advantage over the Mean Absolute Error (MAE; Eq. 7) that the gradient is easier to compute, which is important for the optimisation process of machine learning models.

$$Mean\ Squared\ Error = \frac{1}{N}\sum_{j=1}^{N}(y_i - \hat{y}_j)^2 \quad \text{Eq. 6}$$

$$Mean\ Absolute\ Error = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{y}_j\right| \quad \text{Eq. 7}$$

Often, the quoted metric for machine learning models is Root Mean Squared Error (RMSE), which is the square root of MSE and allows for evaluation of model error at a scale that is generally proportional to the output values. RMSE and MAE have units that

correspond to the output variable units and are best understood in context of the given data.

Another common metric for evaluating machine learning models in the literature is the coefficient of determination (R^2). This can be understood by the proportion of variance of the dependent variable that is predicted by a regression model, or a 'goodness of fit'. It is calculated by Eq. 8, where y_i is the actual observation for sample i, $\hat{y}_i$ is the prediction for sample i, and $\bar{y}$ is the mean of all actual y values. R^2 values vary between 0 and 1 where higher values mean the model better predicts the outcome.

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad \text{Eq. 8}$$

This metric is often useful for comparing models that are predicting different tasks and the output range is not directly comparable, since it is essentially a percentage score of model fit.

1.4.4 Generalisation error and bias-variance trade-off

The purpose of machine learning models is to perform well on new data, and the ability to do so is called 'generalisation'. Generalisation error is estimated by evaluating model performance on an unseen 'test' dataset and we need to minimise this (Goodfellow et al., 2016, pp 107).

The generalisation error in machine learning performance is generally distinguished by two properties: bias and variance, which are related and between which a trade-off exists (see Figure 1.7). Bias refers to systematic model error, or in other words, is a measure of how close a model can capture the mapping function between inputs and outputs. Meanwhile, the variance refers to the stochastic error of the parameters, or, the

amount model performance would change when fit on different training data (Pichler and Hartig, 2023; pp 10-11, Brownlee, 2021; pp 34, James et al., 2021). Either can be minimised – for example, bias can be reduced with increasing model complexity – however, at the cost of the other increasing. Underfitting occurs when bias is high but variance is low (i.e. further left on the x-axis in (a) in Figure 1.7) and is characterised by high error on the training dataset. Meanwhile, overfitting occurs when bias is reduced but variance inevitably increases (i.e. further right on the x-axis in (a) in Figure 1.7), learning the specific details of the training data and noise too well, and this results in low training error but high test (generalisation) error (Goodfellow et al., 2016, pp 108, 126). Thus, a sweet spot must be found which minimises both.

Since the variance component of the error generally increases with increasing model complexity (i.e. parameters), the question is: why do machine learning algorithms which contain many parameters demonstrate impressively good performance when classical models of similar complexity (number of parameters) would fail? The answer is that practically all ML approaches are designed with the ability to perform implicit complexity adjustments that limit flexibility and avoid overfitting (Pichler and Hartig, 2023). This can be via internal algorithmic (self-adapting) complexity optimisation, e.g. manifested in ensemble model learning, or external adaptation of model complexity controlled by hyperparameters which control and can constrain model architecture and learning behaviour. The former of which is explained further in section 1.4.5 and the latter in section 1.4.6. Generally, deep learning models can exhibit somewhat surprising behaviour whereby the generalisation error can continue to increase even with increasing model complexity, as shown in Figure 1.7. This phenomenon is further discussed in section 1.4.6.

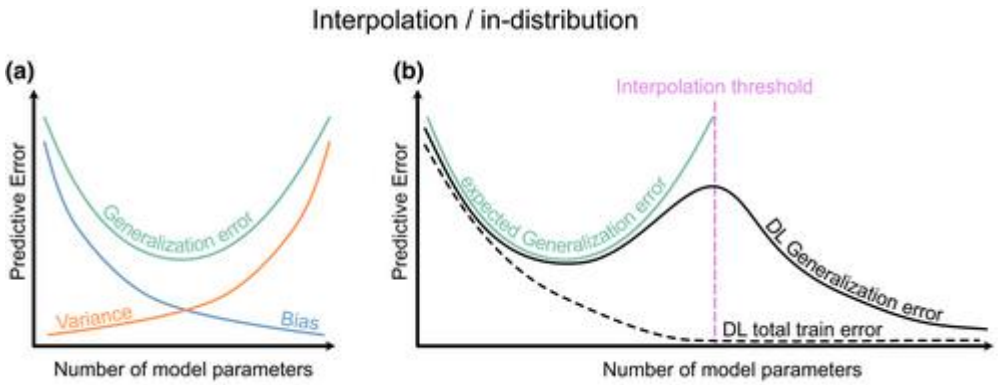


Figure 1.7: *Generalisation error and bias-variance trade-off relevant to machine learning for in-distribution data and the unusual behaviour of deep learning (DL) where generalisation error can continue decreasing even with increasing model complexity. Adapted from Pichler and Hartig (2023).*

1.4.5 Ensemble Methods and Random Forests

Random Forest (RF) is a type of ensemble machine learning algorithm, based on 'forests' of many decision trees, often used for supervised learning. It is a popular technique since it has proved powerful for both classification and regression problems with relatively little need of hyperparameter tuning (i.e. it is a good 'out-of-the-box' method).

Ensemble methods are based on the idea of gathering 'opinions' from skilled (i.e. can achieve better than random chance) learners and making a final conclusion based on the collective opinions. This relies on members having diverse and independent opinions based on their unique knowledge and interpretation of information given to them which, when aggregated together can be decentralised and unbiased (pp 22, Rokach, 2010). Ensemble machine learning algorithms involve the training of many 'weak learner' models (of the same or different types), which have all learned differently in

some way, and then combining their predictions for the final decision. This results in increased robustness and, usually, the averaged predictive performance of the ensemble will surpass that of any one member. This is due to the decrease in variance of prediction errors by aggregation of many ensemble members, but without significant increase in bias (pp 1, Zhang and Ma, 2012). (In some cases it may also reduce bias – see, e.g., Rokach, 2009, pp 39.) Although, formally, the total number of parameters across all ensemble members is large, the collective complexity is not dependant on the number of members but rather on average complexity and differences between ensemble members (Dietterich, 2000; Dormann et al., 2018; Ganaie et al., 2022).

Random forests are formed by ensembles of many decision tree predictors, such that each tree depends on random, independently sampled information (from the same distribution/training dataset). With an increasing number of trees within the forest, the generalisation error converges to a limit that depends on the strength of the individual trees and the correlations between them (Breiman, 2001). The individual trees are decorrelated by bootstrap aggregation ('bagging'), where each tree is based on slightly different samples of the training dataset with replacement (so instances may be selected more than once, i.e. bootstrap sampling). Furthermore, the random forest algorithm also selects a random subset of input features at each node (split point), where the feature values are evaluated in the process of deciding the best split point parameters while constructing each tree (James et al., 2014, pp 320; Kuhn and Johnson, 2013, pp 199). Trees may also be under-pruned so that they reach depths (or numbers of splits) that cause individuals to overfit. All these factors contribute to the individual trees making more different predictions based on different experience and thus with less correlated prediction errors (Brownlee, 2021). In general, random forests are very unlikely to overfit, since adding further trees does not cause this effect (Hastie et al., 2016, pp 596).

Hyperparameters that can be adjusted for random forest models include number of trees (or estimators) in the forest, maximum tree depth, maximum number of features to consider when looking for the best split and number of samples to take from the training data to train each tree. These can be implemented in machine learning libraries such as Scikit-learn in Python (see Pegregosa et al., 2011). These and other hyperparameters allow flexible adjustment of collective model complexity and correlation between tree estimators, which in turn will impact on the bias-variance trade-off and generalisation (test dataset) error.

There are empirical rules of thumb on optimal ranges to trial for these hyperparameters - for example, a number of estimators between 100 (Scikit-learn's default) and 1000 or more (recommended by Kuhn and Johnson, 2013), tree depth of up to 10 (Brownlee, 2021) and number of features selected per split at 1/3 and the square root of the total number of input features respectively for regression and classification problems (Kuhn and Johnson, 2013, pp 387). However, ultimately, optimal hyperparameter configurations can be tuned via a grid (or random) search trial and error approach to reach optimal performance (Goodfellow et al., 2016, 420-420). This is easier to implement for machine learning algorithms such as random forests which have fewer hyperparameters, as opposed to deep learning algorithms that have a diverse array of many hyperparameters to consider.

1.4.6 Neural Networks and Deep Learning

Neural Networks (NN) generally refer to a whole family of machine learning algorithms which were loosely inspired by the neural structure of the brain. The predecessor of neural networks was the perceptron, described by Rosenblatt in 1958, which mapped inputs to a binary classifier output via a single node but this was not flexible enough to

solve non-linear tasks. It was only decades later that artificial neural networks were developed and their potential realised and this developed into deep learning. The term ‘deep learning’ encompasses a variety of different architectural model types which are essentially neural networks, but importantly they comprise of multiple hidden layers which process the data between input and output layers and learn useful representations (LeCun et al., 2015).

These can be simply neural networks of densely (fully) connected layers of neurons, called multilayer perceptrons (MLPs), or can comprise more complex architectures such as convolutional neural network (CNN) or recurrent neural network (RNN) blocks. These more complex architectures can be useful when addressing particular tasks, for example CNNs are good at processing image data and building feature maps layer upon layer, while RNNs are particularly good for sequenced or time series data.

Deep learning neural networks are generally used for supervised learning tasks, with labelled data, but in some cases can be structured in a way that supports unsupervised (or ‘self-supervised) learning. These architectures are called Autoencoders and are tasked with reconstructing the given input onto the output but via a bottleneck of limited dimension, forcing the model to learn a limited but useful representation of the original data (Goodfellow et al., 2016, pp 493-494).

The family of deep learning algorithms are set somewhat apart from other machine learning algorithms since they generally exhibit greater complexity, often with far greater numbers of parameters within the model and external hyperparameters. As with other machine learning models, deep learning models demonstrate impressive performance despite formally having high numbers of parameters and complexity due to internal algorithmic mechanisms and external hyperparameters that can adjustment

effective complexity according to the given task. Deep learning models even go beyond simpler machine learning models and can exhibit a second decrease in generalisation error (after the formal bias-variance trade-off 'sweet spot' – see Figure 1.7) with increasing model number of parameters and complexity (Pichler and Hartig, 2023).

There are contemporary discussions in the literature as to why this is, but it is generally attributed to the automatic complexity adjustments the models can make during the learning process due to inherent redundancy and flexibility within the network and hyperparameters. Among the hyperparameters, there are many that can be applied for regularisation in various ways, to constrain the network and its ability to over-learn the data, and these can cause some interesting effects. One example is to add 'dropout' layers by applying a hyperparameter which makes the network ignore, or 'drop', a certain proportion of nodes in each layer as the algorithm iterates through the learning process (see Srivastava et al., 2014). However, for each iteration the nodes that are dropped in each layer are different, meaning that the network architecture and corresponding parameters are slightly different each time (visualised in Figure 1.8). This can be thought of as similar to the ensemble effect described in the previous section. Effectively, many slightly differing model architectures (or individual estimators) are being combined which can successfully reduce generalisation error. To facilitate effectively having many different pathways to learn from within the same overall model architecture, it can be beneficial to over-parameterise the model and allow for redundancy and flexible learning (not relying on too much on one pathway nor every single node present) (Pichler and Hartig, 2023).

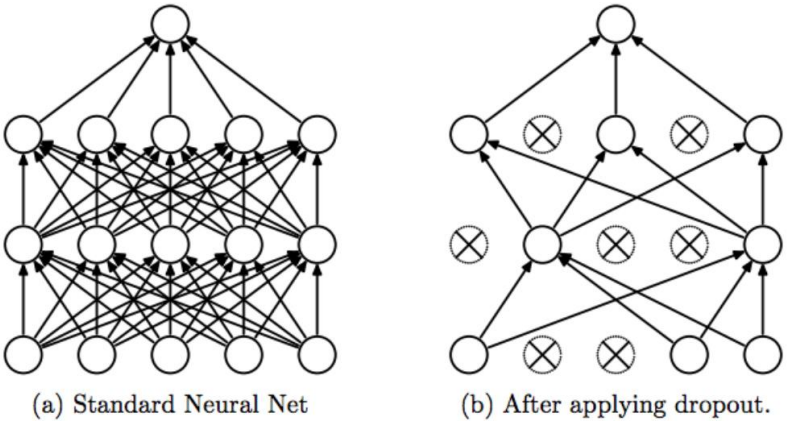


Figure 1.8: *Diagram of implementation of dropout in neural networks. Taken from Srivastava et al. (2014).*

The various hyperparameters that can be configured in (MLP) neural networks include number of layers, number of nodes per layer, learning rate and optimiser (dictating how the model parameters are updated on each iteration), batch size (number of samples to go through before updating model parameters), and activation function (dictating how each node transforms the input data). Furthermore, there are numerous techniques that can regularise the model in different ways; these include early stopping, dropout layers, weight penalties, weight constraints, and adding noise. Model performance can also be developed further by various ensemble methods for neural networks (see, e.g., Brownlee, 2018).

1.5 Research gaps and thesis objectives

This introduction has discussed the importance of studying bioaerosols and pollen for human health and climate, and has described the current technologies we have available to measure them. There is currently a significant limitation in the ability to study atmospheric bioaerosols with automated, real-time data collection and high time resolution. Where this possibility is now emerging, instrument costs are extremely expensive. If any can be bought, they certainly cannot be bought and deployed numerously to cover a wide variety of contexts with flexibility and high spatial resolution. While there are now various types of commercially available low-cost sensors for measuring PM, there is no such option for bioaerosols.

This is in part due to the great complexity of bioaerosols present in the atmosphere, which vary significantly in sizes and shape depending simply on their species but also on the physical conditions they are exposed to. Furthermore, their composition is complex and bioaerosols can release or break down into smaller particles that can still be relevant, if not more so, to human health. Our understanding of the changes that bioaerosols undergo in the atmosphere and the subsequent implications for us is severely limited.

Thus, the objectives that we set out to achieve with this work are as follows:

- To develop low-cost methods of studying pollen using currently commercially available equipment that can give useful information to fill the current knowledge gaps.
- To test the application of these methods in unique contexts and demonstrate their ability to provide novel, previously inaccessible information on pollen.

- To evaluate the potential of such methods, the limitations of their application, and advise how they could be taken forwards.

The focus of this work was primarily pollen, but with the thought that the ideas presented could be taken further to apply to other bioaerosols, particularly fungal spores since they remain within the range of micrometres.

1.6 Thesis overview

This thesis comprises a collection of studies that address the objectives outlined in section 1.5. Chapter 2 uses data collected from a pollen monitoring intercomparison campaign (described in Maya-Manzano, 2023) and presents a few different methods trialled for extracting estimated pollen concentrations from commercially available optical particle counter (OPC) data. It establishes the comparative success of the machine learning methods that are explored further in Chapters 3 and 4. This study is already published in the journal Science of the Total Environment (Mills et al., 2023a).

Chapter 3 explores in detail how the machine learning methods of processing OPC data can be improved upon to produce more accurate information on pollen concentrations, including more pollen types. It covers methodical hyperparameter tuning, discussing the effect of various model hyperparameters on the results, to inform those training such models in the future. Explainable AI techniques are applied to interpret the input-output feature relationships and understand how the model is learning to distinguish different pollen types. This study is also published as a sequel in the journal Science of the Total Environment (Mills et al., 2023b).

Chapter 4 presents a study where the methods developed in the previous two chapters are applied to new data collected from OPCs in a different context. It investigates what novel information can be gained from an array of low-cost sensors with high-spatial resolution through the canopy of a mature forest site at the Birmingham Institute of Forest Research (BIFoR) Free Air CO_2 Enrichment (FACE) facility.

Chapter 5 presents and evaluates a novel method for studying atmospherically relevant physical properties of pollen, using a commercially available acoustic levitator and based on visual analysis using a macroscope and basic computer vision techniques. The

visual information provided is contrasted with conventional techniques that measure mas measurements with expensive equipment. This study is published in the journal Atmospheric Measurement Techniques (Mills et al., 2023c).

The results presented here are just a small portion of a great deal of fieldwork and data that was collected at the BIFoR FACE site. Various fieldwork was conducted between the years 2020 and 2022, initially in the process of developing and testing the sensors, but also covering some fungal spore seasons at different elevations and pollen seasons across different CO_2 treatment arrays in the forest. Meanwhile, catkin surveying and sampling fieldwork was undertaken at the BIFoR FACE site to complement the automatic sensor measurements. Further fieldwork was also carried out, deploying the sensors in other sites in Birmingham and Worcester, and under deliberate, direct exposure to pollen. Collectively, this has contributed to a wealth of experience using these sensors and interpreting their data. However, due to the constraints of this PhD and in the interest of keeping the narrative streamlined, we only present the tip of the iceberg of results from fieldwork in this thesis.

At the end of each of each technical chapter is an additional comment on the study in the context of the whole thesis. In Chapter 6, we summarise what has been learned from these studies and their importance within the scientific community. Finally, we evaluate the limitations that remain and the directions that would be most beneficial for future focus.

1.6 References

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers. Aerobiologia. https://doi.org/10.1007/s10453-021-09724-5.

Alarcón, M., Periago, C., Pino, D., Mazón, J., Casas-Castillo, M.C., Ho-Zhang, J.J., Linares, C., Rodríguez-Solà, R., Belmonte, J., 2022. Potential contribution of distant sources to airborne Betula pollen levels in northeastern Iberian Peninsula, Sci. Total. Environ., 818, 151827. https://doi.org/10.1016/j.scitotenv.2021.151827.

Albertine, J.M., Manning, W.J., DaCosta, M., Stinson, K.A., Muilenberg, M.L., Rogers, C.A., 2014. Projected Carbon Dioxide to Increase Grass Pollen and Allergen Exposure Despite Higher Ozone Levels. PLoS ONE, 9(11), e111712. https://doi.org/10.1371/journal.pone.0111712.

Albrecht, B.A.:, 1989. Aerosols, Cloud Microphysics, and Fractional Cloudiness, Science, 245(4923), 1227-1230, https://doi.org/10.1126/science.245.4923.1227.

Alfano, B., Barretta, L., Del Giudice, A., De Vito, S., Di Francia, G., Esposito, E., Formisano, F., Massera, E., Miglietta, M.L., Polichetti, T., 2020. A Review of Low-Cost Particulate Matter Sensors from the Developers' Perspectives. Sensors (Basel), 20(23), 6819. https://doi.org/10.3390/s20236819.

Anderegg, W.R.L., Abatzoglou, J.T., Anderegg, L.D.L, Bielory, L., Kinney, P.L., Ziska, L., 2021. Anthropogenic climate change is worsening North American pollen seasons. PNAS, 118(7), e2013284118. https://doi.org/10.1073/pnas.2013284118.

Andreae, M.O., Jones, C.D., Cox, P.M., 2005. Strong present-day aerosol cooling implies a hot future, Nature, 435(7046), 1187-1190, https://doi.org/10.1038/nature03671.

Andrews, E., Ogren, J.A., Bonasoni, P., Marinoni, A., Cuevas, E., Rodrígues, S., Sun, J.Y., Jaffe, D.A., Fischer, E.V., Baltensperger, U., Weingarter, E., Collaud Coen, M., Sharma, S., Macdonald, A.M., Leaitch, W.R., Lin, N.-H., Laj, P., Arsov, T., Kalapov, I., Jefferson, A., Sheridan, P., 2011. Climatology of aerosol radiative properties in the free troposphere, Atmos. Res., 102(4), 365-393, https://doi.org/10.1016/j.atmosres.2011.08.017.

Ariya, P.A., Amyot, M., 2004. New Directions: The role of bioaerosols in atmospheric chemistry and physics, Atmos. Env., 38(8), 1231-1232, https://doi.org/10.1016/j.atmosenv.2003.12.006.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K.S., Christensen, M., Daniau, A.-L., Dufresne, J.-L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J.M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D.T., Myhre, G., Mülmenstädt, J., Neubauer, D., Possner, A., Rugenstain, M., Sato, Y., Schulz, M., Schwartz, S.E., Sourdeval, O., Storelvmo, T., Toll, V., Winker, D., Stevens, B., 2019. Bounding Global Aerosol Radiative Forcing of Climate Change, Reviews of Geophysics, 58(1), e2019RG000660. https://doi.org/10.1029/2019RG000660.

Boucher, O., Haywood, J., 2001. On summing the components of radiative forcing of climate change, Climate Dynamics, 18, 297-302, https://doi.org/10.1007/s003820100185.

Bradley, R.S., 2015. Paleoclimatology: Chapter 12 – Pollen, 3rd ed., Elsevier, Massachusetts, USA, https://doi.org/10.1016/B978-0-12-386913-5.00012-0.

Breiman, L., 2001. Random Forests, Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Brownlee, J., 2018. Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions. Available from https://machinelearningmastery.com.

Brownlee, J., 2021. Ensemble Learning Algorithms With Python. Available from https://machinelearningmastery.com

Burge, H.A., Rogers, C.A., 2000. Outdoor allergens. Environ. Health Perspect., 108(4), 653-659, https://doi.org/10.1289/ehp.00108s4653.

Burkart, J., Gratzl, J., Seifried, T.M., Bieber, P., Grothe, H., 2021. Isolation of subpollen particles (SPPs) of birch: SPPs are potential carriers of ice nucleating macromolecules, 18, 5751-5765. https://doi.org/10.5194/bg-18-5751-2021.

Buters, J.T.M., Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clinical and Translational Allergy, 8(1). https://doi.org/10.1186/s13601-018-0197-8.

Buters, J.T.M., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.M., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Seliger, A.K., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto, A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia. https://doi.org/10.1007/s10453-022-09750-x.

Carslaw, K.S., Boucher, O., Spracklen, D.V., Mann, G.W., Rae, J.G.L., Woodward, S., Kulmala, M., 2010. A review of natural aerosol interactions and feedbacks within the Earth system, Atmos. Chem. Phys., 10, 1701-1737. https://doi.org/10.5194/acp-10-1701-2010.

Chappuis, C., Tummon, F., Clot, B., Konzelmann, T., Calpini, B., Crouzy, B., 2020. Automatic pollen monitoring: first insights from hourly data. Aerobiologia, 36, 159-179. https://doi.org/10.1007/s10453-019-09619-6.

Choi, Y.-J., Oh, H.-R., Oh, J.-W., Kim, K. R., Kim, M.-J., Kim, B.-J., Baek, W.-G., 2018. Chamber and Field Studies demonstrate Differential Amb a 1 Contents in Common Ragweed Depending on CO_2 Levels. Allergy Asthma Immunol. Res., 10(3), 278-282. https://doi.org/10.4168/aair.2018.10.3.278.

Choi, Y.-J., Lee, K. S., Oh, J.-W., 2021. The Impact of Climate Change on Pollen Season and Allergic Sensitization to Pollens. Immunology and Allergy Clinics of North America, 41(1), 97-109. https://doi.org/10.1016/j.iac.2020.09.004.

Christianini, N., 2023. The Shortcut: Why Intelligent Machines Do Not Think Like Us. CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781003335818.

Christner, B.C., Cai, R., Morris, C.E., Sands, D.C., 2008. Geographic, seasonal, and precipitation chemistry influence on the abundance and activity of biological ice nucleators in rain and snow, PNAS, 105(48), 18854-18859. https://doi.org/10.1073/pnas.0809816105.

Contini, D., Lin, Y.-H., Hänninen, O., Viana, M., 2021. Contribution of Aerosol Sources to Health Impacts, Atmosphere, 12, 730. https://doi.org/10.3390/atmos12060730.

Crilley, L.R., Shaw, M.D., Pound, R.J., Kramer, L.J., Price, R., Young, S., Lewis, A., Pope, F.D., 2017. Evaluation of low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. Atmos. Meas. Tech. Discuss.. https://doi.org/10.5194/amt-2017-308.

Crilley, L.R., Singh A., Kramer, L.J., Shaw, M.D., Alam, M.S., Apte, J.S., Bloss, W.J., Ruiz, L.H., Fu, P., Fu, W., Gani, S., Gatari, M., Ilyinskaya, E., Lewis, A.C., Ng'ang'a, D., Sun, Y., Whitty, R.C.W., Yue, S., Young, S., Pope, F.D., 2020. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos. Meas. Tech., 13(3), 1181-1193. https://doi.org/10.5194/amt-13-1181-2020.

Crouzy, B., Stella, M., Konzelmann, T., Calpini, B., Clot, B., 2016. All-optical automatic pollen identification: Towards an operational system. Atmos. Environ., 140, 202-212. https://doi.org/10.1016/j.atmosenv.2016.05.062.

Després, V.R., Huffman, J.A., Burrows, S.M., Hoose, C., Safatov, A.S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M.O., Pöschl, U., Jaenicke, R.: Primary biological aerosol particles in the atmosphere: a review, Tellus B: Chemical and Physical Meteorology, 64(1), 15598. https://doi.org/10.3402/tellusb.v64i0.15598.

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wüest, R.O., Hartig, F., 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs, 88(4), 485-504. https://doi.org/10.1002/ecm.1309.

Douwes, J., Thorne, P., Pearce, N., Heederik, D., 2003. Bioaerosol Health Effects and Exposure Assessment: Progress and Prospects, The Annals of Occupational Hygiene, 47(3), 187-200. https://doi.org/10.1093/annhyg/meg032.

Ervens, B., Turpin, B.J., Weber, R.J., 2011. Secondary organic aerosol formation in cloud droplets and aqueous particles (aqSOA): a review of laboratory, field and model studies, Atmos. Chem. Phys., 11, 11069-11102. https://doi.org/10.5194/acp-11-11069-2011.

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions, Atmos. Res., 182, 346-376. https://doi.org/10.1016/j.atmosres.2016.07.018.

Ganaie, M.A., Hu., M., Malik, A.K., Tanveer, M., Suganthan, P.N., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151. https://doi.org/10.1016/j.engappai.2022.105151.

Ghosh, B., Lal, H., Srivastava, A., 2015. Review of bioaerosols in indoor environment with special reference to sampling, analysis and control mechanisms, Environ. Int., 85, 254-272. https://doi.org/10.1016/j.envint.2015.09.018.

Giordano, M.R., Malings, C., Pandis, S.N., Presto, A.A., McNeill, V.F., Westervelt, D.M., Beekmann, M., Subramanian, R., 2021. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. J. Aerosol Sci., 158, 105833. https://doi.org/10.1016/j.aerosci.2021.105833.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA. ISBN: 9780262035613.

Grewling, Ł., Bogawski, P., Kryza, M., Magyar, D., Šikoparija, B., Skjøth, C.A., Udvardy, O., Werner, M., Smith, M, 2019. Concomitant occurrence of anthropogenic air pollutants, mineral dust and fungal spores during long-distance transport of ragweed pollen. Environmental Pollution, 254, 112948. https://doi.org/10.1016/j.envpol.2019.07.116.

Griffiths, Ps.T., Borlace, J.-S., Gallimore, P.J., Kalberer M., Herzog, M., Pope, F.D., 2012. Hygroscopic growth and cloud activation of pollen: a laboratory and modelling study. Atmos. Sci. Lett., 13(4), 289-295. https://doi.org/10.1002/asl.397.

Griffiths, W.D., DeCosemo, G.A.L., 1994. The Assessment of Bioaerosols: A Critical Review, J. Aerosol Sci., 25(8), 1425-1458. https://doi.org/10.1016/0021-8502(94)90218-6.

Groma, V., Alföldy, B., Börcsök, E., Czömpöly, O., Füri, P., Kéri, A.H., Kovács, G., Török, S., Osán, J., 2022. Sources and health effects of fine and ultrafine aerosol particles in an urban environment, Atmospheric Pollution Research, 13(2), 101302. https://doi.org/10.1016/j.apr.2021.101302.

Hampton, V., Deepak, A., 1992. The International Global Aerosol Program (IGAP) plan. ISBN: 0937194298.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed.. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7.

Hendrickson, B.N., Alsante, A.N., Brooks, S.D., 2023. Live oak pollen as a source of atmospheric particles. Aerobiologia, 39, 51-67. https://doi.org/10.1007/s10453-022-09773-4.

Hirst, J.M., 1952. An Automatic Volumetric Spore Trap, Annals of Applied Biology, 39(2), 257-265. https://doi.org/10.1111/j.1744-7348.1952.tb00904.x.

Hoose, C., Kristjánsson, J.E., Burows, S.M., 2010. How important is biological ice nucleation in clouds on a global scale? Environ. Res. Lett., 5, 024009. https://doi.org/10.1088/1748-9326/5/2/024009.

Huffman, J.A., Perring, A.E., Savage, N.J., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M.A., Crawford, I., Gallagher, M., Topping, D., Doughty, D.C., Hill, S.C., Pan, Y., 2019. Real-time sensing of bioaerosols: Review and current perspectives, Aerosol Science and Technology, 54(5), 465-495. https://doi.org/10.1080/02786826.2019.1664724.

Hughes, D.D., Mampage, C.B.A., Jones, L.M., Liu, Z., Stone, E.A., 2020. Characterisation of Atmospheric Pollen Fragments during Springtime Thunderstorms. Environ. Sci. Technol. Lett., 7, 6, 409-414. https://doi.org/10.1021/acs.estlett.0c00213.

Humbal, C., Gautam, S., Trivedi, U., 2018. A review on recent progress in observations, and health effects of bioaerosols, Environ. Int., 118, 189-193. https://doi.org/10.1016/j.envint.2018.05.053.

Ito, K., Weinberger, K.R., Robinson, G.S., Sheffield, P.E., Lall, R., Mathes, R., Ross, Z., Kinney, P.L., Matte, T.D., 2015. The associations between daily spring pollen counts, over-the-counter allergy medication sales, and asthma syndrome emergency department visits in New York City, 2002-2012. Environ. Health., 14, 71. https://doi.org/10.1186/s12940-015-0057-0.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An Introduction to Statistical Learning with Applications in R, 2nd Ed., Springer, New York. https://doi.org/10.1007/978-1-0716-1418-1.

Jiang, C., Wang, W., Du, L., Huang, G., McConaghy, C., Fineman, S., Liu, Y., 2022. Field Evaluation of an Automated Pollen Sensor, Int. J. Environ. Res. Public Health, 19(11), 6444. https://doi.org/10.3390/ijerph19116444.

Jones, A.M., Harrison, R.M., 2004. The effects of meteorological factors on atmospheric bioaerosol concentrations – a review, Sci. Total Environ., 326(1-3), 151-180. https://doi.org/10.1016/j.scitotenv.2003.11.021.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260. https://doi.org/10.1126.science.aaa8415.

Juprasong, Y., Sirirakphaisarn, S., Siriwattanakul, U., Songnuan, W., 2022. Exploring the effects of seasons, diurnal cycle, and heights on airborne pollen load in a Southeast Asian atmospheric condition. Front. Public Health, 10, 1067034. https://doi.org/10.3389/fpubh.2022.1067034.

Kim, K. R., Oh, J.-W., Woo., S.-Y., Seo, Y. A., Choi, Y.-J., Kim, H. S., Lee, W. Y., Kim, B.-J., 2018. Does the increase in ambient CO_2 concentration elevate allergy risks posed by oak pollen? Int. J. Biometeorol., 62(9), 1587-1594. https://doi.org/10.1007/s00484-018-1558-7.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modelling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3.

Lake, I.R., Jones, N.R., Agnew, M., Goodness, C.M., Giorgi, F., Hamaoui-Laguel, L., Semenov, M.A., Solomon, F., Storkey, J., Vautard, R., Epstein, M.M., 2017. Climate Change and Future Pollen Allergy in Europe. Environ. Health Perspect., 125(3), 385-391. https://doi.org/10.1289/EHP173.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539.

Leru, P.M., Anton, V.F., Eftimie, A.M., Stefanut, S., 2022. Biologic Pollution Due to *Ambrosia* (Ragweed) Pollen in Urban Environment of Bucharest. Int. J. Environ. Res. Public Health, 19(17), 10613. https://doi.org/10.3390/ijerph191710613.

Li, Q., Wu, B., Liu, J., Zhang, H., Cai, X., Song, Y., 2020. Characteristics of the atmospheric boundary layer and its relation to $PM_{2.5}$ during haze episodes in winter in the North China Plain. Atmos. Environ. 223, 117265. https://doi.org/10.1016/j.atmosenv.2020.117265.

Li, W., Liu, L., Xu, L., Zhang, J., Yuan, Q., Ding, X., Hu, W., Fu, P., Zhang, D., 2020. Overview of primary biological aerosol particles from a Chinese boreal forest: Insight into morphology, size, and mixing state at microscopic scale. Sci. Total Environ., 719, 137520. https://doi.org/10.1016/j.scitotenv.2020.137520.

Lim, D.V., Simpson, J.M., Kearns, E.A., Kramer, M.F., 2005. Current and Developing Technologies for Monitoring Agents of Bioterrorism and Biowarfare, Clin. Microbiol. Rev., 18(4), 583-607. https://doi.org/10.1128/CMR.18.4.583-607.2005.

Lindemann, J., Constantinidou, H.A., Barchet, W.R., Upper, C.D., 1982. Plants as sources of airborne bacteria, including ice nucleation-active bacteria, Appl. Environ. Microbiol., 44(5), 1059-1063. https://doi.org/10.1128/aem.44.5.1059-1063.1982.

Matthews, B.H., Alsante, A.N., Brooks, S.D., 2023. Pollen Emissions of Subpollen Particles and Ice Nucleating Particles, ACS Earth Space Chem. https://doi.org/10.1021/acsearthspacechem.3c00014.

Maya-Manzano, J.M., Smith, M., Markey, E., Clancy, J.H., Sodeau, J., O'Connor, D.J., 2020. Recent developments in monitoring and modelling airborne pollen, a review. Grana, 60(1). https://doi.org/10.1080/00173134.2020.1769176.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, L., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavulj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2023. Towards European automatic bioaerosol monitoring: Comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ., 866, 161220. https://doi.org/10.1016/j.scitotenv.2022.161220.

McDuffie, E., Martin, R., Yin, H., Brauer, M., 2021. Global Burden of Disease from Major Air Pollution Sources (GBD MAPS): A Global Approach. Res. Rep. Health Eff. Inst., 2021, 210. PMC9501767.

McMurry, P.H., 2000. A review of atmospheric aerosol measurements, Atmos. Environ., 34, 19-59-1999. https://doi.org/10.1016/S1352-2310(99)00455-0.

Miguel, A.G., Taylor, P.E., House, J., Glovsky, M.M., Flagan, R.C., 2006. Meteorological influences on respirable fragment release from Chinese elm pollen, Aerosol Sci. Technol., 40(9), 690-696. https://doi.org/10.1080/02786820600798869.

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023a. Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests. Sci. Total Environ., 871, 161969. https://doi.org/10.1016/j.scitotenv.2023.161969.

Mills, S.A., Maya-Manzano, J.M, Tummon, F., MacKenzie, A.R., Pope, F.D., 2023b. Machine Learning Methods for Low-Cost Pollen Monitoring – Model Optimisation and Interpretability. Sci. Total Environ., 903, 165853. https://doi.org/10.1016/j.scitotenv.2023.165853.

Mills, S.A., Milsom, A., Pfrang, C., MacKenzie, A.R., Pope, F.D., 2023c. Acoustic levitation of pollen and visualisation of hygroscopic behaviour. Atmos. Meas. Tech., 16, 4885-4898, https://doi.org/10.5194/amt-16-4885-2023.

Molina, R.T., Rodríguez, A.M., Palaciso, I.S., López, F.G., 1996. Pollen production in anemophilous trees. Grana, 35(1), 38-46. https://doi.org/10.1080/00173139609430499.

Mushtaq, Z., Sharma, M., Bangotra, P., Gautam, A.S., Gautam, S., 2022. Atmospheric Aerosols: Some Highlights and Highlighters, Past to Recent Years, Aerosol Science and Engineering, 6, 135-145. https://doi.org/10.1007/s41810-022-00133-w.

Myhre, G., Myhre, C.E.L., Samset, B.H., & Storelvmo, T., 2013. Aerosols and their Relation to Global Climate and Climate Sensitivity, Nature Education Knowledge, 4(5), 7. https://www.nature.com/scitable/knowledge/library/aerosols-and-their-relation-to-global-climate-102215345/.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., Zhang, H., 2013. Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the First Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Nasir, Z.A., Hayes, E., Williams, B., Gladding, T., Rolph, C., Khera, S., Jackson, S., Bennett, A., Collins, S., Parks, S., Attwood, A., Kinnersley, R.P., Walsh, K., Alcega, S.G., Pollard, S.J.T., Drew, G., Coulon, F., Tyrrel, S., 2018. Scoping studies to establish the capacity and utility of a real-time bioaerosol sensor to characterise emissions from environmental sources. Sci. Total Environ., 648, 25-32. https://doi.org/10.1016/j.scitotenv.2018.08.120.

O'Connor, D.J., Healy, D.A., Hellebust, S., Buters, J.T.M., Sodeau, J.R., 2013. Using the WIBS-4 (Waveband Integrated Bioaerosol Sensor) Technique for the On-Line Detection of Pollen Grains. Aerosol Sci. Technol., 48(4), 341-349. https://doi.org/10.1080/02786826.2013.872768.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and Online Pollen Monitoring, Int. Arch. Allergy Immunol., 167(3), 158-166. https://doi.org/10.1159/000436968.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition, Environ. Res., 191, 110031. https://doi.org/10.1016/j.envres.2020.110031.

Pazmandi, K., Kumar, B.V., Szabo, K., Boldogh, I., Szoor, A., Vereb, G., Veres, A., Lanyi, A., Rajnavolgyi, E., Bacsi, A., 2012. Ragweed Subpollen Particles of Respirable Size Activate Human Dendritic Cells. PLoS ONE, 7(12), e52085. https://doi.org/10.1371/journal.pone.0052085.

Peccia, J., Hospodsky, D., Bibby, K., 2011. New Directions: A revolution in DNA sequencing now allows for the meaningful integration of biology with aerosol science, Atmos. Environ., 45(10), 1896-1897. https://doi.org/10.1016/j.atmosenv.2010.11.037.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python, J. Machine Learning Res., 12(85), 2825-2830. Accessible via https://scikit-learn.org

Pichler, M., Hartig, F., 2023. Machine learning and deep learning – A review for ecologists. Methods in Ecology and Evolution, 14(4), 994-1016. https://doi.org/10.1111s/2041-210X.14061.

Plaza, M.P., Kolek, F., Leier-Wirtz, V., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2022. Detecting Airborne Pollen Using an Automatic, Real-Time Monitoring System: Evidence from Two Sites. Int. J. Environ. Res. Public Health, 19(4), 2471. https://doi.org/10.3390/ijerph19042471.

Pope, F.D., 2010. Pollen grains are efficient cloud condensation nuclei. Environ. Res. Lett., 5, 044015. https://doi.org/10.1088/1748-9326/5/4/044015.

Prenni, A.J., Petters, M.D., Kreidenweis, S.M., 2009. Relative roles of biogenic emissions and Saharan dust as ice nuclei in the Amazon Basin, Nature Geoscience, 2, 402-405. https://doi.org/10.1038/ngeo517.

Reponen, T., 2011. Encyclopedia of Environmental Health: Methodologies for Assessing Bioaerosol Exposures, Elsevier, Cincinnati, USA. https://doi.org/10.1016/B978-0-12-409548-9.11822-6.

Ríos, B.E., Torres-Jardón, R., Ramírez-Arriaga, E., Martínez-Bernal, A., Rosas, I., 2016. Diurnal variations of airborne pollen concentration and the effect of ambient temperature in three sites of Mexico city. Int. J. Biomet., 60(5), 771-787. https://doi.org/10.1007/s00484-015-1061-3.

Rodrigues, A., Sardinha, R.A., Pita, G., 2021. Fundamental Principles of Environmental Physics: General Characteristics of the Atmospheric Boundary Layer. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-69025-0.

Rokach, L., 2009. Pattern Classification Using Ensemble Methods, Series in Machine Learning Perception and Artificial Intelligence: Volume 75. https://doi.org/10.1142/7238.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev., 65(6), 386-408. https://doi.org/10.1037/h0042519.

Ruske, S., Topping, D.O., Foot, V.E., Morse, A.P., Gallagher, M.W., 2018. Machine learning for improved data analysis of biological aerosol using the WIBS. Atmos. Meas. Tech., 11(11), 6203-6230. https://doi.org/10.5194/amt-11-6203-2018.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech., 12, 3435-3452. https://doi.org/10.5194/amt-12-3435-2019.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2019. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech., 427. https://doi.org/10.5194/amt-2019-427.

Savage, N.J., Krentz, C.E., Könemann, T., Han, T.T., Mainelis, G., Pöhlker, C., Huffman, J.A., 2017. Systematic characterization and fluorescence threshold strategies for the wideband integrated bioaerosol sensor (WIBS) using size-resolved biological and interfering particles. Atmos. Meas. Tech., 10(11), 4279-4302. https://doi.org/10.5194/amt-10-4279-2017.

Schaefer, J., Milling, M., Schuller, B.W., Bauer, B., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2021. Towards automatic airborne pollen monitoring: From commercial devices to operational by mitigating class-imbalance in a deep learning approach. Sci. Total Environ., 796, 148932. https://doi.org/10.1016/j.scitotenv.2021.148932.

Sedghy, F., Varasteh, A.-R., Sankian, M., Moghadam, M., 2018. Interaction Between Air Pollutants and Pollen Grains: The Role on the Rising Trend in Allergy. Rep. Biochem. Mol. Biol., 6(2), 219-224. PMC5941124, PMID: 29766006.

Shaffer, R.M., Sellers, S.P., Baker, M.G., de Buen Kalman, R., Frostad, J., Suter, M.K., Anenberg, S.C., Balbus, J., Basu, N., Bellinger, D.C., Birnbaum, L., Brauer, M., Cohen, A., Ebi, K.L., Fuller, R., Grandjean, P., Hess, J.J., Kogevinas, M., Kumar, P., Landrigan, P.J., Lanphear, B., London, S.J., Rooney, A.A., Stanaway, J.D., Trasande, L., Walker, K., Hu, H., 2019. Improving and Expanding Estimates of the Global Burden of Disease Due to Environmental Health Risk Factors. Environ. Health Perspect., 127, 10, 105001. https://doi.org/10.1289/EHP5496.

Shammi, M., Rahman, M.M., Tareq, S.M., 2021. Distribution of Bioaerosols in Association With Particulate Matter: A Review on Emerging Public Health Threat in Asian Megacities, Front. Environ. Sci., 9. https://doi.org/10.3389/fenvs.2021.698215.

Sheffield, P.E., Weinberger, K.R., Ito, K., Matte, T.D., Mathes, R.W., Robinson, G.S., Kinney, P.L., 2011. The Association of Tree Pollen Concentration Peaks and Allergy Medication Sales in New York City: 2003-2008. ISRN Allergy, 2011, 537194. https://doi.org/10.5402/2011/537194.

Shiraiwa, M., Ueda, K., Pozzer, A., Lammel, G., Kampf, C.J., Fushimi, A., Enami, S., Arangio, A.M., Fröhlich-Nowoisky, J., Fujitani, Y., Furuyama, A., Lakey, P.S.J., Lelieveld, J., Lucas, K., Morino, Y., Pöschl, U., Takahama, S., Takami, A., Tong, H., Weber, B., Yoshino, A., Sato, K., 2017. Aerosol Health Effects from Molecular to Global Scales, Environ. Sci. Technol., 51(23), 13545-13567. https://doi.org/10.1021/acs.est.7b04417.

Skjøth, C.A., Sommer, J., Stach, A., Smith, M., Brandt, J., 2007. The long-range transport of birch (*Betula)* pollen from Poland and Germany causes significant pre-season concentrations in Denmark, Clin. Expt. Allergy, 37(8), 1204-1212. https://doi.org/10.1111/j.1365-2222.2007.02771.x.

Smith, M., Matavulj, P., Mimić, G., Panić, M., Grewling, Ł., Šikoparija, B., 2022. Why should we care about high temporal resolution monitoring of bioaerosols in ambient air? Sci. Total Environ., 826, 154231. https://doi.org/10.1016/j.scitotenv.2022.154231.

Sofiev, M., Sofieva, S., Palamarchuk, J., Šaulienė, I., Kadantsev, E., Atanasova, N., Fatahi, Y., Kouznetsov, R., Kuula, J., Noreikaite, A., Peltonen, M., Pihlajamäki, T., Saarto, A., Svirskaite,

J., Toiviainen, L., Tyuryakov, S., Šukienė, L., Asmi, E., Bamford, D., Hyvärinen, A.-P., Karppinen, A., 2022. Bioaerosols in the atmosphere at two sites in Northern Europe in spring 2021: Outline of an experimental campaign. Environ. Res., 214, 113798. https://doi.org/10.1016/j.envres.2022.113798.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Machine Learning Res., 15(56), 1929-1958. https://dl.acm.org/doi/10.5555/2627435.2670313.

Stanley, R.G., Linskins, H.F., 1974. Pollen: biology, chemistry and management, Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-642-65905-8.

Stone, E.A., Mampage, C.B.A., Hughes, D.D., Jones, L.M., 2021. Airborne sub-pollen particles from rupturing giant ragweed pollen, Aerobiologia, 37, 625-632. https://doi.org/10.1007/s10453-021-09702-x.

Taylor, P.E., Flagan, R.C., Miguel, A.G., Valenta, R., Glovsky, M.M., 2004. Birch pollen rupture and the release of aerosols of respirable allergens, Clin. Exp. Allergy, 34(10), 1591-1596. https://doi.org/10.1111/j.1365-2222-2004.02078.x.

Tryner, J., Mehaffy, J., Miller-Lionberg, D., Volckens, J., 2020. Effects of aerosol type and simulated aging on performance of low-cost PM sensors, J. Aerosol Sci., 150, 105654. https://doi.org/10.1016/j.jaerosci.2020.105654.

Tummon, F., Arboledas, L.A., Bonini, M., Guinot, B., Hicke, M., Jacob, C., Kendrovski, V., McCairns, W., Petermann E., Peuch, V.-H., Pfaar, O., Sicard, M., Sikoparija, B., Clot, B., 2021. The need for Pan-European automatic pollen and fungal spore monitoring: A stakeholder workshop position paper. Clin. Trans. Allergy, 11(3), e12015. https://doi.org/10.1002/clt2.12015.

Tummon, F., Bruffaerts, N., Celenk, S., Choël, M., Clot, B., Crouzy, B., Galán, C., Gilge, S., Hajkova, L., Mokin, V., O'Connor, D., Rodinkova, V., Sauliene, I., Sikoparija, B., Sofiev, M., Sozinova, O., Tesendic, D., Vasilatou, K., 2022. Towards standardisation of automatic pollen and fungal spore monitoring: best practices and guidelines. Aerobiologia. https://doi.org/10.1007/s10453-022-09755-6.

Twomy, S., 1977. The Influence of Pollution on the Shortwave Albedo of Clouds, Journal of the Atmospheric Sciences, 34(7), 1149-1152. https://doi.org/10.1175/1520-0469.

Verhoeff, A.P., Burge, H.A., 1997. Health risk assessment of fungi in home environments, Ann. Allergy Asthma Immunol., 78(6), 544-554. https://doi.org/10.1016/S1081-1206(10)63214-0.

Wang, P., 2008. What Do You Mean by "AI"? Frontiers in Artificial Intelligence and Applications, 171(1), 362-373.

Wayne, P., Foster, S., Connolly, J., Bazzaz, F., Epstein, P., 2002. Production of allergenic pollen by ragweed (Ambrosia artimisiifolia L.) is increased in CO_2-enriched atmospheres. Ann. Allergy Asthma Immunol., 88(3), 279-282. https://doi.org/10.1016/S1081-1206(10)62009-1.

Weger, L.A., Molster, F., Raat, K., Haan, J., Romein, J., Leeuwen, W., Groot, H., Mostert, M., Hiemstra, P.S., 2020. A new portable sampler to monitor pollen at street level in the

environment of patients, Sci. Total Environ., 741, 140404. https://doi.org/10.1016/j.scitotenv.2020.140404.

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications, Springer, New York. https://doi.org/10.1007/978-1-4419-9326-7.

Zhang, Y., Steiner, A.L., 2022. Projected climate-driven changes in pollen emission season length and magnitude over the continental United States. Nature Communications, 13, 1234. https://doi.org/10.1038/s41467-022-28764-0.

Zhao, F., Elkelish, A., Durner, J., Lindermayr, C., Winkler, J.B., Ruëff, F., Behrendt, H., Traidl-Hoffmann, C., Holzinger, A., Kofler, W., Braun, P., von Toerne, C., Hauck, S.M., Ernst, D., Frank, U., 2015. Common ragweed (*Ambrosia artimisiifolia* L.): allergenicity and molecular characterization of pollen after plant exposure to elevated NO_2. Plant, Cell & Environment, 39(1), 147-164. https://doi.org/10.1111/pce.12601.

Ziska, L.H., Caulfield, F.A., 2000. Rising CO_2 and pollen production of common ragweed (*Ambrosia artemisiifolia*), a known allergy-inducing species: implications for public health. Aust. J. Plant Physiol., 27(10), 893-898. https://doi.org/10.1071/PP00032.

Ziska, L.H, 2021. Climate, Carbon Dioxide, and Plant-Based Aero-Allergens: A Deeper Botanical Perspective. Front. Allergy, 2. https://doi.org/10.3389/falgy.2021.714724.

2. CHAPTER 2: Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests

This study is published in the journal *Science of the Total Environment*:

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023. Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests. Sci. Total Environ., 871, 161969. https://doi.org/10.1016/j.scitotenv.2023.161696.

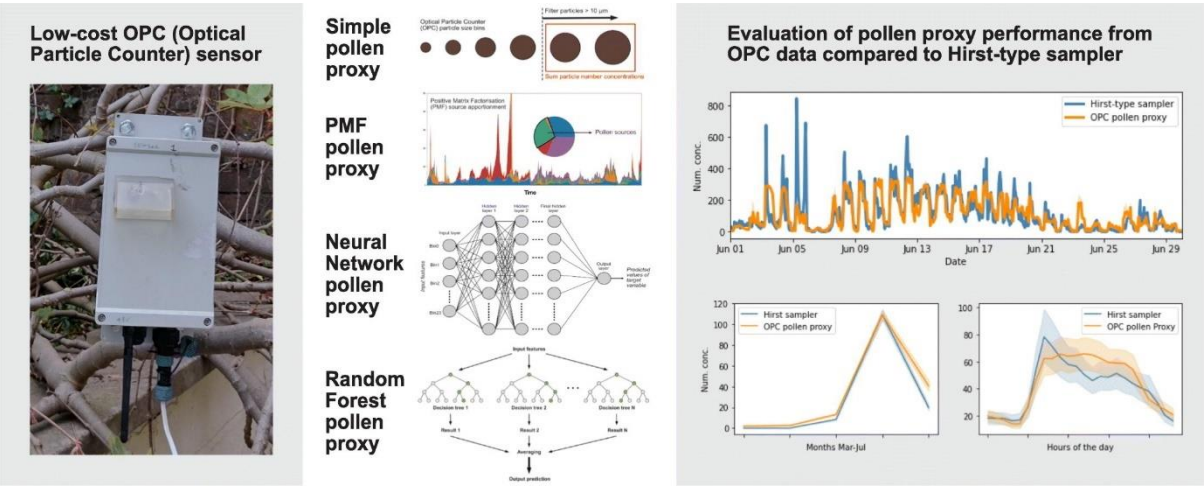


Figure 2.0: *Graphical abstract for Chapter 2.*

2.1 Abstract

Pollen allergies affect a significant proportion of the global population, and this is expected to worsen in years to come. There is demand for the development of automated pollen monitoring systems. Low-cost Optical Particle Counters (OPCs) measure particulate matter and have attractive advantages of real-time high time resolution data and affordable costs. This study asks whether low-cost OPC sensors can be used for meaningful monitoring of airborne pollen.

We employ a variety of methods, including supervised machine learning techniques, to construct pollen proxies from hourly-averaged OPC data and evaluate their performance, holding out 40 % of observations to test the proxies. The most successful methods are supervised machine learning Neural Network (NN) and Random Forest (RF) methods, trained from pollen concentrations collected from a Hirst-type sampler. These perform significantly better than using a simple particle size-filtered proxy or a Positive Matrix Factorisation (PMF) source apportionment pollen proxy. Twelve NN and RF models were developed to construct a pollen proxy, each varying by model type, input features and target variable. The results show that such models can construct useful information on pollen from OPC data. The best metrics achieved (Spearman correlation coefficient = 0.85, coefficient of determination = 0.67) were for the NN model constructing a Poaceae (grass) pollen proxy, based on particle size information, temperature, and relative humidity. Ability to distinguish high pollen events was evaluated using $F1$ Scores, a score reflecting the fraction of true positives with respect to false positives and false negatives, with promising results ($F1 \leq 0.83$). Model-constructed proxies demonstrated the ability to follow monthly and diurnal pollen trends. We discuss the suitability of OPCs for monitoring pollen and offer advice

for future progress. We demonstrate an attractive alternative for automated pollen monitoring that could provide valuable and timely information to the benefit of pollen allergy sufferers.

2.2 Introduction

Pollen constitutes a significant proportion of atmospheric bioaerosols (primary biological aerosol particles, PBAP) and affects as much as 40 % of the population in industrialised countries with pollen allergies (Fröhlich-Nowoisky et al., 2016) causing health issues such as allergic rhinitis and asthma. Pollen is released from plants as part of their reproductive cycle, carrying the male gamete, and thus are vital for the survival and spread of terrestrial ecosystems. Anemophilous species - including almost all cone-bearing plants (gymnosperms) and many flower-bearing plants (angiosperms) - depend on wind to transport their pollen grains through the atmosphere where the grains can travel long distances at significant concentrations during pollen seasons (Skjøth et al., 2007, Alarcón et al., 2022, Jochner et al., 2015, Grewling et al., 2016, Siljamo et al., 2008, Manninen et al., 2014).

Individual pollen grains can vary in size between 10 and 100 μm (see, e.g., Després et al., 2012, Reponen, 2011, pp. 723, Bradley, 2015, pp. 408–409), though many anemophilous species fall towards the lower end of this range, often below 40 μm. On rupturing in the atmosphere, pollen grains can release smaller subpollen particles which can also contain allergenic proteins relevant to human health (Bacsi et al., 2006), and reside in a more easily respirable size fraction compared to the intact pollen grains (Taylor et al., 2004). Subpollen particle size ranges have been reported between 20 nm and 6.5 μm for giant ragweed (Stone et al., 2021) and <0.25–3.2 μm (aerodynamic diameter) for birch (Burkart et al., 2021). Other work reports subpollen particles containing specific pollen allergens of a similar size range of a few to several micrometres (Miguel et al., 2006). The category of subpollen particles is still loosely defined in the literature and can encompass a wide range of particles, including starch granules and other cytoplasmic

debris. As with pollen grains, subpollen particles can be expected to vary in size range across different taxa and species of pollen. However, the wide range of particles included within this term likely makes the distinction between species somewhat difficult.

Pollen grains and subpollen particles have been considered to have potential cloud condensation nucleation (CCN) activity (Pope, 2010; Griffiths et al., 2012; Steiner et al., 2015; Mikhailov et al., 2019), as well as ice nucleation (IN) properties (Diehl et al., 2001, Diehl et al., 2002, Pummer et al., 2012, Tong et al., 2015, Dreischmeier et al., 2017, Gute and Abbatt, 2020, Burkart et al., 2021) which can be affected by atmospheric processing (Gute et al., 2020). This means that these particles could affect the process of water condensation or ice crystal formation in clouds, and thus precipitation and ultimately climate.

Whether in the interest of public health, ecosystem regeneration and recolonisation, or climate change, there is consensus that further focused studies of pollen, and other bioaerosols, are required (Huffman et al., 2019). Yet there are significant limitations to current methods available for studying airborne pollen dynamics. Conventional methods for measuring pollen generally use volumetric Hirst-type samplers which are time consuming and labour intensive, requiring trained staff to identify and count pollen concentrations from collected slides. Data is therefore often only available in daily, or for limited periods up to hourly, time resolution, is not available in real-time, and suffers from relatively large uncertainties, easily up to 30 % (Adamov et al., 2021). An overview of the current state of pollen monitoring stations globally can be found by Buters et al. (2018), which highlights the need for technological advancement.

Attention is now turning towards the use of automated monitoring systems, with an aim to standardise such methods to replace conventional manual techniques (Tummon et al.,

2022). Recent automated pollen monitoring instruments have been developed such as PollenSense's PS-300 (Jiang et al., 2022), Hund-Wetzlar's BA500 pollen monitor (Oteros et al., 2015, Oteros et al., 2020; Plaza et al., 2022), the Rapid-E (Šaulienė et al., 2019; Smith et al., 2022), and the Swisens Poleno (Sauvageat et al., 2020, Sofiev et al., 2022). The first two use an impactor to capture grains on a slide, similarly to conventional manual traps, but the slide imaging and pollen grain classification are automated. The Rapid-E is a cytometer which identifies particles using classification algorithms based on laser scattering signal, fluorescence spectrum and fluorescence lifetime data. Meanwhile the Poleno presents a recent advancement whereby pollen grains are identified in an automated system utilising light scattering, fluorescence, and holographic imaging techniques, trained by machine learning. Fluorescence spectroscopy methods have been utilised to characterise pollen using the Wideband Integrated Bioaerosol Sensor, WIBS (e.g. O'Connor et al., 2014; Savage et al., 2017; Ruske et al., 2018), the WIBS has also been used to identify pollen fragments (Hughes et al., 2020). These and new technologies are in continuous development to meet the present need for more advanced pollen monitoring systems. A comprehensive review, including the use of various machine learning models for pollen modelling, can be found in Maya-Manzano et al. (2020) and in Buters et al. (2022).

These methods do however bring limitations of time resolution in some cases and heavy cost limitations in general. With each instrument costing thousands or tens of thousands of U.S. dollars, such methods become unavailable for the global majority and smaller institutions, let alone for personal use. It would also be infeasible to utilise such sensors in monitoring networks that could offer high spatial resolution data. While high performance sensors with high accuracy have their value, there is also a need for more

affordable and portable options, that could be used in networks to provide high spatiotemporal resolution information throughout dynamic pollen seasons.

Low-cost optical particle counter (OPC) sensors have already been explored extensively to monitor particulate matter (PM) – i.e., all generic, internally and externally mixed, solid and liquid particles suspended in the atmosphere - primarily in the context of air quality and anthropogenic pollution (for example: Crilley et al., 2020; Giordano et al., 2021; Dubey et al., 2022; Narayana et al., 2022). These OPC sensors take in a continuous flow of particles from ambient air, direct laser light at the stream of particles and from the scattering pattern of the light intercepted by the particles determines, via Mie theory, counts for particles in different size range bins. Pollen, being an aerosol, should also be detectable by such methods, though it is generally larger in size than most monitored anthropogenic particulate matter (i.e. larger than PM10, which is PM with an aerodynamic diameter of 10 μm or less). There has been very limited previous research and reported success in using a laser optics monitoring system to measure pollen concentrations (Kawashima et al., 2017).

The Alphasense OPC-N3 (Alphasense, Braintree, UK) has capability to measure particles up to 40 μm in size (i.e., optically measured diameter) thereby covering a large proportion of anemophilous pollen species. Costing a few hundred U.S. dollars, it is significantly more affordable than other current pollen monitoring options. Additionally, it comes with advantages of high time resolution (down to 10 s) and flexibility on deployment and functionality. It can also be functionalised to give real-time and remotely accessible data. The challenge, however, is isolating a meaningful pollen signal from the sized particle count information provided by the OPC sensors and this is what

this study addresses. Previously, our group has used the OPC-N2, the precursor to the OPC-N3 to monitor for airborne fungal spores in a forest (Baird et al., 2022).

Our aim is to assess whether the data available from low-cost OPC sensors such as these have potential to be used for monitoring pollen, to an extent that provides useful information for public health. We have taken commercially available Alphasense OPC-N3 sensors, functionalised them to log independently outdoors for extended periods, and to deliver their sized particle data online via a mobile phone network in real-time. Through the EUMETNET Autopollen ADOPT - COST Action Intercomparison Campaign in 2021 (Maya-Manzano et al., 2022), we acquired data from our sensors in parallel with baseline pollen data from Hirst-type volumetric samplers. Using these data, we investigated potential methods to determine pollen proxies from OPC data and evaluated for each method its accuracy and potential value for monitoring airborne pollen concentrations.

2.3 Materials and methods

2.3.1 Instrumentation

Three commercially manufactured Alphasense optical particle counter (OPC-N3) devices were used in this study. These are small (mass < 105 g), and cost in the range of a few hundred US dollars. Their predecessor, the OPC-N2, has been described previously by Sousan et al. (2016) and Crilley et al. (2018) with the main difference being that the OPC-N3 units have an extended measurable size range up to larger particle sizes. While the OPC-N2s measure within a size range of 0.38–17 μm (divided into 16 software bins), the OPC-N3s output raw particle counts segregated into 24 bins of different particle size ranges between 0.35 and 40 μm (see Table 7.1 in the Appendices for individual bin size ranges). This means that, unlike its predecessor, its detection size range overlaps with the typical size range of many anemophilous pollen taxa, including oak (*Quercus*), birch (*Betula*), grass (*Poaceae*), nettle (*Urtica*). Though some taxa are not in this range, such as pine (*Pinus*) pollen, which are typically over 40 μm in diameter (Song et al., 2012).

The OPC device actively draws in a stream of air, at around 5 L min−1, which passes through a 658 nm-wavelength laser beam that is scattered by the incoming particles. Measurements of scattered light intensity are calculated based on Mie scattering theory and used to determine number counts for particles that fall within each bin size range. A refractive index of 1.5 and particle density of 1.65 g mL−1 is assumed, and the devices have a maximum particle count of up to 10,000 particles s−1 (equivalent to number concentrations of 1.2 × 108 particles m−3 at a 5 L min^{-1} flow rate).

A custom-built system was developed with an Arduino MKR GSM1400 microprocessor (Arduino S.r.l., Via Andrea Appiani 25, Monza, 20,900, Italy) to facilitate the independent logging of the OPC-N3 data, both onto internal SD card memory and online via the GSM

(Global System for Mobile Communications) network to a webpage where the raw data could be accessed in real-time. The lowest time resolution for the logged data in this case was 1 min, averaged from measurements set to be taken by the OPC every 10 s. Each OPC device and microprocessor-controlled circuit board system were encased inside a plastic weatherproof box and also included BME280 sensors (Bosch Sensortec GmbH, Gerhard-Kindler-Strasse 9, 72770 Reutlingen, Germany) to measure temperature and RH (relative humidity). These were positioned externally to the OPC device but inside the weatherproof box for this study. We found the measurements from the BME280 sensors to be more accurate than the temperature and relative humidity measurements supplied by OPC devices themselves, which are more influenced by the heat generated by the OPC-N3 sensors and surrounding electronics.

Figure 7.1 in the Appendices shows the time series of RH and temperature measured by the BME280 sensors against reference data, obtained from the German National Meteorological Service (DWD) at the München-City station located 7.5 km away from the sampling site. While there is a systematic error for both variables due to the positioning of the BME280 sensors inside the weatherproof box, it is evident they are consistent with the ambient variations over time.

While the OPCs output a number of different variables, including calculated PM mass concentrations, the main data collected that was relevant to this study were the raw particle number counts from each of the 24 size bins, alongside variables that would influence the bin counts, including the laser status, sample period and flow rate.

The benchmark pollen data for this study were obtained from the mean of four collocated Hirst-type samplers (all Burkard Manufacturing Co Lid, Rickmansworth, UK) provided by the coordinators of the EUMETNET AutoPollen – ADOPT COST Action

(CA18226) Intercomparison Campaign 2021. Further details on this can be found in the initial overview paper (Maya-Manzano et al., 2022). These data consisted of hourly averages and included pollen number concentrations in grains/m^3 for 16 different taxa: *Alnus* (alder), *Ambrosia* (ragweed), *Artemisia* (mugwort), *Betula* (birch), *Carpinus* (hornbeam), *Fagus* (beech), *Fraxinus* (ash), *Picea* (spruce), *Pinus* (pine), *Plantago* (plantai n), *Poaceae* (grass), *Populus* (poplar), *Quercus* (oak), *Taxaceae*
Cupress (yew), *Tilia* (lime), *Urtica* (nettle), as well as a total pollen concentration.

2.3.2 Context

The three OPCs ('OPC1', 'OPC2' and 'OPC3') were placed within a few metres of each other, facing the same direction, on the building roof site used for the AutoPollen Intercomparison Campaign at the Center of Allergy & Environment (ZAUM) in Munich, Germany, alongside all the other sensors participating in the campaign. The campaign ran from early March until July 2021. Hirst pollen concentrations were recorded for each hour of every day between 3rd March and 19th July. The OPCs were deployed in tandem from 9th March until 7th July, however, due to technical issues, not all the datasets were complete for the entire period. OPC1 recorded the most complete dataset, covering the whole period without significant gaps. Meanwhile, OPC2 recorded from 20th May until 7th July and OPC3 from 9th March until 29th June, with only one significant gap of under 48 h between 13th–15th March). Figure 7.2 in the Appendices displays the data availability of each OPC.

2.3.3 Data pre-processing

Initial data processing was performed in RStudio using R version number 4.1.2. As standard procedure, timestamps where the laser status variable (which should be around 620 W cm^{-2}) was below 570 or above 670 W cm^{-2} were omitted, as anomalies in

laser intensity could make the corresponding particle counts unreliable. This resulted in no data points being omitted for OPCs 1 and 3 and <0.001 % of data points omitted for OPC 2. Subsequently, the raw particle counts were converted to particle number concentrations (PNC) in particles m^{-3} by the following Eq. 1:

$$PNC\ (particles\ m^{-3}) = \frac{raw\ counts}{sample\ period\ (s) \times flow\ rate\ (L\ min^{-1}) \times \frac{0.001\ (m^3\ L^{-1})}{60\ (s\ min^{-1})}} \quad \text{Eq. 1}$$

The sample period and flow rate for each timestamp was used in the calculation but the mean sample period for all OPCs during this period was 5.0 s while the mean flow rate was 5.2 L min^{-1}. OPC datasets were averaged over each hour for comparison with the hourly manual pollen data.

2.3.4 Simple (large particle) pollen proxy

The first method trialled as a pollen proxy, which we refer to here as the 'Simple Pollen Proxy', used the sum of all detected particles in size bins above 10 µm. Pollen grains are generally greater in diameter than 10 µm (Reponen, 2011, pp. 723, Bradley, 2015, pp. 408–409), which means they are also larger than most reported PM. It was observed that the OPC bins between 10 and 40 µm generally showed very little or no activity at normal times when deployed away from sources of coarse dust (e.g., quarries, construction sites, ocean spray, etc). We hypothesised that, when the sensors were placed in the vicinity of significant pollen sources during an active pollen season, particle concentrations present in this particular size range would be dominated by pollen. Thus the total number concentrations greater in diameter than 10 µm could be taken as a proxy for pollen.

2.3.5 PMF pollen proxy

The second method was a Positive Matrix Factorisation (PMF) source apportionment technique used to isolate a signal for pollen from all OPC bins. This is a multivariate data analysis method developed by Paatero and Tapper (1994), which attempts to find patterns in the variables of a dataset. It then assigns these patterns into factors according to their unique features and provides a relative contribution for each one of them for each timestep of the dataset. This technique is generally used to isolate specific pollutant sources from environmental pollution data that is collectively comprised of background and other various component sources (e.g., Sun et al., 2020). We hypothesised that pollen could be isolated in OPC time series via this method, as has been demonstrated previously for typical pollutant sources (Bousiotis et al., 2022).

This PMF method took the particle counts of the 24 bins as input and output a corresponding time series of the relative contributions of five different factors at each timestamp. Five factors were chosen on the basis of previous experience with atmospheric aerosol samples. In theory, each factor should equate to a particular source of aerosols picked up by the sensors. For each factor, size distribution information is also available which can be used to identify which factor corresponds with which source. We investigated correlations of each of the factors with total pollen and individual taxa concentrations, and also used the particle size distributions to inform our decision on which factors were most likely to be associated with pollen. Two factors were selected for each OPC dataset that demonstrated the highest correlations with various pollen species and the ‘PMF Pollen Proxy’ was calculated from the sum of these two factors.

2.3.6 Supervised machine learning methods

Two further methods – Neural Network (NN) and Random Forest (RF) - were employed in parallel, with identical data for training and testing, in an attempt to construct a pollen proxy from the OPC data. These methods were implemented using Python (version 3.9.7, Anaconda distribution) in Jupyter Notebook.

Both NN and RF are supervised learning methods where a target variable from reference data is used to train and evaluate the model. This target variable was chosen from the Hirst data. Due to the limitations of the OPCs, measuring up to just 40 μm, it was considered that the data may not effectively capture all taxa included within the Hirst 'total pollen' variable. For the purposes of this study, two different target variables were selected to train and evaluate the models: total pollen and *Poaceae* (grass) pollen. The total pollen target variable was the sum of all pollen taxa concentrations collected by the Hirst. The *Poaceae* taxon was chosen here as the Hirst data showed a substantial active season that the OPCs could have a good chance of detecting, and provide sufficient information for the models to learn from. It is also generally ubiquitous and affects many allergy sufferers.

Both the NN and RF require an appropriate selection of input variables or features for the model to extract information from for the learning process. Here we prepared three sets of input variables for the models: number concentrations (particles m^{-3}) from all 24 OPC bins, all 24 OPC bin number concentrations plus meteorological variables relative humidity (RH) (%) and temperature (°C) from the BME280 sensor, and finally, just meteorological variables RH and temperature. As mentioned previously, these values from the BME280 sensor do have a systematic error from ambient conditions, however

the variation over time is consistent with ambient conditions (see Figure 7.1 in the Appendices), and so should serve the purpose sufficiently for these models.

The data for the input features were collated from all three OPC-N3s present at the campaign and split into 'train' and 'test' datasets simultaneously with the target variable using the *train_test_split* function from the Scikit-learn library (Pedregosa et al., 2011). Data was shuffled before being split and 40 % of all the data (6220 data points) was taken for the test dataset. (Other split percentages were trialled but this was selected as optimal for this dataset based on convergence of NN training and validation learning curves.) The same train and test datasets were used for all models of both methods. Before implementing the Neural Network method, the input feature train and test datasets were normalised using Scikit-learn *MinMaxScaler* function, with the scaler fit to the train dataset so the test dataset was kept 'unseen'. This was not considered necessary for the Random Forest method.

2.3.7 Neural Network method

The Neural Network (NN) model was a simple multilayer feedforward Neural Network (Svozil et al., 1997; Sazli, 2006) constructed using a Sequential model from the TensorFlow library (Abadi et al., 2015). The Sequential model structure facilitates the construction of linear stacks of layers, each with a chosen number of nodes and an activation function through which the data passes in the training process. The models used here all had five Dense (fully connected) hidden layers, each with 5 times as many nodes as input features (i.e. 130 or 120, with or without meteorological variables) and a ReLU (Rectified Linear Unit) activation function. After each hidden layer a Dropout layer was implemented with a frequency rate of 0.2 (i.e. 20 % of nodes randomly dropped in each layer) to reduce overfitting and improve generalisation. Each hidden layer had a

kernel constraint, often used in combination with Dropout layers to manage overfitting, constraining the max norm of incident weights to less than or equal to 1. A Dense single node output layer, also with a ReLU activation function, followed the last hidden layer to produce the output. A generalised model, simplified without visible dropout layers, can be found in Figure 7.4 in the Appendices.

Models were compiled with the Adam optimiser and mean squared error (MSE) loss metric and fit on the training data subset, using the test data subset for validation, with a batch size of 128. Early stopping was implemented so the training process stopped automatically when the loss in MSE had not improved any further within a given patience parameter of 100 epochs. A model checkpoint saved subsequent models of best MSE so the last to be saved was that which reached the lowest MSE, rather than the final one wherever training stopped. Training generally stopped between 400 and 800 epochs (runs through the whole training dataset).

For each model, training and validation MSE loss curves were plotted to assess the model learning process. Root mean squared error (RMSE), mean absolute error (MAE), explained variance (R^2) and Spearman's rank correlation coefficient (ρ) values were calculated between the predicted target values from the test dataset and the corresponding real target values as metrics to evaluate model performance. Relative root mean squared (RRMSE) error - RMSE divided by the mean of the reference target variable - was also calculated.

2.3.8 Random Forest method

The Random Forest model was implemented using the RandomForestRegressor model from the Scikit-learn library. A generalised diagram of the random forest framework used for regression can be found in Figure 7.5 in the Appendices. An initial grid search

cross validation was used to trial different values for the number of estimators (number of decision trees) and max depth (the number of splits allowed in each decision tree). For the final models, the number of estimators used was 600, the max depth was 10 and the maximum number of features (to be used for each tree) was the square root of the total number of input features. Similarly, RMSE, RRMSE, MAE, R^2 and ρ were calculated between predicted test target values and the real target values to evaluate model performance on unseen data.

The RandomForestRegressor model from the Scikit-learn library enables calculation of feature importance score for each of the variables input into the model. These scores are measured by Gini Importance, or, Mean Decrease in Gini Impurity (MDG) (Breiman, 1984; Menze et al., 2009). The Gini Impurity function determines how useful a node in a decision tree, with a given split rule based on a particular feature, was at separating the observations being passed through. A higher decrease in Gini impurity, which ranges between 0 and 0.5, means that node was more useful. Decrease in Gini impurity values are averaged across all nodes using the same feature, producing MDG, which measures the contribution of each feature across the whole Random Forest in separating the observations according to the labels or target variable. The higher the MDG importance score, the greater effect or predictive power this feature has over the Random Forest model.

This technique was applied to all Random Forest models after training to ascertain the relative influences of each of the OPC size bins, also RH and temperature where applicable, on the predictions output from the model. Due to their more complex nature, this is process is not as easily applied Neural Network models and so such scores were not calculated here. To accompany this, we also calculated Spearman correlation

coefficients between each of the input features and target variables. This does not provide a true measure of feature importance on each model since the machine learning methods can learn from more complex variable relationships. However, we still considered it useful to investigate since it can provide information the feature importances do not, i.e. whether the correlation between input feature and target variable is positive or negative.

2.3.9 Evaluating the spatial footprint of pollen vs pollen proxy

As a further test of the models' ability to accurately identify pollen, wind direction (°) and wind speed (m/s) measurements, also from the DWD München-City meteorology station, were used to assess the spatial footprint of the pollen detected in this campaign. Using the R Openair package (Carslaw and Ropkins, 2012), bivariate polar plots showing pollen concentration by wind direction and speed were constructed for both Hirst and OPC model concentrations. From these plots we can ascertain the direction and approximate distance of the particle sources relative to the sampling site. By comparing Hirst and model concentration plots, we can gauge how accurately the models are identifying particles from the same source of pollen.

2.3.10 Evaluating model ability to distinguish high pollen events

Rather than producing highly accurate pollen concentrations, as is the aim with instruments specifically designed to measure bioaerosols, we hope to be able to use the low-cost OPCs to produce a sufficiently accurate and precise estimate for pollen concentrations to make it possible to determine when high pollen events are occurring. Simple information such as this could be useful for allergy sufferers, if suitable thresholds for ‘high’ concentrations are set.

We evaluated the models using further metrics which test their ability to distinguish high pollen concentration events. The threshold for high *Poaceae* pollen events was set at 50 grains m^{-3}, following standards stated by the UK Met Office (https://www.metoffice.gov.uk/weather/warnings-and-advice/seasonal-advice/health-wellbeing/pollen/what-is-the-pollen-count, Jan 2023), and the threshold was scaled up for total pollen by the ratio of the 99th percentiles. The manual and model-constructed pollen proxy test datasets were split into Positive or Negative categories - equal to/greater than or less than the threshold value respectively. For each data point it was determined whether the constructed pollen proxy was a True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN) compared to the manual baseline (see Figures 7.5 and 7.6 in the Appendices for more detail). The precision, recall and F1 score (the harmonic mean of precision and recall) were calculated from Eqs. (2), (3), (4). The closer the F1 score to 1, the better the ability of the model to distinguish successfully high pollen events from the OPC data.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{Precision + Recall} \tag{4}$$

2.4 Results and discussion

2.4.1 Simple pollen proxy

Figure 2.1 shows a comparison of the Simple Pollen Proxy for all sensors, engineered from OPC bins >10 µm in particle diameter, in orange against the manual total pollen concentration in blue. To facilitate visual comparison, the Simple Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables (7.1). During the summer months of June and July the Simple Pollen Proxy suggests a greater frequency of high particle concentration events, coinciding approximately with the *Poaceae* (grass) and *Urtica* (nettle) pollen seasons. However, the resemblance is not close enough to draw significant conclusions. The Spearman correlation coefficients between the Simple Pollen Proxy and each of the total pollen and *Poaceae* pollen target variables were negligible (between −0.1 and 0.1).

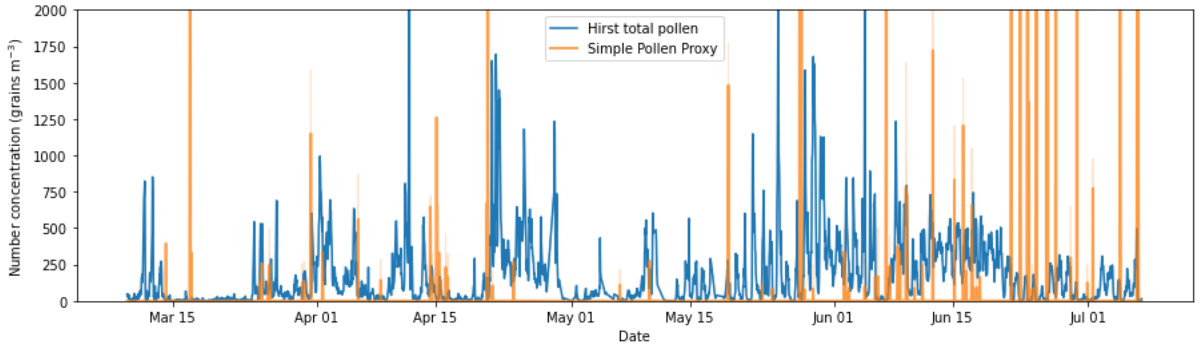


Figure 2.1: *Time series for Simple Pollen Proxy over the whole campaign duration. Simple Pollen Proxy concentrations from OPC data in orange and total pollen concentrations from Hirst data in blue. The Simple Pollen Proxy has been scaled up by the ratio of the 99th percentile of the two variables to facilitate comparison.*

While most of the time the Simple Pollen Proxy concentrations are relatively small or equal to zero, there are peaks of very high concentrations which do not appear to coincide with the manual pollen counts. It is possible that some of these peaks are not indicative of pollen grains but rather of other non-biological particles or meteorological conditions at the time such as precipitation, as this method has no way to distinguish from other particles in this size range. Some peaks, particularly in late June and in July, coincide with rainfall events (see Figure 7.3 in the Appendices), suggesting that rain droplets may have been detected in this size range by the OPCs. Maya-Manzano et al. (2022) reported that some construction work was taking place near the measurement site during the campaign period, which may have affected sensors such as the OPCs. The concomitant occurrence of anthropogenic air pollutants, mineral dust and fungal spores with pollen has been studied by Grewling et al. (2019) and demonstrates the possibility of such particles interfering. Unfortunately, however, we lacked the means to investigate further here.

The significant lack of non-zero concentrations in this size range suggests that the OPC sensors may in fact struggle to capture larger sized particles from ambient air where the sources are varied and potentially some distance away. Larger particles typically are not transported over as long distances as smaller particles, and do not follow air streamlines around sensor inlets (e.g., Hinds, 1999), which can cause sample losses ahead of the detection laser.

2.4.2 PMF pollen proxy

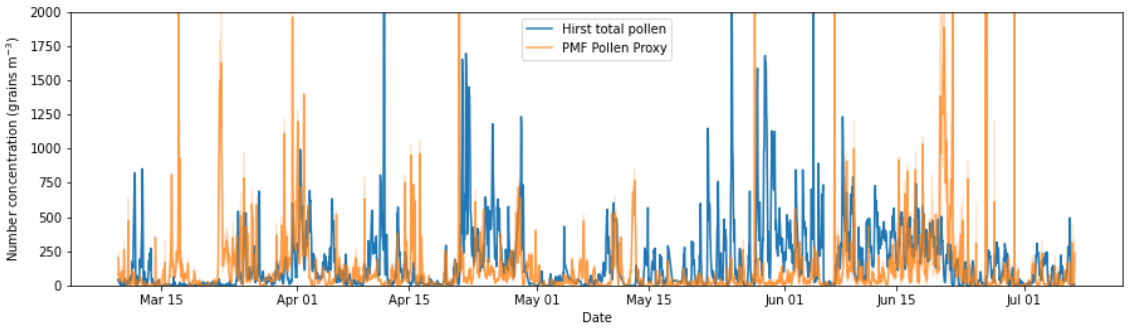


Figure 2.2: *Time series for PMF Pollen Proxy over the whole campaign duration. PMF Pollen Proxy constructed from OPC bin data in blue and Hirst data total pollen in orange. PMF Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables to facilitate comparison.*

Figure 2.2 shows in orange the PMF Pollen Proxy constructed from all OPC bins of all sensors against the blue Hirst total pollen observations. Again, the PMF Pollen Proxy values have been scaled up by the ratio of the 99th percentiles of the two variables (51.8) to facilitate comparison. Visually, the PMF Pollen Proxy appears to demonstrate closer resemblance to the manual total pollen counts than the Simple Pollen Proxy. Some activity also coincides with the *Taxaceae Cupressus* (March-early April), *Fraxinus* (April), *Betula* (April), *Quercus* (May), *Poaceae* (June), *Urtica* (June–July) pollen seasons, suggesting that these taxa may be better detected by the OPCs. The PMF Pollen Proxy achieved Spearman correlation coefficients of 0.18 and 0.10 with the target variables total pollen and *Poaceae* respectively.

While the PMF Pollen Proxy, with its more sophisticated processing of the data and inclusion of all bins, performed better than the Simple Pollen Proxy, it still does not

provide a significantly accurate indicator for pollen concentrations when compared with the manual pollen measurements. Correlations were also explored between the PMF Pollen Proxy and individual pollen taxa, with certain taxa exhibiting higher correlations with the proxy than other taxa. This suggests that some taxa were more successfully detected by the OPC instruments, due to their grain size and higher concentrations.

A possibility to improve this method further would be to attempt resolving a greater number of factors from the data, which might better isolate some pollen sources from other particle sources. However, this has not been taken further within the scope of this study.

2.4.3 Neural Network and Random Forest methods

2.4.3.1 Input feature correlations

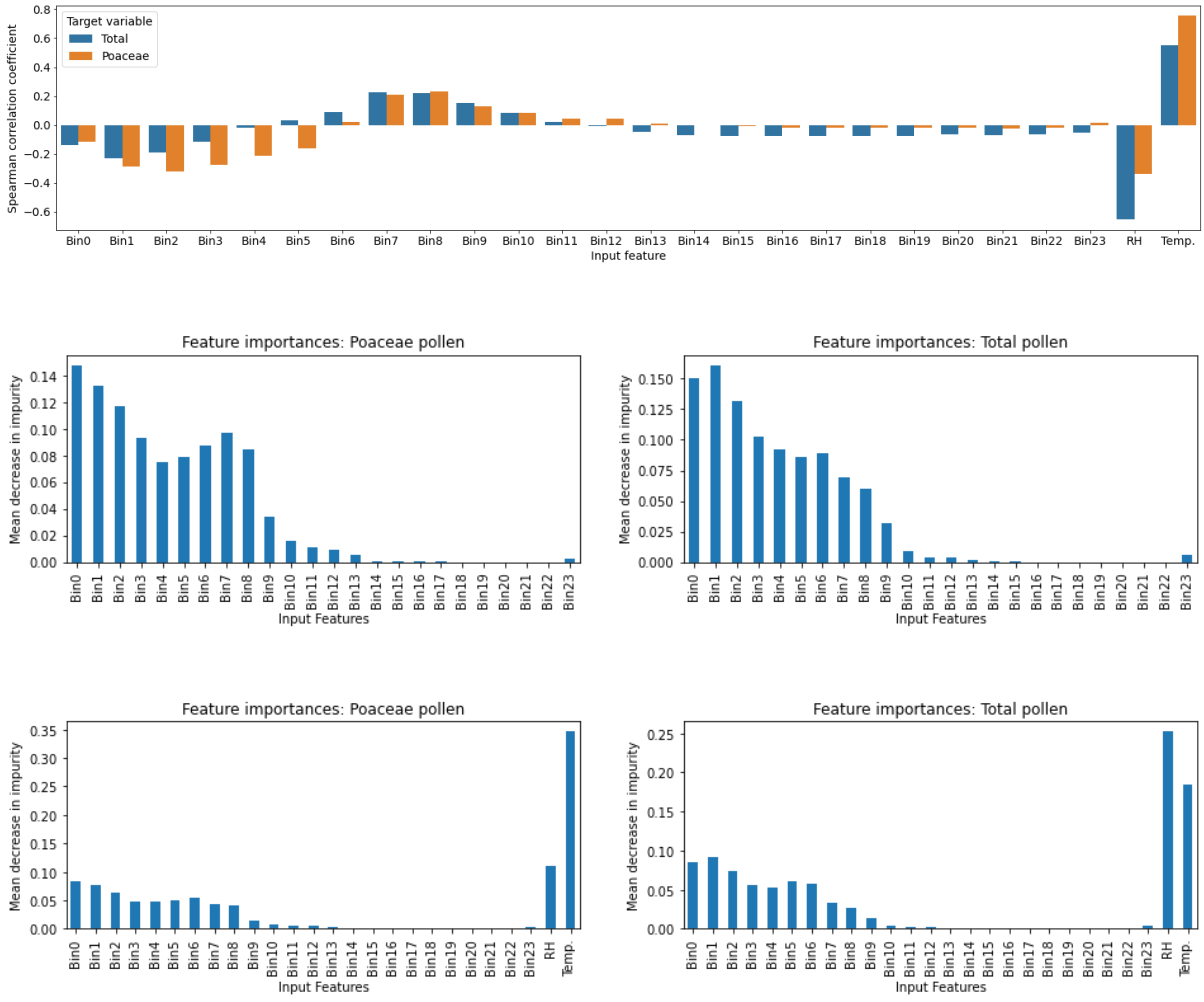


Figure 2.3: *Bar plots of correlations between OPC input features and Hirst target variables and Random Forest feature importances. Top: Spearman correlation coefficient values of each input feature (24 OPC bins, RH and temperature) with each target variable (*Poaceae *and total pollen). Middle: Feature importances of Random Forest models with only OPC bins. Bottom: Feature importances of Random Forest models with relative humidity (RH) and temperature (Temp.) input features as well as 24 OPC bins. Corresponding particle size ranges for each OPC bin can be found in Table 7.1 in the Appendices.*

The Spearman correlation coefficients between input and target variables are displayed as a bar chart at the top of Figure 2.3. Below this, also in Figure 2.3, are the Random Forest model feature importances – those with only OPC bin input features above and those including meteorological variables below. These plots provide an idea of which input features have more influence on the model learning process. It may seem surprising that the larger size bins, equivalent to the size of intact pollen grains, do not show significant correlation with the target variables. However, as discussed in Section 2.4.1, it seems that the OPCs struggle to detect significant particle counts in this range and this could be the reason for the negligible correlations.

Meanwhile, the most positively correlated OPC bins with the target variables are bins 7 and 8, corresponding to a size range of 3.0–5.2 μm. This is a possible size range for pollen fragments and some subpollen particles judging by the measurements of a few different species by Taylor et al. (2004), Miguel et al. (2006), Stone et al. (2021) and Burkart et al. (2021). Interestingly, we then see negative correlations for the smaller size ranges, especially bins 1 and 2 which correspond to 0.46–1.0 μm. It should be noted that the particle size ranges determined by the OPC optical Mie scattering measurements may not correspond absolutely with accurate geometric diameters, but it should be consistent across OPC-N3 instruments in general. Nevertheless, we conclude from this that the models in fact may not be learning primarily from the signal of intact pollen grains. Instead, the models are likely learning from the particle size information primarily associated with pollen fragments and subpollen particles released from ruptured pollen grains. We assume here that such particles coming directly from pollen grains are present simultaneously with pollen in the atmosphere. Yet such particles would be far more numerous than pollen grains (Stone et al., 2021), while likely suspended for increased time periods and transported greater distances compared to

intact pollen grains. They would likely be much more readily detected by our OPC instruments. Other particles such as fungal spores and anthropogenic particles may also be present but, since the machine learning methods are supervised using Hirst target data, pollen-correlated particles should be distinguished from these. Thus, we assume that this information can serve as a useful pollen proxy.

Both correlations and calculated importances for meteorological variables are more significant than those of each of the particle bins. RH shows negative correlation while temperature has a positive correlation with pollen concentrations. This is likely a result of diurnal variation where temperature peaks during the day and RH during the night in general, as pollen concentrations tend to be higher in the daytime. The positive temperature correlation could also be due to the onset of significant pollen seasons, including *Pinus*, *Poaceae* and *Urtica*, towards the later summer months, or a more general correlation between temperature and pollen season start (Van Vliet et al., 2002). Negative correlation with RH could also be a result of precipitation occurring when humidity is high, as this could effectively wash airborne pollen grains out of the atmosphere and reduce concentrations. Pollen grains are hygroscopic (Pope, 2010) and therefore swell at high humidity which could also cause the grains to settle more quickly under gravity even without precipitation.

2.4.3.2 Neural Network and Random Forest models: overview

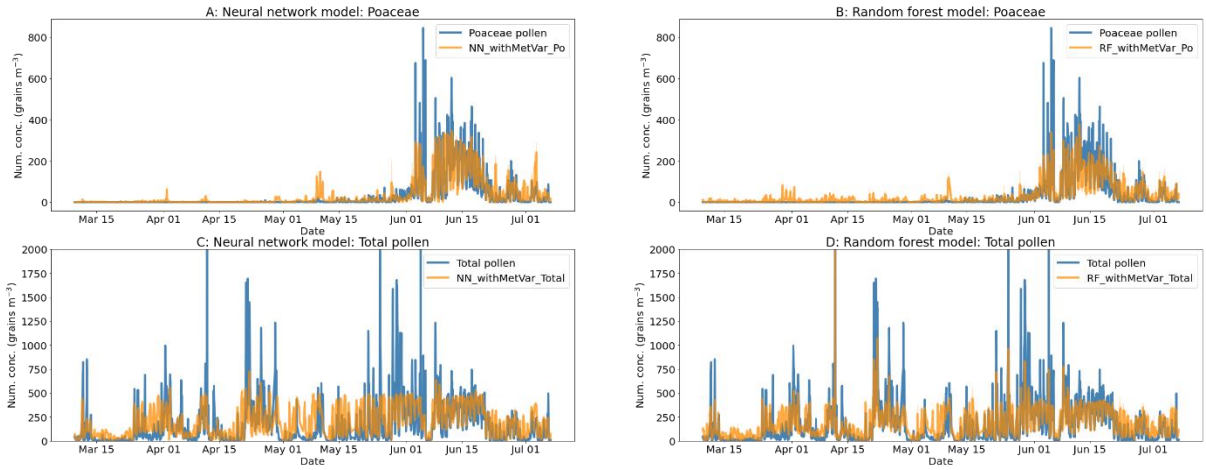


***Figure 2.4**: Time series for NN and RF pollen proxies over the whole campaign duration. Hirst data target variable are in blue and model predictions for models with RH and temperature as input features are in orange. Plots on the left (A and C) are from Neural Network models while those on the right (B and D) are from Random Forest models. A and B are predictions for* Poaceae *pollen and C and D for total pollen number concentrations in grains* m^{-3}.

Table 2.1: *Summary of Neural Network (NN) and Random Forest (RF) models and evaluation metrics. The 8 supervised machine learning models tested to predict pollen concentrations from OPC data, differentiated by model method (NN /RF), input features (with/without meteorological variables RH and temperature) and target variable (Poaceae/total pollen). The best NN and RF results have been emboldened.*

Model name	**Method**	**Input features**	**Target variable**	**RMSE (grains m^{-3})**	**RRMSE**	**MAE (grains m^{-3})**	**ρ**	**R^2**	**F1 Score**
NN_Po	Neural Network	24 OPC bins	*Poaceae*	69.5	1.7	34.2	0.62	0.34	0.65
NN_Total	Neural Network	24 OPC bins	Total pollen	240.3	1.3	156.1	0.54	0.15	0.65
NN_withMetVar_Po	**Neural Network**	**24 OPC bins + RH + T**	***Poaceae***	**49.1**	**1.2**	**20.7**	**0.85**	**0.67**	**0.83**
NN_withMetVar_Total	Neural Network	24 OPC bins + RH + T	Total pollen	215.8	1.1	118.0	0.77	0.3	0.78
RF_Po	Random Forest	24 OPC bins	*Poaceae*	71.8	1.8	39.4	0.61	0.29	0.60
RF_Total	Random Forest	24 OPC bins	Total pollen	232.1	1.2	146.5	0.55	0.19	0.67
RF_withMetVar_Po	**Random Forest**	**24 OPC bins + RH + T**	***Poaceae***	**54.3**	**1.4**	**26.5**	**0.79**	**0.59**	**0.80**
RF_withMetVar_Total	Random Forest	24 OPC bins + RH + T	Total	208.5	1.1	113.6	0.78	0.34	0.78

In total, 8 models were trained and can be categorised by method (Neural Network/Random Forest), input features (with/without RH and temperature), and target variable (*Poaceae*/total pollen). Figure 2.4 displays the time series for each of these models for the duration of the campaign, with the reference target variable in blue and the model-constructed proxy in orange. This figure includes only those models which included meteorological input features, whereas an alternative figure also displaying the models without meteorological input features can be found in Figure 7.6 in the Appendices. Table 2.1 lists all the models, their defining categories, and the metrics - RMSE, RRMSE, MAE, ρ, R^2 and F1 score - attained by each for the test datasets.

The time series and metrics tables for further models, trained on just RH and temperature input features, can be found in Figure 7.7, Figure 7.8, Table 7.2 and Table 7.3 in the Appendices.

From the time series plots in Figure 2.4, it is evident that all NN and RF model-constructed proxies performed significantly better than the Simple Pollen Proxy and PMF Pollen Proxy, and pollen season variations are to some extent successfully reconstructed. There are false positive signals occurring at times when the manual baseline shows low or zero pollen concentrations, for example *Poaceae* pollen being predicted outside of its season. However, this was observed for most instruments that took part in the campaign (Maya-Manzano et al., 2022). Particularly for the OPCs, non-biological particles - or other biological particles - which are of a similar size to the particles of interest may easily be confused with them causing such false positives. Many different pollen taxa overlap in size ranges, and so may limit the ability of the model to distinguish individual species with the data available. The model-constructed proxies in general appear to underestimate when very high pollen concentrations are observed in

the manual counts. This may suggest greater physical limitations for the particles being sampled through the inlet and detected by the OPC than for the Hirst-type sampler.

Table 2.2: *Mean-averaged metrics for models by category. Categorised by input feature in top table A, by model type in middle table B, and by target variable in bottom table C.*

A) Input features	**RMSE**	**RRMSE**	**MAE**	**ρ**	**R^2**	**F1 score**
OPC bins	153.4	1.49	94.1	0.58	0.24	0.64
OPC bins + RH + T	131.9	1.20	69.7	0.80	0.48	0.80

B) Model type	**RMSE**	**RRMSE**	**MAE**	**ρ**	**R^2**	**F1 score**
NN	143.7	1.33	82.3	0.70	0.37	0.73
RF	141.7	1.36	81.5	0.68	0.35	0.71

C) Target variable	**RMSE**	**RRMSE**	**MAE**	**ρ**	**R^2**	**F1 score**
Poaceae	61.2	1.52	30.2	0.72	0.47	0.72
Total pollen	224.2	1.17	133.6	0.66	0.25	0.72

2.4.3.3 Difference in model performance due to input variables

The most apparent difference observed among the models is that those with RH and temperature added as input features perform significantly better than those without. This can be seen in Figure 7.6 in the Appendices. The metrics (see Table 2.2a) support this where the models without RH and temperature have average RMSE, RRMSE, ρ and R2 values of 153 grains m^{-3}, 1.5, 0.58 and 0.24 while the models with these features have

values of 132 grains m^{-3}, 1.2, 0.80 and 0.48 respectively. The F1 score is also improved from 0.64 to 0.80 demonstrating that RH and temperature significantly add to the accuracy of the models. The likely reasons for the effect of RH and temperature as input features were discussed in Section 2.4.3.1.

Models trained only on meteorological input features RH and temperature (results shown in Figures 7.7 and 7.8 and Tables 7.2 and 7.3 in the Appendices) scored higher metrics than models with only particle size information, yet significantly below models with both particle size and meteorological information. This demonstrates, as the input feature correlations in Figure 7.3 suggested, that RH and temperature add valuable predictive power to the models. Meanwhile it proves that the combination of both particle size and meteorological information produces the most accurate results.

2.4.3.4 Difference in model performance due to model type

The difference between the Neural Network and Random Forest methods is less distinct. From visual inspection, it appears that the Random Forest models adhere a little closer to the Hirst benchmark and may capture more detail of individual peak events. The metrics (see Table 2b) are very close between NN and RF models with values of 144 grains m−3, 1.3, 0.70, 0.37 and 0.73, and 142 grains m^{-3}, 1.4, 0.68, 0.35 and 0.71 for RMSE, RRMSE, ρ, R^2 and F1 Score values respectively.

While this demonstrates presently that both model types are similarly useful for this purpose, we note that the NN methods deployed here have the potential for further development and accuracy improvement in the future. A grid search cross validation method was performed on the RF models first to decide on optimal hyperparameters. Meanwhile the NN method is more complex with many different hyperparameters and possible architecture variations. While different variations were trialled in the process

of achieving the final models presented here, this is by no means exhausted or optimised to its full potential.

2.4.3.5 Difference in model performance due to target variable

For the choice of target variable, the single taxon *Poaceae* models performed best by all metrics (see Table 2c), except for RRMSE. This is because total pollen had significantly greater range resulting in inevitably greater error margins. The RRMSE metric however corrects for this and allows for some comparison across different variables. Total pollen models had the lower RRMSE at 1.2 compared to *Poaceae* models at 1.5.
Meanwhile *Poaceae* and total pollen models achieved joint best F1 scores at 0.72.

The question of which target variable is preferable to use is not simple, and for future studies no doubt depends on the context, which taxa are present and of interest. While targeting individual taxa like *Poaceae* can result in smaller error, the limitations of the OPC's ability to distinguish between taxa based solely on size information should be kept in mind and false positives are to be expected. Including input features that correlate in some way to season (including temperature) will likely reduce this error yet make the model less generalisable to use in different locations, where seasonal patterns are different or irrelevant. A key point from this study is that using target variables at both the individual taxa and collective pollen levels are worth investigating for further studies. While in many cases the accuracy of precise number concentrations may still be lacking, the results here indicate that reducing the information conveyed to simply the occurrence of high pollen events offers better accuracy across different target variables.

2.4.3.6 Individual model performances

Of all 12 individual models, the best performing across most metrics was the NN_withMetVar_Po model at RMSE, MAE, ρ, R^2 values and F1 score of 49 grains m^{-3}, 21

grains m^{-3}, 0.85, 0.67 and 0.83 respectively. The RF alternative, RF_withMetVar_Po, performed similarly, also reaching an R^2 value above 0.5 and F1 score of 0.80. While difficult to compare across different target variables, the total pollen models with meteorological variables (both NN and RF) performed best in terms of RRMSE while also demonstrating high F1 scores at 0.78. Meanwhile, the error margins and R^2 scores for these and the other models would be ideal to improve upon.

In the campaign overview (Maya-Manzano et al., 2022), 9 out of 18 of the automated pollen measurement systems (including different classification algorithms for the same instrument) were reported to reach R^2 values above 0.5 for both 3 hourly and daily averaged data. These were high quality instruments, specifically designed to measure pollen with sophisticated algorithms trained directly on pollen samples. In comparison, 2 out of 8 of our models here achieved R^2 values above 0.5, at hourly time resolution. When averaged to daily resolution, all R^2 values improve significantly – 3 further models (including total pollen proxy models) are at or very close to R^2 values of 0.5. The best models – *Poaceae* pollen proxies with meteorological variables – reach R^2 values of 0.75 and 0.86.

It is however interesting that the automated instruments in the intercomparison campaign in general achieved the lowest metrics for *Poaceae* pollen compared to other taxa (with 6 out of 18 systems achieving $R^2 > 0.75$ for daily resolution data) while we achieved good results for our *Poaceae* models. The comparatively worse performance for *Poaceae* in the campaign overview (Maya-Manzano et al., 2022) was thought to be due to greater variation in size among different species of the *Poaceae* family (Frenguelli et al., 2010), which could also be expected to affect the ability of the OPC and machine learning algorithm to distinguish this taxon. Nevertheless, we did not study models

targeting other taxa here, largely due to considered data limitations for model training, and it is possible they may be able to perform even better given ample training data.

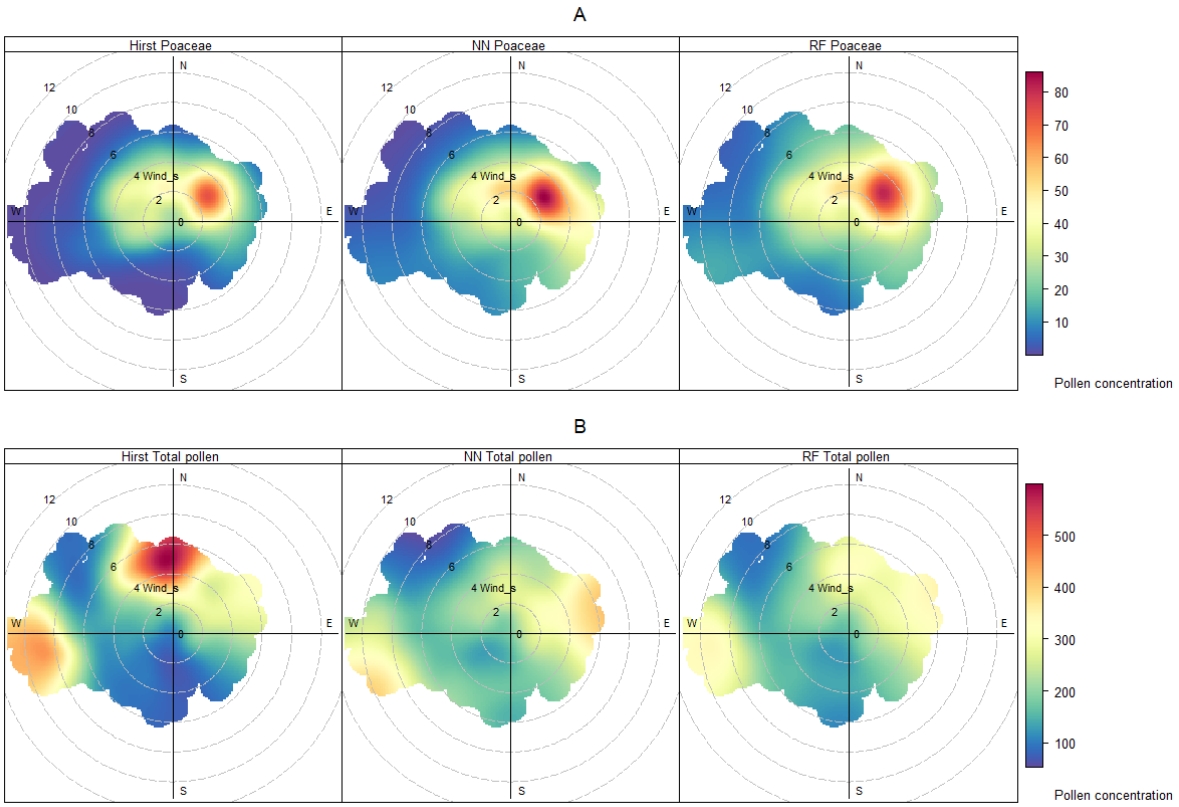


Figure 2.5: *Polar plots of detected pollen source concentrations as a function of wind direction and speed. Data for A)* Poaceae *and B) total pollen for each of the Hirst measurements, NN model pollen proxy and RF model pollen proxy from left to right (models are those which included RH and temperature input variables).*

Figure 2.5 shows polar plots - using the R Openair package (Carslaw and Ropkins, 2012) - of pollen concentrations plotted according to wind direction and speed, which provides further information on pollen sources. The meteorological data here was sourced from a weather station situated 7.5 km away from the intercomparison campaign site and so relates to regional-scale flow rather than local eddies and currents. Hirst, NN model and RF model (only those which included temperature and RH input

variables) pollen concentrations are presented (left to right) for each target variable (top to bottom).

The *Poaceae* models appear to have isolated the same source as the Hirst for this pollen taxa with remarkable success, strengthening the evidence that the models are effective at constructing a pollen proxy. The total pollen models appear less accurate in terms of absolute pollen concentrations, but the source areas are generally consistent with the Hirst. There seems to be a source from the North that the Hirst has collected at high concentrations but the OPCs have been less effective at detecting. This could be due to pollen taxa which are out of the measurable size range of the OPC, likely *Pinus* in the context of this campaign, in the case of total pollen. It may also be that sources brought by the wind from the direction the OPC inlet was facing were detected more efficiently and this may be important to consider for future work.

2.4.3.7 Monthly and diurnal variation

To answer the question of whether these model-constructed proxies can give meaningful information about real-life pollen trends, we compared monthly and diurnal variations between proxy and manual benchmark. Figure 2.6 displays plots of the averaged number concentrations by month on the left and by hour of the day on the right for each target variable – *Poaceae* and total pollen from top to bottom. In each plot, the blue line shows the target variable reference while the other colours show the other models for each model type with and without meteorological input features.

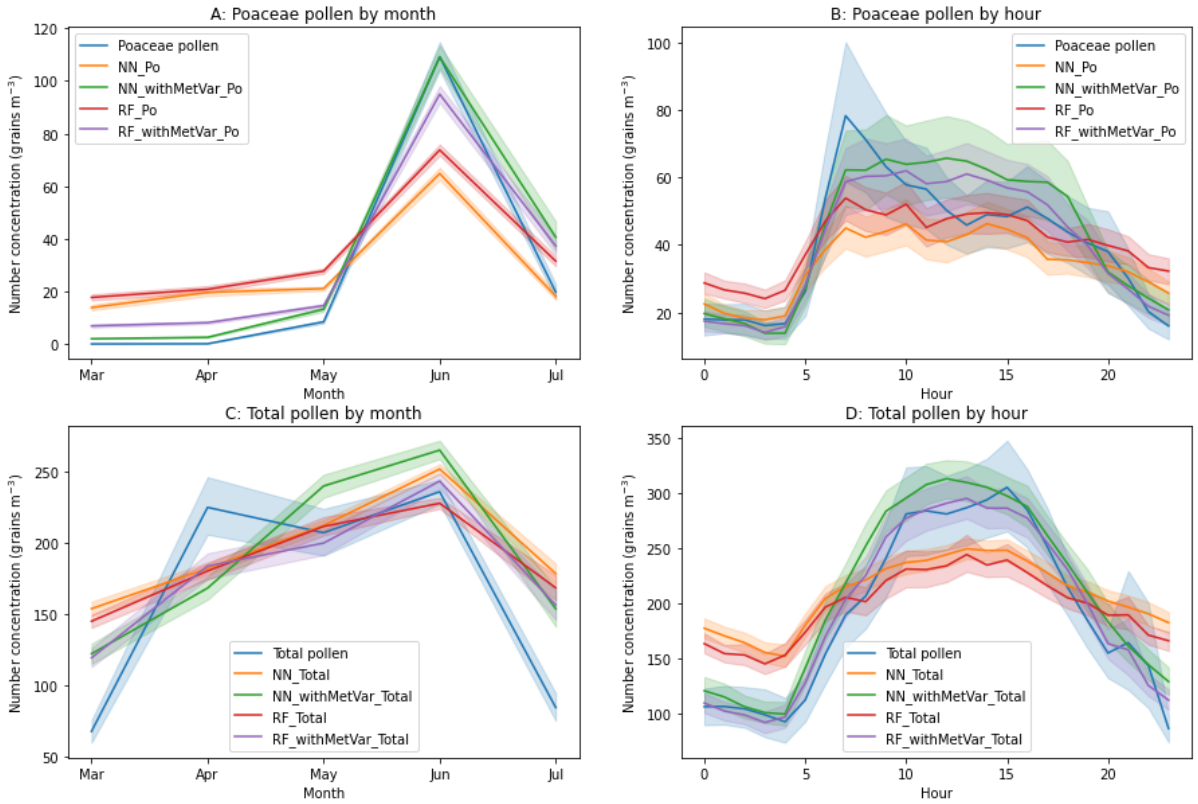


Figure 2.6: *Monthly and diurnal trends for model-constructed pollen proxies compared to Hirst baseline. Mean-averaged pollen concentrations by month, from March to July, on the left (A and C) and by hour of the day on the right (B and D) for each of the NN and RF models. Top (A and B):* Poaceae *models; Bottom (C and D): Total pollen models. Blue lines denote the target variable from Hirst data, green and orange lines denote NN models with and without meteorological input variables respectively, purple and red lines denote RF models with and without meteorological input variables.*

All models adhere approximately to the monthly and diurnal trends (for example increasing in daylight hours), even those without meteorological variable inputs. This demonstrates that the model does not simply rely on RH and temperature to create these trends, but that the particle concentrations also provide useful information. Nevertheless, as proven by the time series and calculated metrics, those models with meteorological variables perform markedly better. These models in fact seem to even

overestimate pollen concentrations at the height of daylight hours while the manual concentrations lie in between the model-constructed proxies with and without meteorological variables.

Furthermore, as seen in Figure 2.6, there is some evidence that the RF models may be better at detecting some details in temporal variation. For example, the total pollen baseline counts show a small peak in the evening after 8 pm and the RF models follow this noticeably better than the NN models.

2.4.3.8 Implications for public health

The utility of the methods tested here for public health rely on their fidelity in identifying high pollen events; that is, being able to detect high pollen events with a minimum of false positives and false negatives. Detailed results from the high pollen event thresholding test can be found from Figures 7.9 and 7.10 in the Appendices. Figure 7.10 shows the percentage of datapoints belonging to each group of the confusion matrix, i.e., true positive, false positive, true negative and false negative. NN and RF models perform very similarly with respect to true positives; the better performance of the RF model is due mostly to better discrimination between true negatives and false positives (Figure 7.10).

The false negative category for all models is below 5 % of all data points (except for the Neural Network *Poaceae* model without meteorological input features which was 6 %). From a public health perspective, this false negative category is the most import to keep minimal. As an instrument that informs allergy sufferers whether to anticipate severe symptoms on going outside, the false positive category would result in an inconvenient false alarm. Meanwhile, the false negative category could result in people at risk setting out when they believe it safe and could potentially result in more severe consequences

including hospitalisations. Therefore, with a false negative rate generally under 5 %, we suggest that the capability of the OPC sensors to determine high pollen events above a set threshold is sufficient and can provide valuable information for public health.

2.5 Conclusions

In this study we have explored four different potential methods to monitor pollen using low-cost OPC instruments, and we have made progress in constructing useful pollen proxies from OPC data. This step towards developing affordable pollen monitoring techniques and forecasting capabilities builds upon that already produced from the EUMETNET AutoPollen ADOPT - COST Action Intercomparison campaign which brought together various automated pollen sensors to facilitate further advancement in this area.

This is, to our knowledge, the first study implementing such techniques to construct a pollen proxy from low-cost OPC sensors conventionally used to measure particulate matter. While inferior in terms of accuracy to other instruments specifically developed to measure pollen, such low-cost sensors can, when coupled with suitable machine learning algorithms, estimate pollen concentrations to a useful degree of accuracy. They would thus supply an attractive alternative that has the potential provide automated, high temporal and spatial resolution data with fleets of sensors deployed in networks. This has not been possible before due to the high cost and labour demands of conventional manual instruments.

We presented here our investigation into finding and demonstrating an appropriate method of processing OPC general particle size data to gain useful information on airborne pollen concentrations. We show that NN and RF methods demonstrate the most success and can predict pollen concentrations to R^2 values reaching above 0.5 when compared to a Hirst-type sampler as a baseline, when meteorological variables RH and temperature are included as training features. The models also demonstrate an ability to distinguish high concentration pollen events above a certain threshold, achieving F1 scores (for reliable event detection) between 60 and 83 %.

We recommend for further investigations that meteorological variables including RH and temperature are used for model training and target variables for both individual and collective pollen taxa should be studied, though the limitations of the OPC should be kept in mind. We find that NN and RF methods achieve similar results, but there is still potential for optimisation of these methods (particularly NN variations) to improve accuracy even further. Future work should include extending the method to other locations and environments and assessing the generalisability across sensors and different environments.

Furthermore, it would be beneficial to investigate application across networks of sensors to see if the information from the models can discern environmental variations on a local scale, as this is where the affordability of the sensors will have their edge. Hybrid networks would likely be required, with high quality instruments integrated across a few of the low-cost network sites, to provide accurate references for the low-cost sensors. High performance automated instruments, such as the Swisens Poleno or Plair Rapid-E, may become increasingly prevalent in populating pollen monitoring networks. Therefore, studies using pollen data from instruments such as these (in particular those that detect particles in-flight similarly to the OPC) as a reference, to provide the target variables in model training, would also be valuable. From here, future possibilities of high spatiotemporal resolution pollen measurement via sensor networks, or even personal pollen monitors, would look promising. This could ultimately provide valuable information to improve pollen forecasting models and help many allergy sufferers.

2.6 Contributors

This study using the low-cost OPCs was conceived and planned by FDP, who gave guidance and contributed to the final manuscript, and SAM who developed the methodology for the machine learning models, performed the analyses and prepared the first draft. ARM also gave guidance and contributed to the final manuscript. DB performed the PMF source apportionment analysis and contributed to the final manuscript. JMM and FT planned and organised the intercomparison campaign in Munich and facilitated the running of the sensors there, providing the context for the data used in this study. JMM and FT also contributed to the final manuscript.

2.7 Acknowledgements

We thank Andrew Tanner from the University of Birmingham, UK, for his technical work on functionalising the Alphasense OPC sensors for the purposes of this study. We thank the ZAUM (Cordula Ebner von Eschenback, Gudrun Pusch and Christine Weil) and MeteoSwiss (Nina Burgdorfer and Sophie Erb) teams for their assistance facilitating the EUMETNET AutoPollen – COST ADOPT intercomparison campaign, as well as the Helmholtz Zentrum München as the campaign host and their assisting team (Daphne Kolland, George Matuscheck and Benjamin Schnautz). The manual pollen analysts for the campaign Łukasz Kostecki and Agata Szymanska are greatly appreciated for their work facilitating the vital reference data for this campaign. We thank Robert Gebauer, Gisela Nagy and Anton Pointner for their IT support during the campaign.

The work here involving the OPC sensors and related analysis was funded by the Natural Environment Research Council (NERC) through its Central England NERC Training Alliance (CENTA) doctoral research training consortium, at the University of

Birmingham, and the grant “Quantification of Utility of Atmospheric Network Technologies (QUANT)” (NE/T001968/1). The intercomparison campaign where the OPC sensors were deployed and accompanying Hirst reference data was obtained for context of this study was funded by the Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) and EUMETNET AutoPollen Programme. Financial support was also received for this from the COST Action CA18226 ADOPT – *New approaches in detection of pathogens and aeroallergens*.

2.8 Data availability

Data supporting this publication are openly available from the UBIRA eData repository at doi: https://doi.org/10.25500/edata.bham.00000898.

2.9 Comment

This chapter presents an initial investigation into using low-cost OPC sensors for monitoring pollen and, consequentially, the conception of the original machine learning methodology that becomes a large part of this thesis. It is demonstrated how this methodology outperforms other methods and assessed what useful information can be offered from this novel method from the perspective of public health. The limitations of accuracy and reliability that remain are discussed, as well as suggestions for improving performance. As with machine learning methods in general, the models trained here are crucially dependent on data quantity and representativeness. Sufficiently addressing these data requirements would require extensive fieldwork with appropriate benchmark instruments. Without the means to arrange such fieldwork, the subsequent chapters follow further development and application of the methodology to document evidence of its usefulness and potential to take further.

2.10 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467v2 [cs.DC]. https://doi.org/10.48550/arXiv.1603.04467.

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers. Aerobiologia, https://doi.org/10.1007/s10453-021-09724-5.

Alarcón, M., Periago, C., Pino, D., Mazón, J., Casas-Castillo, M.C., Ho-Zhang, J.J., Linares, C., Rodríguez-Solà, R., Belmonte, J., 2022. Potential contribution of distant sources to airborne *Betula* pollen levels in Northeastern Iberian Peninsula. Sci. Total Environ., 818, 151827. https://doi.org/10.1016/j.scitotenv.2021.151827.

Bacsi, A., Choudhury, B.K., Dharajiya, N., Sur, S., Boldogh, I., 2006. Subpollen particles: carriers of allergenic proteins and oxidases. J. Allergy Clin. Immunol. 118 (4), 844-850. https://doi.org/10.1016/j.jaci.2006.07.006.

Baird, A.B., Bannister, E.J., MacKenzie, A.R. and Pope, F.D., 2022. Mass concentration measurements of autumn bioaerosol using low-cost sensors in a mature temperate woodland free-air carbon dioxide enrichment (FACE) experiment: investigating the role of meteorology and carbon dioxide levels. Biogeosciences, 19 (10), 2653-2669. https://doi.org/10.5194/bg-19-2653-2022

Bousiotis, D., Beddows, D.C.S., Singh, A., Haugen, M., Diez, S., Edwards, P.M., Boies, A., Harrison, R.M., Pope, F.D., 2022. A study on the performance of low-cost sensors for source apportionment at an urban background site. Atmos. Meas. Tech., 15, 4047-4061. https://doi.org/10.5194/amt-15-4047-2022.

Bradley, R.S., 2015. Paleoclimatology: Chapter 12 – Pollen. 3rd ed., Elsevier, Massachusetts, USA, pp. 408-409. https://doi.org/10.1016/B978-0-12-386913-5.00012-0.

Breiman, L., 1984. Classification and Regression Trees. 1st ed, Routledge, New York, USA. https://doi.org/10.1201/9781315139470.

Burkart, J., Gratzl, J., Seifried, T.M., Bieber, P., Grothe, H., 2021. Isolation of subpollen particles (SPPs) of birch: SPPs are potential carriers of ice nucleating macromolecules. Biogeosciences, 18, 5751-5765. https://doi.org/10.5194/bg-18-5751-2021.

Buters, J.T.M., Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy 8, 9. https://doi.org/10.1186/s13601-018-0197-8.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.A., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Kofol Seliger, A., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto,

A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia, https://doi.org/10.1007/s10453-022-09750-x.

Carslaw, D.C., Ropkins, K., 2012. *Openair* – An R package for air quality data analysis. Environ. Model. Softw., 27-28, 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008.

Crilley, L.R., Shaw, M., Pound, R., Kramer, L.J., Price, R., Young, S., Lewis, A.C., Pope, F.D., 2018. Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. Atmos. Meas. Tech., 11, 709-720. https://doi.org/10.5194/amt-11-709-2018.

Crilley, L.R., Singh, A., Kramer, L.J., Shaw, M.D., Alam, M.S., Apte, J.S., Bloss, W.J., Hildebrandt Ruiz, L., Fu, P., Fu, W. and Gani, S., 2020. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos. Meas. Tech., 13, 1181-1193. https://doi.org/10.5194/amt-13-1181-2020

Després, V.R., Huffman, J.A., Burrows, S.M., Hoose, C., Safatov, A.S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M.O., Pöschl, U., Jaenicke, R., 2012. Primary biological aerosol particles in the atmosphere: a review. Tellus B, 64. http://dx.doi.org/10.3402/tellusb.v64i0.15598.

Diehl, K., Quick, C., Matthias-Maser, S., Mitra, S.K., Jaenicke, R., 2001. The ice nucleating ability of pollen: Part I: Laboratory studies in deposition and condensation freezing modes. Atmos. Res. 58 (2), 75-87. https://doi.org/10.1016/S0169-8095(01)00091-6.

Diehl, K., Matthias-Maser, S., Jaenicke, R., Mitra, S.K., 2002. The ice nucleating ability of pollen: Part II: Laboratory studies in immersion and contact freezing modes. Atmos. Res. 61 (2), 125-133. https://doi.org/10.1016/S0169-8095(01)00132-6.

Dreischmeier, K., Budke, C., Wiehemeier, L., Kottke, T., Koop, T., 2017. Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides. Sci. Rep., 7, 41890. https://doi.org/10.1038/srep41890.

Dubey, R., Patra, A.K., Joshi, J., Blankenberg, D., Sekhara, S., Kolluru, R., Madhu, B., Raval, S., 2022. Evaluation of low-cost particulate matter sensors OPC N2 and PM Nova for aerosol monitoring. Atmos. Pollut. Res., 13 (3), 101335. https://doi.org/10.1016/j.apr.2022.101335.

Frenguelli, G., Passalacqua, G., Bonini, S., Fiocchi, A., Incorvaia, C., Marcucci, F., Tedeschini, E., Canonica, G.W., Frati, F., 2010. Briging allergologic and botanical knowledge in seasonal allergy: A role for phenology. Ann. Allergy, Asthma and Immunol., 105, 223-227. https://doi.org/10.1016/j.anai.2010.06.016.

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. Atmos. Res., 182, 346-376. http://dx.doi.org/10.1016/j.atmosres.2016.07.018.

Giordano, M.R., Malings, C., Pandis, S.N., Presto, A.A., McNeill, V.F., Westervelt, D.M., Beekmann, M., Subramanian, R., 2021. From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors. Journal of Aerosol Science, 158, 105833. https://doi.org/10.1016/j.jaerosci.2021.105833.

Grewling, Ł., Bogawski, P., Jenerowicz, D., Czarnecka-Operacz, M., Šikoparija, B., Skjøth, C.A., Smith, M., 2016. Mesoscale atmospheric transport of ragweed pollen allergens from infected to uninfected areas. International Journal of Biometeorology, 60, 1493-1500. https://doi.org/10.1007/s00484-016-1139-6.

Grewling, Ł., Bogawski, P., Kryza, M., Magyar, D., Šikoparija, B., Skjøth, C.A., Udvardy, O., Werner, M., Smith, M., 2019. Concomitant occurrence of anthropogenic air pollutants, mineral dust and fungal spores during long-distance transport of ragweed pollen. Environmental Pollution 254 (Part A), 112948. https://doi.org/10.1016/j.envpol.2019.07.116.

Griffiths, P.T., Borlace, J.-S., Gallimore, P.J., Kalberer, M., Herzog, M., Pope, F.D., 2012. Hygroscopic growth and cloud activation of pollen: a laboratory and modelling study. Atmospheric Science Letters, 13 (4), 289-295. https://doi.org/10.1002/asl.397.

Gute, E., Abbatt, J.P.D., 2020. Ice nucleating behaviour of different tree pollen in the immersion mode. Atmos. Environ., 231, 117488. https://doi.org/10.1016/j.atmosenv.2020.117488.

Gute, E., David, R.O., Kanji, Z.A., Abbat, J.D.P., 2020. Ice Nucleation Ability of Tree Pollen Altered by Atmospheric Processing. ACS Earth Space Chem., 4 (12), 2312-2319. https://doi.org/10.1021/acsearthspacechem.0c00218.

Hinds, W.C., 1999. Aerosol Technology: Properties, Behaviour, and Measurement of Airborne Particles. 2nd edn, John Wiley & Sons Inc., New York, USA.

Huffman, J.A., Perring, A.E., Savage, N.J., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M.A., Crawford, I., Gallagher, M., Topping, D., Doughty, D.C., Hill, S.C., Pan, Y., 2019. Real-time sensing of bioaerosols: Review and current perspectives. Aerosol Science and Technology, 54 (5), 465-495. https://doi.org/10.1080/02786826.2019.1664724.

Hughes, D.D., Mampage, C.B.A., Jones, L.M., Liu, Z., Stone, E.A., 2020. Characterisation of Atmospheric Pollen Fragments during Springtime Thunderstorms. Environ. Sci. Technol. Lett. 7 (6), 409-414. https://doi.org/10.1021/acs.estlett.0c00213.

Jiang, C., Wang, W., Du, L., Huang, G., McConaghy, C., Fineman, S., Liu, Y., 2022. Field Evaluation of an Automated Pollen Sensor. Int. J. Environ. Res. Public Health, 19 (11), 6444. https://doi.org/10.3390/ijerph19116444.

Jochner, S., Lüpke, M., Laube, J., Weichenmeier, I., Pusch, G., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., Menzel, A., 2015. Seasonal variation of birch and grass pollen loads and allergen release at two sites in the German Alps. Atmos. Environ., 122, 83-93. https://doi.org/10.1016/j.atmosenv.2015.08.031.

Kawashima, S., Thibaudon, M., Matsuda, S., Fujita, T., Lemonis, N., Clot, B., Oliver, G., 2017. Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen. Aerobiologia, 33, 351-362. https://doi.org/10.1007/s10453-017-9474-6.

Manninen, H.E., Bäck, J.K., Sihto-Nissilä, S.-L., Huffman, J.A., Pessi, A.-M., Hiltunen, V., Aalto, P.P., Hidalgo, P.J., Hari, P., Saarto, A., Kulmala, M., Petäjä, T., 2014. Patterns in airborne pollen and other primary biological aerosol particles (PBAP), and their

contribution to aerosol mass and number in a boreal forest. Boreal Environment Research, 19, 383-405. http://hdl.handle.net.10138/165208.

Maya-Manzano, J.M., Smith, M., Markey, E., Clancy, J.H., Sodeau, J., O'Connor, D., 2020. Recent developments in monitoring and modelling airborne pollen, a review. Grana, 60 (1), 1-19. https://doi.org/10.1080/00173134.2020.1769176.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, L., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavulj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2023. Towards European automatic bioaerosol monitoring: Comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ., 866, 161220. https://doi.org/10.1016/j.scitotenv.2022.161220.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10, 213. https://doi.org/10.1186/1471-2105-10-213.

Miguel, A.G., Taylor, P.E., House, J., Glovsky, M.M., Flagan, R.C., 2006. Meteorological Influences on Respirable Fragment Release from Chinese Elm Pollen. Aerosol Sci. Technol., 40 (9), 690-696. https://doi.org/10.1080/02786820600798869.

Mikhailov, E.F., Ivanova, O.A., Nebosko, E.Y., Vlasenko, S.S., Ryshkevich, T.I., 2019. Subpollen Particles as Atmospheric Cloud Condensation Nuclei. Izv. Atmos. Ocean. Phys., 55, 357-364. https://doi.org/10.1134/S000143381904008X.

Narayana, M.V., Jalihal, D., Nagendra, S.M.S., 2022. Establishing A Sustainable Low-Cost Air Quality Monitoring Setup: A Survey of the State-of-the-Art. Sensors (Basel), 22 (1), 394. https://doi.org/10.3390/s22010394.

O'Connor, D.J., Healy, D.A., Hellebust, S., Buters, J.T. and Sodeau, J.R., 2014. Using the WIBS-4 (Waveband Integrated Bioaerosol Sensor) technique for the on-line detection of pollen grains. Aerosol Sci. Technol., 48 (4), 341-349. https://doi.org/10.1080/02786826.2013.872768

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and Online Pollen Monitoring. Int. Arch. Allergy. Immunol., 167, 158-166. https://doi.org/10.1159/000436968.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition. Env. Res., 191, 110031. https://doi.org/10.1016/j.envres.2020.110031.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: A Non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 5, 111-126. https://doi.org/10.1002/env.3170050203.

Pedregosa, F., Varoquaux, G., Granfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12 (85), 2825-2830.

Plaza, M.P., Kolek, F., Leier-Wirtz, V., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2022. Detecting Airborne Pollen Using an Automatic, Real-Time Monitoring System: Evidence from Two Sites. Int. J. Environ. Res. Public Health, 19 (4), 2471. https://doi.org/10.3390/ijerph19042471.

Pope, F.D., 2010. Pollen grains are efficient cloud condensation nuclei. Environ. Res. Lett., 5, 044015. https://doi.org/10.1088/1748-9326/5/4/044015.

Pummer, B.G., Bauer, H., Bernardi, J., Bleicher, S., Grothe, H., 2012. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. Atmos. Chem. Phys., 12, 2541-2550. https://doi.org/10.5194/acp-12-2541-2012.

Reponen, T., 2011. Encyclopedia of Environmental Health: Methodologies for Assessing Bioaerosol Exposures. Elsevier, Cincinnati, USA, pp. 723. https://doi.org/10.1016/B978-0-12-409548-9.11822-6.

Ruske, S., Topping, D.O., Foot, V.E., Morse, A.P. and Gallagher, M.W., 2018. Machine learning for improved data analysis of biological aerosol using the WIBS. Atmos. Meas. Tech., 11 (11), 6203-6230. https://doi.org/10.5194/amt-11-6203-2018

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech., 12, 3435-3452. https://doi.org/10.5194/amt-12-3435-2019.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2020. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech., 13, 1539-1550. https://doi.org/10.5194/amt-13-1539-2020.

Savage, N.J., Krentz, C.E., Könemann, T., Han, T.T., Mainelis, G., Pöhlker, C. and Huffman, J.A., 2017. Systematic characterization and fluorescence threshold strategies for the wideband integrated bioaerosol sensor (WIBS) using size-resolved biological and interfering particles. Atmos. Meas. Tech., 10 (11), 4279-4302. https://doi.org/10.5194/amt-11-6203-2018.

Sazli, M.H., 2006. A brief review of feed-forward neural networks. Commun. Fac. Sci. Univ. Ank. Series A2-A3, 50 (1), 11-17. https://doi.org/10.1501/commua1-2_0000000026.

Siljamo, P., Sofiev, M., Severona, E., Ranta, H., Kukkonen, J., Polevova, S., Kubin, E., Minin, A., 2008. Sources, impact and exchange of early-spring birch pollen in the Moscow region and Finland. Aerobiologia, 24, 211-230. https://doi.org/10.1007/s10453-008-9100-8.

Skjøth, C.A., Sommer, J., Stach, A., Smith, M., Brandt, J., 2007. The long-range transport of birch (*Betula*) pollen from Poland and Germany causes significant pre-season concentration in Denmark. Clin. Exp. Allergy, 37 (8), 1204-1212. https://doi.org/10.1111/j.1365-2222.2007.02771.x.

Smith, M., Matavulj, P., Mimić, G., Panić, M., Grewling, Ł., Šikoparija, B., 2022. Why should we care about high temporal resolution monitoring of bioaerosols in ambient air? Sci. Total Environ., 826, 154231. https://doi.org/10.1016/j.scitotenv.2022.154231.

Sofiev, M., Sofieva, S., Palamarchuk, J., Šauliene, I., Kadantsev, E., Atanasova, N., Fatahi, Y., Kouznetsov, R., Kuula, J., Noreikaite, A., Peltonen, M., Pihlajamäki, T., Saarto, A., Svirskaite, J., Toiviainen, L., Tyuryakov, S., Šukienė, L., Asmi, E., Bamford, D., Hyvärinen, A.-P., Karppinen, A., 2022. Bioaerosols in the atmosphere at two sites in Northern Europe in spring 2021: Outline of an experimental campaign. Environ. Res., 214 (Part 2), 113798. https://doi.org/10.1016/j.envres.2022.113798.

Song, U., Park, J., Song, M., 2012. Pollen morphology of *Pinus* (Pinaceae) in northeast China. For. Sci. Technol., 8 (4), 179-186. https://doi.org/10.1080/21580103.2012.704973.

Sousan, S., Koehler, K., Hallet, L., Peters, T.M., 2016. Evaluation of the Alphasense Optical Particle Counter (OPC-N2) and the Grimm Portable Aerosol Spectrometer (PAS-1.108). Aerosol Sci. Technol., 50 (12), 1352-1365. https://doi.org/10.1080/02786826.2016.1232859.

Steiner, A.L., Brooks, S.D., Deng, C.,Thornton, C.O., Pendleton, M.W., Bryant, V., 2015. Pollen as atmospheric cloud condensation nuclei. Geophys. Res. Lett., 42 (9), 3596-3602. https://doi.org/10.1002/2015GL064060.

Stone, E.A., Mampage, C.B.A., Hughes, D.D., Jones, L.M., 2021. Airborne sub-pollen particles from rupturing giant ragweed pollen. Aerobiologia, 37, 625-632. https://doi.org/10.1007/s10453-021-09702-x.

Sun, X., Wang, H., Guo, Z., Lu, P., Song, F., Liu, L., Liu, J., Rose, N.L., Wang, F., 2020. Positive matrix factorisation on source apportionment for typical pollutants in different environmental media: a review. Environ. Sci.: Processes Impacts, 22, 239-255. https://doi.org/10.1039/C9EM00529C.

Svozil, D., Kvasnička, V., Pospíchal, J., 1997. Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst., 39, 43-62. https://doi.org/10.1016/S0169-7439(97)00061-0.

Taylor, P.E., Flagan, R.C., Miguel, A.G., Valenta, R. and Glovsky, M.M., 2004. Birch pollen rupture and the release of aerosols of respirable allergens. Clinical & Experimental Allergy, 34 (10), 1591-1596. https://doi.org/10.1111/j.1365-2222.2004.02078.x

Tong, H.-J., Ouyang, B., Nikolovski, N., Lienhard, D.M., Pope, F.D., Kalberer, M., 2015. A new electrodynamic balance (EDB) design for low-temperature studies: applicationto immersion freezing of pollen extract bioaerosols. Atmos. Meas. Tech., 8, 1183-1195. https://doi.org/10.5194/amt-8-1183-2015.

Tummon, F., Bruffaerts, N., Celenk, S., Choël, M., Clot, B., Crouzy, B., Galán, C., Gilge, S., Hajkova, L., Mokin, V., O'Connor, D., Rodinkova, V., Sauliene, I., Sikoparija, B., Sofiev, M., Sozinova, O., Tesendic, D., Vasilatou, K., 2022. Towards standardisation of automatic pollen and fungal spore monitoring: best practices and guidelines. Aerobiologia, https://doi.org/10.1007/s10453-022-09755-6.

Van Vliet, A.J.H., Overeem, A., De Groot, R.S., Jacobs, A.F.G., Spieksma, F.T.M., 2002. The influence of temperature and climate change on the timing of pollen release in the Netherlands. Int. J. Climatol., 22, 1757-1767. https://doi.org/10.1002/joc.820.

3. CHAPTER 3: Machine learning methods for low-cost pollen monitoring – model optimisation and interpretability

This study is published in the journal *Science of the Total Environment*:

Mills, S.A., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023. Machine Learning Methods for Low-Cost Pollen Monitoring – Model Optimisation and Interpretability, 903, 165853. https://doi.org/10.1016/j.scitotenv.2023.165853.

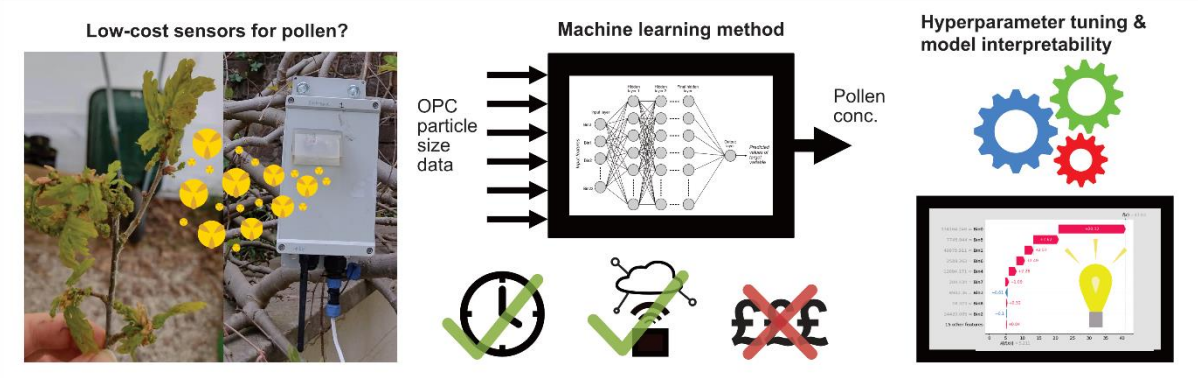


Figure 3.0: *Graphical abstract for Chapter 3.*

3.1 Abstract

Pollen is a major issue globally, causing as much as 40% of populations of industrialised countries to suffer from hay fever and other allergic conditions. Current techniques for monitoring pollen are either laborious and slow or very expensive, thus alternative methods are needed to provide timely and more localised information on airborne pollen concentrations. We have demonstrated previously that low-cost Optical Particle Counter (OPC) sensors can be used to estimate pollen concentrations when machine learning methods are used to process the data and learn the relationships between OPC output data and conventionally measured pollen concentrations.

Here we perform a thorough study which demonstrates how methodical hyperparameter tuning can be employed to significantly improve model performance. We present the results of a range of models based on tuned hyperparameter configurations trained to predict *Poaceae* (Barnhart), *Quercus* (L.), *Betula* (L.), *Pinus* (L.) and total pollen concentrations. We found 6/10 of models trained on hourly resolution data and 8/10 of those trained on daily resolution data achieved R^2 scores > 0.5. The results achieved here are a significant improvement on results we previously reported: the average R^2 scores for the total pollen models have at least doubled compared to using previous parameter settings.

Furthermore, we employ the explainable Artificial Intelligence (XAI) technique, SHAP, to interpret the models and understand how each of the input features (i.e. particle sizes) affect the estimated output concentration for each pollen type. In particular, we found that *Quercus* pollen has a strong positive correlation with particles of optical diameter 1.7-2.3 μm, which distinguishes it from other pollen types such as *Poaceae* and may suggest that type-specific subpollen particles are present in this size range.

There is much further work to be done, especially in training and testing models on data obtained across different environments to evaluate the extent of generalisability. Nevertheless, this work demonstrates the potential this method can offer for low-cost monitoring of pollen and the valuable insight we can gain from what the model has learned.

3.2 Introduction

Pollen is a major issue globally causing as much as 40% of industrialised country populations to suffer from hay fever and other allergic conditions (Fröhlich-Nowoisky et al., 2016). Only recently, Japan reportedly 'declared war' on pollen to combat the serious health and quality of life impacts it is having on the population (McCurry, 2023). An important step in addressing this issue is first being able to effectively monitor pollen, and afterwards to map and forecast its spatial distribution as a bioaerosol in the atmosphere. However, the current technology available for doing this is limited and in great need of technological advancements and new innovative methods (Buters et al., 2018; Oteros et al., 2020).

Pollen grains are produced as plant male microgametophytes (containing both vegetative and reproductive cells in varying numbers). Pollen that is most atmospherically relevant, since it is likely to travel the furthest as a free aerosol, would be from anemophilous (wind-pollinated) species. Pollen grains from such species are generally smaller than, for example, insect-pollinated species and have strong association with pollen allergies affecting the public. This includes tree types, such as birch (*Betula*, L.) and oak (*Quercus*, L.), grass (*Poaceae*, Barnhart), and weed types, such as ragweed (*Ambrosia*, L.).

This pollen dispersion accomplishes an important task to ensure successful reproduction when a large portion of it is released into the atmosphere, because they can reach new areas and disseminate their genes. As such, pollen is vital for terrestrial ecosystem regeneration but, during the transport process, it becomes atmospherically relevant too as a class of Primary Biological Aerosol Particle (PBAP). Among PBAPs, which includes fungal spores, bacteria, viruses, and other plant debris (see, e.g.,

Huffman et al., 2019), pollen grains are generally large in diameter (10-100 μm) and vary widely in shape and size across taxa and species (see, e.g., Després et al., 2012; Reponen, 2011, pp. 723, Bradley, 2015, pp. 408-409).

In the atmosphere, pollen grains can release subpollen particles, for example at high relative humidity conditions or during thunderstorms (Subba et al., 2023). These are often starch granules or other cytoplasmic debris with size ranges reported between 20 nm and 6.5 μm (Stone et al., 2021). Hendrickson et al. (2023) and Matthews et al. (2023) measured wind-stimulated emission factors of subpollen particles from oak (*Quercus*), ryegrass (*Lolium*, L.) and giant ragweed (*Ambrosia*). They reported emissions of 10^3-10^5 particles/pollen grain and 10^{13}-10^{15} particles/m^2 (for a given sample area populated by the species), for particles between 0.010-1.0 μm.

Subpollen particles often also carry allergens that affect human health (Bacsi et al., 2006; Smiljanic et al., 2017), while their size also allows them to penetrate more easily into the respiratory system. Meanwhile, pollen and subpollen particles have potential for cloud condensation nucleation (CCN) activity (Pope, 2010; Griffiths et al., 2012; Steiner et al., 2015; Mikhailov et al., 2019) and ice nucleation (IN) activity (Diehl et al., 2001; Diehl et al., 2002; Pummer et al., 2012; Tong et al., 2015; Dreischmeier et al., 2017; Gute and Abbatt, 2020; Burkart et al., 2021) which means they can affect cloud processes, weather and climate.

The current standard methodology for measuring airborne pollen is based on a manual Hirst-type trap to sample particles from the air, then for the particle-impacted melinex tape to be taken to the lab and viewed under the microscope by experienced scientists who count the number of pollen grains present for each type. This method is laborious and time-consuming, and data is not automatically nor quickly available. Other

instruments have more recently been developed to automate the process, such as the BAA500 (Oteros et al., 2015), the RapidE (Šaulienė et al., 2019), PollenSense (Jiang et al., 2022) and the Swisens Poleno (Chappuis et al., 2019; Huffman et al., 2019) (see Buters et al., 2022, for an overview). Recent work has focused on comparing their performance with the Hirst reference to address the current limitations in pollen monitoring capabilities (Crouzy et al., 2016, Oteros et al., 2020; Maya-Manzano et al., 2023).

Generally, these automated techniques involve machine learning methods, since they are designed to learn from many pollen samples the characteristics of each pollen type through whichever data medium they use. These can include optical light scattering or fluorescence data, but high-end instruments rely largely on image or holographic image data (e.g. Oteros et al., 2020; Sauvageat et al., 2020), i.e. training an algorithm to recognise the individual grain shapes as conventionally experienced scientists do by eye under the microscope for the Hirst. Naturally this requires certain hardware demands and complex computer vision models with stringent sample demands for training. Thus, these instruments can demonstrate good performance with an ability to accurately distinguish pollen grains, however they can also be bulky and expensive.

We recently demonstrated an alternative method for monitoring pollen that is low-cost and convenient in terms of mobility and real-time, remote data accessibility using Alphasense OPC-N3 sensors (see Mills et al., 2023). This method employed supervised machine learning (neural network and random forest) models which were trained, using Hirst data as a benchmark, to produce pollen concentrations from OPC sized particle concentrations. We demonstrated that this method showed promising results and could so far achieve coefficients of determination between model-predicted and Hirst-observed pollen concentrations of up to 0.67, but also highlighted how in

particular the neural network methods presented further opportunities for performance improvement.

Machine learning models are optimised to make the best predictions within the feature space representation it is given by the training data. While often demonstrating impressive performance, they can be described as 'black boxes' since it is difficult to truly understand why exactly the predicted outputs are being produced from the input data. This difficulty in human comprehension of the model is because: (a) the model is often too large (i.e., contains too many parameters) to be easily read and interpreted by humans; and (b) the model is not constrained by causal relationships, and so may appear to get the right answer for the 'wrong' reasons (see, e.g., Christianini, 2020, Cristianini, 2023). Being too large to read means that, even though the best description of a model is by definition the model itself, it is frequently useful to have smaller, simpler 'explanation models' that are an interpretable approximation of the original model (Lundberg and Lee, 2017). Having not correctly learned the causal relationships could potentially be an issue when applying a fitted model to a context or environment outside that from which the training data came. A comprehensive review explaining these points in an environmental (ecological) context, can be found by Pichler and Hartig (2023). Thus, for the sake of making decisions based on such models, it appears important to investigate the relationships that the models are learning (we return to this point in the concluding discussion). This can be valuable for deciding how much to trust a certain model, for investigating potentially unknown variable relationships, or diagnosing models as to which features are serving usefully or perhaps unhelpfully to the model.

There are recently developed techniques for explaining machine learning models and interpreting outputs (often called 'explainable artificial intelligence', or XAI), an

overview of which can be found by Gohel et al. (2021). The XAI technique that is utilised in this work is 'SHAP' (SHapley Additive exPlanations), developed by Lundberg and Lee (2017). This is a method inspired by game theory, where Shapley values are calculated to explain the output of a machine learning model by considering all possible combinations ('coalitions') of input features ('players'), quantifying and aggregating the marginal contributions of each feature towards the final output.

In this study, we continue to develop the neural network method detailed in our previous work (Mills et al., 2023) for low-cost monitoring of pollen. While the machine learning methods yielded the best results out of those trialled in our study, these methods are not without limitations and should be used only with appropriate consideration of the context. As discussed, these methods provide a trade-off, casting aside interpretability of causal inference in favour of optimal predictive performance based solely on the statistical patterns the model learns (Chistianini, 2010). The learning capability of these algorithms, especially deep learning neural networks, depend heavily on data quantity and representative quality. The dataset used in this study is relatively small (6220 data points) for deep learning purposes and covers various pollen seasons that each are absent, emerge, disappear, and overlap for different segments of the time series. Meanwhile, the sensors we use are designed to count general aerosols (i.e. particulate matter) and the task we give the algorithm is to isolate and measure pollen, a small fraction of total ambient aerosols. A further limitation is that the benchmark concentrations we use to teach the algorithm have their own associated uncertainties (Adamov et al., 2021), yet is the best standard we currently have for determining pollen concentrations.

Despite these limitations, our previous work (Mills et al., 2023) showed that machine learning methods are able to learn useful information from this data for public health purposes. In particular, we demonstrated that, on classifying pollen concentrations into 'high' and 'low' categories, we can achieve promising accuracy (F1 scores) with a low false negative rate.

Machine learning has become such a popular and broad field that yields impressive results because the algorithms are generic and can be chosen and applied for a huge variety of tasks, yet come with great flexibility and potential for external optimisation of hyperparameters (on top of the intrinsic optimisation of parameters the model performs itself on training). Control of hyperparameters can determine whether the model underfits, overfits or finds the sweet spot in between. A model underfits when it has low complexity and does not learn effective patterns from training data. It is quite possible for machine learning algorithms with high complexity to overfit, by learning the training data too well that it generalises poorly for unseen (i.e. test) data. Hyperparameters can be used to improve model learning capability but also for regularisation. There are various established regularisation techniques, which can be controlled by hyperparameters and constrain model complexity in a flexible way that adjusts to the input data. Further explanation on the roles of specific hyperparameters relevant to this study can be found in the Appendices extended methods section 8.2).

In this work, we apply a methodological approach to tune various hyperparameters, including regularisation techniques, with an aim to improve predictive performance and generalisability on unseen data. (In this case, we measure generalisability as being able to make accurate predictions on a hold-out test dataset from the same experimental context. We do not test generalisability of the method on data from a different campaign

context.) Selecting the optimal model configuration, we demonstrate the significant improvement in performance metrics this yields. We also extend the method to further pollen types, including *Quercus* (oak), *Betula* (birch) and *Pinus* (L., pine), beyond those reported in the previous work ('total' pollen and *Poaceae* (grass)). The XAI SHAP method is used to interpret and diagnose models for each pollen target variable, to assess which particle size ranges from the OPC are responsible for 'pushing up' or 'pulling down' predicted outputs. We investigate the relationship between OPC bin particle concentrations (input feature values), impact on model (SHAP values) and events of high and low pollen concentration when the model performed with high and low accuracy. We conclude by revisiting the rationale for ML in this domain and outline a pathway for further improvement.

3.3 Methods

3.3.1 Context and Instrumentation

The context and data used for this study are the same as detailed in previous works, Maya-Manzano et al. (2023) and Mills et al. (2023), from the EUMETNET AutoPollen – ADOPT COST Action (CA18226) Intercomparison Campaign 2021 at the Centre of Allergy & Environment (ZAUM) in Munich, Germany. The data input given to the machine learning models from which to predict pollen concentrations was obtained collectively from three Alphasense OPC-N3 sensors logging data between 9th March and 7th July 2021.

These commercially available OPC sensors count and categorise the size of particles as they pass through a 685 nm-wavelength laser beam, making measurements based on scattered light intensity and Mie scattering theory. A refractive index of 1.5 and particle density of 1.65 g mL^{-1} is assumed. Particle counts are monitored for 24 different bins of varying size ranges between 0.34 and 40 μm. The sensors were functionalised for outdoor and automatic use, as described in our previous work (Mills et al., 2023), and particle number concentrations (grains m^{-3}) were calculated from raw particle counts for each bin using measured sample period (mean: 5.0 s) and flow rate (mean: 5.2 L min^{-1}) values. The corresponding size ranges for each of the OPC bins can be found in Table 8.1 in the Appendices.

The target data given to the models, to calculate the error of predicted outputs from 'actual' values, was averaged from four Hirst-type samplers (Burkard Manufacturing Co Lid, Rickmansworth, UK), collocated with the OPC sensors. The Hirst data obtained for the intercomparison campaign included number concentrations for 17 pollen types - *Alnus* (Mill., alder), *Ambrosia* (L., ragweed), *Artemisia* (L., mugwort), *Betula* (L., birch),

Carpinus (L., hornbeam), *Corylus* (L., hazel), *Fagus* (L., beech), *Fraxinus* (Tourn. ex L., ash), *Picea* (A. Dietr., spruce), *Pinus* (L., pine), *Plantago* (L., plantain), *Poaceae* (Barnhart, grass), *Populus* (L., poplar), *Quercus* (L., oak), *Taxaceae-Cupressaceae* (Gray, yew), *Tilia* (L., lime), *Urtica* (L., nettle) – and varia (other pollen that were not included in the previous list or unknown pollen) and for 'total' pollen which was the sum of all types present at any given time. Full details on the context of the whole campaign can be found in Maya-Manzano et al. (2023) and further details on the Hirst observations are available in Triviño et al. (2023).

3.3.2 Neural network hyperparameter tuning

Scripts for external (hyperparameter) optimisation were run in Python 3.9.6 using the University of Birmingham's BlueBEAR Supercomputing facility. The module versions Scikit-learn 1.0.1, Tensorflow 2.8.4, Matplotlib 3.4.3, Numpy 1.21.3 and Pandas 1.3.4 were utilised. Model architecture was initially constructed with the same hyperparameters as demonstrated in our previous work (Mills et al., 2023) and then a range of values for selected hyperparameters were tested. However, this time the original dataset was split into training, validation and test datasets (as opposed to just training and test datasets). The training dataset was used for the model to directly train on. The validation dataset was used to determine a validation score by the loss metric (Mean Squared Error (MSE)) for a hold-out dataset for each training iteration (epoch) alongside a training score. The validation score was used to inform the model when to stop updating the parameters, i.e. when the mean squared error (between predictions and observations) had stopped decreasing further. The test dataset was used to calculate metrics, including Root Mean Squared Error (RMSE) and R^2 score (coefficient of determination), to evaluate the various models and effect of varying

hyperparameters. The external optimisation procedure was computationally expensive, involving the training of many models, and therefore was only applied to one input feature-target variable pairing as a demonstration for this work. Thus, the specific optimisation results reported here apply to the model which is trained to predict total pollen concentrations from the 24 particle size bins available from the OPC data (i.e. not including temperature and relative humidity). We assume going forward that the model tasks are generally similar enough to benefit from hyperparameters tuned for just one of the models.

The tested hyperparameters were as follows: number of layers and number of nodes in each layer (i.e. model capacity), batch size, optimisation algorithm, initial learning rate of best optimisation algorithm, activation function at hidden layer nodes, scaling of target variable, train:validation:test datasets split size, proportion of dropout nodes in each layer, weight constraints applied in each layer, and implementation of noise at different configurations within the network. A summary of the values trialled for each hyperparameter is presented in Table 3.1.

An extended methods section containing further details on each hyperparameter and the configurations chosen for this study is provided in the Appendices (section 8.2). Further explanations on the effect each hyperparameter has on a given model and other general information can be found in the book 'Deep Learning' by Goodfellow et al. (2017) and ebook 'Better Deep Learning' by Brownlee (2018) as well as other resources by Bengio (2012), Masters and Luschi (2018), and Reed and Marks (1999).

For testing, 5 distinct splits of training, validation and test data were created randomly (with the same random seeds for each hyperparameter test) from the whole dataset. Models were then trained and evaluated 5 times on each of the 5 sets of train-validation-

test data for each hyperparameter configuration. Thus, each hyperparameter configuration was sampled 25 different times in total, over 5 different combinations of training, validation and test datasets (visualised in Panel 1). Due to the stochastic nature of machine learning algorithms they can produce different results on multiple runs, and they are also sensitive to the provided train/test data. We followed this method to allow for variation caused by differences in train/test datasets and the stochastic process to truly assess the effect of varying each hyperparameter.

***Table 3.1**. Summary table of hyperparameter configurations trialled for this study. *Initial learning rate was only trialled for chosen optimisation algorithm (*Adam*).*

Hyperparameter	Trialled values
Number of layers	3; 4; 5; 6; 7
Number of nodes	50; 100; 200; 400
Batch size	32; 64; 128; 256; 512
Optimisation algorithm	*RMSProp* (Root Mean Square Propagation); *Adagrad* (Adaptive Gradient Algorithm); *Adam* (Adaptive Moment Estimation);
Initial learning rate*	0.0001; 0.0005; 0.001; 0.005
Activation function	ReLU; ELU; SELU
Scaling of target variable	True; False
Train:Validation:Test dataset ratio	50:25:25; 60:20:20; 70:15:15; 80:10:10
Dropout node proportion	0.1; 0.2; 0.3; 0.4; 0.5
Weight constraint	Max-norms 1.0; 2.0; 3.0; 4.0; 5.0 and unit norm.
Noise implementation	None; input layer; hidden layers
Noise standard deviation	0.001; 0.0001; 0.00001

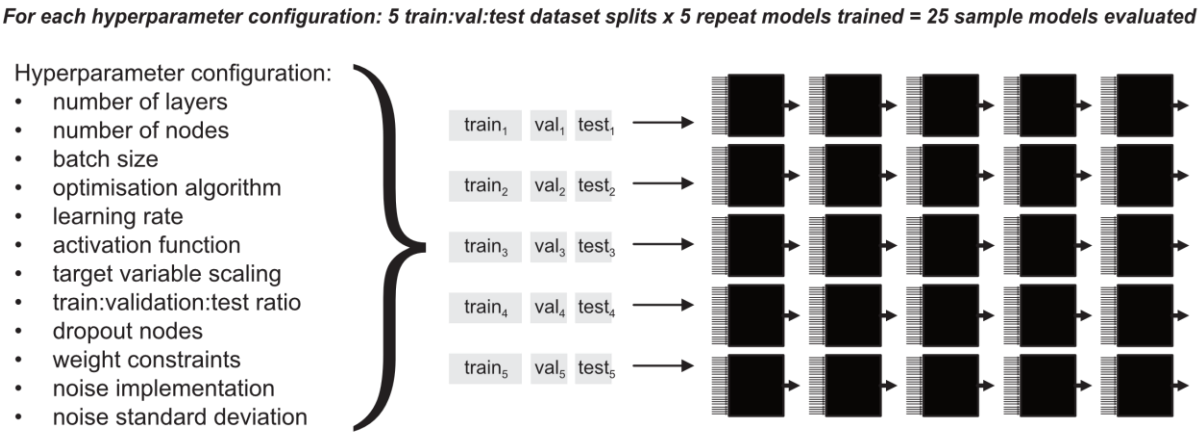


***Figure 3.1:** Methodology by which various hyperparameter configurations were tested. Five repeat models were trained for five different train:validation:test dataset splits, resulting in a total of 25 sample models for each hyperparameter configuration.*

A metrics dataset was produced from each hyperparameter experiment – including RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Spearman correlation coefficient values and R^2 scores for model-predicted values vs Hirst-measured pollen concentrations based on the test dataset in each case. These results were analysed and visualised using Python 3.9.7 in Jupyter Notebook from the Anaconda Distribution with module versions Numpy 1.24.1, Pandas 1.5.3, Matplotlib 3.6.3 and Seaborn 0.12.2. Box-and-whisker plots were produced to summarise the R^2 scores achieved across all 25 sample models for each hyperparameter configuration for each experiment. R^2 scores were chosen as the primary metric for comparison, since they provide better comparison across models with different target value ranges than RMSE. This was also the metric used to assess the performance of all instruments that took part in the AutoPollen intercomparison campaign (see Maya-Manzano et al., 2023).

3.3.3 Final models

The results from the hyperparameter tuning experiments were used to inform the configurations applied to the final optimised models. Final models were trained for 3 sets of input features – OPC bins only, OPC bins, RH and temperature, and RH and temperature only – and 5 different target variables – total pollen, *Poaceae*, *Quercus*, *Betula* and *Pinus* pollen categories. These pollen types were chosen for their relevance to public health (see, e.g., Darrow et al., 2012) and their prevalence in the dataset of this environment. The same hyperparameter configurations were used for all models, with the exception of train:validation:test ratio, which was 80:10:10 for total pollen and *Poaceae* models and 60:20:20 for the rest. This was because, although the 80:10:10 split yielded optimal performance for the total pollen model that underwent hyperparameter tuning, for *Quercus* and other less prevalent types in this case the test dataset seemed

not to be large enough to be representative. A greater test dataset size reduces the proportion that can be used for training but ensures that the model is evaluated fairly on a representative sample.

This combination of input features and target variables produced 15 different model sets, each of which was trained for 5 sample models within each set. The model hyperparameters within each set were identical however, due to the stochastic nature of the learning process, the 'optimal' model weights (parameters) settled on in each case will be different depending on the statistical patterns each model has learned. R^2 scores were calculated for all models between relevant target variables and model predictions from the unseen test dataset. The resulting spread of R^2 scores for each model set was visualised and the mean and best scores recorded. We can say that models which achieved high R^2 scores generalised well in terms of the hold-out test dataset here, but we cannot comment on how well they would generalise in other contexts. For example, the models with the best R^2 scores here may not be those with the best R^2 (and therefore generalisability) in another context, though this is unlikely to be models with very low scores here. If the spread of scores among a model set is narrow, it suggests any model with these hyperparameter configurations will result in similar performance and hence these hyperparameter configurations may be more robust. If there is great variance among R^2 scores, it may suggest the hyperparameter configuration is not very robust and learning outcome depends heavily on the stochastic learning process.

This process was also applied to produce an equivalent set of models, predictions and R^2 scores for models trained on daily (24h) time resolution data. This was to observe the difference in performance between models trained on hourly and daily resolution data. It was also in preparation to apply for future comparison purposes, since pollen

concentration data from the standard Hirst-type samplers is often only available in daily time resolutions.

3.3.4 Model interpretation using SHAP values

To explain the models, understand the input feature-output variable relationships and why certain predictions are produced, we employed the SHAP (Shapley Additive exPlanations) method developed by Lundberg and Lee (2017). SHAP uses Shapley values from game theory to explain the output of any machine learning model. It does this by considering all possible combinations, or 'coalitions', of input features ('players' from a game theory perspective), quantifying the marginal contributions each feature makes towards the final output when added to coalitions, then aggregating the marginal contributions. The sum of SHAP values across all features for any given observation equates to the difference between the model prediction and the observation mean, or 'null model'. This process involves training models for each distinct feature coalition to calculate differences in predicted output (marginal contribution) and so is computationally expensive. SHAP values have limitations when feature dependencies are present and cannot simply be taken for causal inference; it tells us how important the feature is to the model. Nevertheless, it is a powerful method for interpreting machine learning models and can be implemented efficiently using Lundberg's SHAP library (https://github.com/slungberg/shap).

We took the best performing models (highest R^2 score on test dataset) for each of our model sets and applied the SHAP method to investigate how each of the given features were affecting the output of the model. 'Bee swarm' plots were constructed in each case, which give information on feature importance, magnitude of the affect each feature has on the model output, and the corresponding magnitude of the input feature value in

each case. The information from these plots was condensed into a matrix plot which scored input feature and target variable pairings by colour for strong positive correlation, weak positive correlation, strong negative correlation, weak negative correlation and no correlation respectively. Correlations were nonlinear and decided for each case based on the visual distribution of data points on the plot demonstrating feature value and SHAP value (i.e. impact on output prediction of target variable).

To investigate further, we isolated data points that fit into four categories: high pollen low error (HPLE), low pollen low error (LPLE), high pollen high error (HPHE) and low pollen high error (LPHE) events - after excluding all zero values for the target variable. High and low pollen was defined as all data points above the 75th quantile and below the 25th quantile for the Hirst-measured target pollen variable and high and low error was defined as above the 75th and below the 25th quantile for the absolute difference between Hirst-observed and model-predicted target variable concentrations. The only exception for this was that for *Betula* the low error threshold was raised to the 40th quantile for the HPLE category because there were no data points available satisfying the 25th quantile threshold for high (> 75th quantile) *Quercus* concentrations. The percentage of non-zero values for each target variable in the whole and test datasets, the number of data points that were filtered by each category threshold, and the specific quantile threshold values calculated in each case can all be found in the Appendices in Tables 8.3-8.5.

The original feature (bin) values and SHAP values were averaged over all data points present in each group. Bar plots were constructed to show the relative averaged feature values for each category, and a colour scale added corresponding to SHAP value sign and magnitude.

3.4 Results and Discussion

3.4.1 Hyperparameter tuning

The results of the optimisation experiments are summarised in Figure 3.2, based on the R^2 scores calculated from the test datasets across the 25 samples for each configuration. This visualises the effects of varying the model capacity, by the number of layers and nodes per layer in the neural network architectures, and model learning behaviour by varying batch size, optimisation algorithm, learning rate and node activation functions. Also visualised are the effects of certain data preparation considerations, such as scaling of the target variable and the test-train split size, as well as the effects of varying regularisation methods to improve generalisation such as dropout percentage, weight constraints and addition of noise.

In theory, there would be an optimal capacity of the model architecture, defined by the number of layers and nodes, with enough complexity to address the given task yet not so much as to learn the training dataset too meticulously (i.e., to be fitted but not overfitted). Increasing the number of layers is computationally cheaper than increasing the number of nodes in a layer, yet the optimal capacity in practice can only be found empirically, keeping computational limitations also in mind. The results from our empirical experiments here in plot A of Figure 3.2 demonstrate the general increase in model performance with increasing number of layers and nodes, but the increase with additional layers diminishes after 5 layers for a large number of nodes ($N \geq 200$). The configuration that achieved the highest mean R^2 score (0.24) across all model samples was the combination of 5 layers and 400 nodes; thus this combination was chosen to proceed. (Other combination means also achieved close to this (>0.22), e.g. 6 or 7 layers with 200 or 400 nodes, but these were also more computationally expensive.)

Batch size dictates how often the weights of the model are updated during training, thus affecting training dynamics. No clear trend was observed across different batch sizes. The highest mean across models (see plot B in Figure 3.2) was achieved with a batch size of 256 (0.22), however, the difference among other batch size means was not large (<0.11). The results using the commonly used batch size of 32 showed the smallest range and highest minimum, and therefore was chosen to proceed for further models.

The *Adam* optimiser (R^2 = 0.17) (see plot C in Figure 3.2) clearly performed better than the *RMSprop* (R^2 = 0.06) or *Adagrad* (R^2 = 0.04) alternatives, when used on default settings. When the initial learning rate was varied for the *Adam* optimiser (see plot D in Figure 3.2) an optimal value of 0.0005 was found (R^2 = 0.22). The standard, and less computationally expensive, ReLU activation function applied at each hidden layer in the network proved to yield generally higher R^2 scores (0.19) than either of ELU (0.11) or SELU (0.15) (see plot E in Figure 3.2). This suggests that the extra nuance these activation functions can add (e.g. decreasing bias shift, self-normalisation, and robustness – see Pedamonti et al., 2018; Marchisio et al., 2018) did not offer an advantage. On the contrary, it appears that the characteristic of the ReLU activation function to equate all negative incoming values to zero (while ELU and SELU continue to propagate negative values) has benefited the algorithm in this case. The best performing configuration in each case was taken to proceed.

Previous work during the development of these models (in Mills et al., 2023) showed that scaling (normalisation) of the particle concentration input data is important for successful model training with good predictive performance. However, it was not previously conclusive whether scaling (normalising) the pollen concentration target variable was useful or not. The results here (see plot F in Figure 3.2) demonstrate that

scaling the target variable in this case is not useful (R^2 = -0.24), but rather leaving it unscaled results in generally better (0.20) and less variable performance.

While some testing had been tried before, we decided to investigate more thoroughly the potential effect that the train-test split size had on estimated model performance. We considered this useful to test as the available data here is limited and therefore wanted to make best use of the available data to provide effectively for both model training and evaluation. The varied proportions of the whole dataset taken for the training dataset are labelled in the x-axis of plot G in Figure 3.2. It is important to note that the remaining proportion of the dataset was then split once more, in equal parts, into validation (used to settle on the optimal model over training epochs) and test datasets. The latter of which was used for final model evaluation.

From the results, we found that using 80% for the training dataset, and therefore 10% each for validation and test datasets, yielded the best performance (R^2 = 0.25). This split was used for further models predicting total pollen and *Poaceae* pollen, however it was observed that this did not generally result in best performance for models targeting *Quercus* pollen. This is likely because the representation (i.e. season length) of the *Quercus* pollen season in the given dataset is relatively small, at least compared to 'total' or *Poaceae* pollen, so a smaller dataset proportion used for evaluation would less likely be representative. After testing splits of 80:10:10 and 60:20:20 (train:validation:test) it was decided that the latter was best when applied to *Quercus*, *Betula* and *Pinus* models. (This split ratio also achieved the second best R^2 score (0.20) for the original total pollen hyperparameter testing experiment.)

Dropout layers and weight constraints are both effective methods for model regularisation and reducing overfitting, but again the optimal tuning of dropout node

percentage and threshold values for weight constraints can generally only be determined empirically. We found for the purposes of our task that a dropout percentage of 20%, see plot H in Figure 3.2, and weights constrained by a max-norm (upper value threshold) of 4.0, see plot I in Figure 3.2, at each hidden layer yielded optimal performance (R^2 scores of 0.20 and 0.33 respectively). These configurations were applied to further models.

Finally, the addition of a small amount of noise at input or hidden layers was trialled to see if it affected the ability of the model to accurately predict the unseen test dataset (in theory by limiting dependency on training data and improving generalisability). Our results found that the addition of noise to the data at the input layer with a standard deviation of 0.001 improved model performance on the test dataset ($R^2 = 0.24$), compared to no noise being added ($R^2 = 0.19$), and among all other tested noise configurations. Thus, this addition of noise to input data was also applied to further models.

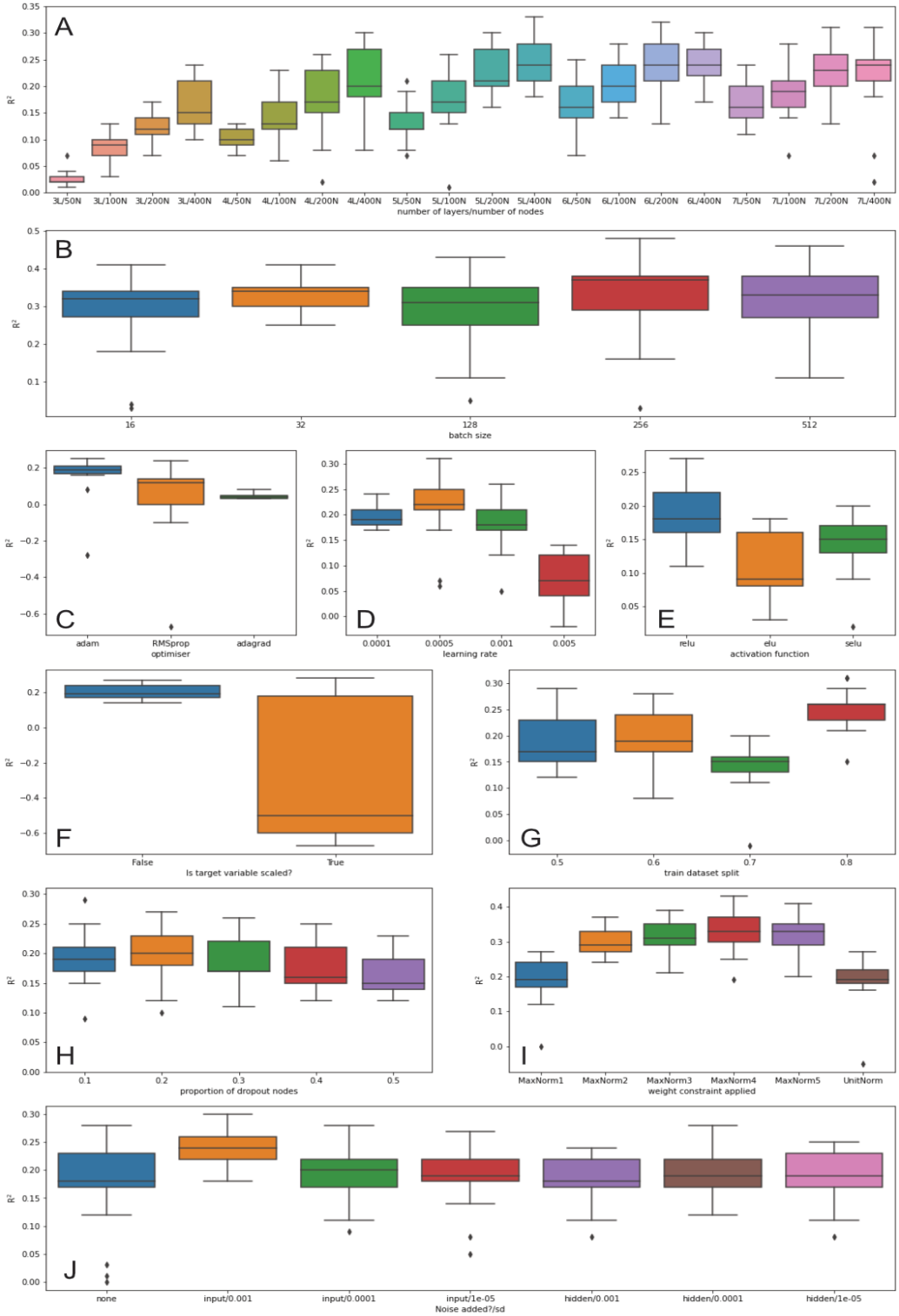


Figure 3.2: *Effect of varying hyperparameters on model performance, quantified by the R^2 score. A: model capacity, L = number of layers and N = number of nodes. B: batch size. C: optimisation algorithm. D: initial learning rate. E: activation function. F: target variable scaling. G: train:validation:test split ratio – x-axis value corresponds to the fraction of the whole dataset used for training, while validation and test datasets where taken in equal parts from the remaining fraction. H: dropout node proportion. I: weight constraint. J: noise implementation (none/input layer/hidden layers) and standard deviation applied. Colours are unique for each x-axis/hyperparameter value.*

3.4.2 Improved models from chosen hyperparameter values

The plots in Figure 3.3 present the results of the final model evaluations after the optimisation process for hourly and daily resolution models respectively, while the mean and best R^2 scores achieved for each model set are summarised in the table in Figure 3.3. We evaluate, in particular, the models that used OPC particle size bins data since it is that sensor-ML combination that we are primarily concerned with. The models based solely on meteorological variable input feature are provided for comparison to demonstrate the importance of the particle size bin data for the models.

The results achieved here are a significant improvement to the initial results reported by Mills et al. (2023), denoted by the horizontal white dotted lines. The average R^2 scores for the total pollen models have at least doubled - from 0.15 to 0.49 without meteorological variables and 0.3 to 0.65 with them - and *Poaceae* models improved from 0.34 to 0.48 without, and 0.67 to 0.87 with, meteorological input features. The greater improvement in total pollen models demonstrates the gain of hyperparameter tuning for specific tasks. These results also demonstrate the method being applied to a greater variety of pollen types, which appears to show similar promise provided appropriate data availability and model optimisation.

From the hourly and daily model sets (blue and orange only), 6/10 and 8/10 respectively produced a model that achieved an R^2 score ≥ 0.5 and there are even some models that have performed extremely well ($R^2 \geq 0.8$) when evaluated on the test dataset. The daily resolution models generally outperformed the hourly resolution models. It has been reported previously (Maya-Manzano et al., 2023) that comparison alongside the Hirst-type sampler generally yields better results at coarser time resolutions, since the temporal averaging results in less deviation between instruments.

Models targeting total and *Poaceae* pollen generally performed better than others. This is likely due to the better representation of these target variables within the time period for which data was collected and the fact that a larger proportion of the dataset was used for training (80% as opposed to 60% for the rest). In general, as more data is fed to machine learning models to learn from their prediction performance increases. However, we had a limited supply of data here. While total pollen models demonstrate the most improvement, since this target variable was used for hyperparameter tuning, it is interesting that the task of predicting *Poaceae* pollen also benefited from this process. The other types may also have benefitted to an extent, yet these models could benefit further by tuning hyperparameters for each pollen type.

There were some instances, in particular for the *Betula* model with meteorological input features, where the R^2 score was small or negative, implying that these models did not learn reliably useful information. This may be possible to improve upon by applying hyperparameter tuning for this specific model case, or it may simply be due to limitations of the available data. However, since the equivalent daily-resolution trained model appears to have performed considerably well, it does suggest there is useful information from which to learn the *Betula* type.

In the overview study which summarised the blind performance of all instruments participating in the intercomparison campaign (Maya-Manzano et al., 2023), an R^2 of >0.5 was considered good while >0.75 was considered excellent agreement with the benchmark Hirst instruments. Measuring total pollen, 9/18 of the automated pollen monitoring systems achieved $R^2 > 0.5$ for 3-hourly and daily resolution, while 3/18 achieved $R^2 > 0.75$ for daily resolution. Those instruments that performed best included the BAA500 and the Swisens Poleno. These instruments (depending on the algorithm

applied) also demonstrated R^2 scores above 0.5 for daily and above 0.75 for 3-hourly *Poaceae* pollen concentrations. R^2 scores of above 0.75 were observed for 3-hourly and daily measurements of *Quercus* and *Betula* pollen, the latter with the most success.

Our results from this study applying low-cost sensors seem impressive, with over half of our models achieving $R^2 > 0.5$ and some even > 0.75. However, it should be noted that these algorithms are not blind and have been specifically trained on the data in this context. Meanwhile, the results reported in the overview paper are for instruments that use machine learning algorithms trained independently from the Hirst data, on specific pollen samples previously provided to them.

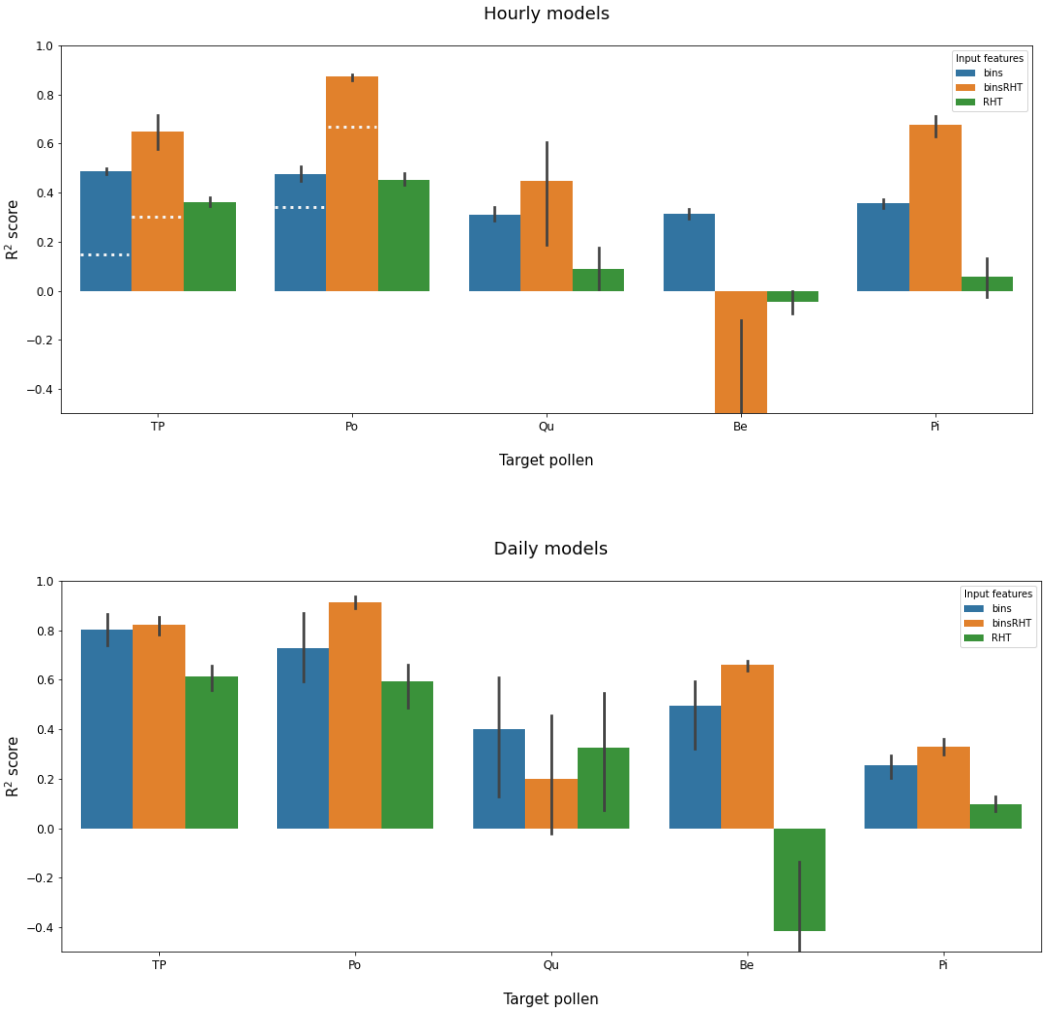


Figure 3.3: *Summary of R^2 scores achieved for final hourly (above) and daily (below) resolution models for each target pollen type. TP = Total pollen; Po =* Poaceae *(Grass); Qu =* Quercus *(Oak); Be =* Betula *(Birch); Pi =* Pinus *(Pine). Horizontal white dotted lines denote scores achieved in previous study for models tested there (Mills et al., 2023). Error bars denote the 95% confidence interval among the 5 model samples for each set, constructed as described in the Seaborn documentation (https://seaborn.pydata.org/tutorial/error_bars.html, accessed June, 2023).*

Table 3.2: *Summary of R^2 scores achieved for all final models, both hourly and daily time resolution. TP = total pollen; Po =* Poaceae*; Qu =* Quercus*; Be =* Betula*; Pi =* Pinus*. Blue = models with only OPC bin input features; orange = models with OPC bin & meteorological input features; green = models with only meteorological input features.*

Model set	Hourly mean R^2 score	Hourly best R^2 score	Daily mean R^2 score	Daily best R^2 score
binsTP	0.49	0.50	0.80	0.90
binsRHTTP	0.65	0.73	0.82	0.87
RHTTP	0.36	0.39	0.61	0.68
binsPo	0.48	0.53	0.73	0.95
binsRHTPo	0.87	0.89	0.91	0.95
RHTPo	0.45	0.50	0.59	0.67
binsQu	0.31	0.36	0.40	0.67
binsRHTQu	0.45	0.62	0.20	0.66
RHTQu	0.09	0.22	0.33	0.66
binsBe	0.31	0.36	0.50	0.61
binsRHTBe	-0.53	-0.05	0.66	0.68
RHTBe	-0.05	0.04	-0.42	-0.01
binsPi	0.36	0.38	0.25	0.31
binsRHTPi	0.68	0.73	0.33	0.38
RHTPi	0.06	0.16	0.10	0.15

3.4.3 Model interpretation and input-output relationships

Figure 3.4 shows bee swarm plots from the Python SHAP package for the test datasets for each pollen type. The relationships learned by the model are generally not linear and are complex to interpret (see Introduction, above, and Lundberg and Lee, 2017), but our aim here is to discern which particle size ranges generally increased when a given pollen type was present (i.e., find and interpret an 'explanation model' for that purpose). Strong SHAP signals may relate to the size ranges of subpollen particles but we have no way of demonstrating this. We also aim to distinguish differences among the types that the model has learnt, in order to assess the feasibility of differentiating between types.

The general observed correlations based on the bee swarm plots are summarised in the matrix plot in Figure 3.4. Corresponding bee swarm plots for models based on OPC bins and meteorological variables are displayed in Fig. 8.1 and partial dependence plots for each input bin and target pollen type can be found in Fig. 8.2-8.6 in the Appendices.

While the meteorological variables were generally very useful for predicting pollen concentration, for models where they were included the correlations among the OPC bin ranges appear less clear. Therefore, for investigating the impact of the different bin sizes in each case we have focused on the models trained on OPC bins only.

***Figure 3.4**: SHAP 'bee swarm' plots for all pollen types bins-only models. The features are ranked in order of general importance and each point corresponds to a datapoint in the test dataset. The x axis corresponds to the impact each datapoint and feature had on the model output (i.e. target pollen concentration), relative to the mean output at 0. The colour corresponds to the actual value of that feature – red for high particle concentrations in a given bin and blue for low values. Therefore, feature rows which generally have more red further to the right and blue further to the left on the x axis likely have some general positive correlation between feature value and SHAP values (impact on the model) and vice versa for negative correlation. Corresponding bee swarm plots for models based on OPC bins and meteorological variables are displayed in Figure 8.1 and partial dependence plots for each input bin and target pollen types can be found in Figures 8.2-8.6 in the Appendices.*

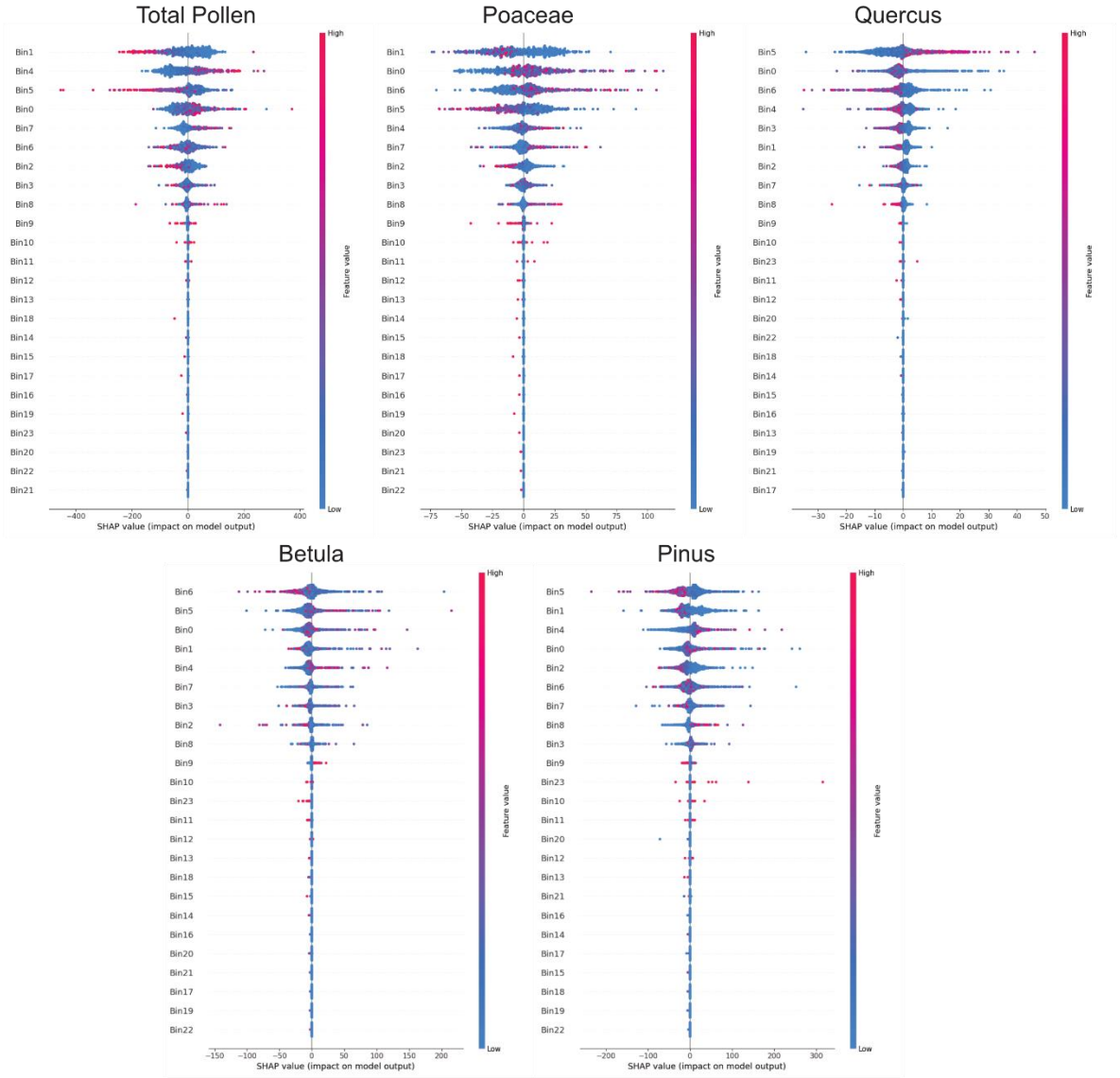


Figure 3.5: *Matrix plot summarising observable nonlinear correlations between input features and target pollen variables based on SHAP bee swarm plots in Figure 3.4.*

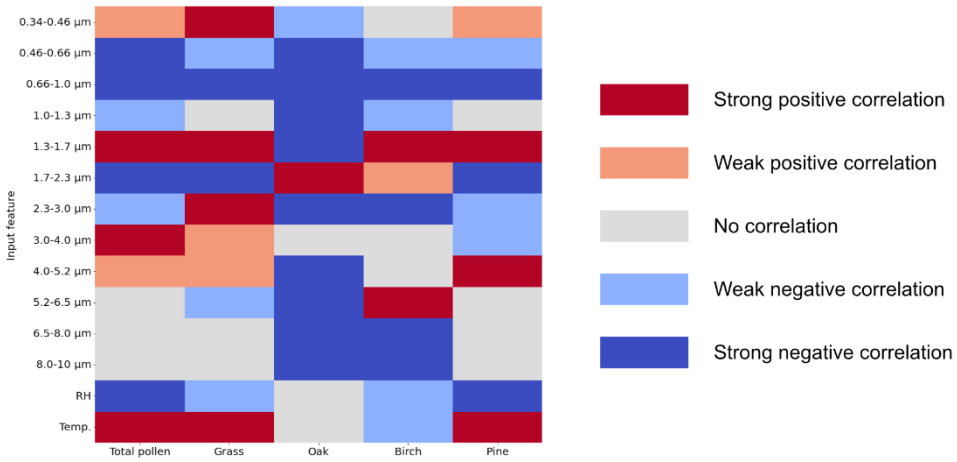
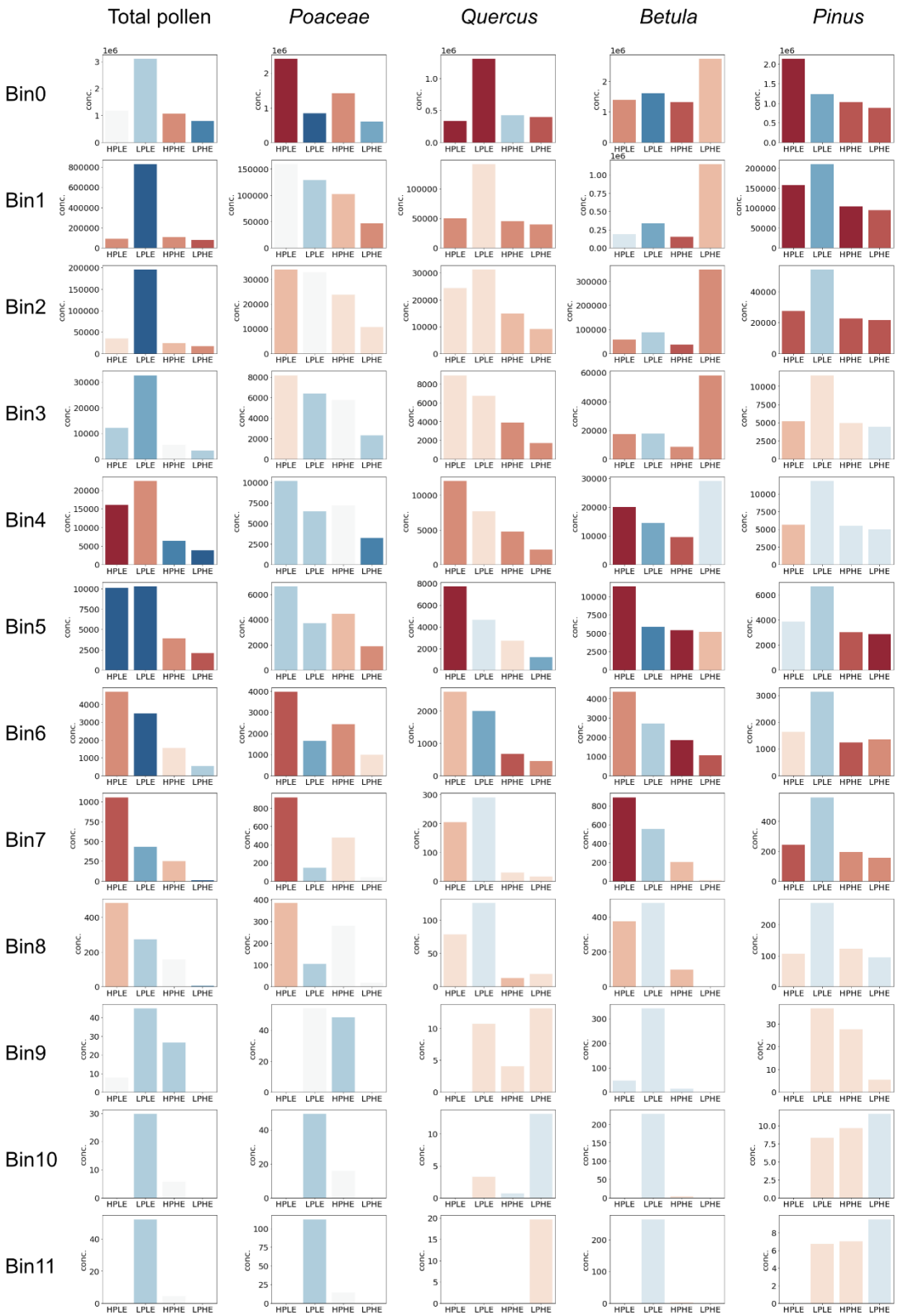


Figure 3.6 shows bar plots visualising relative averaged particle concentrations for each OPC bin for four categories - high pollen low error (HPLE), low pollen low error (LPLE), high pollen high error (HPHE) and low pollen high error (LPHE) events. The colour corresponds to the size and magnitude of the averaged SHAP value. Bins beyond bin 11 (> 10 µm) have not been included since few particles were detected in that size range and the bee swarm plots did not provide useful information. Generally, when particles were observed in these larger-size bins, it pushed the predicted output of the model down, as can be seen for larger-size bins for the total pollen, *Poaceae*, and *Betula* plots.

These plots can give further evidence about which features are affecting the model and are correlated with certain events such as high pollen. For example, a taller bar of a strong red colour for the HPLE category (e.g. bin 0 for *Poaceae*) compared beside a shorter bar of a blue colour for LPLE suggests that high values of this feature contributed greatly to low error high pollen events while comparatively smaller values were also indicative of low error low pollen events. We may also be able to suspect

features that might be confounding the model and be responsible for inaccurate predictions. In particular, strong blue bars for the HPHE category may suggest this feature is pulling the model predictions down and creating a large error from the high output value it should be. Likewise, strong red bars for the LPHE category (e.g. bin 5 for *Pinus*) may suggest the feature is pushing the model predictions up even when the actual pollen concentrations are low, potentially confusing the model.

Figure 3.6: *Bar plots showing feature value for bins 0-11 for HPLE (high pollen low error), LPLE (low pollen high error), HPHE (high pollen high error) and LPHE (low pollen high error) events for each target. Bar colour represents an arbitrary scale for the corresponding SHAP value (i.e. impact on the model output), with a stronger red meaning the given feature value pushed the model output up and vice versa for blue. The particle sizes sensed by each bin are tabulated in Table 8.1: the particle size range increases with bin number.*



3.4.4 Total Pollen

From the bee swarm and matrix plots in Figures 3.4 and 3.5, it appears that total pollen had some positive correlation with bins 0, 4, 7 and 8 (corresponding particle sizes can be found in Table 8.1 in the Appendices). From the bar plots in Figure 3.6, comparing high/low pollen events with high/low error it appears that bins 0-3, 5 and 9-11 had large responsibility for pushing the predicted pollen concentration down when particles in these bins were at high concentrations. Meanwhile bins 6-8 are responsible for pushing the pollen concentrations up with high particle concentrations present in these bins. We conclude that total pollen here is largely correlated with particles, measured optically by the OPC, in the size range 2.3-5.2 μm.

The model appears to have associated bins 1 and 5 with a strong negative correlation to total pollen. For example, when these particle concentrations are high they have contributed strongly to predicting low pollen events (i.e. strong blue bars for LPLE). However, the issue is that when the pollen concentrations have been low it has contributed strongly to predicting that pollen is high when it is actually not (i.e. strong red bars for LPHE). This suggests that these bins are causing the model to learn unhelpful correlations and it might be beneficial to remove them as input features.

The total pollen target variable encompasses a wide range of pollen types which may be associated with varying particle sizes and would be very dependent on the representation of 'total pollen' in this context (e.g. majority of *Poaceae* representation in this context). Therefore it can be expected that there may be conflicting correlations with different types within this target variable and the model may not generalise well in other contexts where the representation of 'total pollen' is different. It could be expected to generalise less well than other specific pollen type models in other contexts.

3.4.5 *Poaceae*

From the bee swarm and matrix plots, the *Poaceae* type appears to have positive correlation with bins 0, 4, and 6-8. From the bar plots, bins 0 and 6-8 are particularly congruent with this positive correlation. This suggests that particles in size ranges 0.35-0.46 and 2.3-5.2 μm are largely co-existent with *Poaceae* pollen in this context, similar to the observed size range for total pollen. Bins 1 and 5 also show a similar effect as described for total pollen, being responsible for the high error low pollen (LPHE) events. These similarities are due to the fact that the total pollen model in this context is largely represented by the *Poaceae* type.

3.4.6 *Quercus*

The *Quercus* model specifically shows strong correlation with bin 5 in the bee swarm plot, while most other bins suggest negative correlation. While it appears many bins have contributed to pushing the *Quercus* model output up generally, the bar plots (Figure 3.6) show that specifically bins 4-6 demonstrate a stronger impact on the model (i.e. darker red colour) with higher particle concentrations in these bins (i.e. taller bars) for high pollen low error events (HPLE; 1st bar) when compared to low pollen low error events (LPLE; 2nd bar). This implies that the *Quercus* type is particularly associated with particles in the size range of 1.3-3.0 μm.

The bar plots suggest that bin 0 is responsible for pushing the predicted output concentration up when actual pollen concentrations are low (strong red LPHE bar) and is responsible for the high error in the HPHE category (blue bar, i.e. bringing predictions down when actual pollen is high). Other bins including bins 1-4 and 6 seem also responsible for causing high error for LPHE events. Bin 5 is an example that the model has learned the relationship well, since the bars are red for high pollen events (HPLE

and HPHE) and blue for low pollen events (LPLE and LPHE) but the overall output may be confounded by other less helpful bins. Here, for example, it may be beneficial to remove bins 0-3 as input features and allow the model to learn to make better use of bins 4-6.

3.4.7 *Betula*

The bee swarm plots suggest that the *Betula* type has positive correlations with bins 4,5 and 9. Meanwhile, the bar plots suggest that accurate high pollen events are largely positively correlated with bins 4-7 (i.e. darker red colour and taller bars for HPLE particularly when compared to LPLE), but not bin 9. This suggests that *Betula* can be associated with particles in size ranges 1.3-2.3 µm (bins 4-5), possibly extending up to 4 µm (bins 6-7). This size range is similar to *Quercus* and the two pollen types have similar size ranges - 17-26 µm for *Betula* (Mäkelä, 2009) and 20-42 µm for *Quercus* (Wrońska-Pilarek, 2016) – while generally smaller than *Poaceae* (22-46 µm, Radaeski et al., 2016) or *Pinus* (28-97 µm, Song et al., 2012). This may suggest that not only are the intact pollen grains of similar sizes but so are the associated subpollen particles that come from each.

3.4.8 *Pinus*

Pinus appears to have positive correlations with bins 0, 4 and 7 from the bee swarm plots. This positive correlation is only reinforced for bin 0 in the bar plots, as this is the only case where the bin particle concentration is higher for HPLE events compared to LPLE events while the SHAP values (impact on the predicted *Pinus* concentration) were strongly positive. This corresponds with particle sizes between 0.35-0.46 µm however, being the lowest size range measurable by the OPC instrument there may be subpollen particles associated with *Pinus* below this size range. There appear to be some

similarities in positively correlated bins when compared with *Poaceae* according to the bee swarm plots which are not reinforced by the high pollen event bar plots. This may be due to the fact that the *Pinus* season overlapped with *Poaceae*, so the two types may often have co-occurred leading to learned associations related to the wrong type, i.e. *Pinus* learning from *Poaceae* subpollen particles.

The most unhelpful feature for the *Pinus* model appears to be bin 5 since it has pulled the predicted output down when it should be high (i.e. blue bar for HPLE) and strongly pushed the output up when it should be low (i.e. strong red bar for LPHE). Meanwhile, even bins that have been helpful for distinguishing HPLE and LPHE events (e.g. bins 0-2) have also contributed to the high error when pollen should be low (red for LPHE). This model may benefit from removing certain features such as bin 5 but also by training on data in a different context that is not confounded by co-occurring pollen types. This would in fact be desirable for each pollen type, however, may be hard to realise in practice.

3.4.9 Summary

The conclusions of the results discussed in sections 3.4.3-3.4.8, taking into account the SHAP analyses in Figures 3.4-4.6, are summarised visually in Figure 3.7. It shows which bin size ranges contributed positively to the detection of each pollen type, and demonstrates the similarities and differences in terms of the associated bins. For example, total and *Poaceae* pollen are similar, due to total pollen being largely represented by *Poaceae* in this dataset, as well as *Quercus* and *Betula* pollen, likely because the pollen grains or associated particles are similar in size.

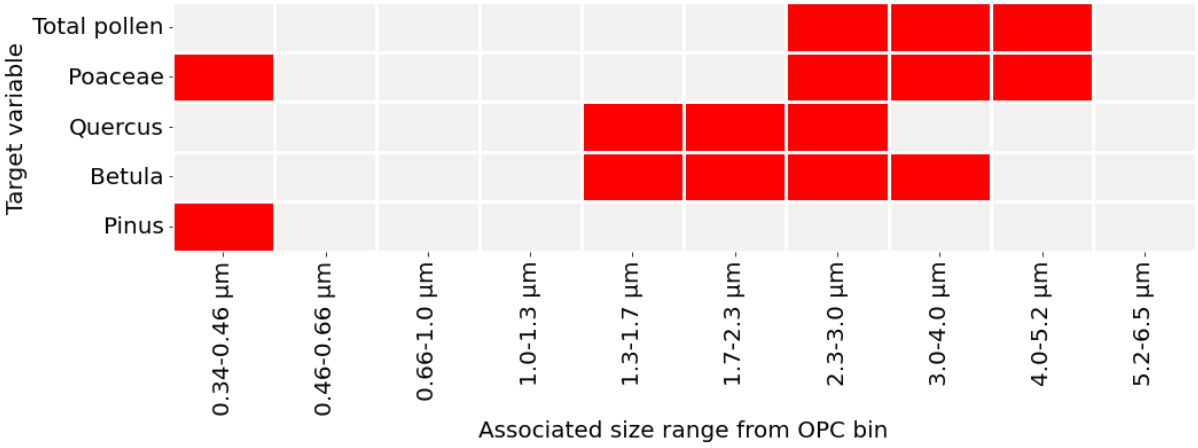


Figure 3.7: *Summary of particle sizes from OPC bins that demonstrate some positive correlation (shaded red) with each target variable according to further SHAP analysis (including the information from bar plots in Figure 3.5). Note: The different bins/cells correlate positively to varying extents (e.g. 1.7-2.3 µm has relatively stronger correlation than adjacent bins for* Quercus*) but this is difficult to quantify and is not taken into account in this figure.*

3.4.10 Importance of input features and context

We proceeded to test the effect of removing bins that may have confounded the model in each case according to the discussed observations. The resulting R^2 scores for each model set with certain bins removed are displayed in Figure 8.7 in the Appendices. Removing those above bin8 (>5.2 µm) did not significantly affect performance. However, removing other specific bins considered potentially 'unhelpful' in each case (detailed in the caption of Figure 8.7) resulted in decreased performance, generally the more bins removed the worse the score. The models do not seem to gain much information from bins above 5.2 µm, which may be in part due to limitations of the OPC instrument when measuring larger particles. However, the results suggest that the models make use of all

other bins for optimal performance in this context, as they are able to learn even very complex nonlinear relationships that are difficult for us to comprehend or visualise.

This may include learned statistical relationships with some size ranges that are collinear with the given pollen target only in this context, and which may become irrelevant and unhelpful if applied to other contexts. If the bins that appear positively correlated with a given target pollen taxon are directly associated with subpollen particles within those size ranges, then these bins may generalise well in other environments where the same taxon is present. However, bins that have simply been used by the model because they have some collinearity with the given target taxon (in particular, this may be relevant for negatively correlated bins) may cause the model to generalise less well in other environments and hence their removal may bring higher scores when applied to different contexts.

Ultimately, it must be emphasised how sensitive such models are to the environment and context from which training data is taken. A model that performs best on the reserved test dataset here may not necessarily perform best on a test dataset in another context. In other words, a model with a lower R^2 score here could potentially generalise better in other environments. Thus, we emphasise the need to further validate these models with more data in different contexts. For such methods to be put to purpose, for example among pollen monitoring networks for public health, it may be necessary to produce trained models for a particular context and scale within which there is some consistency of the local pollen environment. For a model to be used across a large scale among different contexts, i.e. different regions or countries, it would likely need a lot of varied data from within these regions for training.

Nevertheless, we have also demonstrated the effect that methodical hyperparameter tuning can have on increasing model performance and that this works best on a case-by-case basis, depending on the task and available data. The method we have presented in this study could be further applied to optimise models that take fewer or different input features, which may be able to achieve even better results.

3.5 Conclusions

In this study, we have demonstrated how the methodology to detect pollen using low-cost OPC sensors presented previously in Mills et al. (2023) can be improved by hyperparameter tuning and regularisation techniques for better performance. We have also demonstrated its application on an increased range of pollen types. For hourly and daily time resolution models, 6/10 and 8/10 respectively were able to produce R^2 scores above 0.5 for the test dataset. Maximum R^2 scores achieved were 0.89 and 0.95 for hourly and daily models respectively, once again for the models targeting *Poaceae* pollen with particle size and meteorological input data.

Using the SHAP explainable AI (XAI) method, we visualised and described the observable relationships between input features and target pollen concentration and assessed which particle size ranges were responsible for pushing up or pulling down predicted outputs. We further investigated the relationship between specific bin particle concentrations and impact on model output under four selected circumstances of high/low pollen and high/low error respectively. In particular, *Quercus* pollen displayed strong positive correlation with particles in the range 1.3-3.0 µm, which could be evidence of associated subpollen particles in this size range. *Quercus* and *Betula* pollen showed some similarity with each other but difference from *Poaceae* (which largely contributed to total pollen), while *Pinus* pollen showed mild similarity with *Poaceae*. Thus, this method may be more effective at isolating some types (e.g. *Quercus*) than others, and at differentiating between certain types (e.g. *Quercus* vs *Poaceae*) based on particle size measurements alone.

This work demonstrates the potential that this method shows for low-cost monitoring of pollen and provides deeper explanation as to how the models are learning and the

influence each input feature has on predicting output pollen concentrations. While caution must always be taken when implying causal inference from machine learning models, useful information is obtainable that we can pair with our current scientific understanding. Currently, our scientific understanding of subpollen particles is limited and further studies characterising these particles for different pollen types would be an important step forward to provide scientific verification for these methods. Meanwhile, as with most machine learning methods, this technique would greatly benefit from training and testing on more data representing more varied environments to truly assess the extent of generalisability.

3.6 Contributors

The initial concept of this study to use low-cost OPCs for monitoring pollen was conceived by FDP. JMM and FT planned and organised the intercomparison campaign in Munich and facilitated the running of the sensors there, providing the context for the data used in this study. The machine learning concept and methodology was conceived and developed by SAM who curated the data, planned the specific content of this study, performed analyses and prepared the first draft. FDP and ARM provided supervision and guidance during this process. All authors contributed to editing the final manuscript.

3.7 Acknowledgements

We thank Andrew Tanner from the University of Birmingham, UK, for his technical work on functionalising the Alphasense OPC sensors for the purposes of this study. We thank the ZAUM (Jeroen Buters, Carsten B. Schmidt-Weber, Cordula Ebner von Eschenback, Gudrun Pusch, Marina Triviño and Christine Weil) and MeteoSwiss (Bernard Clot, Nina Burgdorfer and Sophie Erb) teams for their assistance facilitating the EUMETNET AutoPollen – COST ADOPT intercomparison campaign, as well as the Helmholtz Zentrum München as the campaign host and their assisting team (Daphne Kolland, George Matuscheck and Benjamin Schnautz)s. The manual pollen analysts for the campaign Łukasz Kostecki and Agata Szymanska, from Łukasz Grewling's laboratory at Adam Mickiewicz University (Poznan, Poland) are greatly appreciated for their work facilitating the vital reference data for this campaign. We thank Robert Gebauer, Gisela Nagy and Anton Pointner for their IT support during the campaign.

We acknowledge the funding support of the Natural Environment Research Council (NERC) CENTA2 graint NE/S007350/1 via the University of Birmingham, and the grant "Quantification of Utility of Atmospheric Network Technologies (QUANT)" (NE/T001968/1). The intercomparison campaign where data was obtained for this study was funded by the Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) and EUMETNET AutoPollen Programme. Financial support was also received for this from the COST Action CA18226 ADOPT – *New approaches in detection of pathogens and aeroallergens*.

3.8 Data availability

Data supporting this publication are openly available from the UBIRA eData repository at doi: https://doi.org/10.25500/edata.bham.00000898.

4.9 Comment

This chapter follows on from the work established in Chapter 2, focusing in particular on the neural network machine learning method. It draws on machine learning best practices and investigates the potential for improving model performance by tuning various hyperparameters for specific model cases. Developing deep learning methods to their full potential and the interpreting the outputs can be a complicated art. This chapter demonstrates a methodical approach, assessing the relevance and impact of the various hyperparameters that can be utilised by machine learning practitioners. It also demonstrates the application of an XAI method which can be applied to any machine learning model. Machine learning models are generally seen as 'black boxes' and questions are often raised about what exactly they are learning from, their transparency and reliability. This work should benefit those who are attempting a deep learning approach for such purposes in the future, and provide some guidelines for best practices in terms of developing and explaining their models. These are important steps for making such techniques a viable option for future public or commercial use.

4.10 References

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers. Aerobiologia. https://doi.org/10.1007/s10453-021-09724-5.

Bacsi, A., Choudhury, B.K., Dharajiya, N., Sur, S., Boldogh, I., 2006. Subpollen particles: carriers of allergenic proteins and oxidases. J. Allergy Clin. Immunol., 118(4), 844-850. https://doi.org/10.1016/j.jaci.2006.07.006.

Bengio, Y., 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In: Montavon, G., Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg.

Bengio, Y., 2012. Practical for gradient-based training of deep architectures. Neural Networks: Tricks of the Trade: Second Edition, pp437-478. https://doi.org/10.48550/arXiv.1206.5533.

Bradley, R.S., 2015. Paleoclimatology: Chapter 12 – Pollen. 3rd ed., Elsevier, Massachusetts, USA, pp. 408-409. https://doi.org/10.1016/B978-0-12-386913-5.00012-0.

Brownlee, J., 2018. Better Deep Learning. Edition: v1.8. https://machinelearningmastery.com/better-deep-learning/.

Burkart, J., Gratzl, J., Seifried, T.M., Bieber, P., Grothe, H., 2021. Isolation of subpollen particles (SPPs) of birch: SPPs are potential carriers of ice nucleating macromolecules. Biogeosciences, 18, 5751-5765. https://doi.org/10.5194/bg-18-5751-2021.

Buters, J.T.M, Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy, 8, 9. https://doi.org/10.1186/s13601-018-0197-8.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.A., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Kofol Seliger, A., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto, A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia, https://doi.org/10.1007/s10453-022-09750.

Chappuis, C., Tummon, F., Clot, B., Konzelmann, T., 2020. Automatic pollen monitoring: first insights from hourly data. Aerobiologia, 36, 159-170. https://doi.org/10.1007/s10453-019-09619-6.

Christianini, N., 2010. Are we there yet? Neural Networks, 23(4), 466-470. https://doi.org/10.1016/j.neunet.2010.01.006.

Cristianini, N. 2023, *The shortcut: Why Intelligent Machines Do Not Think Like Us*. CRC Press. ISBN 9781032305097.

Crouzy, B., Stella, M., Konzelmann, T., Calpini, B., Clot, B, 2016. All-optical automatic pollen identification: Towards an operational system. Atmos. Environ., 140, 202-212. https://doi.org/10.1016/j.atmosenv.2016.05.062.

Darrow, L.A., Hess, J., Rogers, C.A., Tolbert, P.E., Klein, M., Sarnat, S.E., 2012. Ambient pollen concentrations and emergency department visits for asthma and wheeze. J. Allergy Clin. Immunol., 130(3), 630-638. https://doi.org/10.1016/j.jaci.2012.06.020.

Després, V.R., Huffman, J.A., Burrows, S.M., Hoose, C., Safatov, A.S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M.O., Pöschl, U., Jaenicke, R., 2012. Primary biological aerosol particles in the atmosphere: a review. Tellus B, 64. http://dx.doi.org/10.3402/tellusb.v64i0.15598.

Diehl, K., Quick, C., Matthias-Maser, S., Mitra, S.K., Jaenicke, R., 2001. The ice nucleating ability of pollen: Part I: Laboratory studies in deposition and condensation freezing modes. Atmos. Res. 58 (2), 75-87. https://doi.org/10.1016/S0169-8095(01)00091-6.

Diehl, K., Matthias-Maser, S., Jaenicke, R., Mitra, S.K., 2002. The ice nucleating ability of pollen: Part II: Laboratory studies in immersion and contact freezing modes. Atmos. Res. 61 (2), 125-133. https://doi.org/10.1016/S0169-8095(01)00132-6.

Dreischmeier, K., Budke, C., Wiehemeier, L., Kottke, T., Koop, T., 2017. Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides. Sci. Rep., 7, 41890. https://doi.org/10.1038/srep41890.

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the Earth system: climate, health, and ecosystem interactions. Atmos. Res., 182, 346-376. https://doi.org/10.1016/j.atmosres.2016.07.018.

Gohel, P., Singh, P., Mohanty, M., 2021. Explainable AI: current status and future directions. IEEE Access, arXiv:2107.07045. https://doi.org/10.48550/arXiv.2107.07045.

Goodfellow, I., Bengio, Y., Courville, A., 2017. Deep Learning. The MIT Press, Cambridge. ISBN: 9780262035613.

Griffiths, P.T., Borlace, J.-S., Gallimore, P.J., Kalberer, M., Herzog, M., Pope, F.D., 2012. Hygroscopic growth and cloud activation of pollen: a laboratory and modelling study. Atmospheric Science Letters, 13 (4), 289-295. https://doi.org/10.1002/asl.397.

Gute, E., Abbatt, J.P.D., 2020. Ice nucleating behaviour of different tree pollen in the immersion mode. Atmos. Environ., 231, 117488. https://doi.org/10.1016/j.atmosenv.2020.117488.

Hendrickson, B.N., Alsante, A.N., Brooks, S.D., 2023. Live oak pollen as a source of atmospheric particles. Aerobiologia, 39, 51-67. https://doi.org/10.1007/s10453-022-09773-4.

Huffman, J.A., Perring, A.E., Savage, N.J., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M.A., Crawford, I., Gallagher, M., Topping, D., Doughty, D.C., Hill, S.C., Pan, Y., 2019. Real-time sensing of bioaerosols: Review and current perspectives. Aerosol Science and Technology, 54 (5), 465-495. https://doi.org/10.1080/02786826.2019.1664724.

Jiang, C., Wang, W., Du, L., Huang, G., McConaghy, C., Fineman, S., Liu, Y., 2022. Field Evaluation of an Automated Pollen Sensor, Int. J. Environ. Res. Public Health, 19(11), 6444. https://doi.org/10.3390/ijerph19116444.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30, 4765-4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Mäkelä, E.M., 2009. Size distinctions between *Betula* pollen types – A review. Grana, 35(4), 248-256. https://doi.org/10.1080/00173139609430011.

Marchisio, A., Hanif, M.A., Rehman, S., Shafique, M., 2018. A Methodology for Automatic Selection of Activation Functions to Design Hybrid Deep Neural Networks, arXiv:1811.03980v1. https://doi.org/10.48550/arXiv.1811.03980.

Masters, D., Luschi, C., 2018. Revisiting Small Batch Training for Deep Neural Networks. arXiv:1804.07612. https://doi.org/10.48550/arXiv.1804.07612.

Matthews, B.H., Alsante, A.N., Brooks, S.D., 2023. Pollen Emissions of Subpollen Particles and Ice Nucleating Particles. ACS Earth Space Chem., 7(6), 1207-1218. https://doi.org/10.1021/acsearthspacechem.3c00014.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, Ł., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavilj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2023. Towards European automatic bioaerosol monitoring: Comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ., 866, 161220. https://doi.org/10.1016/j.scitotenv.2022.161220.

McCurry, J., 2023. "*Japan declares war on pollen as hay fever epidemic grips the nation*", The Guardian. https://www.theguardian.com/world/2023/apr/06/japan-declares-war-on-pollen-as-hay-fever-epidemic-grips-the-nation, (accessed 10/05/2023).

Mikhailov, E.F., Ivanova, O.A., Nebosko, E.Y., Vlasenko, S.S., Ryshkevich, T.I., 2019. Subpollen Particles as Atmospheric Cloud Condensation Nuclei. Izv. Atmos. Ocean. Phys., 55, 357-364. https://doi.org/10.1134/S000143381904008X.

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023. Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests. Sci Total Environ., 871, 161969. https://doi.org/10.1016/j.scitotenv.2023.161969.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and Online Pollen Monitoring. Int. Arch. Allergy. Immunol., 167, 158-166. https://doi.org/10.1159/000436968.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on

automatic image recognition. Environ. Res., 191, 110031. https://doi.org/10.1016/j.envres.2020.110031.

Pedamonti, D., 2018. Comparison of non-linear activation functions for deep neural networks on MNIST classification task. arXiv:1804.02763. https://doi.org/10.48550/arXiv.1804.02763.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning – A review for ecologists. Methods in Ecology and Evolution, 14(4), 994-1016. https://doi.org/10.1111/2041-210X.14061.

Pope, F.D., 2010. Pollen grains are efficient cloud condensation nuclei. Environ. Res. Lett., 5, 044015. https://doi.org/10.1088/1748-9326/5/4/044015.

Pummer, B.G., Bauer, H., Bernardi, J., Bleicher, S., Grothe, H., 2012. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. Atmos. Chem. Phys., 12, 2541-2550. https://doi.org/10.5194/acp-12-2541-2012.

Radaeski, J.N., Bauermann, S.G., Pereira, A.B., 2016. Poaceae Pollen from Southern Brazil: Distinguishing Grasslands (Campos) from Forests by Analysing a Diverse Range of Poaceae Species. Front. Plant Sci., 7. https://doi.org/10.3389%2Ffpls.2016.01833.

Reed, R., Marks, R.J., 1999. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. The MIT Press. ISBN: 9780262527019.

Reponen, T., 2011. Encyclopedia of Environmental Health: Methodologies for Assessing Bioaerosol Exposures. Elsevier, Cincinnati, USA, pp. 723. https://doi.org/10.1016/B978-0-12-409548-9.11822-6.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech., 12(6), 3435-3452. https://doi.org/10.5194/amt-12-3435-2019.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2020. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech., 13, 1539-1550. https://doi.org/10.5194/amt-13-1539-2020.

Smiljanic, K., Apostolovic, D., Trifunovic, S., Ognjenovic, J., Perusko, M., Mihajlovic, L., Burazer, L., van Hage, M., Cirkovic Velickovic, T., 2017. Subpollen particles are rich carriers of major short ragweed allergens and NADH dehydrogenases: quantitative proteomic and allergomic study. Clin. Exp. Allergy, 47(6), 815-828. https://doi.org/10.1111/cea.12874.

Song, U., Park, J., Song, M., 2012. Pollen morphology of *Pinus* (*Pinaceae*) in northeast China. Forest Science and Technology, 8(4), 179-186. https://doi.org/10.1080/21580103.2012.704973.

Steiner, A.L., Brooks, S.D., Deng, C.,Thornton, C.O., Pendleton, M.W., Bryant, V., 2015. Pollen as atmospheric cloud condensation nuclei. Geophys. Res. Lett., 42 (9), 3596-3602. https://doi.org/10.1002/2015GL064060.

Stone, E.A., Mampage, C.B.A., Hughes, D.D., Jones, L.M., 2021. Airborne sub-pollen particles from rupturing giant ragweed pollen. Aerobiologia, 37, 625-632. https://doi.org/10.1007/s10453-021-09702-x.

Subba, T., Zhang, Y., Steiner, A.L., 2023. Simulating the Transport and Rupture of Pollen in the Atmosphere. Journal of Advances in Modelling Earth Systems, 15(3), e2022MS003329. https://doi.org/10.1029/2022MS003329.

Tong, H.-J., Ouyang, B., Nikolovski, N., Lienhard, D.M., Pope, F.D., Kalberer, M., 2015. A new electrodynamic balance (EDB) design for low-temperature studies: applicationto immersion freezing of pollen extract bioaerosols. Atmos. Meas. Tech., 8, 1183-1195. https://doi.org/10.5194/amt-8-1183-2015.

Triviño, M.M., Maya-Manzano, J.M., Tummon, F., Clot, B., Schmidt-Weber, C. and Buters, J., 2023. Resistance free flow adjustment of Hirst-Type pollen traps reduces variability between traps. Aerobiologia. (*Submitted)*

Wrońska-Pilarek, D., Danielewicz, W., Bocianowski, J., Malinski, T., Janyszek, M., 2016. Comparing Pollen Morphological Analysis and Its Systematic Implications on Three European Oak (Quercus L., Fagaceae) Species and Their Spontaneous Hybrids. PLoS ONE, 11(8): e0161762. https://doi.org/10.1371/journal.pone.0161762.

4. CHAPTER 4: Can low-cost sensors be used to observe airborne pollen at a forest site at varying heights through the canopy?

This study is prepared for submission to a journal.

4.1 Abstract

Conventional techniques for monitoring pollen currently have significant limitations in terms of labour, cost and the spatiotemporal resolution that can be achieved. Pollen monitoring networks across the world are generally sparse and are not able to fully represent the detailed characteristics of airborne pollen. There are few studies that observe concentrations on a local scale, and even fewer that do so in ecologically-rich rural areas and close to emitting sources. Better understanding of these would be relevant to occupational risk assessments for public health, as well as ecology, biodiversity, and climate.

We present a study using low-cost optical particle counters (OPCs) and applying pre-trained machine learning models, described in detail in our previous works, to monitor particulate matter and pollen within a mature oak forest in the UK. We characterise the observed pollen concentrations, first during an OPC colocation period for calibration purposes, then for a period when the OPCs were distributed on an observational tower at different heights through the canopy. We assess the efficacy and usefulness of this method and discuss directions for future development, including the requirements for training data.

The results show promise, with the derived pollen concentrations following the expected diurnal trends and interactions with meteorological variables. While PM_{10} concentrations were suppressed within the canopy height, *Quercus* pollen concentrations appeared greatest at this height (20-30 m). *Quercus* pollen concentrations were lowest at the greatest height that is above the canopy (40 m), which is congruent with previous studies of background pollen in urban environments. The attenuation of pollen concentrations as sources are depleted is also observed across

the season and at different heights, with some evidence that the pollen concentrations persist later at the lowest level beneath the canopy (10 m) where catkins mature latest.

4.2 Introduction

At present, the UK has just 11 regulatory pollen monitoring station across the whole country. The UK has a land area of 244 km^2 and population of over 67 million. This means that there is only one regulatory pollen station per 6 million people. For example, there is just one station in the whole of Scotland (see https://www.worcester.ac.uk/about/academic-schools/school-of-science-and-the-environment/science-and-the-environment-research/national-pollen-and-aerobiology-research-unit/pollen-calendar.aspx, accessed June 2023). Other industrialised countries have similar densities of monitoring networks, perhaps somewhat better across Europe. However, outside of these industrialised areas, there are vast regions that have very scarce or non-existent monitoring networks, particularly in tropical forested regions of South America, Africa and South Asia (see Buters et al., 2018).

The lack of regulatory monitoring stations is a function of cost since they have significant capital expenditure (CAPEX) and operating expenditure costs (OPEX), since they are labour intensive. This cost reduces the spatial resolution of data available. The current 'best' benchmark we have, which occupies most functional pollen monitoring stations globally, is the manual Hirst-type (Hirst, 1952) (e.g. Burkard, https://burkard.co.uk/) or Rotorod (https://www.pollen.com/help/rotorod) sampler. These instruments deposit aerosols from the ambient air onto a film. The film is taken back to the lab and trained experts, under the microscope, identify and count the number of pollen grains for a range of different taxa. Hence, the process to accumulate pollen count datasets, which can then be used to inform the public, takes considerable time and at considerable costs. Data is often only available after a delay of at least

several days and is reported in coarse daily time resolutions (Tummon et al., 2021). This limitation reduces the temporal resolution available from these types of techniques.

Recently, automated technologies for monitoring pollen have been under development. These use techniques such as automatic imaging, laser scattering, fluorescence and holographic imaging, paired with machine learning algorithms trained to classify different pollen types from the data they are given. An overview of these emerging technologies can be found in Buters et al. (2022) and other references discussed in our previous work Mills et al. (2023a) and Mills et al. (2023b). However, these automated instruments, while showing promising results with high accuracy, can cost tens to hundreds of thousands of pounds. They are not an affordable option to populate monitoring networks more densely and provide high spatial resolution data and across varied and unique environments.

Pollen is a major health issue, as up to 40% of populations can suffer from pollen allergies, including hay fever and asthma (Fröhlich-Nowoisky et al., 2016). Climate change and air pollution can also have impacts that increase exposure and sensitivity to allergens (see, e.g., Sedghy et al., 2018; Damialis et al., 2019; Anderegg et al., 2020; Berger et al., 2021; Singh and Kumar, 2022; Zhang and Steiner, 2022). It is expected that, by 2050, at least one person out of two will suffer from allergies due to the accelerating issues of climate change and air pollution (Dbouk et al., 2022). We need to understand the spatial distribution and emission characteristics from sources to assess risk for public health. To do this, we need greater spatial and temporal coverage of pollen concentrations than that are currently available via conventional monitoring techniques.

In the absence of being able to directly measure pollen concentrations, studies tend to rely on physical modelling approaches that use fluid dynamics and atmospheric

transport modelling (Dbouk et al., 2022) or remote sensing imaging techniques to identify vegetation types and coverage (Xu et al., 2016) to assess allergenic risk. Each method comes with its own set of advantages and limitations.

As well as the horizontal spatial distribution of pollen, another unknown is the vertical spatial distribution within the atmosphere. This vertical information is key information to parameterise meteorology and aerosol transport models. Pollen (and subpollen particles released from pollen grains) have been discussed in the literature to have potential to act as cloud condensation nuclei (CCN) (see, e.g., Pope, 2010; Griffiths et al., 2012; Steiner et al., 2015; Mikhailov et al., 2019) and ice nuclei (IN) (see, e.g., Diehl et al., 2001; Diehl et al., 2002; Pummer et al., 2012; Dreischmeier et al., 2017; Gute and Abbatt, 2020; Burkart et al., 2021), which affect the formation and lifetime of clouds and hence precipitation. As such, it is certainly of interest to investigate the vertical behaviour of pollen bioaerosols.

In the interest of public health, it is important to observe near-ground pollen concentrations for risk assessment purposes, at the height that people go about daily activity. There are a few studies which focus on this: Xiao et al. (2013), Wagle et al. (2018), Rojo et al. (2019) and Hugg et al. (2020) – which themselves discuss the present scarcity of such datasets. Rojo et al. (2019) studied the vertical concentration profile of pollen between 1.5 and 50 m across various countries, but predominantly in Europe, while Xiao et al. (2023) studied at 1.5, 35 and 70 m in China - both using Hirst/Burkard samplers. Generally, pollen concentrations were found to decrease with height, though with much variability depending on pollen type, season, meteorology and area environment. Heights above 10 m were found to be representative of background near-ground levels (Rojo et al., 2019). Meanwhile, Wagle et al. (2018) and Hugg et al. (2020)

investigated heights between 1.5 and 5.3 m, with Wagle et al. (2018) focusing on Plantago pollen in the US and Hugg et al. (2020) on *Poaceae* pollen in Finland. Wagle et al. (2018) observed higher Plantago cumulative counts at 5.3 vs 1.7 m, but some minor peaks were observed earlier at 1.7 m. Hugg et al. (2020) generally observed greater *Poaceae* concentrations at 1.5 vs 4 m, however, environment urbanisation had a significant effect – i.e., a greater degree of urbanisation resulted in greater concentrations at 4 vs 1.5 m.

Generally, these studies were performed within urban areas of major cities because of the proximity of the majority of receptors (i.e. people) to the pollen. Little is known of pollen concentrations in rural areas and actual emission profiles close to the source. This is interesting from an ecological perspective, to understand the mechanism of pollen release from within specific ecological populations and potentially the environmental impacts that dictate the time and magnitude of pollen emissions. It may also inform us of implications that may be relevant for biodiversity, future species habitat colonisation and survival, and, ultimately, climate and human health.

Recently, we have developed a novel methodology for monitoring pollen with automatic low-cost sensors (Mills et al., 2023a; Mills et al., 2023b) and demonstrated that, with machine learning techniques to process the data, good quality pollen concentration data can be produced. While these sensors have limitations in terms of accuracy when compared with the manual Hirst or more expensive instruments, they also offer important advantages of real-time, high time resolution data for affordable costs. They are small and easy to deploy – just about anywhere – currently with the limitation that they need a power supply, but we have also developed models that run on battery power alone. Crucially, this means that arrays of multiple sensors can be deployed to provide a

network of high spatial resolution, supplying coverage that has simply not been possible up until now. While the potential is great, first it is necessary to assess the reliability of this technique in different environments and whether they can effectively discern pollen concentration variations on relevant temporal and spatial scales.

We present here a study which investigates the near-ground vertical profile of *Quercus* (oak) pollen up to 40 m from within a mature, predominantly oak, forest environment during the oak pollen season. We test the effectiveness of our automated low-cost optical particle counter (OPC) pollen sensors, paired with pre-trained machine learning models to process the particle size data, for measuring near-source pollen concentrations and discerning the spatial variation on this scale. This is, to our knowledge, the first study of this kind which places multiple sensors through the canopy of a forest to monitor pollen. This has also never been possible before with such high time resolution.

4.3 Methods

4.3.1 Context

The data for this study was obtained at the Birmingham Institute of Forest Research (BIFoR) Free Air CO_2 Enrichment (FACE) experiment facility in Staffordshire, UK (latitude: 52.7996, longitude: −2.3039). The site is mature temperate woodland of predominantly English oak (*Quercus robur*), with an average canopy height of 25 m, and hazel (*Corylus avellana*) understory. The mature oak trees were planted around 150 years old while other species such as sycamore (*Acer pseudoplatanus*), silver birch (*Betula pendula*) and holly (*Ilex aquifolium*) have self-seeded, giving rise to a diverse canopy (Norby et al., 2015).

A map of the site, and its location in the UK, can be found in Fig. 1 (taken from Hart et al., 2019). Within the forest is a long-term ecological experiment that has been running since 2016, where there are nine 30 m diameter plots (or 'arrays') with three different treatments. There are 'elevated CO_2' plots (orange in Fig. 1) which are fumigated by surrounding infrastructure with air of 150 ppm of CO_2 above current ambient levels. Meanwhile there are two types of control plots: 'ambient CO_2' plots (blue in Fig. 1) which are fumigated by surrounding infrastructure by air that contains current ambient levels of CO_2 and 'ghost' plots (red in Fig. 1) which have no infrastructure nor fumigation. Also denoted in Fig. 1 is the location of the flux tower at the site, which stands above the canopy at 40 m tall and various meteorological and other measurements are recorded, including a 'PhenoCam' which captures images of the forest canopy each hour of every day throughout the year.

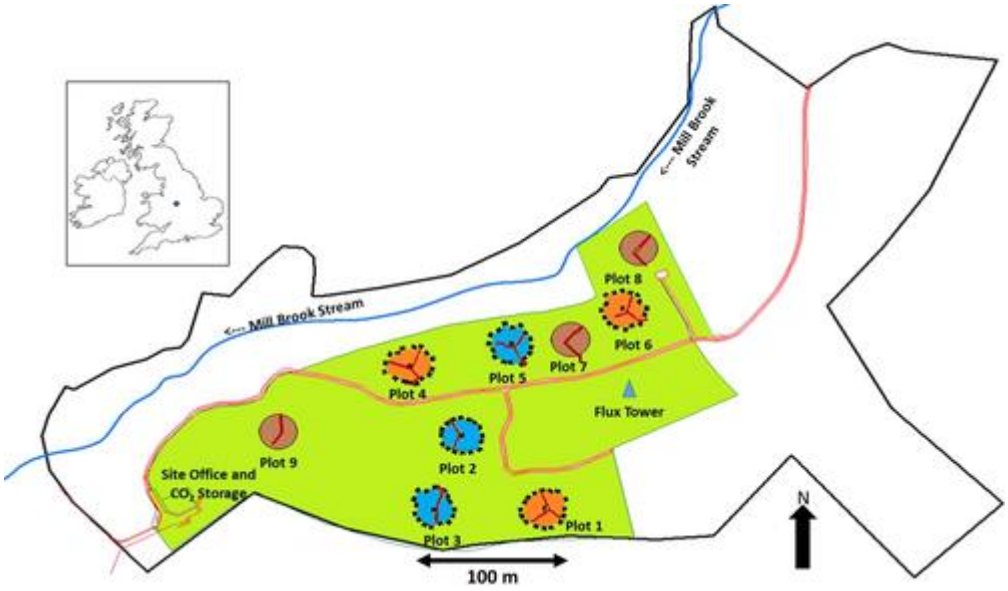


Figure 4.10: *Map and location of the Birmingham Institute of Forest Research (BIFoR) Free Air CO_2 Enrichment (FACE) facility, taken from Hart et al. (2019).*

This site presented a unique opportunity to monitor pollen using our recently developed low-cost pollen sensor methodology, since it had the infrastructure (such as the flux tower and mains power onsite) to support the study and a wealth of accompanying meteorological, ecological and biological data simultaneously being collected from the site.

4.3.2 Instrument deployment

This study reports data collected from 4 Alphasense (https://www.alphasense.com/) OPC-N3 sensors, described previously in Maya-Manzano et al. 2023 and Mills et al. 2023. The data from these sensors was transmitted via the mobile network and was remotely accessible online, via IoT services and platform provided by Soracom (https://www.soracom.co.uk/). Initially, the sensors were placed side by side at the base of the flux tower (at ~4m) for a colocation calibration period between 30th March – 5th April 2022. Subsequently, the sensors were placed individually at platforms of different

heights – 10, 20, 30, and 40 m - up the flux tower, all facing the same direction (see Figure 2). This 'tower' period continued from 5th April until 8th May 2022, however the sensor at 40 m stopped logging data from 12th April. This time period was chosen since it coincides with the oak (*Quercus*) pollen season which generally appears around April-May. Some other tree pollen season also occur around this time, for example birch (*Betula*), while the grass (*Poaceae*) pollen season may start within this period but tends to intensify later towards June and July.

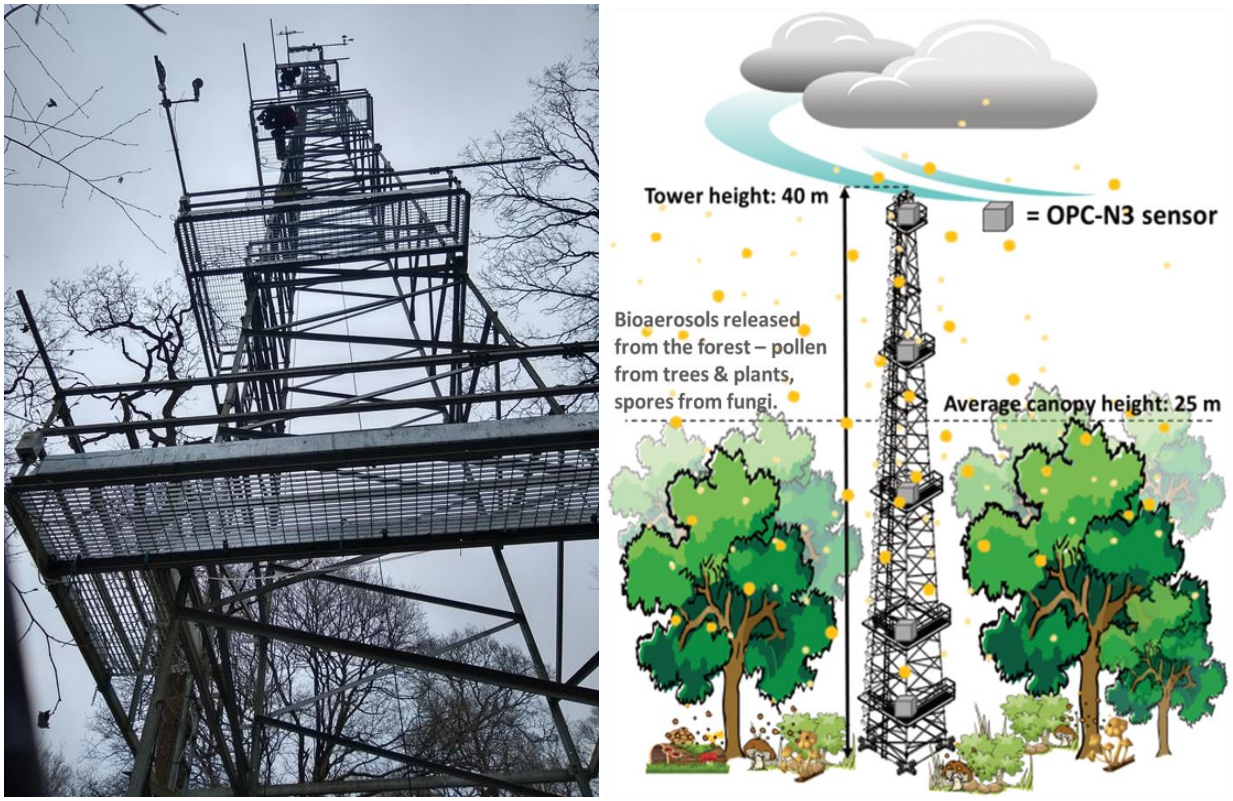


Figure 4.11: *Picture and diagram of Alphasense OPC-N3 sensors on the flux tower at BIFoR FACE. (Note there was also a 5th sensor at a 5 m platform but this did not log sensible data so has not been included in this study.)*

The Alphasense OPC-N3 sensors provide particle number counts across 24 bins of varying particle size ranges between 0.35 and 40 μm, as well as calculated PM_1, $PM_{2.5}$ and PM_{10} values (μg m^{-3}). The particle counts in each bin were converted to number concentrations (grains m^{-3}) using the sample period and flow rate (as described previously in Mills et al. (2023). The sensors were also equipped with BME280 sensors (Bosch Sensortec GmbH, Gerhard-Kindler-Strasse 9, 72770 Reutlingen, Germany) which recorded temperature (°C) and relative humidity (RH) (%). These were attached on the outside of the weatherproof sensor casing under the hood sheltering the inlet, so as to monitor ambient conditions as accurately as possible.

Wind velocity data was monitored by the facility's control program, using a two-dimensional sonic anemometer (WMT700, Vaisala Oyj, Vantaa, Finland), measured from ~50 m away (at Plot 6 in Fig. 1) at a height of approximately 20 m within the canopy. Detailed information on the infrastructure of the facility can be found in Hart et al. (2020) and a specific study of wind characteristics and the residence times of air at the site is available in Bannister et al. (2023).

4.3.3 Model training

Quercus, *Poaceae* and total pollen concentrations were estimated using methodology we have previously established in Mills et al. (2023a) and developed further in Mills et al. (2023b). Previously, we employed supervised machine learning techniques using data from the EUMETNET AutoPollen-ADOPT COST Action intercomparison campaign in Germany, March-July 2021, (described in detail in Maya-Manzano et al., 2023) to train models to learn the relevant relationships to output pollen concentrations from the 24 OPC bin number concentrations. We performed hyperparameter tuning on the neural network models to further improve performance. An explainable artificial intelligence

(XAI) technique, SHAP (SHapley Additive exPlanations), to interpret the models and investigate the effect of each input feature on the output pollen concentration (Mills et al., 2023b).

The pollen concentration time series used in this study were obtained by applying the fitted models on the new data obtained from the OPC sensors at the BIFoR FACE facility. Specifically, the neural network models that performed best for each model set (i.e. *Quercus*, *Poaceae* and total pollen) in Mills et al. (2023b) were used. This excluded models that took additional meteorological variables (temperature and relative humidity) as we wanted to focus on models that purely used the particle size bin data. The reason being that these models may, in theory, generalise better than those that depend heavily on meteorological features and have the potential to overfit. The environment that provides the context of this study is different to that from which the data was obtained to train the model (in Munich, Germany), though the general climate would be similar.

4.3.4 General analyses

All scripts, unless stated otherwise, were written in Python (version 3.9.7) under the Anaconda distribution (https://docs.anaconda.com/). The libraries numpy (1.23.5), scipy (1.10.0), pandas (1.5.3), matplotlib (3.6.3) and seaborn (0.12.2) were used for data curation, analysis and visualisation.

The time series for PM_{10}, model-estimated *Quercus*, *Poaceae* and total pollen, as well as temperature and RH, are presented as both raw and corrected results. Data from the OPCs originally had a time resolution of 1 minute but was averaged to hourly data, since the models were trained on hourly data. Correction was applied to make all sensors consistent with each other and therefore comparable, but does not serve as an absolute

calibration. Corrected values were calculated for each variable by multiplying the raw values by a scaling factor which was calculated by dividing the mean variable value across all sensors by the mean for the individual sensor (Eq. 1):

$$Corrected\ value\ =\ Raw\ value\ \times\ \frac{Mean\ across\ all\ sensors}{Mean\ of\ indivual\ sensor} \qquad \text{Eq. 1}$$

An further RH correction was applied to PM_{10} mass concentrations to adjust for the hygroscopic effect which has been shown to inflate mass concentrations measured by low-cost OPC sensors under high RH conditions (Crilley et al., 2018). The correction factor, C, was calculated from Eq. 2 (explained in Crilley et al., 2018) and applied to the PM_{10} concentrations, already corrected by the sensor means, as shown in Eq. 3. The hygroscopicity parameter, κ, was set at 0.3, since this has been used in a previous similar study (Baird et al., 2022) and was found empirically to be the most suitable in this case. All further mention of corrected PM_{10} refers to values that have undergone both the calibration across sensors and the hygroscopicity correction with κ = 0.3.

$$C\ =\ 1\ +\ \frac{\frac{\kappa}{1.65}}{-1\ +\ \frac{100}{RH}} \qquad \text{Eq. 2}$$

$$PM_{hygroscopicity\ corrected}\ =\ \frac{PM_{sensor\ mean\ corrected}}{C} \qquad \text{Eq. 3}$$

Diurnal trends were investigated by plotting the variable averages across each hour of the day with 95% confidence interval error bars, constructed as described in the

Seaborn documentation (https://seaborn.pydata.org/tutorial/error_bars.html, accessed June, 2023). Bar plots were also constructed using the mean values and 95% confidence intervals for each variable to compare across the different sensors and colocation vs separated 'tower' period. Pearson correlations were calculated between each sensor pairing and presented as heatmap matrix plots for colocation and tower periods to show the relative congruency between sensors. Bar and matrix plots were only constructed from data between 30th March until 11th April since one sensor stopped logging after this date while another developed apparently anomalous measurements for PM_{10} after this date.

Wind velocity data was used to construct polar plots showing concentration of each variable (i.e. PM_{10} or pollen concentrations) as a function of wind speed and direction. These plots were constructed using R (4.1.2) Openair (2.17.0) package described in Carslaw and Ropkins (2012). Similar plots were also constructed using data imported via Openair from the NOAA Integrated Surface Database (ISD), from a meteorology station in Shawbury approximately 24 km west of the site of study. There was little discrepancy so the wind data from the BIFoR FACE facility was chosen for displayed results.

4.4 Results and discussion

4.4.1 Results summary

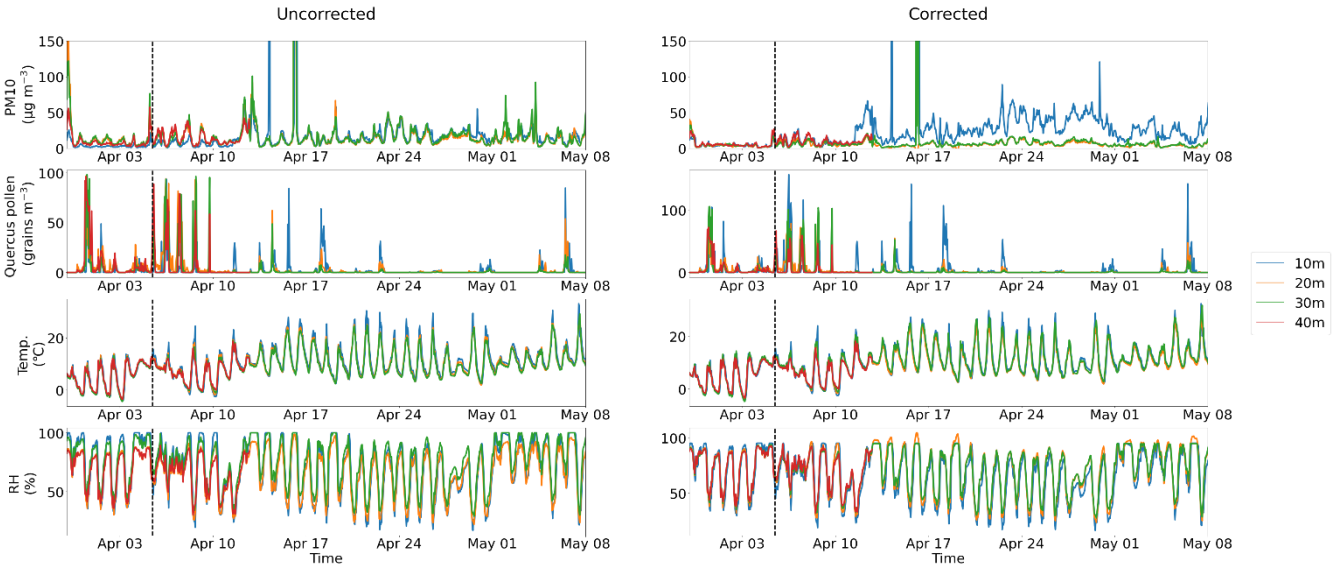


Figure 4.3: *Time series of PM_{10},* Quercus *pollen number concentration, temperature, and RH from top to bottom. On the left are uncorrected time series, while those on the right are corrected. (An equivalent set of plots can be found in Figure 9.1 in the Appendices which also includes* Poaceae *and total pollen time series.)*

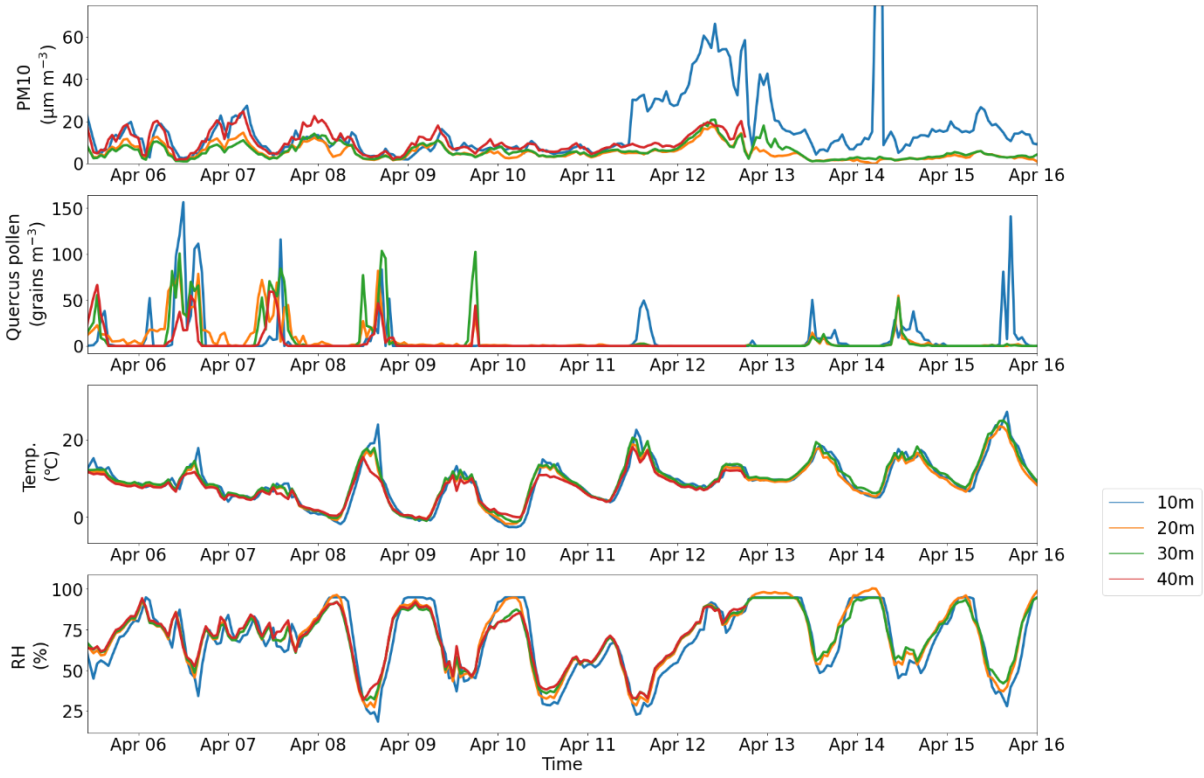


Figure 4.4: *Zoomed in time series from Figure 4.3 with only corrected data and from the start of the 'tower' period (5th April) up to 16th April 2022. (An equivalent set of plots can be found in Figure 9.3 in the Appendices which also include* Poaceae *and total pollen.)*

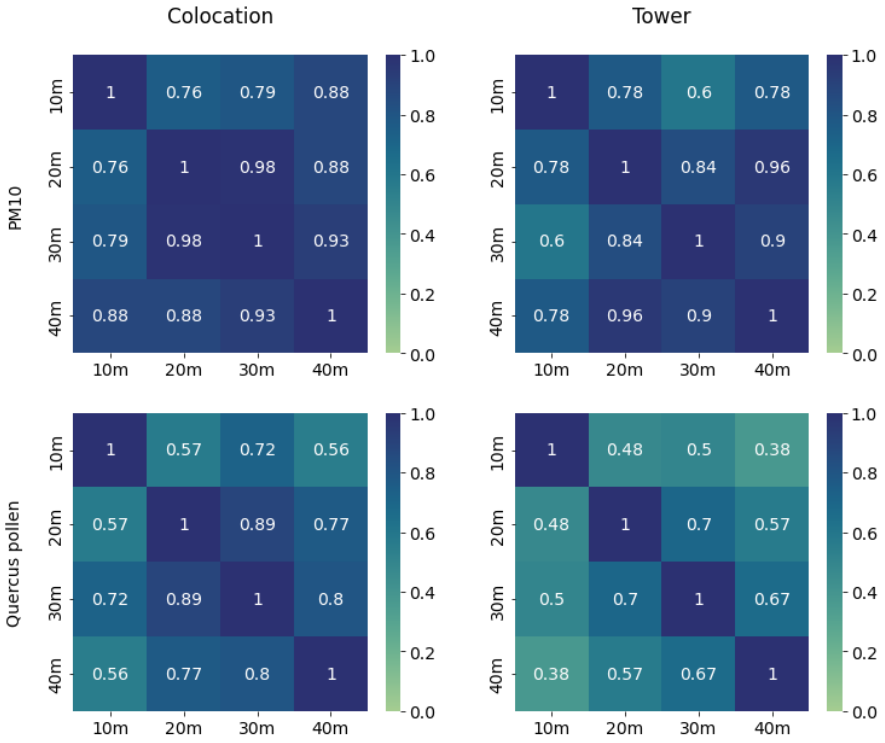


Figure 4.5: *Pearson correlation coefficients visualised as matrix plots for each sensor pairing for corrected PM_{10} (above) and* Quercus *pollen number concentration (below) for each of colocation (left) and tower (right) periods, respectively. Values for the tower period were only included up until 11th April to exclude the later period when the 40 m sensor stopped logging and the 10 m sensor PM_{10} values became anomalous compared to the other sensors.*

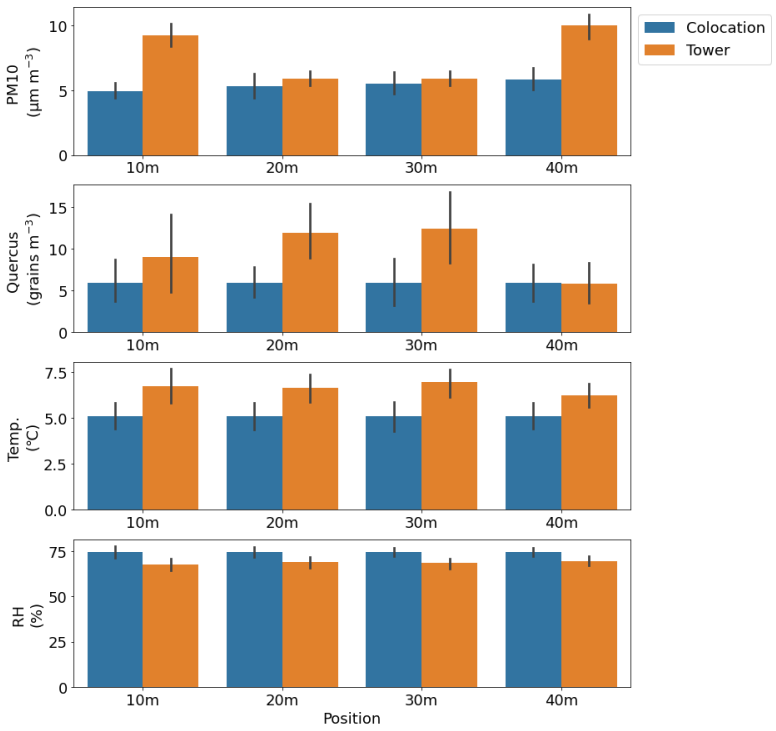


Figure 4.6: *Bar plots of mean values for each variable – PM_{10},* Quercus *pollen number concentration, temperature and RH - with 95% confidence intervals for each sensor (x-axis) and colocation/tower period (colour) respectively. Blue = colocation period; orange = tower period. Values for the tower period were only included up until 11th April to exclude the later period when the 40 m sensor stopped logging and the 10 m sensor PM_{10} values became anomalous compared to the other sensors. (An equivalent set of plots can be found in Figure 9.2 in the Appendices which also include* Poaceae *and total pollen.)*

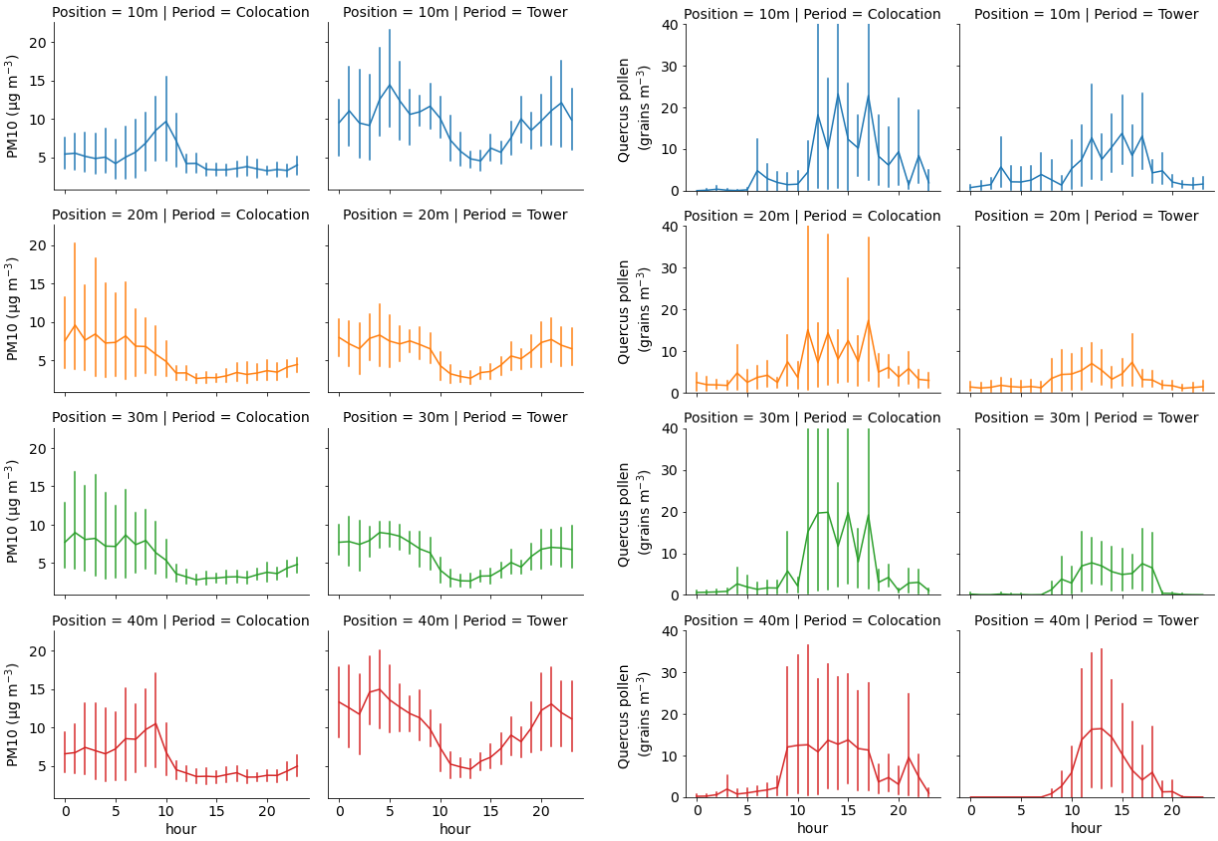


Figure 4.7: *Diurnal trends of PM_{10} (left grid) and* Quercus *pollen (right grid).* Quercus *pollen number concentrations averaged over each hour of the day, for each sensor (row & colour), for both colocation (left column) and tower (right column) periods. The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May). (Corresponding diurnal plots for* Poaceae*, total pollen, temperature, RH and wind speed can be found in Figures 9.4-9.8 in the Appendices.)*

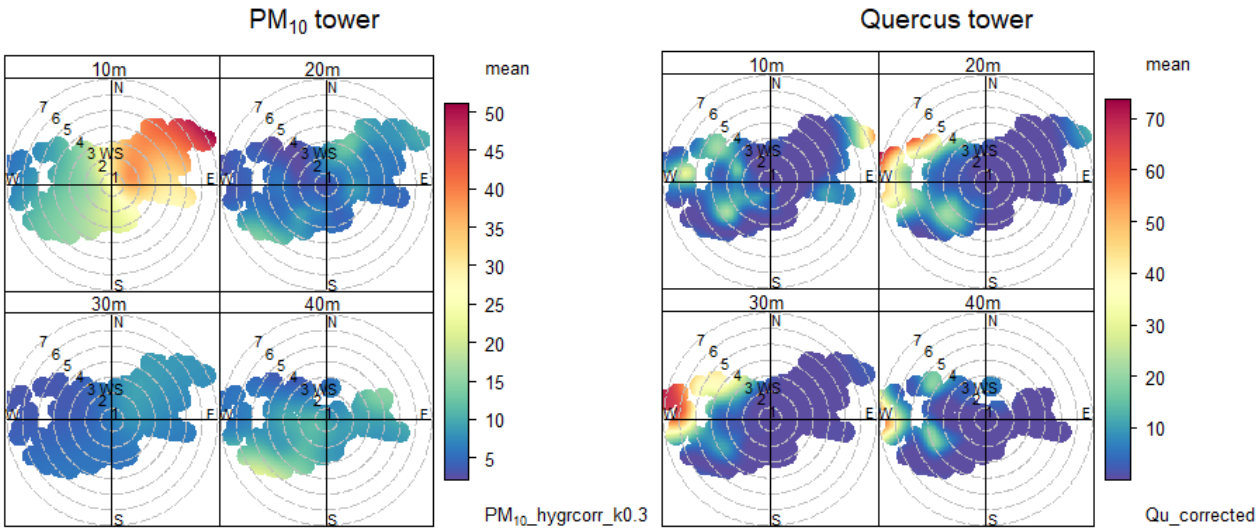


Figure 4.8: *Polar plots showing corrected PM_{10} mass concentration (left) and* Quercus *pollen proxy number concentration (right) as a function of wind speed and direction for each sensor height during the tower period. Similar plots for* Poaceae *pollen, total pollen, temperature and RH can also be found in Figures 9.9-9.12 in the Appendices.*

Figure 4.3 displays the time series, from 30th March until 8th May 2022, for each of the variables PM_{10} (µg m^{-3}), *Quercus* pollen number concentration (grains m^{-3}), temperature (°C) and RH (%) with uncorrected and corrected data in the left and right columns respectively. The colocation and distributed 'tower' periods are split by the black dashed line in each plot at 10:30 am (GMT) on 5th March. Meanwhile, Figure 4.4 displays a zoomed in section of the same time series presented in Figure 4.3, from the beginning of the tower period until 16th April. This is to facilitate comparison between sensors at different heights during the active oak pollen season. Equivalent plots which also include *Poaceae* and total pollen time series can be found in Fig. 9.1 and 9.2 in the Appendices.

Figures 4.3 and 4.4 immediately make two things visible. First, that the 40m sensor (red lines) ceased logging sometime during 12th April and, secondly, that the 10m sensor (blue lines) reports corrected PM_{10} mass concentrations with distinct discrepancy from the other sensors from 11th April onwards (5 days after the tower period began). Since this is a sudden change that occurred days after the sensors were repositioned (and up till the point the corrected concentrations are sensible), it was considered that the PM_{10} readings for the 10m sensor were anomalous from this point onwards. Further reasons for this are discussed in section 4.4.3. However, it is important to note that the following results and figures were calculated excluding the period after 11th April, in the interest of fair comparison among all four sensors.

The Pearson correlation coefficients between each sensor pairing during both periods, colocation on the left and tower period on the right, are visualised as matrix plot in Figure 4.5. The top two grids present the congruency among the sensors in terms of measured PM_{10} mass concentrations, while the two lower grids present the congruency regarding *Quercus* number concentrations. The difference in values between grids on the left and right demonstrate the decrease in correlation when the sensors were moved apart and positioned on the tower platforms.

Figure 4.6 presents bar plots of mean values for each variable and the 95% confidence intervals for each sensor, bars in blue for the colocation period and in orange for the tower period. The sensor at 40 m stopped logging after 12th April and the PM_{10} values of the 10 m sensor appear somewhat anomalous after 11th April so, for fair comparison, data only up until 11th April were included for the construction of these plots. Equivalent plots also including *Poaceae* and total pollen are in Figure 9.3 in the Appendices.

Figure 4.7 summarises the diurnal trend of PM_{10} mass concentration and estimated *Quercus* pollen number concentration for each sensor during each of the colocation and tower periods. Once again, the vertical lines represent 95% confidence intervals, which vary greatly since pollen concentrations vary considerably over time, affected by phenology, the finite supply of nearby sources, and meteorology. Diurnal plots for *Poaceae* pollen, total pollen, temperature, RH and wind speed can be found in Figures 9.4-9.8 in the Appendices.

Figure 4.8 shows the corrected concentrations of PM_{10} (left) and *Quercus* pollen (right) as a function of wind speed and direction. Similar plots for each other variables (*Poaceae*, total pollen, temperature and RH) are presented in Figure 4.9-4.12 in the Appendices.

4.4.2 Sensor congruency

The four sensors during the colocation period generally show good congruency, i.e. consistency among sensors, for PM_{10} measurements, visualised in the top left matrix plot in Figure 4.5. All sensor pairings achieved congruency above 0.75, with a minimum of 0.76 for the 10/20m pairing). When the sensors were moved apart for the tower period (top right plot in Figure 4.5), most correlation coefficients were reduced (by 0.03-0.14) except for the 10/20m and 20/40m pairings which increased slightly. The sensor at 10 m generally had the lowest correlation coefficients with the other sensors, with the largest discrepancy observed between the 10/30m and 10/30m pairings. These results suggest that the low-cost sensors do experience some variation in PM_{10} concentrations due to differences in height.

Meanwhile, the Pearson correlations for the *Quercus* pollen proxy between sensors during the colocation period (bottom left in Figure 4.5) are less consistent and generally

demonstrate lower congruencies than for PM_{10}. The pairing with the highest correlation coefficient is 20/30m at 0.89 while that with the lowest is 10/40m at 0.56. The correlation coefficients decreased by 0.09-0.22 when the sensors were moved apart for the tower period (bottom right plot in Figure 4.6) and the greatest disparity between tower platforms was observed between the pairing of greatest distance (10/40m).

4.4.3 Observed PM_{10} characteristics

Observed PM_{10} concentrations in the forest were generally low, averaging around 5-10 $\mu g\ m^{-3}$, (Figures 4.6 and 4.7), with diurnal maxima reaching approximately 20 $\mu g\ m^{-3}$ (Figures 4.3, 4.4 and 4.7). Since this is a rural area, located far from major urban areas or roads, it is expected that the air would be relatively clean and PM levels low. The sensor at the lowest height (10 m; blue line) appears to record significantly higher concentrations after 11th April (see Figure 4.3) but this was considered an anomalous systematic error; the low-cost OPC sensors can sometimes shift in accuracy over time or due to disrupting influences at the inlet (e.g., caused by insects or dirt). On further investigation, the anomalously high PM_{10} concentrations were found to be due to a specific OPC bin (Bin 9) corresponding to particle sizes between 5.2-6.5 μm (see Figure 9.13 in the Appendices). This suggests that there was specific anomalous activity near the sensor inlet which produced such a high concentration of particles in this size range; from our previous observations in this forest environment, an insect of some kind may likely be responsible.

From Figure 4.7, PM_{10} concentrations demonstrate a diurnal trend of peaking in the early hours of the morning between midnight and dawn (~6 am) before decreasing during daylight hours. This has generally been observed previously and is due to the boundary layer forming, or increased boundary layer height, which occurs after sunrise

accompanied by increases in solar radiation and wind speeds (Liu et al., 2015). The corresponding diurnal trends of air temperature and wind speed can be seen in Figures 9.6 and 9.8 in the Appendices. The increased height of the boundary layer and increased winds provide greater volume and energy for PM to concentrations to mix and be diluted.

It seems from Figure 4.6 (and Figure 4.4) that PM_{10} concentrations are generally greater below and above the canopy, at 10 and 40 m, while being suppressed within the canopy at heights of 20 and 30 m. This is likely due to increased obstruction and deposition caused by the denser vegetation of leaves and branches at this level.

The polar plot on the left in Figure 4.8 shows how PM_{10} concentration varied generally with wind speed and direction. The lowest sensor at 10 m detected relatively greater PM_{10} concentrations coming from the northeast, especially when wind speeds were high. While the sensors at 20 and 30 m do not show any significant relationship with wind speed or direction, the sensor at 40 m shows increased concentrations when wind blows from particularly southwest and northeast at relatively high speeds. This distinct difference between the sensors at lowest and highest points on the tower suggests that they are sometimes detecting PM_{10} from different sources. Note also that the 40 m sensor stopped logging before the campaign end while the 10 m sensor showed anomalously high concentrations after this time. It appears this period was responsible for the particularly high PM_{10} concentrations observed from the northeast. It is difficult to say if the 10 m encountered some problem or if there was some unusual (perhaps human) activity which caused such high concentrations at this height beneath the canopy. The sensor at 40 m is likely detecting general background sources of PM_{10} from further away (for example, from the nearest urban areas).

4.4.4 Observed *Quercus* pollen characteristics

Quercus is the specific pollen taxa we are interested in observing in this context, since the site is a mature, predominantly oak, forest and the time period coincides with the oak pollen season. From the time series estimated by our pre-trained neural network model in Figures 4.3 and 4.4, we see that there are sharp peaks in daily cycles which reach varying concentrations up to 150 grains m^{-3}. At the troughs between these peaks, concentrations are generally ~0 grains m^{-3}. This suggests that, rather than background pollen levels, it is pollen from nearby sources that are being detected which follow the diurnal cycles of pollen emissions.

The diurnal plots of *Quercus* pollen in Figure 4.7 show a trend that is distinct, even opposite, to that of general PM_{10}: *Quercus* pollen concentrations tend to increase and peak during daylight hours and are lowest at night. This is congruent with diurnal trends demonstrated in Juprasong et al. (2022) and Ríos et al. (2016) and summarised by Jones and Harrison (2004). (The reader is reminded here that the machine learning algorithms were not trained on meteorological conditions, that contain this diurnality.) In general, pollen concentrations increase during daylight hours with an inverse relationship with relative humidity (observed in our results too in Figure 9.7). This is due to surfaces drying out and bioaerosol release mechanisms being more conducive to lower humidity conditions. The diurnal pattern may also be affected by wind speeds (our results in Figure 9.8), downward mixing from higher concentrations above, or source limitations. The sensor at the highest platform on the tower (40 m) appears to detect greater peak concentrations during daylight hours than the sensors on lower platforms. It is likely that the sensor placed highest above the forest canopy is able to

capture the most typical diurnal trend, contributed more generally by surrounding or background pollen sources.

Comparing the time series of *Quercus* pollen and meteorological variables temperature and RH in Figures 4.3 and 4.4, it appears that consecutive days of *Quercus* pollen activity tend to occur when there is some disruption in the typical meteorological diurnal cycles. For example, the burst of significant *Quercus* activity between 5-10th April follows a period when temperature is sustained at relatively high values and the diurnal cycle troughs do not reach as low as previous typical days. After a long process of maturation, pollen release (via anther dehiscence) is finally stimulated when a certain temperature threshold is reached (Van Vliet et al., 2002). Importantly, the pollen release occurs when the RH drops (after a period of sustained high values, see Figure 4.3) and for those days of pollen release the diurnal RH cycle seems somewhat suppressed. The anthers within the catkins that contain pollen open when exterior or epidermal cells dry out and contract in proportion to the inner surface. Hence, low RH and high temperatures are important stimulating factors, as explained by Woodroof (1930) and Yates and Sparks (1993).

4.4.5 Observed *Quercus* pollen characteristics across different heights

Interestingly, from Figure 4.6, *Quercus* pollen concentrations appear to be higher at the platforms within the canopy at 20 and 30 m, compared to those above and below at 10 and 40 m, which is the converse of what is observed for PM_{10}. This could be attributed to the direct release of pollen by surrounding trees at that level before being mixed and diluted vertically. On average, the weakest pollen concentrations are observed at the highest platform at 40 m. This is congruent with observations by Xiao et al. (2013) of decreasing pollen concentration with height and may result from meteorological factors,

as well increased distance from the forest canopy source and greater mixing freely within the boundary layer.

The polar plot on the right in Figure 4.8 gives evidence to suggest that the *Quercus* pollen detected by each sensor is generally coming from the same sources, particularly to the North and West, and is generally detected at higher wind speeds. The impact of wind velocity on detected pollen concentrations decreases with decreasing height (i.e. the red areas in the polar plot of high pollen concentration are less localised at 10 and 20 m compared to 30 and 40 m). This is not surprising as the area beneath the canopy would be sheltered and less affected by wind. Yet, as observed in Figures 4.3, 4.4 and 4.6, *Quercus* concentrations are not diminished at 10 m. From the relative concentrations across the heights, the sensors do not seem to be detecting sources from far afield, but still high wind speeds appear to be conducive for pollen release. This would also be a function of boundary layer dynamics and the diurnal cycle discussed previously.

The *Quercus* pollen season in Figures 4.3 and 4.4 attenuates over time through the month of April for all four sensors, as the sources deplete. However, the highest sensor at 40 m (red) attenuates earlier while the lowest sensor at 10 m (blue) records the most significant peaks later in the season. (Though, note that the 40 m sensor (red) stopped logging after 12th April.) Direct observations in the forest at the time noted that catkins generally matured first from the top of the canopy, where there is more exposure to sunlight, before maturing at lower levels. This could explain the earlier attenuation of pollen concentrations at higher levels and emissions remaining more significant at lower levels later in the season.

Another explanation could be that the detected pollen is being released from different trees across the season and the distance of the actively pollen-releasing tree from the

point of measurement may impact the relative concentrations detected at each height. A recent study was conducted by Baird et al. (2022) which also used Alphasense OPC sensors at the same site (the BIFoR FACE facility) to measure a bioaerosol proxy during winter months (November and December). The proxy measurement in this case was derived from the instrument's intrinsic PM mass concentration measurements (PM_{10} – PM_1) and the main target, given the season and observed size range, was fungal spores. However, they created a bioaerosol dispersal model and demonstrated that, under varying wind speed conditions and release heights, the distance travelled by spores was comparable to the area within the forest. Pollen grains have greater mass than spores so likely descend vertically faster, though they are also generally released from greater heights. Meanwhile, due to the nature of our proxy method, it may also be observing associated subpollen particles which may be more comparable in size to spores. In any case, the Baird et al. (2022) study demonstrates that the distance from the source may significantly impact the vertical height when it reaches the point of measurement, within the scale of this forest.

4.4.6 Limitations

Generally, there is large variance in measured concentration among sensors when pollen peaks occur (Figure 4.4), making it difficult to base solid conclusions of the relative concentrations when sensors are separated on the tower. This is not surprising, however, not only because they are low-cost sensors with limited accuracy and precision. The machine learning model must learn how best to construct pollen concentrations from OPC data using given examples of 'correct' output, called 'labels' in supervised machine learning. In this case, the training data labels were pollen concentrations obtained via the Hirst sampler and manual counting method. While this

currently the conventional benchmark method we have, it is also susceptible to considerable uncertainty and variance across different instruments and people counting. This uncertainty increases greatly for higher time resolutions (i.e. hourly as opposed to daily) and when focusing on individual pollen taxa as opposed to collective 'total' pollen (Adamov et al., 2021).

The methodology still needs progress, in terms of the models' ability to accurate distinguish and estimate pollen concentrations. One current limitation is that we are applying a model trained on limited data and from another environmental context than that used in this study. The dataset used for training, described in Mills et al. (2023a), was relatively small by deep learning standards and the intercomparison campaign from which the data was obtained was based in a suburban background site in Germany. The dataset at the time was also largely dominated by *Poaceae* (grass) and *Pinus* (pine) pollen types. Contrastingly, the data which the pre-trained model was applied to for this study was obtained in a rural forest site in the UK, where we may expect the dominant type to be *Quercus* or other tree species due to the time period and location.

It is also important to note the physical limitations of the current OPC model, which we have found can be less effective at detecting larger particles and lower background concentrations. The current machine learning algorithm learns the best it can despite this limitation, but would certainly benefit from better detection capability in the larger size ranges. This is something we encourage the commercial companies that produce the sensors to address.

Nevertheless, the results presented in this study strongly suggest that the models have worked and can provide meaningful information, which was not previously possible. Importantly, we are able to obtain this information at a much higher spatiotemporal

resolution than is possible using conventional measurements techniques. This study serves as a demonstration for how low-cost sensors in combination with machine learning methodology can be applied. As more training data is collected, the models should further improve.

4.4.7 Comments on *Poaceae* and total pollen for comparison

The time series for *Poaceae* and total pollen proxy number concentrations are shown in Figures 9.1 and 9.2 in the Appendices, as well as bar plots in Figure 9.3. The diurnal trends for *Poaceae* and total pollen are shown in Figures 9.4 and 9.5. These results have been added mainly for comparison beside the *Quercus* pollen proxy. The time period of this study is too early for the *Poaceae* pollen season, so we would not expect to observe *Poaceae* pollen or its proxies, and hence the estimated concentrations should be low. As discussed in Mills et al. (2023a) and Mills et al. (2023b) the total pollen model trained previously would be heavily dependent on the specific partial contributions that comprised 'total pollen' in that context, since it has only been possible to train on one dataset so far. The total pollen model in this case would be strongly influenced by *Poaceae* and *Pinus* taxa since these were predominant in that dataset. To achieve a more generalisable total pollen proxy, it would be necessary to train the models on data across more varied contexts.

Nevertheless, we can observe that the *Poaceae* and total pollen characteristics are distinctly different from *Quercus* during the period of this study. We demonstrated in Mills et al. (2023b) that the *Quercus* model learned relationships from the input features that were distinctly different from the other models. The comparison here in an entirely new environment emphasise that the model has learned to measure something that is distinctly different from PM_{10}, *Poaceae* or total pollen. Alongside the fact that the

Quercus proxy here demonstrates convincing characteristics - i.e. following expected diurnal trends and temporal variation can be explained by meteorology and understanding of source phenology – we conclude that it is likely to be a useful *Quercus* pollen proxy.

4.5 Conclusions

There are significant gaps in our understanding of the characteristics and distribution of airborne pollen and demand for instruments and methodologies that can address this (Buters et al., 2018; Buters et al., 2020). We have applied a methodology that utilises machine learning models to process data from low-cost OPC sensors and output pollen concentrations in a novel environment that has not been possible to investigate before in such detail. Thus, we have been able to characterise PM_{10} mass concentrations and *Quercus* (oak) pollen number concentrations within a mature oak forest, during the oak pollen season, and investigate the impact of height on concentrations through the canopy.

We present evidence that these low-cost sensors, with well-trained machine learning models, can provide detailed information on pollen concentrations released from nearby sources and distinguish some variation across a vertical profile. This information includes period of emission, diurnal patterns and interaction with meteorological variables recorded with high time resolution.

There are current limitations due to training data limitations associated with the machine learning. Further studies would be required to train models and assess accuracy against benchmark pollen concentrations across various environments, though such datasets are difficult to obtain particularly at high time resolutions. It would no doubt be valuable to train the machine learning models using data obtained from more recent automated pollen monitoring instruments, since these have demonstrated better congruency across instruments (Oteros et al., 2020) and suggest superior accuracy (Plaza et al., 2022) than the manual Hirst. Machine learning methods tend to be heavily

dependent on data quantity and quality, so careful curation of suitable data for training would be necessary to develop this method for public or commercial use.

Nevertheless, this study demonstrates the potential of this novel methodology, to supplement regulatory grade pollen sites and collect high spatiotemporal resolution data across unique, local environments where we currently have critically limited understanding of airborne pollen concentrations. This could even be extended to indoor environments since these models use particle size data only and not meteorological variables. Collectively, a wealth of new information could be provided to benefit both the scientific community and public health.

4.6 Contributors

This concept of this study was conceived and planned by FDP, ARM and SAM. SAM managed the sensor development and fieldwork with the help of those acknowledged in Section 4.7, collected and curated the data, developed the methodology of the machine learning models applied, performed analyses and produced the first draft. All authors contributed to editing.

4.7 Acknowledgements

We acknowledge the funding support of the Natural Environment Research Council (NERC) CENTA2 grant NE/S007350/1 via the University of Birmingham, and the grant "Quantification of Utility of Atmospheric Network Technologies (QUANT)" (NE/T001968/1). We thank Andrew Tanner from the University of Birmingham, UK, for his technical work on functionalising the Alphasense OPC sensors for the purposes of this study. We thank the staff at the BIFoR FACE facility – Peter Miles, Kris Hart, Nicholas Harper and Thomas Downes – for facilitating the practical deployment of the sensors in this study and technical assistance. We also acknowledge José Maya-Manzano, Fiona Tummon and all others involved in the EUMETNET AutoPollen – ADOPT COST Action intercomparison campaign. The models used for this study were trained on data obtained from that campaign (described in Maya-Manzano et al., 2023) and thus made this study possible.

Icons used to create the diagram in Figure 4.2 came from the following sources: https://www.iconspng.com/image/118299/colored-oak-tree; https://www.cleanpng.com/png-shrub-free-content-website-clip-art-flower-bush-cl-168629/; https://pixabay.com/vectors/mushroom-big-fungus-brown-29738/;

http://www.clker.com/clipart-768974.html; https://www.freepik.com/premium-vector/log-with-mushrooms-growing-it_1377390.htm; https://creazilla.com/nodes/38722-wind-and-clouds-clipart (last accessed June, 2023).

4.8 Comment

This chapter follows on from the previous Chapters 2 and 3 and applies the improved models from Chapter 3 to data obtained from a novel environmental context. This is an important step to test the generalisability of the trained machine learning models. It also tests the efficacy of deploying such low-cost sensors in a network to capture high spatial resolution variation (vertically in this case, but could also apply horizontally) and the potential wealth of valuable detail they could provide. This is a valuable first study of its kind, using automated low-cost sensors that can provide continuous high time resolution data to monitor near-source pollen emissions through the canopy in a mature forest. A limitation is that it was not possible to measure benchmark pollen concentrations within the same environment to truly assess the accuracy of the results. Nevertheless, the results from this study suggest that the (*Quercus*) pollen concentrations derived from the machine learning model demonstrate convincing promise. This study should be of great interest to people in the field of bioaerosols and should encourage the necessary actions to take this technique further. In particular, facilitating the collection of further comparison data sets to train the models, in different environments and with different instruments, would have huge potential to improve performance and generalisability.

4.9 References

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers, Aerobiologia. https://doi.org/10.1007/s10453-021-09724-5.

Baird, A.B., Bannister, E.J., MacKenzie, A.R., Pope, F.D., 2022. Mass concentration measurements of autumn bioaerosol using low-cost sensors in a mature temperate woodland free-air carbon dioxide enrichment (FACE) experiment: investigating the role of meteorology and carbon dioxide levels. Biogeosciecnes, 19, 2653-2669. https://doi.org/10.5194/bg-19-2653-2022.

Bannister, E.J., Jesson, M., Harper, N.J., Hart, K.M., Curioni, G., Cai, X., MacKenzie, A.R., 2023. Residence times of air in a mature forest: observational evidence from a free-air CO_2 enrichment experiment. Atmos. Chem. Phys., 23(3), 2145-2165. https://doi.org/10.5194/acp-23-2145-2023.

Buters, J.T.M, Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allergy, 8, 9. https://doi.org/10.1186/s13601-018-0197-8.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.A., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Kofol Seliger, A., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto, A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia, https://doi.org/10.1007/s10453-022-09750.

Carslaw, D.C., Ropkins, K., 2012. *Openair* – An R package for air quality data analysis. Environmental modelling & Software, 27-28, 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008.

Crilley, L.R., Shaw, M., Pound, R., Kramer, L.J., Price, R., Young, S., Lewis, L.C., Pope, F.D., 2018. Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring. Atmos. Meas. Tech., 11(2), 709-720. https://doi.org/10.5194/amt-11-709-2018.

Hart, K.M., Curioni, G., Blaen, P., Harper, N.J., Miles, P., Lewin, K.F., Nagy, J., Bannister, E.J., Cai, X.M., Thomas, R.M., Krause, S., Tausz, M., MacKenzie, A.R., 2019. Characteristics of free air carbon dioxide enrichment of a northern temperate mature forest, Global Change Biology, 26(2), 1023-1037. https://doi.org/10.1111/gcb.14786.

Hugg, T.T., Tuokila, M., Korkonen, S., Weckström, J., Jaakkola, M.S., Jaakkola, J.J.K., 2020. The effect of sampling height on grass pollen concentrations in different urban environments in the Helsinki Metropolitan Area, Finland, PLoS ONE. https://doi.org/10.1371/journal.pone.0239726.

Jones, A.M., Harrison, R.M.: The effects of meteorological factors on atmospheric bioaerosol concentrations – a review, Sci. Total Environ., 326(1-3), 151-180, doi: 10.1016/j.scitotenv.2003.11.021, 2004.

Juprasong, Y., Sirirakphaisarn, S., Siriwattanakul, U., Songnuan, W., 2022. Exploring the effects of seasons, diurnal cycle, and heights on airborne pollen load in a Southeast Asian atmospheric condition. Front. Public Health, 10, 1067034. https://doi.org/10.3389/fpubh.2022.1067034.

Liu, Z., Hu, B., Wang, L., Wu, F., Gao, W., Wang, Y., 2015. Seasonal and diurnal variation in particulate matter (PM_{10} and $PM_{2.5}$) at an urban site of Beijing: analyses from a 9-year study. Environ. Sci. Pollut. Res., 22, 627-642. https://doi.org/10.1007/s11356-014-3347-0.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, Ł., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavilj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2023. Towards European automatic bioaerosol monitoring: Comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ., 866, 161220. https://doi.org/10.1016/j.scitotenv.2022.161220.

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023a. Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests. Sci Tot. Environ., 871, 161969. https://doi.org/10.1016/j.scitotenv.2023.161969.

Mills, S.A., Maya-Manzano, J.M, Tummon, F., MacKenzie, A.R., Pope, F.D., 2023b. Machine Learning Methods for Low-Cost Pollen Monitoring – Model Optimisation and Interpretability, 903, 165853. https://doi.org/10.1016/j.scitotenv.2023.165853.

Norby, R.J., Kauwe, M.G.D., Domingues, T.F., Duursma, R.A., Ellsworth, D.S., Goll, D.S., Lapola, D.M., Luus, K.A., MacKenzie, A.R., Medlyn, B.E., Pavlick, R., Rammig, A., Smith, B., Thomas, R., Thonicke, K., Walker, A.P., Yang, X., Zaehle, S., 2015. Model-data synthesis for the next generation of forest free-air CO_2 enrichment (FACE) experiments, New Phytologist, 209(1), 17-28. https://doi.org/10.1111/nph.13593.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition, Environ. Res., 191, 110031. https://doi.org/10.1016/j.envres.2020.110031.

Plaza, M.P., Kolek, F., Leier-Wirtz, V., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2022. Detecting Airborne Pollen Using an Automatic, Real-Time Monitoring System: Evidence from Two Sites. Int. J. Environ. Res. Public Health, 19(4), 2471. https://doi.org/10.3390/ijerph19042471.

Ríos, B.E., Torres-Jardón, R., Ramírez-Arriaga, E., Martínez-Bernal, A., Rosas, I., 2016. Diurnal variations of airborne pollen concentration and the effect of ambient temperature in three sites of Mexico city. Int. J. Biomet., 60(5), 771-787. https://doi.org/10.1007/s00484-015-1061-3.

Rojo, J., Oteros, J., Pérez-Badia, R., Cervigón, P., Ferencova, Z., Gutiérrez-Bustillo, A.M., Bergmann, K.-C., Oliver, G., Thibaudon, M., Albertini, R., Rodríguez-De la Cruz, D., Sánchez-Reyes, E., Sánchez-Sánchez, J., Pessi, A.-M., Reiniharju, J., Saarto, A., Calderón,

M.C., Guerrero, C., Berra, D., Bonini, M., Chiodini, E., Fernández-González, D., García, J., Trigo, M.M., Myszkowska, D., Fernández-Rodríguez, S., Tormo-Molina, R., Damialis, A., Kolek, F., Traidl-Hoffmann, C., Severova, E., Caeiro, E., Ribeiro, H., Magyar, D., Makra, L., Udvardy, O., Alcázar, P., Galán, C., Borycka, K., Kasprzyk, I., Newbigin, E., Adams-Groom, B., Apangu, G.P., Frisk, C.A., Skjøth, C.A., Radišić, P., Šikoparija, B., Celenk, S., Schmidt-Weber, C.B., Buters, J., 2019. Near-ground effect of height on pollen exposure, Environ. Res., 174, 160-169. https://doi.org/10.1016/j.envres.2019.04.027.

Van Vliet, A.J.H., Overeem, A., De Groot, R.S., Jacobs, A.F.G., Spieksma, F.T.M., 2022. The influence of temperature and climate change on the timing of pollen release in the Netherlands. Int. J. Climatol., 22, 1757-1767. https://doi.org/10.1002/joc.820.

Wagle, P., Jones, J., Bielory, L., 2018. A comparison of *Plantago major* pollen counts at two different heights, J. Allergy Clin. Immunol., 141(2), AB31. https://doi.org/10.1016/j.jaci.2017.12.100.

Woodroof, J.G., 1930. Studies of the staminate inflorescence and pollen of Hicoria pecan. Ph.D. Thesis, Michigan State College of Agriculture and Applied Science. https://doi.org/doi:10.25335/M5K51F.

Xiao, X., Fu, A., Xie, X., Kang, M., Hu, D., Yang, P., Liu, Z., 2013. An investigation of airborne allergenic pollen at different heights, Int. Arch. Allergy Immunol., 160(2), 143-151. https://doi.org/10.1159/000339673.

Yates, I.E., Sparks, D., 1993. Environmental Regulation of Anther Dehiscence and Pollen Germination in Pecan. J. Amer. Soc. Hort. Sci., 118(6), 699-706.

5. CHAPTER 5: Acoustic levitation of pollen and visualisation of hygroscopic behaviour

This study has been published in the journal *Atmospheric Measurement Techniques*:

Mills, S.A., Milsom, A., Pfrang, C., MacKenzie, A.R., Pope, F.D., 2023. Acoustic levitation of pollen and visualisation of hygroscopic behaviour. Atmos. Meas. Tech., 16, 4885-4898. https://doi.org/10.5194/amt-16-4885-2023.

5.1 Abstract

Pollen are hygroscopic and so have the potential to act as cloud condensation nuclei (CCN) in the atmosphere. This could have yet uncertain implications for cloud processes and climate, as well as plant biodiversity and human health. Previous studies have investigated the hygroscopic swelling of pollen, linked to CCN activity by the κ-Köhler theory, using methods that follow observed mass increase by electrodynamic balance (EDB) or vapour sorption analyser. This study uses an acoustic levitator to levitate pollen grains in the true aerosol phase and uses a macroscope to image the pollen to investigate hygroscopic behaviour when relative humidity (RH) is changed. Two pollen species were studied in this work: *Lilium orientalis* (oriental lily) and *Populus deltoides* (eastern cottonwood). Both species were successfully levitated, however, the smaller *Populus deltoides* showed greater instability throughout experiments. The quality of images taken by the macroscope, and thus calculations of pollen area and aspect ratio, varied significantly and were sensitive to lighting conditions, as well as levitated pollen grain movement and orientation. Experiments with surface-fixed pollen grains were also conducted. They showed evidence that pollen hygroscopic swelling could be observed by the macroscope. The produced results were comparable with previously reported mass increase values. Although less accurate than methods that measure mass changes, the acoustic levitator and macroscope setup offer an attractive alternative by virtue of being commercial-off-the-shelf, low-cost, and versatile. A key advantage of this method is that it is possible to visually observe particle shape dynamics under varying environmental conditions.

5.2 Introduction

Primary biological aerosol particles (PBAPs, or bioaerosols) are airborne particles derived from biological organisms (Després et al., 2012) and come from a huge variety of sources including plants, animals and microorganisms. They comprise a significant proportion of total atmospheric aerosol mass, calculated to be around 20-30% in urban/rural regions (Matthias-Maser and Jaenicke, 1995; Matthias-Maser et al., 2000) and 10% and 20% in marine and remote continental regions, respectively, (Gruber et al., 1998) for PBAPs greater than 0.2 μm in diameter. As with other aerosol particles, PBAPs are important in the Earth system and while in the atmosphere they undergo transportation, chemical and physical transformations, and cloud processing (Frohlich-Nowoisky et al., 2016; Després et al., 2012). PBAPs have been found to constitute almost 25% of insoluble aerosol particles with diameters greater than 0.2 μm in cloud water (Matthias-Maser et al., 2000) and between 9-50% in rainwater (Diehl et al., 2001), so their potential to influence cloud processes and affect precipitation cannot be neglected (Andreae & Rosenfield, 2008).

Meanwhile, PBAPs play an important biological role, with pollen being responsible for the spread of plant species and colonisation of new ecosystems. Furthermore, many have direct relevance to human health. Pollen are responsible for allergic reactions such as hay fever and asthma. The size and aerodynamic ability of these particles affect their average travel distance and dynamics, so understanding of the physical processes affecting these particles is essential for accurate modelling, spatial concentration and personal exposure estimations.

Pollen particles are among the largest in size of PBAPs, generally measuring between 10 and 100 μm in diameter (Mills et al., 2023a), yet vary considerably in size and shape

between and even within species. While they are not as numerous in the atmosphere as other PBAPs such as bacteria and fungal spores, pollen particles contribute significantly to the global emissions mass, estimated to be around 47-84 Tg yr^{-1} (Després et al., 2012). Most airborne pollen comes from anemophilous trees, shrubs and grasses, whose pollen grains have evolved to be dispersed by wind and are therefore smaller in size, typically between 20-45 μm (Steiner et al., 2015). Such pollen species have characteristic seasons of pollen release so airborne pollen concentrations fluctuate throughout the year, and across the globe depending on local vegetation species. Airborne pollen concentrations are measured to be anywhere between 10-1000 grains m^{-3} (Després et al., 2012), but can reach over 10,000 grains m^{-3}, particularly in the Northern hemisphere spring months of April and May when many deciduous trees peak in pollen emissions (Steiner et al., 2015).

Pollen grains have been shown to reach altitudes of 3000 m and above in the atmosphere and can reside there for considerable lengths of time with favourable meteorology (Diehl et al., 2001). Under such circumstances, pollen grains have the potential to act as cloud condensation nuclei (CCN) - atmospheric particles which provide a surface for water vapour to condense on and cloud droplets to grow. A prerequisite for active CCN particles is that their surface is wettable and they are hygroscopic, and pollen grains have been demonstrated to adsorb moisture from the air and swell under high humidity, displaying moderate hygroscopicity (Pope, 2010; Griffiths et al., 2012; Chen et al., 2019). This has been attributed to the numerous hydrophilic hydroxyl (OH) groups of organic compounds in pollen (Tang et al., 2019). Note that when we refer to hygroscopicity of pollen it should more correctly be 'apparent hygroscopicity', since pollen grains are a mixture of soluble and insoluble components.

Though relatively small in number concentration, the large size of pollen grains may enable them to act as 'giant CCN' (GCCN), which can nucleate cloud droplets at lower supersaturations than other smaller aerosols and facilitate precipitation more rapidly (Després et al., 2012; Pope, 2010; Posselt & Lohmann, 2008). GCCN particularly affect precipitation in clouds of high droplet concentration, i.e. polluted as opposed to pristine clouds, even at concentrations as low as 10^{-3} cm^{-3} (Möhler et al., 2007). Additionally, pollen grains and their extracts have shown potential to act as ice nuclei (IN) and induce heterogenous ice formation in the atmosphere (Diehl et al., 2001 & 2002; Dreischmeier et al., 2017; Pummer et al., 2012; Tong et al., 2015). While there have been increasing discussions in the literature postulating the significance pollen may have for atmospheric cloud processes and climate, there is need for further experiments to measure and better understand the hygroscopic behaviour of pollen and other biological particles (Möhler et al., 2007; Després et al., 2012; Sun & Ariya, 2006). Pollen hygroscopicity may also have important implications for airborne transmission to new areas, and thus the colonisation and biodiversity of various plant species. Additionally, it is important to consider for human health, particularly since the release of respirable-sized subpollen particles from pollen grains can be modulated by humidity conditions (Matthews et al., 2023) and stimulated by events such as thunderstorms (Hughes et al., 2020; Cecchi et al., 2021).

Previous studies have attempted to quantify the hygroscopic water uptake of a variety of different types of pollen under varying RH (relative humidity) conditions by measuring the corresponding mass change. Table 5.1 presents a summary of the relevant findings of each study, including the pollen species studied, the method employed, the RH interval and corresponding mass increase. Some of the studies also report corresponding κ values, derived from the κ-Köhler theory (Petters and Kreidenweis,

2007) which links the hygroscopic size change of a particle with CCN activity, based on Köhler theory (Köhler, 1936). This κ value is determined from fitted hygroscopic growth factor data and quantifies hygroscopic growth, and thus CCN potential (Petters and Kreidenweis, 2007).

A particle is non-hygroscopic at κ = 0 and increases in hygroscopicity and CCN potential with increasing κ value. This κ value is generally found in the range 0.1-0.9 for atmospheric particles, with an upper limit of approximately 1.4 for sodium chloride (sea salt). However, the range for 'slightly to very' hygroscopic organic species is typically found towards the lower end of the scale between 0.01 and 0.5 (Petters and Kreidenweis, 2007). Previous studies have calculated κ values of various pollen species under subsaturated RH to be between 0.05-0.1 which, while suggesting moderate-low hygroscopicity, is not negligible since there is significant CCN potential at critical supersaturations considering their large size (Pope, 2010; Tang et al., 2019).Pollen grains are complex biological structures with varying sizes, shapes and surface features across species. Previous methods have focused on the measurement of mass increase with hydration and have not considered the simultaneous changes that may occur in volume and density. For example, Božič and Šiber (2022) explore the mechanics of pollen grain unfolding, swelling and bursting under humidity changes, which is largely dictated by the soft apertures, or pores, found on the exine (outer surface). The number and size of pores can vary across different pollen taxa but the mechanics involving them are not well understood. Alongside modelling studies such as Božič and Šiber (2022), empirical studies which shed light on the visual changes that occur on the pollen grain surface would bring value. With these considered, we may better understand the complexities of pollen hygroscopic behaviour and how it should be modelled in the atmosphere.

This study presents a novel approach, using an acoustic levitator to suspend pollen and a macroscope camera to capture images of the grains as they undergo changes in RH. As opposed to following the hygroscopic swelling of the pollen grains by the change in mass, this approach investigates the feasibility of following the hygroscopic behaviour visually by the change in pollen size.

Table 5.1: *Previous studies investigating the hygroscopic behaviour of pollen, including the pollen species used for each study, the apparatus used for each method, the RH interval studied, the corresponding mass increase percentage observed (the average or range across all pollen species for each study is generally cited), and the corresponding κ value from the κ-Köhler theory where it has been calculated.*

Reference	Pollen species	Method	RH interval	% mass increase	κ value
Diehl et al. 2001	*Betula alba*, *Alnus incana*, *Populus nigra*, *Quercus rubra*, *Pinus sylvestris*, *Picea alba*, *Agrostis alba*, *Dactylis glomerata*, *Poa pratensis*, *Secale cereale*	Analytical balance	Dry-73% Dry-95%	3-16% ~100-300%	-
Pope 2010	*Narcissus spp.*, *Betula occidentalis*, *Salix caprea*, *Juglans nigra*	Electrodynamic balance	0-75%	16%	0.05-0.10
Griffiths et al. 2012	*Secale cerale, Kochia scoparia, Artemisia tridentata, Iva xantifolia*	Electrodynamic balance	0-75%	18%	0.05-0.22
Bunderson & Levetin 2015	*Juniperus ashei*, *J. monosperma*, *J. pinchotii*	Analytical balance	20-97%	21-25%	-
Tang et al. 2019	*Carya illinoinensis* (pecan) *Populus tremuloides, Populus deltoides, ragweed, corn, paper mulberry*	Vapour sorption analyser	0-30%; 0-60%; 0-90% 0-95% 0-90%	2.3% 6.4% 30.3% ~72% 29-48%	0.034-0.067
Chen et al. 2019	*Pinus massoniana, Pinus tabuliformis, Pinus armandii, Pinus taiwanesis, Pinus bungeana, Typha angustifolia, Pyrus sp., Amygdalus persica, Malus pumila, Prunus salicina, Brassica campestris*	Vapour sorption analyser	0-90%	33-43%	0.036-0.048

5.3 Methods

5.3.1 Pollen samples

Two different types of pollen were used in this study: *Lilium orientalis* (oriental lily) and *Populus deltoides* (eastern cottonwood). The *Lilium orientalis* pollen was collected from fresh lily flowers bought from a UK Sainsbury's supermarket as the anthers ripened. These flowers' pollen is not anemophilous, it is usually insect-pollinated. However, the large grain size made it a good candidate for initial experiments with the acoustic levitator, since larger particles are easier to levitate. *Lilium* pollen grains are generally large and have an ellipsoidal shape with the polar axis generally ranging between 70-120 μm while the equatorial axis is often around 30-70 μm, i.e. with an approximate polar-equatorial ratio of 2:1 (Du et al., 2014).

The *Populus deltoides* pollen was commercially sourced from Merck (Sigma-Aldrich) and is smaller than the *Lilium orientalis* pollen, with optical diameter generally between 23-34 μm as quoted by Rajora (1989). It was used in this study to represent common anemophilous tree pollen species, such as birch and oak, which are generally between 10-40 μm in size and generally more spheroidal in shape (polar-equatorial ratio between 0.88-1.14, as defined by Punt et al., 2006). *Populus deltoides* pollen has a somewhat triangular polar-view outline with usually three apertures like other common tree pollen species including birch and oak.

Pollen grains from the samples used in this study (see Figure 10.1 in the Appendices for microscope images) were measured using a Nikon SPZ1000 stereo microscope equipped with a Nikon DS-Fi1 camera and controlled by NIS-Elements acquisition software at the Birmingham Advanced Light Microscopy (BALM) facility at the University of Birmingham. The *Lilium orientalis* pollen from fresh flowers was measured

to have polar and equatorial diameters ranging between 92-122 μm and 43-67 μm and means of 108.6 μm and 50.9 μm respectively. The *Populus deltoides* pollen used for this study was measured to have diameters across multiple axes ranging between 18 and 32 μm, with a mean of 24.8 μm. There is however considerable variation in size and shape across even pollen grains of the same taxa since these are gametes produced from living organisms with complex biochemistry and physiology.

5.3.2 Acoustic levitation chamber apparatus

The pollen grains were levitated using a modified commercial acoustic levitator (tec5, Oberursel, Germany) in which particles are suspended by the force of a standing acoustic wave between a transducer and opposing reflector. Acoustic levitation has progressed significantly in recent years and presents novel opportunities to study substances in contactless and container-less conditions, i.e. in the true aerosol phase. It has become an opportune, commercially available and low-cost alternative, as opposed to other methods, such as the electrodynamic balance (EDB), and demonstrates huge versatility in being able to levitate a wide variety of objects (Marzo et al., 2017; Andrade et al., 2018).

The setup for this study was based on that previously described by Milsom et al. (2021) and is illustrated in Figure 5.1. The levitator operated at a fixed frequency (100 kHz) with variable power (0.65-5.00 W) and was initially deployed for studies of a wide range of self-assembled materials (e.g. Seddon et al., 2016; Pfrang et al., 2017). The concave reflector, parallel above the transducer, was fitted with a micrometre screw to facilitate adjustment relative to the transducer. The reflector-transducer distance was adjusted accordingly to levitate particles, but generally kept in the range of 20-30 mm, and was kept constant for the duration of each experiment. A custom-built 3D printed chamber

enclosed the levitator and was sealed from ambient air outside except for access ports for the RH sensor and tubing to deliver humidified air. The system for monitoring and controlling RH in real-time was custom-made, using a Raspberry Pi connected to an air pump and an Arduino connected to a DHT22 RH sensor (2–5% RH accuracy between 0-100% RH). 'Dry' (lab ambient) air was pumped through tubing which split into two flows: one was bubbled through water to supply humidified air while the other dry air flow was passed through a manual flow meter to control the dry relative to wet flow. The two flows were re-joined and delivered into the chamber, where the RH sensor was also positioned, and the flow was adjusted accordingly to cover the experiment humidity range. A macroscope with a digital camera attached (Leica MC190 HD) was positioned in front of one of the levitation chamber's transparent microscope slide windows to take snapshots throughout the experiments. A photograph of the acoustic levitator as used in this study can be found in Figure 10.2 in the Appendices.

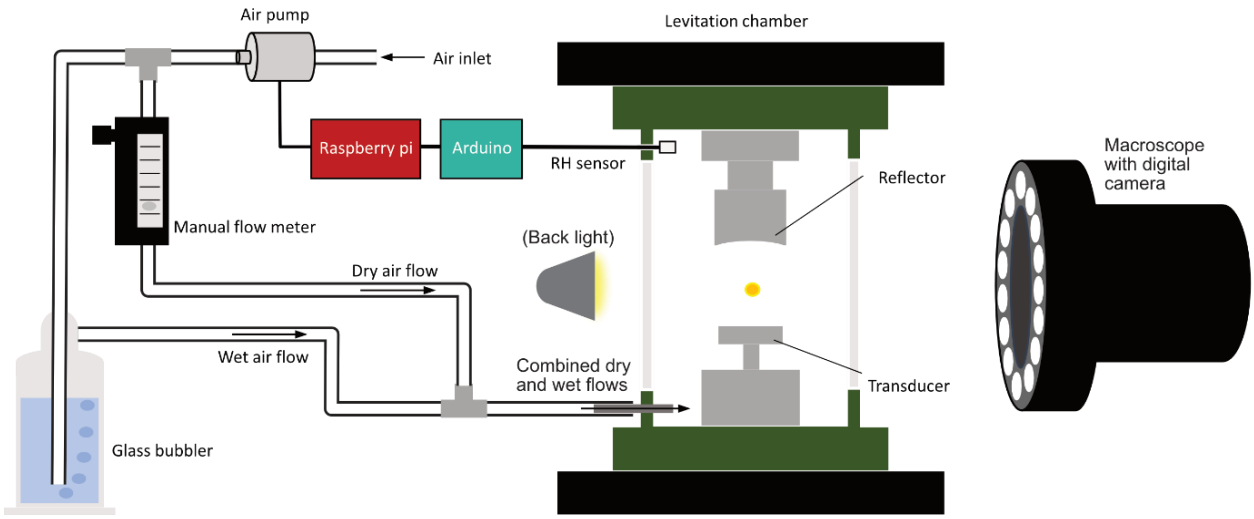


Figure 5.12: *Diagram illustrating setup of acoustic levitation chamber, humidification system and macroscope for experiments investigating the hygroscopic behaviour of pollen.*

5.3.3 Hygroscopicity experiments

For instances where single pollen grains were levitated, these grains were first isolated on a slide under a compound microscope before being introduced to the levitation chamber via pin-tip. This was to make sure that, where possible, single grains were observed. Once levitated, the chamber windows were sealed using tape, lighting applied to illuminate the pollen grain, and the focus of the macroscope adjusted onto the pollen grain. Lighting was directed either from the front, by the lights encircling the macroscope camera lens, or from the back, by a Cole-Parmer Fiber Optic Illuminator (Model 41720), though back-lighting was reverted to as default for reasons discussed in section 3, below. Humidity was increased by increments of generally 5 or 10% RH from ambient room RH up to the maximum RH that could be reached and at each increment 5 snapshots, or in some cases more, were captured. Once maximum RH was reached, the process was repeated for decreasing RH increments back to ambient RH.

Surface-fixed experiments were conducted using a similar RH stepwise procedure, except that pollen grains were placed on a microscope cover slip which was attached to the inside of the levitation chamber window facing the macroscope. These experiments did not investigate the true aerosol phase of airborne pollen grains but ensured the pollen grains were fixed and that the orientation was constant throughout the experiment.

5.3.4 Image and data processing

The images captured by the macroscope camera during each experiment were taken and processed by a bespoke program written using Python 3.9 for the specific purposes of this experiment. This Python program applies a binary threshold, performs Canny edge detection and identifies object contours (i.e. the pollen grain outlines) within the selected image range (see Figure 10.3 in the Appendices for a diagram of the full methodology of this process). An area measurement is recorded for the largest contour found (also checking visually that this is the pollen grain) and an ellipse is fitted to the same contour from which 'polar' and 'equatorial' diameters are extracted in pixel units. Due to the image contrast setting being adjusted as each grain was levitated, the threshold value in the program had to be changed accordingly, so is consistent only within each experiment run.

The raw measurements calculated by the image processing program were processed and visualised using Python 3.9.7 in Jupyter Notebook (Anaconda Distribution) with package versions Numpy 1.24.1, Pandas 1.5.3, Matplotlib 3.6.3 and Seaborn 0.12.2. Pixel area measurements have been converted into average area increase ratios (averaged over 5 repeat snapshot images) relative to the initial size measured at ambient RH. The ellipse-fitted diameter measurements do not necessarily correspond with 'polar' and

'equatorial' axes defined by general pollen terminology, as this cannot be determined with certainty from the images. However, we used the ellipse diameter measurements to calculate an average aspect ratio to evaluate change in shape. The aspect ratio was defined by the following Eq. 1:

$$Aspect\ ratio\ = \frac{smaller\ (\text{equatorial})\ diameter}{larger\ (\text{polar})\ diameter} \qquad \text{Eq. 1}$$

This resulted in values between 0 and 1, with higher values implying more circularity (and thus likely higher sphericity of the whole pollen grain as well). The standard deviation was calculated across the 5 repeats for each averaged value of area and aspect ratios and is presented as error bars in the figures in Results and Discussion sections 3.2 and 3.3.

To facilitate meaningful comparison between experiment results, we fitted a simple least squares linear regression to each pollen grain experiment using the Scikit-learn (version 1.2.1) LinearRegression model. These models fitted area and aspect ratios against RH for the direction of increasing RH only in each case. The linear model is not a good model to describe the behaviour of hygroscopic size change on an incremental level. It was the most parsimonious model to derive the size and shape changes between a consistent RH range while taking all data points fairly into account (i.e. not just taking the start and finish values which could be anomalous). From the fitted models for each instance, we predicted the area and aspect ratio changes between 70 and 95% RH. (These values were chosen as they generally lay within the experimentally measured RH ranges for all instances.) The results were summarised as a box and whisker plot for each pollen type and experiment setup (i.e. surface-fixed and levitated). A paired, two-

sided t-test (using SciPy 1.10.0) was also performed for each experiment group between data points estimated by the linear model at 70 and 95% RH respectively to test if the observed difference was significant.

5.4 Results and discussion

5.4.1 Acoustic levitation and visual imaging of pollen grains

Both *Lilium orientalis* and *Populus deltoides* pollen were successfully levitated in the acoustic levitator, as clusters of multiple pollen grains and single grains, though this was achieved with varying ease and stability. Adjustment of the transducer frequency and transducer-reflector distance were necessary at times when the pollen grains were unstable to the point that snapshot quality was compromised by the erratic motion of the levitated particle.

The stability of an acoustically levitated particle depends on its position relative to nodes created by the pattern of standing waves between transducer and reflector. The acoustic waves produce an upward pressure that counteracts the force of gravity, while levitated objects are most stable when sat at node points where the wave movement and pressure is at a local minimum. The node positions depend on both the frequency of the sound waves being produced and the transducer-reflector distance. The pollen grains may not always manage to settle at a node point and therefore can be prone to instability – moving erratically, spinning, and sometimes dropping out of the trap altogether.

This instability occurred more frequently for the smaller *Populus deltoides* pollen than *Lilium orientalis* pollen, and for single grains as opposed to clusters of multiple grains. These particles, being lighter, were likely more sensitive to fluctuations in their position relative to the acoustic field delicately balancing them, and therefore were more easily thrown out of balance if not securely settled at a node point. Asymmetrical or irregular clusters of pollen grains such as those shown in Figure 5.2 were also less stable, likely due to the juxtaposition of forces exerted by different areas of the acoustic field across

the irregular structure, and this structure was frequently re-organised. Thus, for the following hygroscopicity experiments we always attempted to isolate and levitate single grains, to remove the variability of cluster structure.

Figure 5.13: *Example of multiple Lilium orientalis pollen grains being levitated at once but experiencing instability and fluctuation in structure and orientation. These photos are a representative few of many similar snapshots that were captured in the space of a few minutes.*

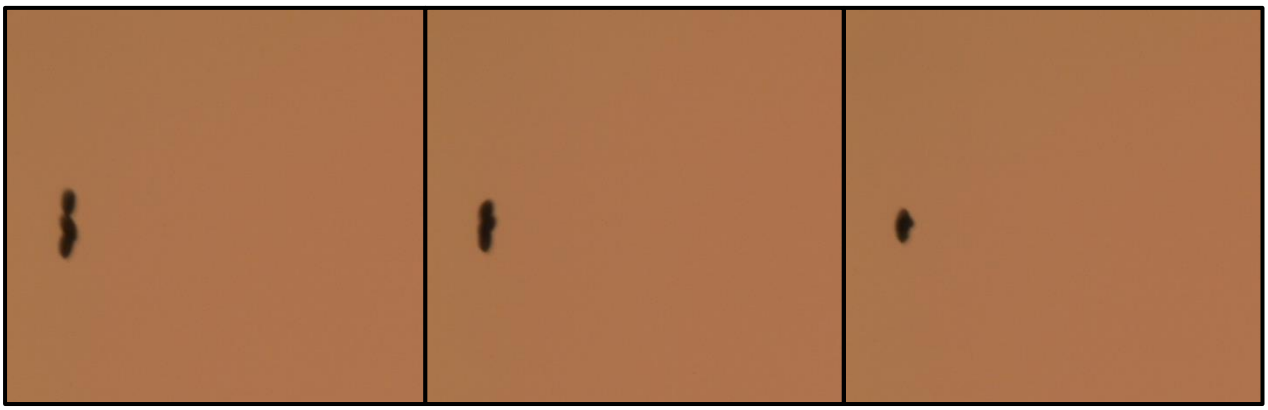
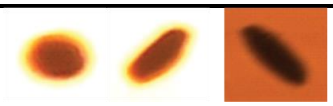
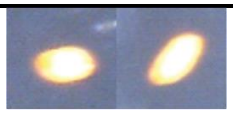
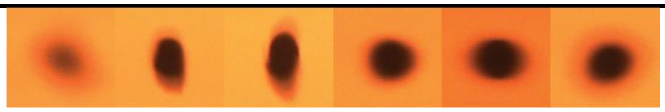
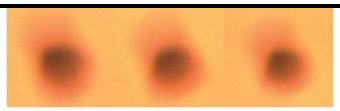


Table 5.2 shows representative snapshots taken for each pollen type and experiment setup to show general image quality and the variation observed among them. Further images showing examples of the pollen grains across the relative humidity range, as well as the program-fitted contour and ellipse used for measurements, can be found in Figures 10.4 and 10.5 in the Appendices. All experiment images used in this study, as well as relevant code and datasets, can be found in the data repository in the Data Availability section.

The macroscope camera contrast was adjusted for each experiment to get the best image with most crisp and easily definable edges; however, this was achieved to varying success for each case. From initial experiments, back-lighting was chosen as a default to proceed with measurements as it resulted in a generally crisper and more easily definable silhouette. Importantly, it also produced a cleaner background devoid of other features that would interfere with the Canny edge detection and contour identification. (Front-lighting resulted in a less-pristine background due to other features picked up from the glass chamber window.)

In the following results, we include for comparison an example where two pollen grains were observed under front-lit conditions. After post-experiment analysis, it could be argued that they achieved more consistent results than the equivalent back-lit experiment. The decision to proceed with back-lighting was made based on observations at the time for the reasons above, in particular with the automated method to distinguish pollen grains from the image background in mind. However, the method could be refined to overcome interference of background features and it may be that front-lit experiments are worth testing again in the future, as this may produce more consistent images that vary less with light brightness and contrast settings.

Table 5.2: *Representative snapshots taken for each pollen type and experiment setup showing the image quality achieved as well as the variation across different instances. Further snapshots showing example pollen grains across the relative humidity range can be found in Figures 10.4 and 10.5 in the Appendices.*

Pollen type	**Levitated or surface-fixed**	**Back or front-lit**	**Example pictures**
Lilium orientalis	Surface-fixed	Back-lit	
Lilium orientalis	Surface-fixed	Front-lit	
Lilium orientalis	Levitated	Back-lit	
Populus deltoides	Surface-fixed	Back-lit	
Populus deltoides	Levitated	Back-lit	

5.4.2 Surface-fixed pollen hygroscopicity experiments

5.4.2.1 Lilium orientalis

The results for the *Lilium orientalis* pollen fixed to the levitation chamber window and under back-lit conditions are shown in Figure 5.3. One pollen grain, the blue line with circle markers in Figure 5. 3 (further images presented in Figure 10.4 in the Appendices), increased in silhouette area to just over 20% between 70-96% RH. Additionally, both directions of increasing and decreasing RH showed a similar trend without hysteresis. Another grain observed under the same conditions (orange with upright triangle in Figure 5.3) showed some evidence of increasing by around 2% but ultimately produced an inconclusive trend for increasing RH. Meanwhile the same grain showed a little decrease in area with decreasing RH but generally only by a few percent. The two pollen grains were different in shape, with different starting aspect ratios of 0.79 and 0.47 - i.e. one more circular and the other more ellipsoidal - for the first and second grain, respectively. The aspect ratio for the first decreased with increasing RH to 0.76 and continued to decrease further down to 0.67 with decreasing RH which seems contradicting. Meanwhile the second grain's aspect ratio increased up to 0.54, then decreased down to 0.45 with increasing and decreasing RH correspondingly.

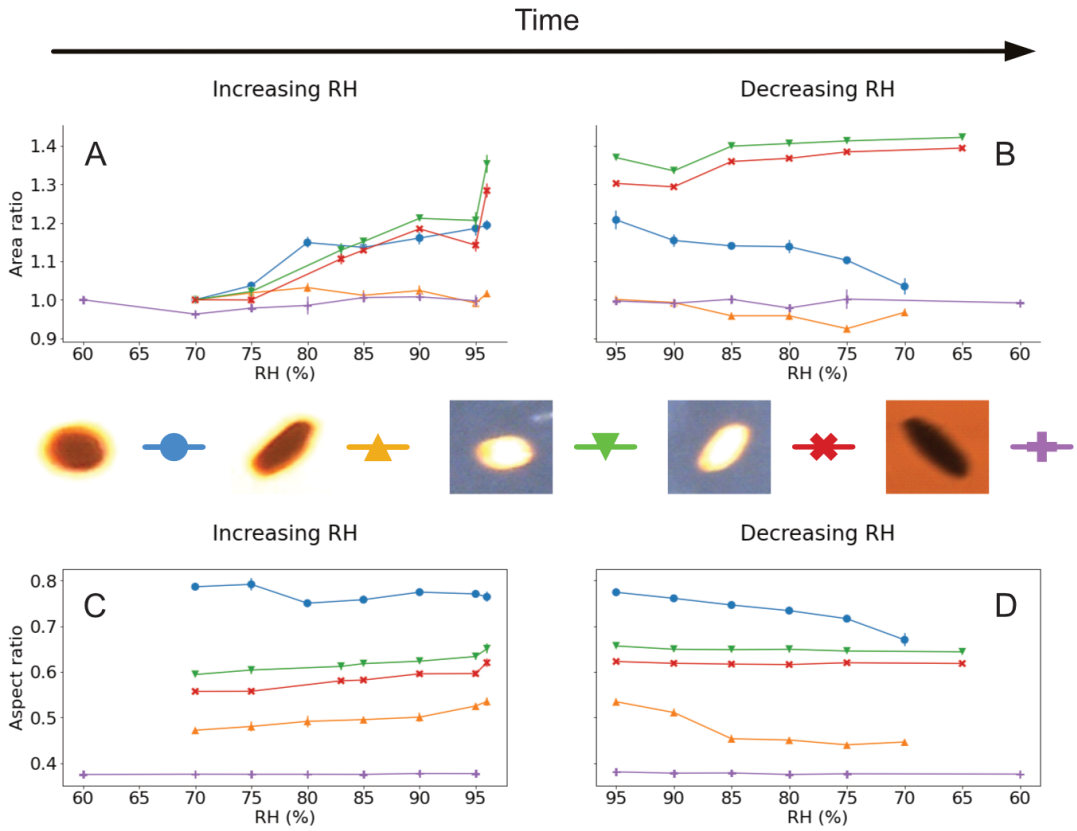


Figure 5.14: *Visualised results for surface-fixed Lilium orientalis pollen grains undergoing incremental RH changes with error bars. A & B: Area ratio against increasing and decreasing RH respectively. C & D: Aspect ratio against increasing and decreasing RH respectively. The colours correspond with the respective pollen grain images shown beside them and lines between the discrete measurements have been included for visual aid.*

The same pollen grains were observed under front-lighting, giving a considerably different picture as can be seen in Table 5.2. Here, on the contrary, the two grains behaved very similarly (green with upside-down triangles and red crossed in Figure 5.3). The first grain appeared more elongated and similar the second grain, and each had an initial aspect ratio of 0.59 and 0.56 respectively. Both aspect ratios increased with increasing RH (to 0.65 and 0.62 respectively) yet showed no significant changed for decreasing RH. Meanwhile, both front-lit grains increased in area with increasing RH by

35 and 28%. They increased even further with decreasing RH, up to 42 and 39%, before starting to decrease a little by the end of the experiment.

A different grain tested once again in a back-lit experiment (purple plus signs in Figure 5.3) showed no increase or significant decrease in area with change in RH. This grain started with a aspect ratio of 0.38 and this too did not fluctuate significantly throughout the experiment.

While these results show some evidence of the hygroscopic size increase we had expected with increasing RH, they also suggest apparent inaccuracy and lack of consistency which must be considered. This variability may be due to limitations of the method in terms of imaging capabilities. Meanwhile there is possibly more convincing evidence of increased circularity – and, by assumption, increased sphericity - of grains as they undergo hygroscopic morphology changes under increased RH. In particular, this trend appears more evident for grains that are initially less circular.

Importantly, these examples show the sensitivity of the experiment to lighting conditions, including contrast settings, of the image taken. The lighting conditions influence the visible pollen grain outline and, consequently, the measurements calculated by the image processing program from these snapshots. It may not be possible to have a one-setting-fits-all for image contrast and it is in turn difficult to relate the observed outline to an absolute value for the pollen grain size. However, it is imperative that the lighting conditions and contrast settings are kept constant throughout each experiment run, as was done for all experiments reported here. This ensures that the change in size and shape of a particular grain silhouette within an experiment run is comparable and meaningful.

5.4.2.2 Populus deltoides

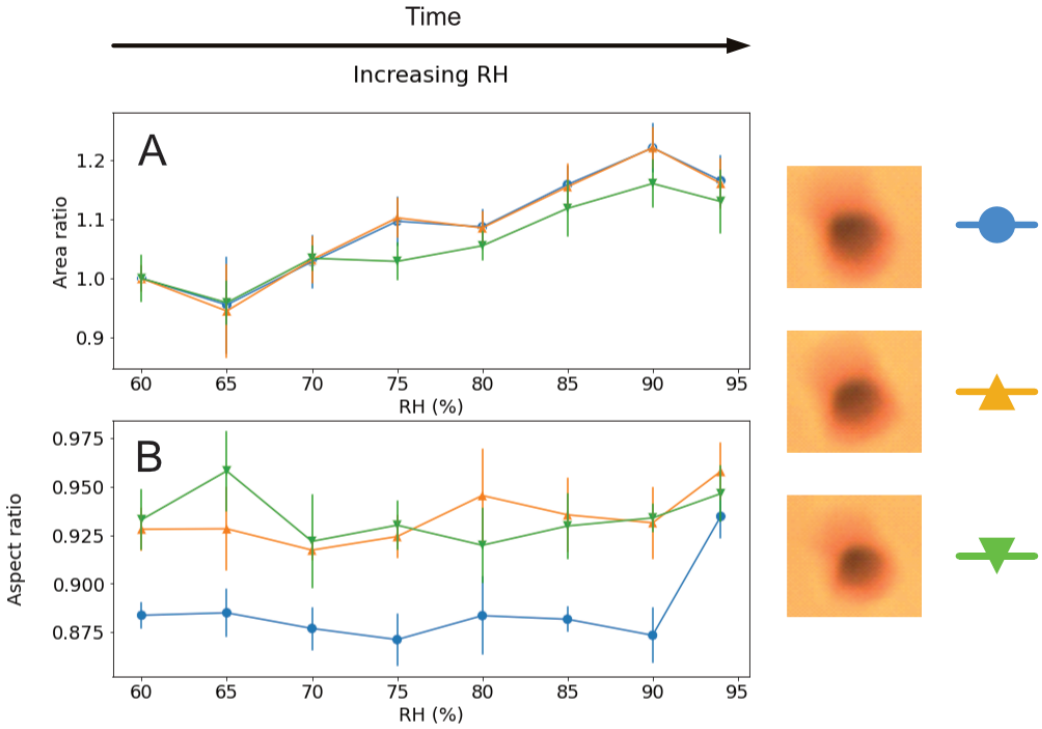


Figure 5.15: *Visualised results for surface-fixed Populus deltoides pollen grains undergoing incremental RH changes with error bars. A: Area ratio against increasing RH. B: Aspect ratio against increasing RH. The colours correspond with the respective pollen grain images shown beside them and lines between the discrete measurements have been included for visual aid. Decreasing RH was not measured in this instance.*

For the surface-fixed *Populus deltoides* pollen experiments (all back-lit), shown in Figure 5.4, three grains were observed simultaneously under the same conditions and all showed similar results. That is, all three grains exhibited a steady increase in area by 13-22% between 60 and 94% RH. Decreasing RH was not recorded for this experiment. The aspect ratios, which were all initially near-circular (aspect ratios between 0.87-0.94), showed some signs of slight increase in circularity to a range of 0.93-0.95 (the grain

with the lowest initial ratio increasing the most) yet the change may not be considered conclusive.

These results show evidence of a hygroscopic size increase for *Populus deltoides* pollen with increasing RH. While there is some evidence of increased grain circularity, these grains are already near-spherical to start with and so a less significant ratio change is to be expected. Nevertheless, since these grains were all simultaneously under the same conditions it would be necessary to conduct more experiment instances before making conclusive assertions from these results.

5.4.3 Levitated hygroscopicity experiments

5.4.3.1 Lilium orientalis

The results for visual size changes of the *Lilium orientalis* pollen while levitated, shown in Figure 5.5, are somewhat inconclusive. Just one out of six experiments showed evidence of area increase, where the area increased beyond 10% with an RH increase from 45 to 99% RH. The other experiments show either no visible trend or even a significant decrease in measured area. The aspect ratio observations generally also did not show a conclusive trend. Initial ratios ranged between 0.45 and 0.92 demonstrating a large variance in observable shape among the grains. Once again, in cases where the aspect ratio was initially close to 1, generally less significant change was seen (including the grain that experienced significant area increase). The instance demonstrating the lowest aspect ratio was that which showed a particularly significant decrease in area with increased humidity. This grain also exhibited the most significant increase in aspect ratio with increasing humidity from 0.45 (at 42% RH) to 0.76 (99% RH) and back down to 0.43 (50% RH).

These results suggest that the measurement accuracy is hindered by the fact that the grains are being levitated freely. The significant but unexpected changes in observed area may be largely affected by their orientation while suspended, as there is no way to guarantee we are always looking at the same angle of the pollen grain. Once again, however, grains which are initially less circular show greater change in aspect ratio with humidity changes.

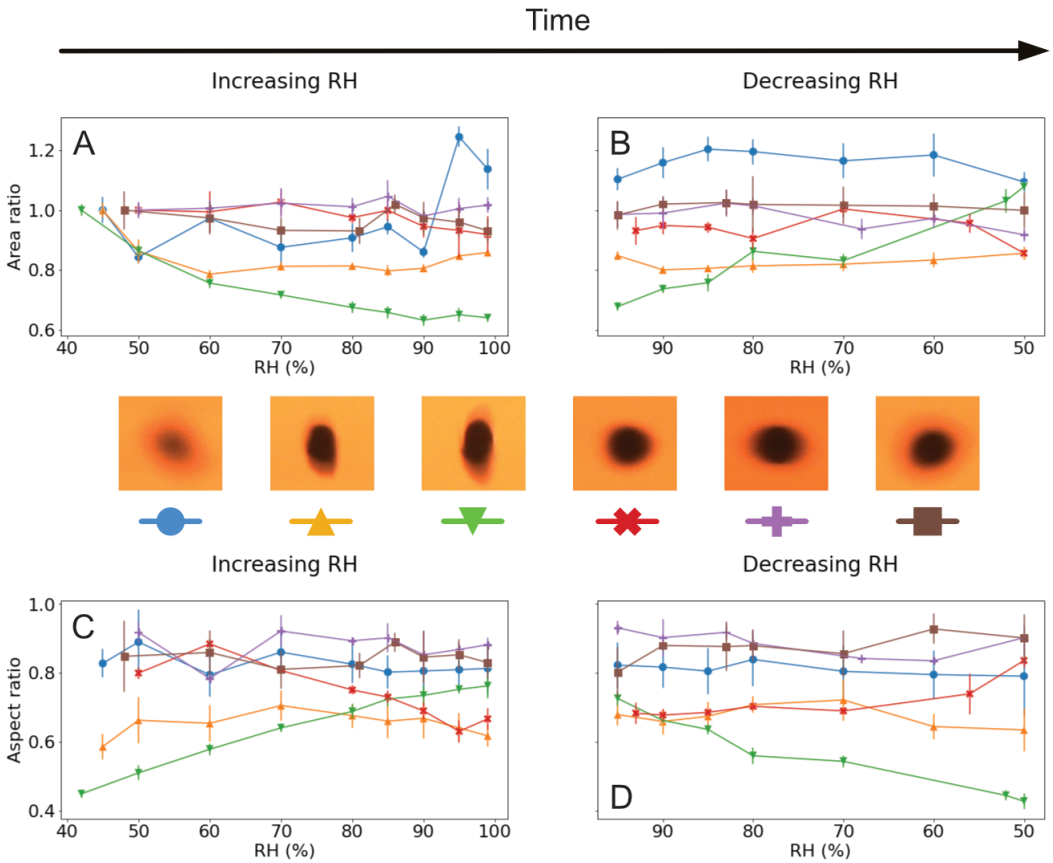


Figure 5.16: *Visualised results for levitated Lilium orientalis pollen grains undergoing incremental RH changes with error bars. A & B: Area ratio against increasing and decreasing RH respectively. C & D: Aspect ratio against increasing and decreasing RH respectively. The colours correspond with the respective pollen grain images shown beside them and lines between the discrete measurements have been included for visual aid.*

5.4.3.2 Populus deltoides

Since the *Populus deltoides* pollen was much more difficult to levitate stably than the *Lilium orientalis* pollen and often fell out of the trap, it was very difficult to collect a complete set of image data across all previous humidity increments. Therefore snapshots were only taken at room RH at 48% and maximum RH at 95% for increasing RH and then 90% and 50% RH in the decreasing direction. The pollen grain decreased by 22% of its original size when RH was increased to 95%. However, at the start of the decreasing RH direction, at 90% RH, it increased to 7% above its original size then to 9% when returned to 50% RH. Meanwhile the aspect ratio increased from 0.33 to 0.49 for increasing RH then decreased back to 0.32 when RH was decreased.

It should be noted that the images for this pollen grain even among the same RH increment displayed considerable variability (demonstrated in Figure 10.5 in the Appendices) implying that the observed orientation was not fixed. Due to the instability of the levitated pollen grain, it was difficult to capture consistent snapshots of the grain even at constant RH. This can also be observed as the data points themselves for area and aspect ratio display considerable variance indicating a large error margin.

5.4.4 Summarised results

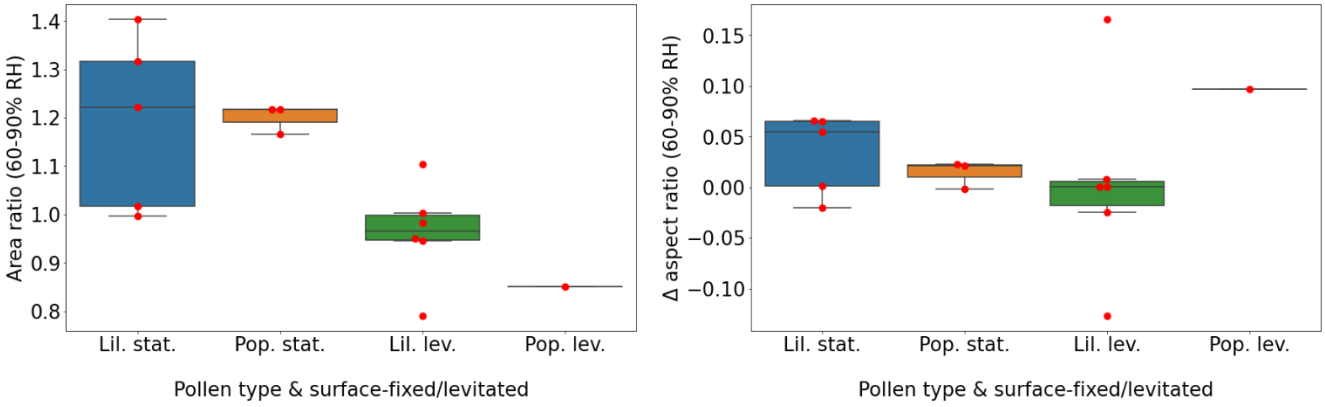


Figure 5.17: *Box-and-whisker plots summarising in area ratios (left) and change in aspect ratios (right) for the RH interval 60-90% for each experiment group with overlaying swarm plot showing individual data points in red. (Lil. = Lilium orientalis; Pop. = Populus deltoides; Stat. = pollen grain was surface-fixed (static); Lev. = pollen grain was acoustically levitated.)*

Table 5.3: *Mean changes in area and aspect ratios within in experiment group and the results of paired t-tests performed between area ratios at 60 and 90% RH respectively.*

	Mean change in area ratio between 60-90% RH	**T statistic of paired t-test**	**p-value of paired t-test**	**Mean change in aspect ratio between 60-90% RH**
***Lil.* Stat.**	0.17	2.4	0.07	0.033
***Pop.* Stat.**	0.19	11.8	0.007	0.014
***Lil.* Lev.**	-0.03	-0.9	0.4	0.004
***Pop.* Lev.**	-0.14	-	-	0.099

Figure 5.6 summarises the area ratios and change ins aspect ratios between 60-95% RH for each experiment type, as estimated by the fitted linear regression models. The overlayed 'swarmplot' shows where the individual data points lie (in red), so as not to overlook that the number of samples composing the plots was varied and relatively sparse. Table 5.3 summarises the mean values for area and aspect ratio change for each experiment type, and the results of a paired t-test compared between the estimated data points at 60 and 90% RH.

When calculated this way, the surface-fixed experiments showed an area increase of 17-19%. The results of the t-test demonstrated a significance level well within 0.05 for *Populus deltoides* while the *Lilium orientalis* only demonstrated significance at the 0.1 level. The proportional size change of the *Lilium orientalis* pollen may have been less evident when measured due to the larger size and different shape. Note that these two pollen species are each evolved for different pollination transport methods, which may affect their hygrscopic behaviour. The levitated pollen experiments could not demonstrate significant area change even at the 0.1 level. The aspect ratios also did not demonstrate signifiant change within the 0.1 level for any experiment type, however, as discussed previously, this seemed to be influenced by initial aspect ratio which is not accounted for here.

5.4.5 Comparison with previous methods

To put our results into context, we can compare them with findings from previous studies, presented in Table 5.1 in section 5.2. Pope (2010) and Griffiths et al. (2012) measured a mass increase of 16-18% for various pollen species when RH was increased from 0 to 75%. When increased from 0 to 90% RH, Tang et al. (2019) and Chen et al.

(2019) reported mass increases between 29-48%, an even further up to ~72% when increased to 95% RH.

However, the experiments in this study did not start at conditions of 0% RH. From observations for *Carya illinoinensis* (pecan) pollen by Tang et al. (2019), mass increases of 6.4%, 30.3% and ~72% were calculated when RH was increased from 0 to 60%, 90% and 95% RH respectively. This means that there was a 22% mass increase between 60-90% RH and a 62% mass increase between 60-95%. These RH ranges are comparable to those tested in this study. Though, the reported literature values for mass increases above 90% RH show considerably large variation. This is predicted by Köhler theory which suggests that the hygroscopic increase increases rapidly at high RH (asymptotically towards 100%), and potentially the variation here is also greater here across different species. It may also suggest reduced accuracy for these methods under such high RH conditions. The RH range 60-90% was chosen for comparison in this study because the literature values appear more consistent.

Assuming the density of the pollen grain is constant and that the grains are near spherical, the relationship between area and mass increase ratios can be described by Eq. 2, where m/m_0 is the ratio of final mass over initial mass and A/A_0 is the ratio of final area over initial area:

$$\frac{m}{m_0} = \left(\frac{A}{A_0}\right)^{\frac{3}{2}} \quad \text{Eq. 2}$$

The average area increases observed between 60-90% RH for the surface-fixed pollen experiments in this study were 17 and 19% for *Lilium orientalis* and *Populus deltoides* respectively, which correspond to mass increases of 27 and 30%. This is a little higher

than the 22% mass increase found between 60-90% RH for *Carya illinoinensis* pollen by Tang et al. (2019).

Considering that the density of the pollen grain would also change while it is swelling with water, the relationship would more accurately be described by Eq. 3, where ρ/ρ_0 is the ratio of the final over the initial density of the pollen:

$$\frac{m}{m_0} = \left(\frac{A}{A_0}\right)^{\frac{3}{2}} \times \frac{\rho}{\rho_0} \qquad \text{Eq. 3}$$

It is difficult to determine exact values for pollen density due to their complex composition as a gamete from a biological organism, and it has not been shown how comparable estimates are between species. Taking an example from the literature, Sosnoskie et al. (2009) estimate the density of dry and fully hydrated *Amaranthus palmeri* pollen to be 1,435 and 1,218 kg m^{-3} respectively. This corresponds to a 15% decrease in density when hydrated, or a ρ/ρ_0 ratio of 0.85. Applying this to equation 2, our area increases of 17 and 19% correspond to mass increases of 8 and 10%. Hence, corresponding mass increase values are considerably smaller when density change is taken into account.

Carya illinoinensis pollen grains generally range between 43-53 µm (Stone, 1963) and so fall into a size range above that of *Populus deltoides* but below that of *Lilium orientalis*, and thus may not be appropriate for comparison. Tang et al. (2019) also investigated one of the species in this study, *Populus deltoides*, and reported an average mass increase of 48% between 0 and 90% RH. For *Carya illinoinensis* the ratio between the mass increases between 60-90% and 0-90% RH was 0.74. Assuming the same ratio of *Populus deltoides* this would mean a mass increase of ~36% between 60-90% RH. By this

comparison, the results from our study would appear to underestimate hygroscopic growth. However, the mass increases calculated by Eq. 2, without considering density changes, produce results that seem plausible in the context of the literature examples discussed here. Table 5.4 summarises these comparisons between the present study and literature examples from Tang et al. (2019).

Table 5.4: *Comparison of the method in this study and vapour sorption analyser method reported in literature, including calculated mass increase values for increasing RH between 60-90% for surface-fixed pollen experiments in this study with reported literature values. The ranges quoted for the examples in this study take into account that the density of the pollen grain may change with hygroscopic swelling. Advantages and disadvantages for each method are also summarised.*

Ref.	Method	Pollen species	RH interval	% mass increase	Advantages	Disadvantages
This study	Macroscope (& acoustic levitator)	*Lilium orientalis*	60-90%	8-27%	Possible to investigate pollen shape dynamics as well as clusters of particles. Lower cost equipment and versatile.	Sensitive to experiment lighting, particle orientation and instability when levitated. Difficult to obtain good quality quantitative results.
		Populus deltoides	60-90%	10-30%		
Tang et al. (2019)	Vapour sorption analyser	*Carya illinoinensis*	60-90%	23%	Currently more robust with better measurement accuracy.	Expensive equipment. Not possible to investigate particle shape.
		Populus deltoides	60-90%	~36%		

However, directly comparing area or volume increase with mass increase does present some potential issues that should be considered. Pollen grains are rarely perfectly spherical, and shape varies across and, to an extent, among species. The grains have features in the exine (outer surface) such as apertures (thinner sections of wall where the pollen tube is designed to break through) which make grains somewhat asymmetrical (Linder and Ferguson, 1985). Under increased humidity conditions it may be over-simplistic to assume that the grains would expand as a uniform sphere in all directions. Volume increase to accommodate the absorbed water may instead occur as an unfolding or inflating mode at places where the surface is creased or deflated, as opposed to an overall uniform grain expansion. Visual examples of this can be seen in ESEM (environmental scanning electron microscope) images presented previously by Pope (2010) and Griffiths (2012) and the mechanics of hygroscopic pollen inflation are studied by Božič and Šiber (2022). Therefore, comparing volume and mass increase ratios may require appropriate consideration of such complexities. It may even be useful to consider alternative metrics to relate pollen state under given RH conditions to CCN ability, such as expansion of wetted surface area, if such attributes can be reliably measured.

The advantages that this method offers, in contrast to previous mass-measuring methods, include the ability to visualise directly particle behaviour and measure aspect ratios, providing quantitative information regarding the dynamic shape of the particle, throughout changing experiment conditions. We have reported attempts at such measurements here for the first time, showing evidence of increased circularity - and by assumption sphericity - with increasing RH.

5.5 Conclusions

This work has, for the first time, explored the hygroscopic size change of pollen grains using acoustic levitation and visual imaging. Results have been compared with previous methods which measured mass change. Pollen was suspended by acoustic levitation, allowing experiments investigating hygroscopicity in the true aerosol phase. The varying stability of the levitation in some cases, particularly for single smaller (< 40 μm) pollen grains, presented some difficulty. The real-time visual imaging of the levitated grains throughout experiments provides a valuable opportunity to understand the response of pollen grains under different conditions such as, but not limited to, varying humidity.

This method of observing pollen grains does have limitations, particularly when it comes to obtaining accurate quantitative measurements. Though an unbiased computer program-based method was employed to extract measurements for particle area and diameters, it is evident that factors such as particle orientation and lighting conditions are important to consider. While lighting conditions can be kept constant for the course of an experiment, particle orientation is something we have little control over and affects the resulting image captured at any point in time. Fluctuations in orientation can give rise to considerable variation among snapshots of the same grain taken under constant conditions and presents an unknown that is difficult to quantify. However, this technique does reveal details that are not possible to observe by previous techniques, such as the non-uniform mode of morphology change of pollen grains under changing humidity conditions. Future studies may benefit from higher resolution image capabilities or even multi-angle imaging to better elucidate particle orientation and consistent 3D changes in size and shape. If possible, it would be beneficial to achieve

greater stability when trapping particles and record the power applied to quantify the acoustic pressure exerted on the particle accurately enough to measure mass changes as well.

In conclusion, this new method may not yet surpass previous mass change-quantifying methods in accuracy and leaves much to be improved upon. Nevertheless, there are novel measurements that can be made from the different perspective it provides, as well as various practical advantages such as lower cost and relative simplicity. We also suggest the importance of shape and surface area beyond simply the mass when considering potential CNN activity. We hope the methodology used in this work introduces a new possibility for further exploration involving the study of freely levitated aerosol or bio-aerosol particles subjected to various environmental conditions.

5.6 Contributors

The idea for this study was conceived and planned by FDP and CP. FDP, ARMK and CP provided general supervision and guidance. The practical experiments were carried out by SAM and AM. Analysis was completed and the first draft prepared by SAM. All authors contributed to the writing and editing of the final manuscript.

5.7 Acknowledgements

SAM acknowledges the funding support of the NERC CENTA2 grant NE/S007350/1. AM acknowledges the support of the NERC SCENARIO DTP (NE/L002566/1). This research has also been supported by the Natural Environment Research Council (grant nos. NE/T00732X/1, NE/G000883/1 and NE/G019231/1) and the Royal Society (grant no. 2007/R2).

5.8 Data and code availability

Data and code supporting this study are openly available from the UBIRA eData repository https://doi.org/10.25500/edata.bham.00000923.

5.9 Comment

This chapter stands apart from the previous chapters and, rather than following the low-cost sensor and machine learning methodology applied for monitoring airborne pollen, instead presents a novel methodology for characterising physical properties of pollen. This method also offers a relatively low-cost and simple alternative that has potential to supply new useful information when compared to conventional methods. Specifically, this method offers previously unavailable visual information on the size and shape of pollen grains as they undergo hygroscopic growth under varying RH conditions. There is also the potential to study pollen grains acoustically levitated in the true aerosol phase, but the methodology would benefit from further advancements before this becomes a viable option. This is, to our knowledge, the first study of its kind to levitate pollen and study it under a macroscope. While there are still many improvements to make, this method could be applied to study the effect of a range of different environmental conditions and chemical exposures on levitated pollen grains and even, perhaps, other bioaerosols.

5.10 References

Božič, A., Šiber, A.: Mechanics of inactive swelling and bursting of porate pollen grains, Biophys. J., 121:5, 782-792, doi: 10.1016/j.bpj.2022.01.019, 2022.

Bunderson, L., Levetin, E.: Hygroscopic weight gain of pollen grains from *Juniperus* species, Int. J. Biometeorol., 59, 533-540, doi: 10.1007/s00484-014-0866-9, 2014.

Cecchi, L., Scala, E., Caronni, S., Citterio, S., Asero, R.: Allergenicity at component level of sub-pollen particles from different sources obtained by osmolar shock: A molecular approach to thunderstorm-related asthma outbreaks, Clin. Exp. Allergy, 51(2), 253-261, doi: 10.1111/cea.13764, 2021.

Chen, L., Chen, Y., Chen, L., Gu, W., Peng, C., Luo, S., Song, W., Wang, Z., Tang, M.: Hygroscopic Properties of 11 Pollen Species in China, ACS Earth Space Chem, 3(12), 2678-2683, doi: 10.1021/acsearthspacechem.9b00268, 2019.

Després, V. R., Huffman, J. A., Burrows, S. M., Hoose, C., Safatov, A. S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M. O., Pöschl, U., Jaenicke, R.: Primary biological aerosol particles in the atmosphere: a review, Tellus Series B: Chemical and Physical Meteorology, 64(1), 15598, doi: 10/3402/tellusb.v64i0.15598, 2012.

Diehl, K., Quick, C., Matthias-Maser, S., Mitra, S. K., Jaenicke, R.: The ice nucleating ability of pollen Part I: Laboratory studies in deposition and condensation freezing modes, Atmospheric Research 58(2), 75-87, doi: 10.1016/S0169-8095(01)00091-6, 2001.

Diehl, K., Matthias-Maser, S., Jaenicke, R., Mitra, S. K.: The ice nucleating ability of pollen Part II: Laboratory studies in immersion and contact freezing modes. Atmos. Res., 61(2), 125-133, doi: 10.1016/S0169-8095(01)00132-6, 2002.

Dreischmeier, K., Budke, C., Wiehemeier, L., Kottke, T., Koop, T.: Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides, Sci. Rep., 7, 41890, doi: 10.1038/srep41890, 2017.

Du, Y., Wei, C., Wang, Z., Li, S., He, H., Jia, G.: Lilium spp. Pollen in China (Liliaceae): Taxonomic and Phylogenetic Implications and Pollen Evolution Related to Environmental Conditions, PLoS ONE, 9(1), e87841, doi: 10.1371/journal.pone.0091543, 2014.

Frohlich-Nowoisky, J., Kampf, C., Weber, B., Huffman, J. A., Pöhlker, C., Andreae, M. O., Lang-Yona, N., Burrows, S. M., Gunthe, S. S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V. R., Pöschl, U.: Bioaerosols in the Earth system: Climate, health, and ecosystem interactions, Atmos. Res., 182, 346-376, doi: 10.1016/j.atmosres.2016.07.018, 2016.

Gruber, S., Brinkmannz, J.: Vertical Distribution of Biological Particles above the North Sea, J. Aerosol Sci., 29, 771-772, 1998.

Hughes, D.D., Mampage, C.B.A., Jones, L.M., Liu, Z., Stone, E.A.: Characterisation of Atmospheric Pollen Fragments during Springtime Thunderstorms, Environ. Sci. Technol. Lett., 7:6, 409-414, doi: 10.1021/acs.estlett.0c00213, 2020.

Köhler, H.: The nucleus IN and the growth of hygroscopic droplets, Transactions of the Faraday Society, 32, 1152, 1936.

Linder, H. P, Ferguson, I. K.: On the morphology and phylogeny of the Restionales and Poales, Grana, 24:2, 65-76, doi: 10.1080/00173138509429917, 1985.

Matthews, B.H., Alsante, A.N., Brooks, S.D.: Pollen Emissions of Subpollen Particles and Ice Nucleating Particles, ACS Earth Space Chem., doi: 10.1021/acsearthspacechem.3c00014, 2023.

Matthias-Maser, S., Jaenicke, R.: The size distribution of primary biological aerosol particles with radii > 0.2 µm in an urban/rural influenced region, Nucleation and Atmospheric Aerosols, 39, 526-529, doi: 10.1016/0169-8095(95)00017-8, 1995.

Matthias-Maser, S., Obolkin, V., Khodzer, T., Jaenicke, R.: Seasonal variation of primary biological aerosol particles in the remote continental region of Lake Baikal/Siberia, Atmos. Environ., 34, 3805-3811, doi: 10.1016/S1352-2310(00)00139-4, 2000.

Matthias-Maser, S., Jaenicke, R.: The size distribution of primary biological aerosol particles in the multiphase atmosphere, Aerobiologia, 16, 207-2010, doi: 10.1023/A:1007607614544, 2000.

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R. and Pope, F.D.: Constructing a pollen proxy from low-cost Optical Particle Counter (OPC) data processed with Neural Networks and Random Forests. Sci. Total Environ., 871, 161969, doi: 10.1016/j.scitotenv.2023.161969, 2023.

Milsom, A., Squires, A. M., Boswell, J. A., Terrill, N. J., Ward, A. D., Pfrang, C.: An organic crystalline state in ageing atmospherc aerosol proxies: spatially resolved structural changes in levitated fatty acid particles, Atmos. Chem. Phys., 21, 15003-15021, doi: 10.5194/acp-21-15003-2001, 2021.

Möhler, O., DeMott, P. J., Vali, G., Levin, Z.: Microbiology and atmospheric processes: the role of biological particles in cloud physics, Biogiosciences, 4, 1059-1071, doi: 10.5194/bg-4-1059-2007, 2007.

Petters, M. D., Kreidenweis, S. M.: A single parameter representation of hygroscopic growth and cloud condensation nucleus activity, Atmos. Chem. Phys., 7, 1961-1971, doi: 10.5194/acp-7-1961-2007, 2007.

Pfrang, C., Rastogi, K., Cabrera-Martinez, E. R., Seddon, A. M., Dicko, C., Labrador, A., Plivelic, T. S., Cowieson, N., Squires, A. M.: Complex three-dimensional self-assembly in proxies for atmospheric aerosols, Nat. Commun., 8, 1724, doi: 10.1038/s41467-017-01918-1, 2017.

Pope, F. D.: Pollen grains are efficient cloud condensation nuclei, Environ. Res. Lett., 5, 044015, doi: 10.1088/1748-9326/5/4/044015, 2010.

Posselt, R., Lohmann, U.: Influence of Giant CCN on warm rain processes in the ECHAM5 GCM, Atmos. Chem. Phys., 8, 3769-3788, doi: 10.5194/acp-8-3769-2008, 2008.

Pummer, B.G., Bauer, H., Bernardi, J., Bleicher, S. and Grothe, H.: Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen, Atmos. Chem. Phys., 12(5), 2541-2550, doi: 10.5194/acp-12-2541-2012, 2012.

Rajora, O. P.: Pollen competition among *Populis deltoides* Marsh., *P. nigra* L. and *P. maximowiczii* Henry in fertilizing *P. deltoides* ovules and siring its seed crop, Sex Plant Reprod., 2, 90-96, doi: 10.1007/BF00191996, 1989.

Seddon, A. M., Richardson, S. J., Rastogi, K., Plivelic, T. S., Squires, A. M., Pfrang, C.: Control of Nanomaterial Self- Assembly in Ultrasonically Levitated Droplets, J. Phys. Chem. Lett., 7, 1341–1345, doi: 10.1021/acs.jpclett.6b00449, 2016.

Sosnoskie, L. M., Webster, T. M., Dales, D., Rains, G. C., Grey, T. L., Culpepper, A. S.: Pollen Grain Size, Density, and Settling Velocity for Palmer Amaranth (*Amaranthus palmeri*), Weed Science, 57(4), 404-409, doi: 10.1614/WS-08-157.1, 2009.

Steiner, A. L., Brooks, S. D., Deng, C., Thornton, D. C. O., Pendelton, M. W., Bryant, V.: Pollen as atmospheric cloud condensation nuclei, Geophysical Research Letters, 42(9), 3596-3602, doi: 10.1002/2015GL064060, 2015.

Stone, D. E., 1963. Pollen Size in Hickories (Carya). Brittonia, 15(3), 208-214, doi: 10.2307/2805161, 1963.

Sun, J., Ariya, P.: Atmospheric organic and bio-aerosols as cloud condensation nuclei (CCN): A review, Atmos. Environ., 40(5), 795-820, doi: 10.1016/j.atmosenv.2005.05.052, 2006.

Tang, M., Gu, W., Ma, Q., Li, Y. J., Zhong, C., Li, S., Yin, X., Huang, R.-J., He, H., Wang, X.: Water adsorption and hygroscopic growth of six anemophilous pollen species: the effect of temperature, Atmos. Chem. Phys., 19, 2247-2258, doi: 10.5194/acp-19-2247-2019, 2019.

Tong, H.J., Ouyang, B., Nikolovski, N., Lienhard, D.M., Pope, F.D. and Kalberer, M.: A new electrodynamic balance (EDB) design for low-temperature studies: application to immersion freezing of pollen extract bioaerosols, Atmos. Meas. Tech., *8*(3), 1183-1195, doi: 10.5194/amt-8-1183-2015, 2015.

6. CHAPTER 6: Summary and future directions

6.1 Thesis summary

This thesis has presented a series of novel studies that make progress in addressing the current limitations we have in studying bioaerosols, outlined in Chapter 1. The methods explored the aim to further our understanding of pollen as a bioaerosol using relatively low-cost, commercially available equipment paired with effective data processing methodologies. Following the objectives outlined in Chapter 1 section 1.5, this work documents the development of these novel methodologies, demonstration in different contexts of the novel information that can be acquired from them, and evaluation of the potential and current limitations for utilising such methods in the future.

Firstly, in Chapter 2, we presented the initial development method to process data from optical particle counters, commercially produced for measuring general PM, processed with machine learning techniques to monitor pollen. Here we demonstrated the superiority of the machine learning models compared to other methods, assessed the temporal detail that could be gained for total and grass pollen concentrations and demonstrated how the data output from this method could be of value to human health.

In the following Chapter 3, we developed the deep learning models further and demonstrated how tuning various hyperparameters collectively could significantly improve (in some cases doubling) the ability of the models to estimate pollen concentrations. This was also extended to models that targeted specific tree pollen taxa. Furthermore, the non-linear relationships between input features and output pollen concentrations were investigated using explainable AI techniques, elucidating the particle size fingerprints that the models associated with each pollen type. This is an important step in assessing the validity of the model, how different environments may

influence the output generalisability, and highlights areas that need further research, particularly, in this case, the characterisation of co-occurring subpollen particles.

While the previous two chapters were based on the same data from a pollen monitoring instrument intercomparison campaign, Chapter 4 made an important step by applying the trained models in a different context to investigate generalisability. This study, and more across a variety of contexts alongside reference data, are crucial to conduct before scaling up or commercialising this method, to evaluate their ability to deliver useful information to aid public health. Furthermore, this study provides a novel insight into pollen emission activity, measured at high time resolution, through the canopy of a mature forest environment during the dominant oak pollen season. This demonstrates the wealth of information could be achieved at high spatial resolution and across numerous unique locations which has not yet been possible.

Lastly, Chapter 5 demonstrated how another commercially available and relatively low-cost instrument, the acoustic levitator, could be paired with appropriate image data processing to study pollen visually grains under changing environmental conditions. While this method needs much improvement before becoming a viable and delivering useful information, it has the potential to provide novel information that other methods could not offer. Pollen and other bioaerosols are often complex structures that can undergo non-uniform changes under varying conditions such as humidity, and methods to study them visually and continuously in flight would be valuable.

6.2 Limitations and future directions

The results presented in this thesis thoroughly demonstrate the potential of the OPC instrument paired with machine learning data processing for monitoring pollen. However, there are still many aspects to address in order to take it to the next step. One aspect may involve physical engineering advancements to the OPC sensors to improve their ability to detect larger particles and lower background concentrations. Development of the machine learning models themselves has been explored extensively in this work. While there are always more things to try, including other types of machine learning or deep learning algorithms (e.g. boosted regression trees, convolutional or recurrent neural networks not explored here), this is likely not the aspect to focus on for future performance improvement beyond that presented here.

Machine learning algorithms depend heavily on data quantity and representation of the contexts they are to be applied to. Due to practical limitations in the scope of this work, the models developed here so far are trained on relatively little data and only from one specific context. To take this methodology further, it is imperative to obtain more data from intercomparison campaigns between the low-cost OPCs and other pollen monitoring instruments (both Hirst-type and recently developed automated instruments) across many varied environments to train the models. It may or may not be possible to develop 'one-size-fits-all' models that generalise well for measuring 'total' pollen or specific pollen taxa across different environments. It may also be beneficial to develop different models trained for different environments, particularly models that take meteorological variables into account when estimating pollen concentrations. There may be a trade-off between accuracy and generalisability. It is important to consider the representativeness of the data that a model has been trained on when

producing output data in each context. Nevertheless, with this in mind, we anticipate there should be huge scope for making this method operational with more data.

To summarise, the following points would be useful to consider going forwards:

- Collection of extensive OPC and reference pollen concentration data from various contexts to train more skilled and nuanced models.
- Evaluation of the generalisability and usefulness of this methodology across a range of contexts and assessment of what array of trained models would be needed in practice (e.g. for a national or international monitoring network).
- Further development of individually optimised machine learning models for each target (perhaps exploring other machine learning methods) and a simple, automated pipeline that processes OPC data to produce pollen concentrations in real-time without need of special expertise.
- Discussions with instrument manufacturers on current instrument limitations so that future instrument models can effectively be used to inform those at risk and aid regulation in the interest of public health.

6.3 Potential value

If developed successfully, this methodology could be employed across regional monitoring networks as well as local environments to supplement the limitations of conventional pollen monitoring methods and provide a wealth of novel, detailed data. It would be valuable to climate, atmospheric and ecological scientists trying to better understand the emission of pollen from vegetation, its transport in the atmosphere and potential to interact with clouds. This would contribute by better informing climate models that aim to include interactions between vegetation and atmosphere, but which currently lack detailed empirical information. It could help bridge the gap between meteorological transport modelling studies for pollen and actual empirical observations.

Above all, information from upgraded monitoring networks and context-specific or occupational studies would be extremely beneficial for the public and allergy sufferers. It could provide them with up-to-date information to make informed decisions and, ultimately, contribute towards relevant policymaking. With effectively implemented regulations, based on a rich evidence base, there would be invaluable gains in terms of reducing public health burden.

Appendices

7. Supporting information for Chapter 2

Table 7.1: *Table of 24 OPC bins with corresponding particle size ranges detected for each.*

OPC bin	Min (µm)	Max (µm)
0	0.35	0.46
1	0.46	0.66
2	0.66	1.0
3	1.0	1.3
4	1.3	1.7
5	1.7	2.3
6	2.3	3.0
7	3.0	4.0
8	4.0	5.2
9	5.2	6.5
10	6.5	8.0
11	8.0	10
12	10	12
13	12	14
14	14	16
15	16	18
16	18	20
17	20	22
18	22	25
19	25	28
20	28	31
21	31	34
22	34	37
23	37	40

Figure 7.1: *Time series of relative humidity and temperature for OPC BME280 sensors (orange) against reference values (blue) obtained from the German National Meteorological Service (DWD), at München-City meteorology station located 7.5 km away from the sampling site. Though the BME280 values show a systematic error, it is evident that the variation over time is consistent.*

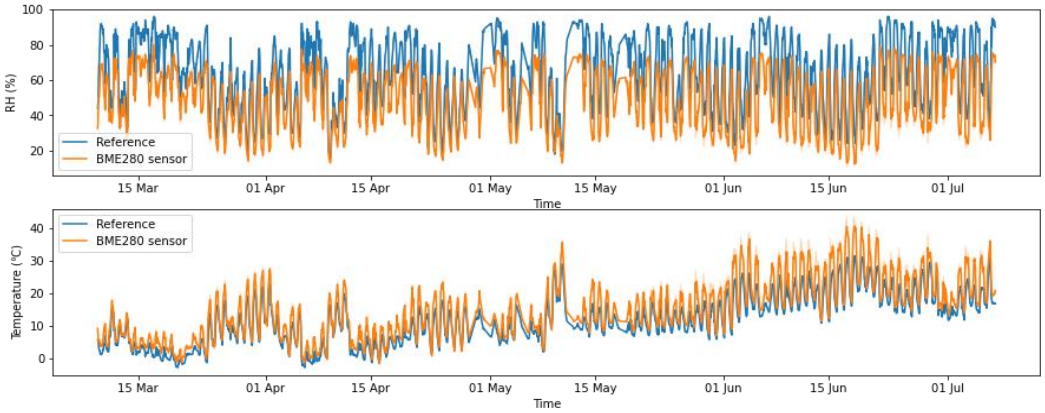


Figure 7.2: *Data availability of each of the OPC instruments in the Munich intercomparison campaign.*

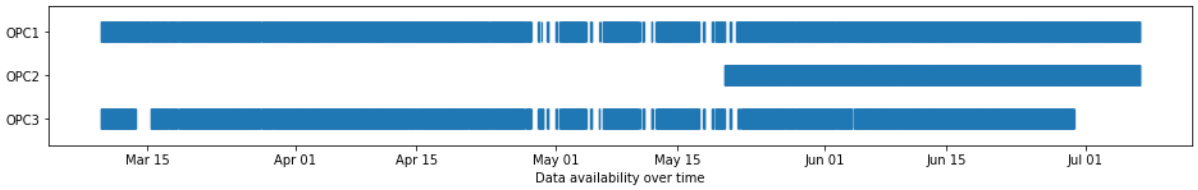


Figure 7.3: *Rainfall time series against the Simple Pollen Proxy for comparison. It can be seen that some peaks in the Simple Pollen Proxy, particularly in late June and July, coincide with rainfall events, suggesting that this can affect the OPCs.*

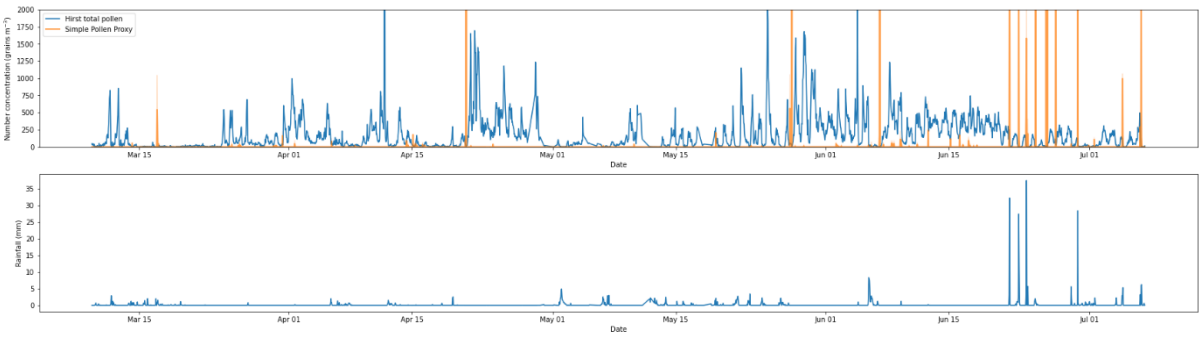


Figure 7.4: *Simplified and generalised diagram of the feedforward neural network, or multilayer perceptron (MLP), architecture used in this study.*

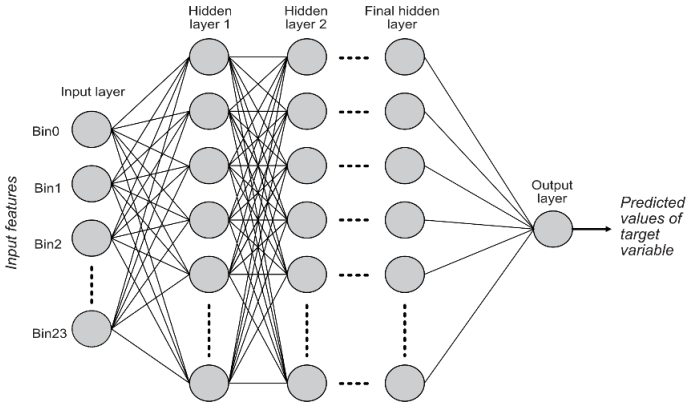


Figure 7.5: *Simplified and generalised diagram of a random forest used for regression machine learning.*

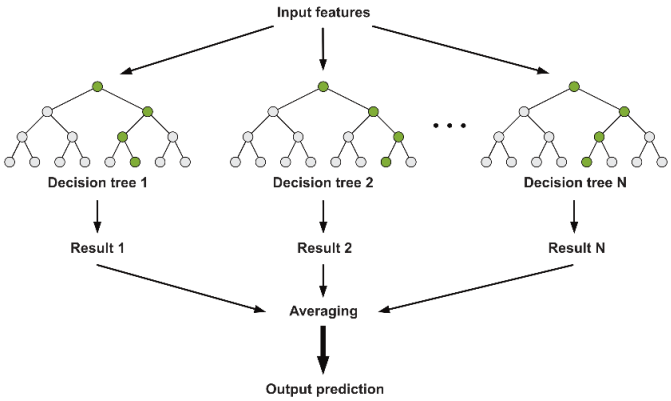


Figure 7.6: *Time series for all Neural Network and Random Forest model-constructed pollen proxies over the whole campaign duration. Hirst data target variable are in blue, model predictions for those with RH and temperature included as input features are in orange, and red for those without meteorological variable input features. Plots on the left (A and C) are from Neural Network models while those on the right (B and D) are from Random Forest models. A and B are predictions for* Poaceae *pollen and C and D for total pollen number concentrations in grains* m^{-3}.

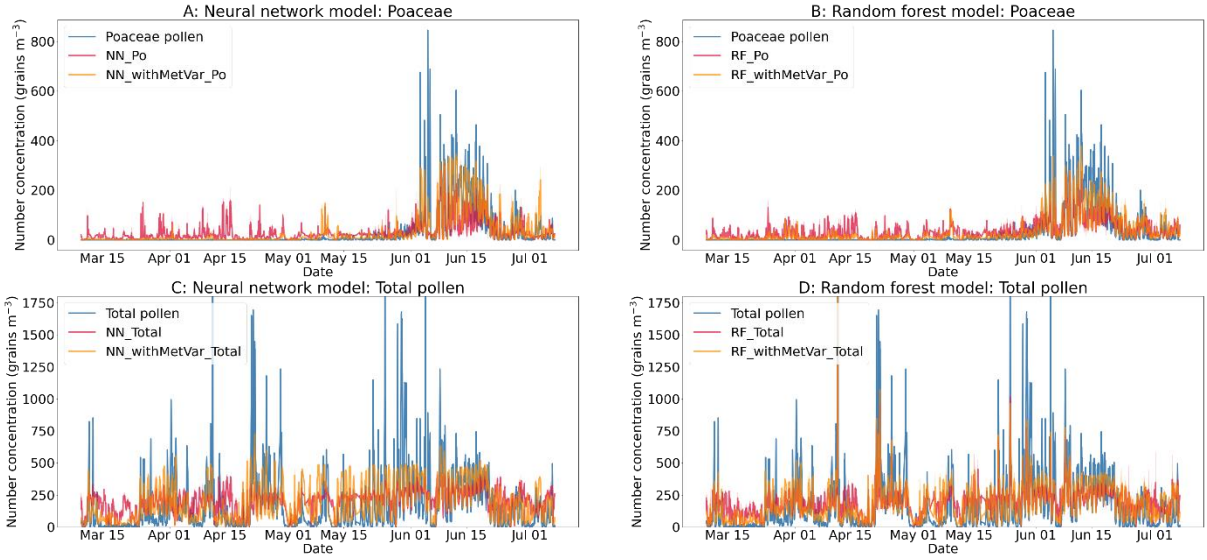


Table 7.2: *Table showing metrics (mean absolute error, Spearman correlation coefficient and coefficient of determination) for Neural Network models trained on just meteorological features relative humidity and temperature.*

NN models	**MAE (grains m^{-3})**	**Spearman**	**R^2**
Poaceae	36.7	0.75	0.36
Total	138.6	0.69	0.18

Figure 7.7: *Pollen concentration time series for Neural Network models trained on just meteorological features RH and temperature, targeting* Poaceae *pollen above and total pollen below.*

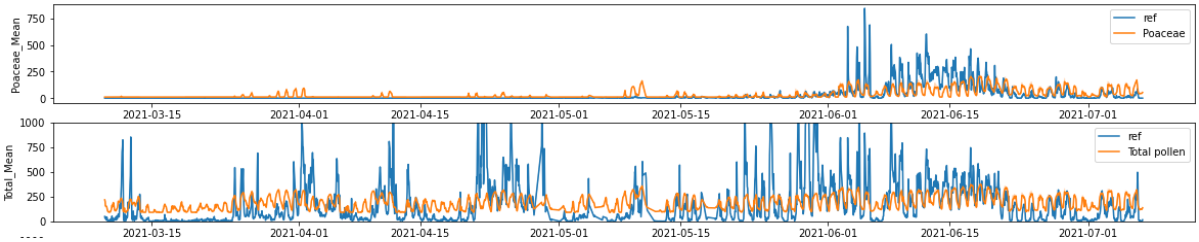


Table 7.3: *Table showing metrics (mean absolute error, Spearman correlation coefficient and coefficient of determination) for Random Forest models trained on just meteorological features relative humidity and temperature.*

RF models	MAE (grains m^{-3})	Spearman	R^2
Poaceae	31.4	0.67	0.44
Total	126.0	0.48	0.22

Figure 7.8: *Pollen concentration time series for Random Forest models trained on just meteorological features RH and temperature, targeting Poaceae pollen above and total pollen below.*

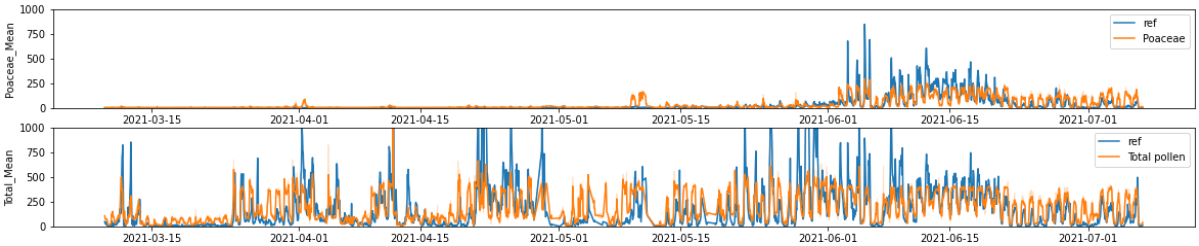


Figure 7.9: *Time series coloured by confusion matrix labels after thresholding for 'high' pollen concentrations. TP = true positive (both Hirst & model-constructed proxy concentrations are above the threshold), FP = false positive (model-constructed pollen proxy concentration is above the threshold while Hirst pollen concentration is below it), TN = true negative (both Hirst and model-constructed proxy are below threshold), FN = false negative (model-constructed proxy is below threshold while Hirst concentration is above it). Data shown here is from the Neural Network model with meteorological input features targeting Poaceae pollen (top) and total pollen (bottom). Includes all data points from the whole dataset, while F1 scores reported were calculated from just the test dataset.*

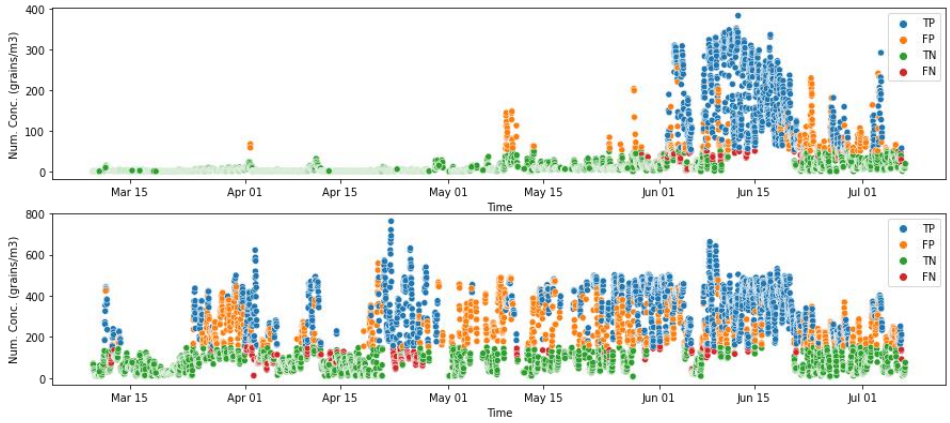
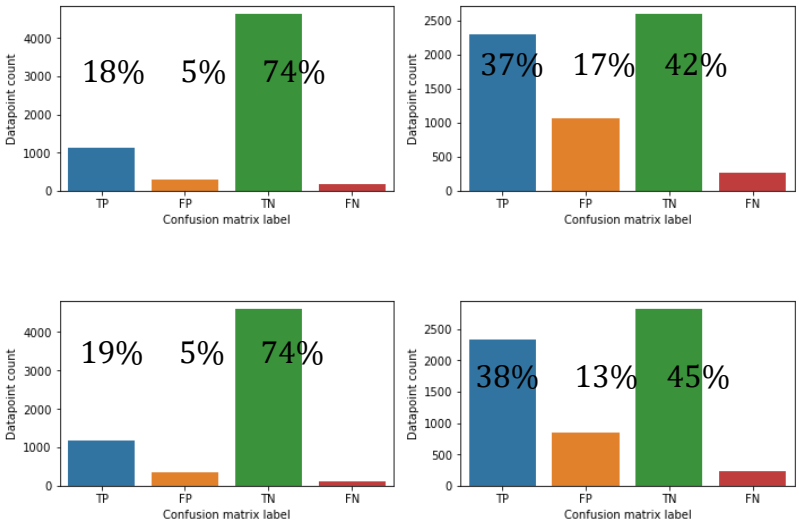


Figure 7.10: *Bar plots of confusion matrix labels on time series data (see Figure 7.5). Confusion matrix label counts across all datapoints (from whole dataset not just test dataset) for Neural Network (above) and Random Forest (below) models with meteorological input features targeting Poaceae pollen (left) and total pollen (right).. TP = true positive (both Hirst & model-constructed proxy concentrations are above the threshold), FP = false positive (model-constructed pollen proxy concentration is above the threshold while Hirst pollen concentration is below it), TN = true negative (both Hirst and model-constructed proxy are below threshold), FN = false negative (model-constructed proxy is below threshold while Hirst concentration is above it).*



8. Supporting information for Chapter 3

Section 1: General OPC bin information

Table 8.1: *Table of 24 OPC bins with corresponding particle optical diameter ranges detected for each.*

OPC bin	Min (µm)	Max (µm)	Width (µm)
0	0.35	0.46	0.11
1	0.46	0.66	0.2
2	0.66	1.0	0.34
3	1.0	1.3	0.3
4	1.3	1.7	0.4
5	1.7	2.3	0.6
6	2.3	3.0	0.7
7	3.0	4.0	1.0
8	4.0	5.2	1.2
9	5.2	6.5	1.3
10	6.5	8.0	1.5
11	8.0	10	2.0
12	10	12	2.0
13	12	14	2.0
14	14	16	2.0
15	16	18	2.0
16	18	20	2.0
17	20	22	2.0
18	22	25	3.0
19	25	28	3.0
20	28	31	3.0
21	31	34	3.0
22	34	37	3.0
23	37	40	3.0

Section 2: Extended methods section

The following sections provide additional detail on tuned hyperparameters - why they were chosen and how they can affect a model's learning and performance.

Number of layers and nodes

The number of layers and the number of nodes within each of those layers in a neural network determines the capacity of the model. Too small a capacity, with too few layers or nodes, causes the model to underfit as it cannot sufficiently learn the training dataset. Too great a capacity, with an excessive number of layers or nodes, can cause a model to overfit and not generalise well for new data (Deep Learning, 2016, p108-109). Therefore, the number of layers and nodes in the model architecture are important hyperparameters that affect performance. For this study, models were tested with for all combinations of 3-7 layers of 50, 100, 200 and 400 nodes in each layer.

Batch size

Batch size is the number of samples from the training dataset used to estimate the error gradient, used to update the weights each iteration as the model learns. Therefore, the batch size affects the accuracy of the error gradient estimate and the learning dynamics, so is an important hyperparameter to consider (Practical recommendations for gradient-based training of deep architectures, 2012, Revisiting Small Batch Training for Deep Neural Networks, 2018). For this study, models were tested with batch sizes of 32, 64, 128, 256 and 512.

Optimisation algorithm and learning rate

Neural networks use stochastic gradient descent, an empirical optimisation method, to decide how much to update the weights while probing the space of possible solutions

for an appropriate set of weights that best maps the inputs onto the target output. The optimisation procedure, or learning rate schedule, controls the resolution at which the space of possible solutions is probed, and thus how fast the model learns a problem. If the learning rate is too small, training may take too long and the model may never converge, or it may get stuck at a suboptimal solution. If too large, the model may miss optimal solutions and the training error may oscillate or even increase over training epochs.

The learning rate can interact with other aspects involved in model optimisation, as such it can be complicated to tune accurately. This has led to the common use of adaptive learning rates, which adjust the learning rate in response to the performance of the model on the training dataset while learning. There are methods that have proven to be robust and generally perform better than badly configured learning rates (pp72, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, 1999). It is however possible for more accurately configured learning rates to perform better. For this study, models were tested with three different methods of optimisation algorithms which apply an adaptive learning rate – RMSProp (Root Mean Square Propagation), Adagrad (Adaptive Gradient Algorithm) and Adam (Adaptive Moment Estimation).

A further experiment was conducted where the initial learning rate was varied for the best achieving optimisation algorithm, Adam. The learning rates trialled were 0.0001, 0.0005, 0.001 and 0.005 (0.001 is the default in Tensorflow).

Activation function

At each node in a neural network, the incoming values are multiplied by the current weights associated with that node, summed together and then transformed by the

activation function that is defined at that node. Thus the activation function defines how the output values of each node are produced before being fed into the nodes of the following layer. The modern default for many multi-layer (or deep) neural network architectures is the Rectified Linear Unit activation function (ReLU) since it demonstrates non-linearity and also removes the vanishing gradient problem that activation functions such as sigmoid and tanh have. Meanwhile it is relatively simple and computationally inexpensive to implement. However, there are other activation functions based on ReLU yet with added complexity which can add more nuance and address the possible problem of dying ReLU. Examples of these are ELU (Exponential Linear Unit) and SELU (Scaled Exponential Linear Unit) which, unlike ReLU, can produce negative outputs. While ELU is similar to ReLU except that it demonstrates exponential behaviour below 0, SELU is relatively more complex since it has an additional parameter and can exhibit self-normalisation by preserving the mean and variance across layers. Still there is no hard certainty for the most appropriate activation function to be used in any given case. So here we test across models applying in turn three different activation functions, ReLU, ELU and SELU, at each hidden layer.

Data scaling

Data scaling is an importance pre-processing step to consider before training neural networks as it can affect their stability during the learning process. Large unscaled input values can cause the model to learn large weights which can bring instability in performance and magnify sensitivity to input values, resulting in worse generalisation (Neural Networks for Pattern Recognition, 1995, p296-8). Unscaled target variables with a large spread may destabilise learning by large error gradient values which change the weight values too dramatically.

Prior to this study it was already established that normalising the input features was necessary. In this study, we trained different models that either did or did not scale the target variable as well to investigate the effect it has.

Test dataset size

Appropriate consideration and preparation of the dataset that is used for learning is evidently important. To obtain separate datasets for training and evaluation of the model we need to split the dataset we have. One may think the larger the training dataset, the more information the model has to learn from and the better it can perform. However, the test dataset used for evaluating the model also has to be large enough to be representative of the whole dataset in order to really test the generalisability of the model (within the conditions of this dataset). This dataset covers a long time period which includes a variety of different pollen seasons, many of which overlap. It is unlikely for any short selected time period to be representative of the whole dataset and be enough for the model to learn across all seasons and taxa present. This is why the dataset is shuffled before being split. This also must be considered when reducing the size of the test dataset which, if too small, may not be representative of all taxa to evaluate the model appropriately.

Therefore, here we also chose to change the size of the portion chosen as the test dataset, to see what effect this had and to what point the training dataset can be maximised without detriment to model evaluation. Here we have tested test dataset proportions of 0.25, 0.2, 0.15 and 0.1, with corresponding training dataset proportions of 0.5, 0.6, 0.7 and 0.8 respectively (the remaining proportion went to the separate validation dataset).

Dropout nodes

A well established and popular method for reducing overfitting in neural networks is using dropout regularisation. With this method, a parameter (α) is chosen which then causes that proportion of nodes in the layer it is applied at to be randomly ignored, or 'dropped out'. It is a computationally cheap and effective regularisation method which can be equivalent to the effect of sampling many slightly different networks, making the overall model more robust and improve generalisation. For this work, dropout layers were implemented after each hidden layer and values of 0.1, 0.2, 0.3, 0.4 and 0.5 were tested for the α parameter (the percentage of nodes dropped at each layer).

Weight constraints

Large weights can cause model instability as well as overfitting of the training dataset, where statistical noise too is learned, at detriment to its generalisation ability. Thus, it is often advantageous to keep the weights in a neural network in check. Weight penalties can be applied which scale down weights depending on their magnitude, thus encouraging weights to remain small. Weight constraints on the other hand force weights to be small by preventing them exceed a given threshold. Here we apply maximum norm constraints which, if a weight magnitude exceeds the specified limit, rescales the weights so they remain within the threshold. We trial threshold values of 1.0, 2.0, 3.0, 4.0 and 5.0 for the max-norm constraint to keep the weight magnitudes within these values at all hidden layers. We also tested applying slightly more aggressive unit norm constraints at all hidden layers which force weights to be scaled within a magnitude of 1.0, regardless of whether they have reached that magnitude or not.

Noise

Adding a small degree of noise to the input data during training is another potential method to increase model robustness. Especially when datasets are relatively small,

complex deep neural network models have a danger of learning to specifically map the training data onto the output too well. With smaller, sparser datasets there is also less opportunity to describe the input space and relate it to the output. In such instances, models cannot learn helpful general mapping from inputs to outputs and so cannot perform well on new data. The whole dataset used for this work in hourly time resolution comprises of 6220 data points (then split into train, validation and test datasets). While not very small, by general deep learning standards this is not a particularly large number. Meanwhile the various pollen seasons for each taxa are only represented in limited portions of this dataset, so particularly for models predicting specific pollen taxa there is limited representative data to learn from.

Noise can be added to parts of the network, as opposed to just the inputs, during training - including to layer outputs (before or after the activation function), to weights, to the gradient updating the weights, or the target variable outputs – with varying degrees of standard deviation. For the purposes of this work, we trial adding noise to either the input data or the outputs from each hidden layer (following the activation function) with standard deviations of 0.001, 0.0001 and 0.00001.

Section 3: Various dataset statistics relevant to the XAI/SHAP analysis on high and low pollen and error categories

Table 8.2: *Table presenting the percentage of non-zero data points from the whole dataset for each target variable, and the fraction and percentage out of the test dataset in each case. Note, the test dataset was 10% in the case of total pollen and* Poaceae *but 20% for the other taxa.*

	% non-zero data points in whole dataset	**Fraction (and %) of non-zero data points in test dataset**
Total pollen	94%	584/622 (94%)
Poaceae	51%	300/622 (48%)
Quercus	22%	263/1244 (21%)
Betula	33%	389/1244 (31%)
Pinus	49%	612/1244 (49%)

Table 8.3: *Table of the number of data points that were filtered by each category thresholds based on pollen concentrations and absolute error between prediction and Hirst observation.*

	HPLE	**LPLE**	**HPHE**	**LPHE**
Total pollen	5	62	77	26
Poaceae	3	29	51	4
Quercus	6	30	51	5
Betula	3	58	69	3
Pinus	2	60	96	22

Table 8.4: *Table of threshold values used to establish high/low pollen concentrations and model prediction error. The high and low thresholds correspond to the 75th and 25th quantiles respectively of pollen concentration and absolute error, above and below which 'high' and 'low' was defined. The only exception is for the low error threshold for* Betula *which was raised to the 40th quantile, since below this there were no data points that met the double thresholds of high pollen and low error.*

	HP threshold (grains m^{-3})	**LP threshold (grains m^{-3})**	**LE threshold (grains m^{-3})**	**HE threshold (grains m^{-3})**
Total pollen	296.0	28.0	20.9	133.0
Poaceae	119.8	10.0	9.6	47.1
Quercus	20.5	4.0	2.4	13.9
Betula	48.0	4.0	3.2*	42.6
Pinus	85.0	6.0	7.2	89.6

Section 4: Extended SHAP analysis figures

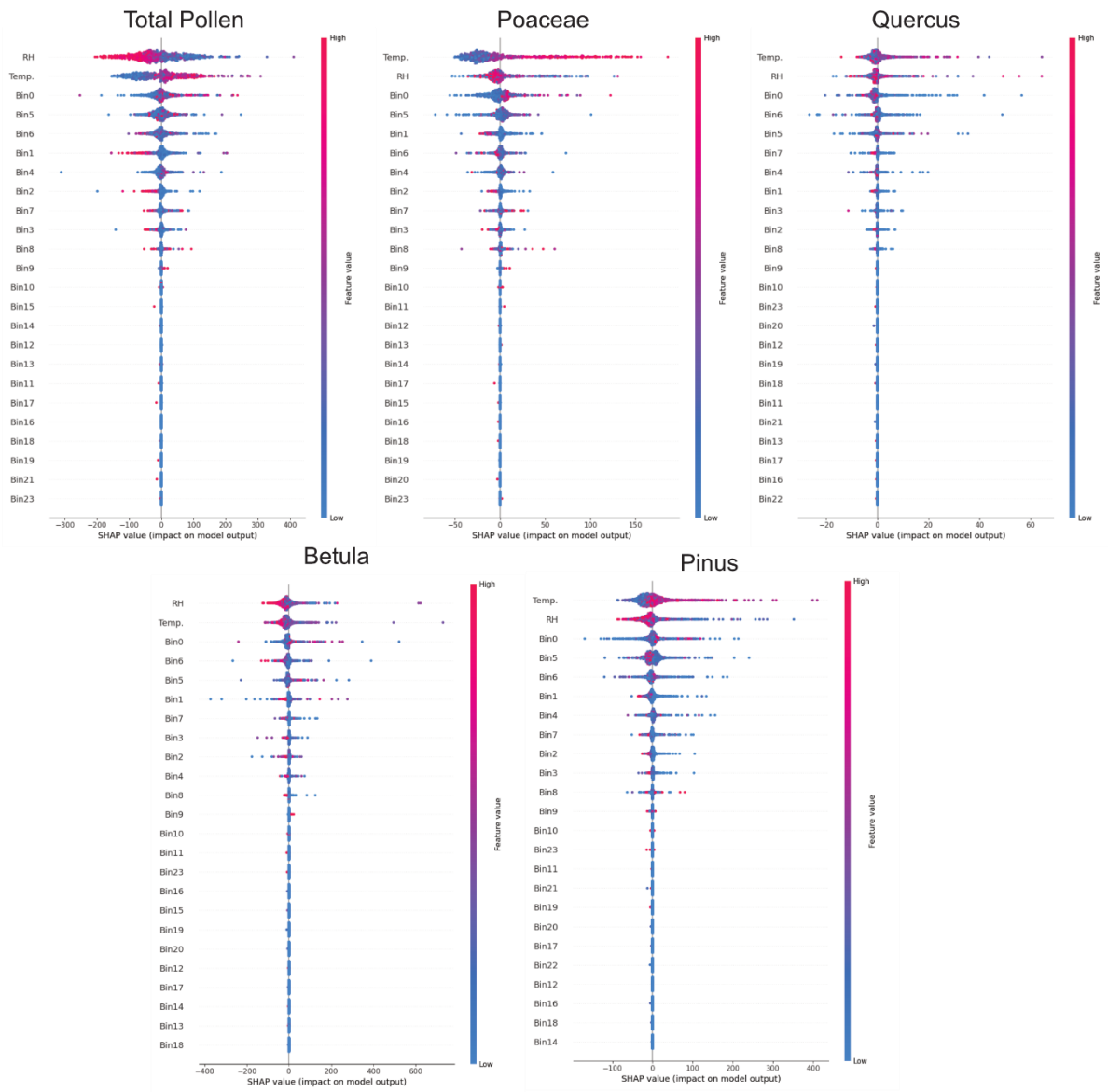


Figure 8.1: *SHAP 'bee swarm' plots for all pollen taxa models based on OPC bins and meteorological variables.*

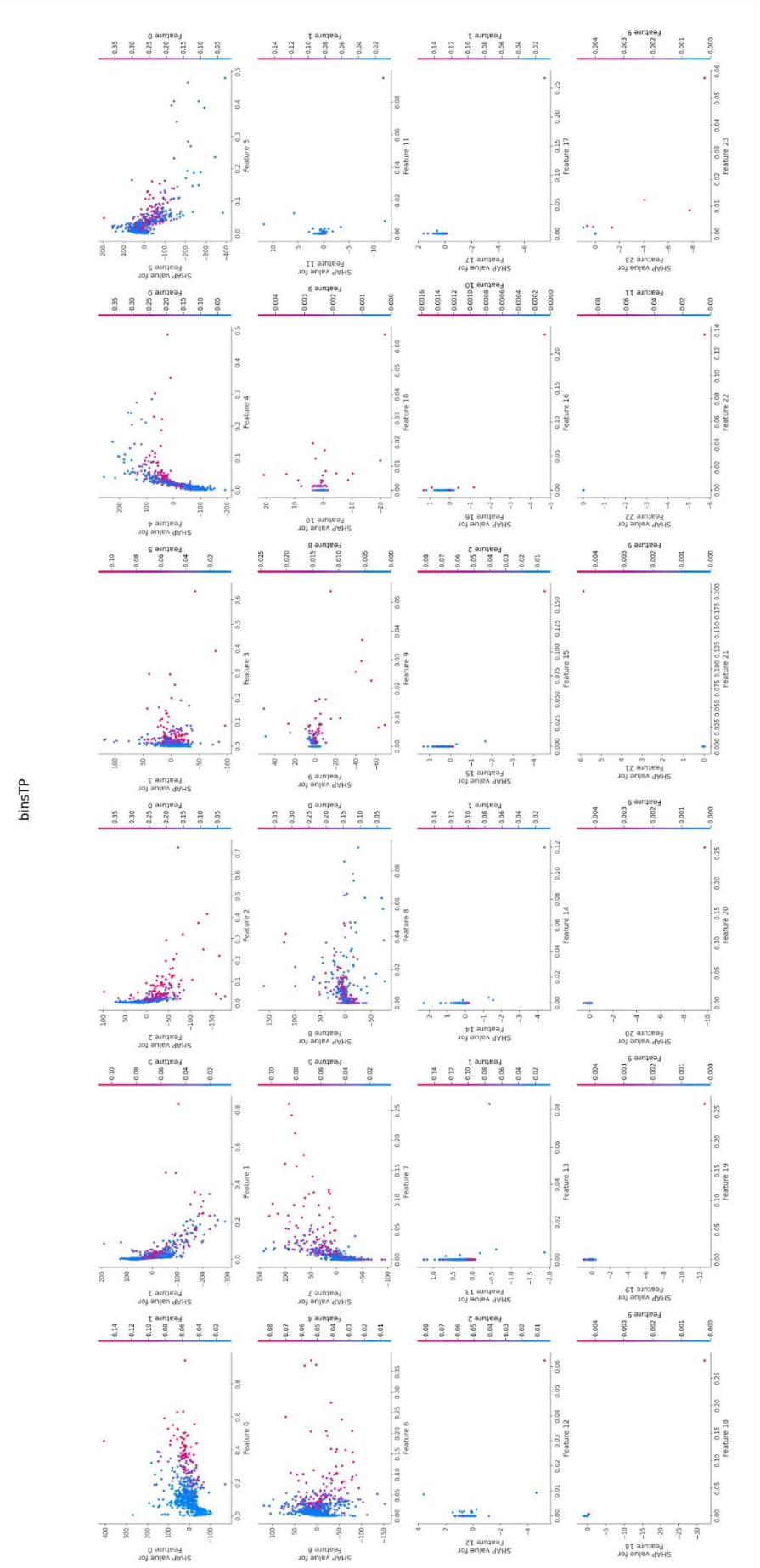


***Figure 8.2**: Partial dependence plots for total pollen model based on OPC bins only. Plot for each input feature/bin (e.g. Bin0, Bin1, …, Bin23) listed across in rows then columns. Y axis: SHAP value (impact on model output), X axis: normalised feature value, colour: value of feature with greatest interaction.*

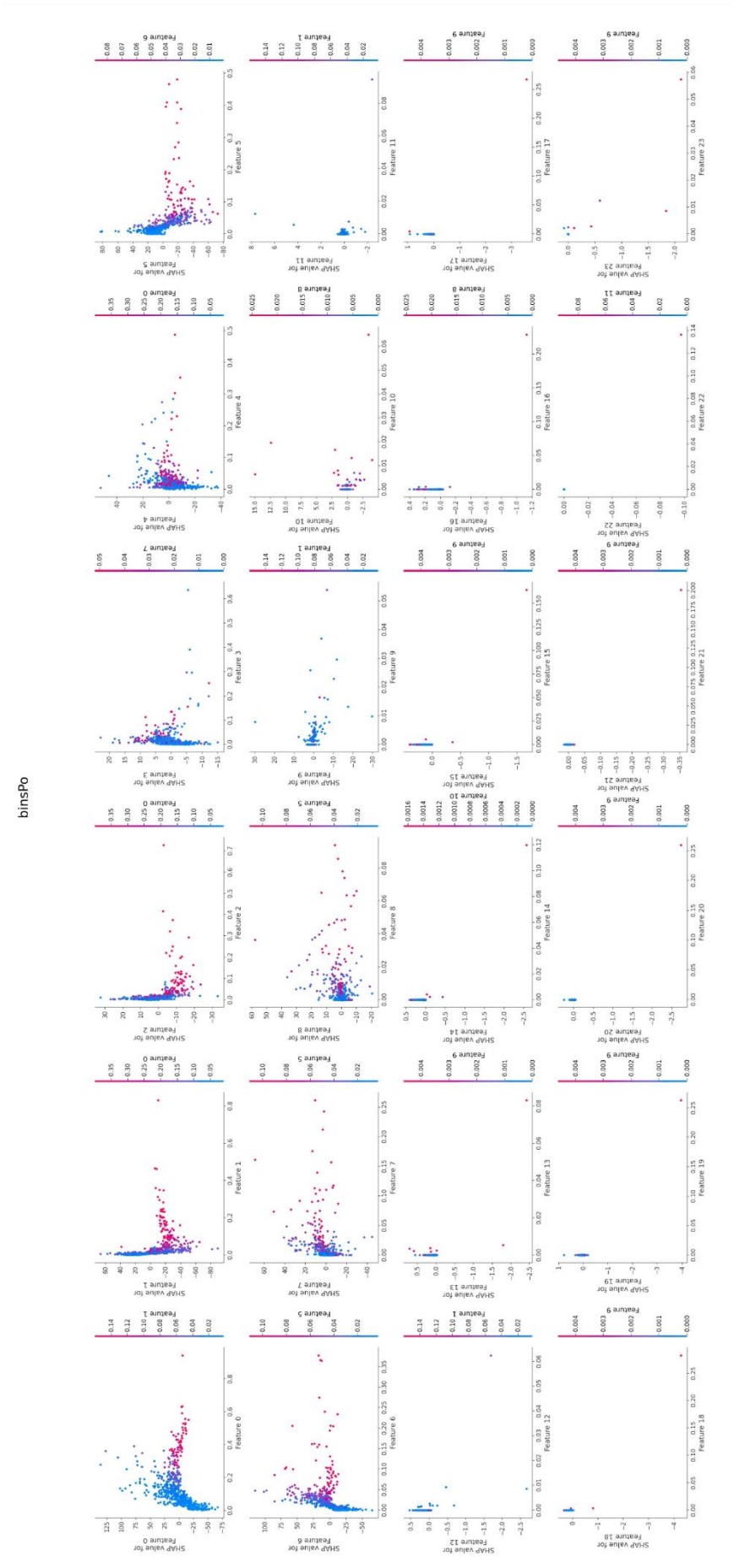


***Figure 8.3**: Partial dependence plots for* Poaceae *pollen model based on OPC bins only. Plot for each input feature/bin (e.g. Bin0, Bin1, …, Bin23) listed across in rows then columns. Y axis: SHAP value (impact on model output), X axis: normalised feature value, colour: value of feature with greatest interaction.*

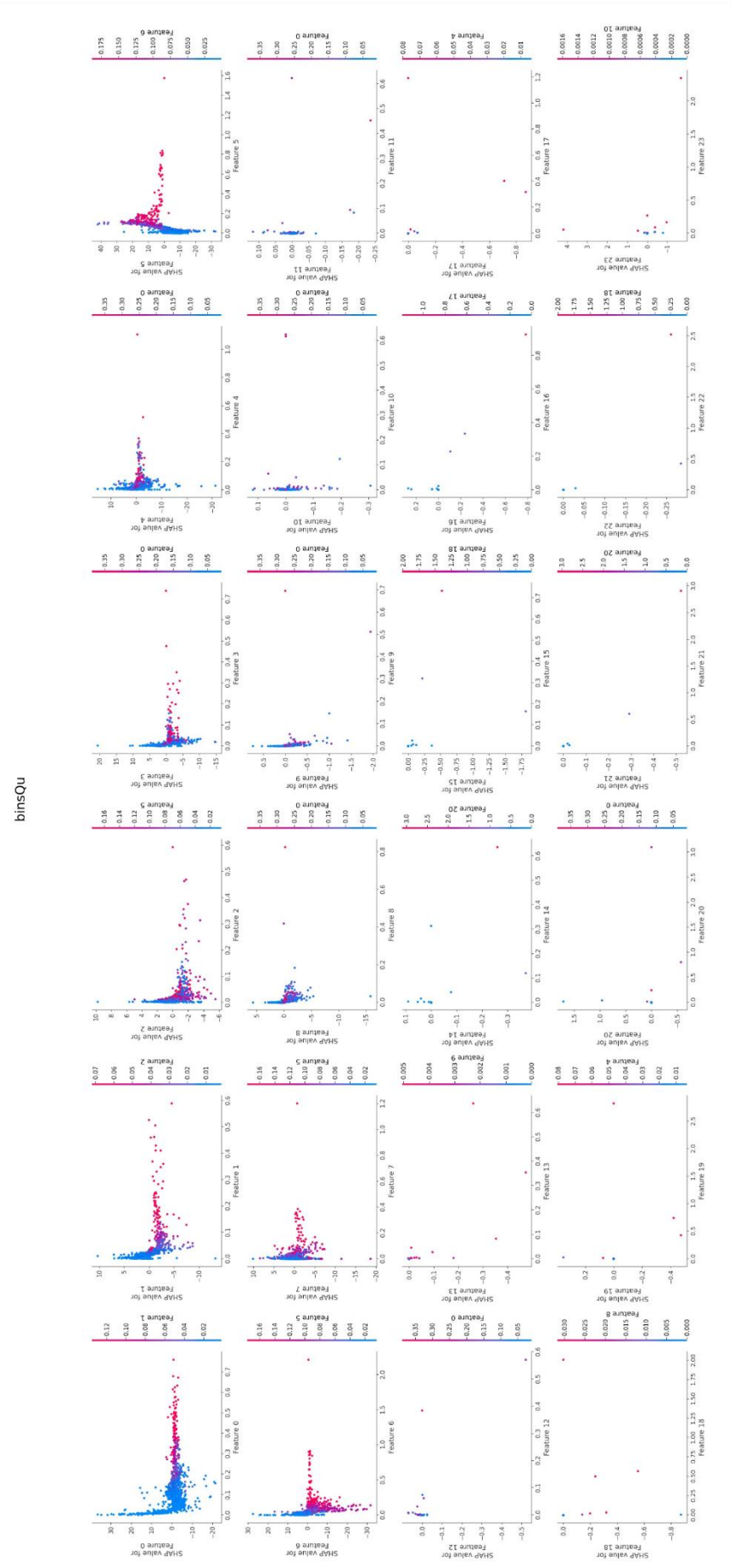


***Figure 8.4**: Partial dependence plots for Quercus pollen model based on OPC bins only. Plot for each input feature/bin (e.g. Bin0, Bin1, …, Bin23) listed across in rows then columns. Y axis: SHAP value (impact on model output), X axis: normalised feature value, colour: value of feature with greatest interaction.*

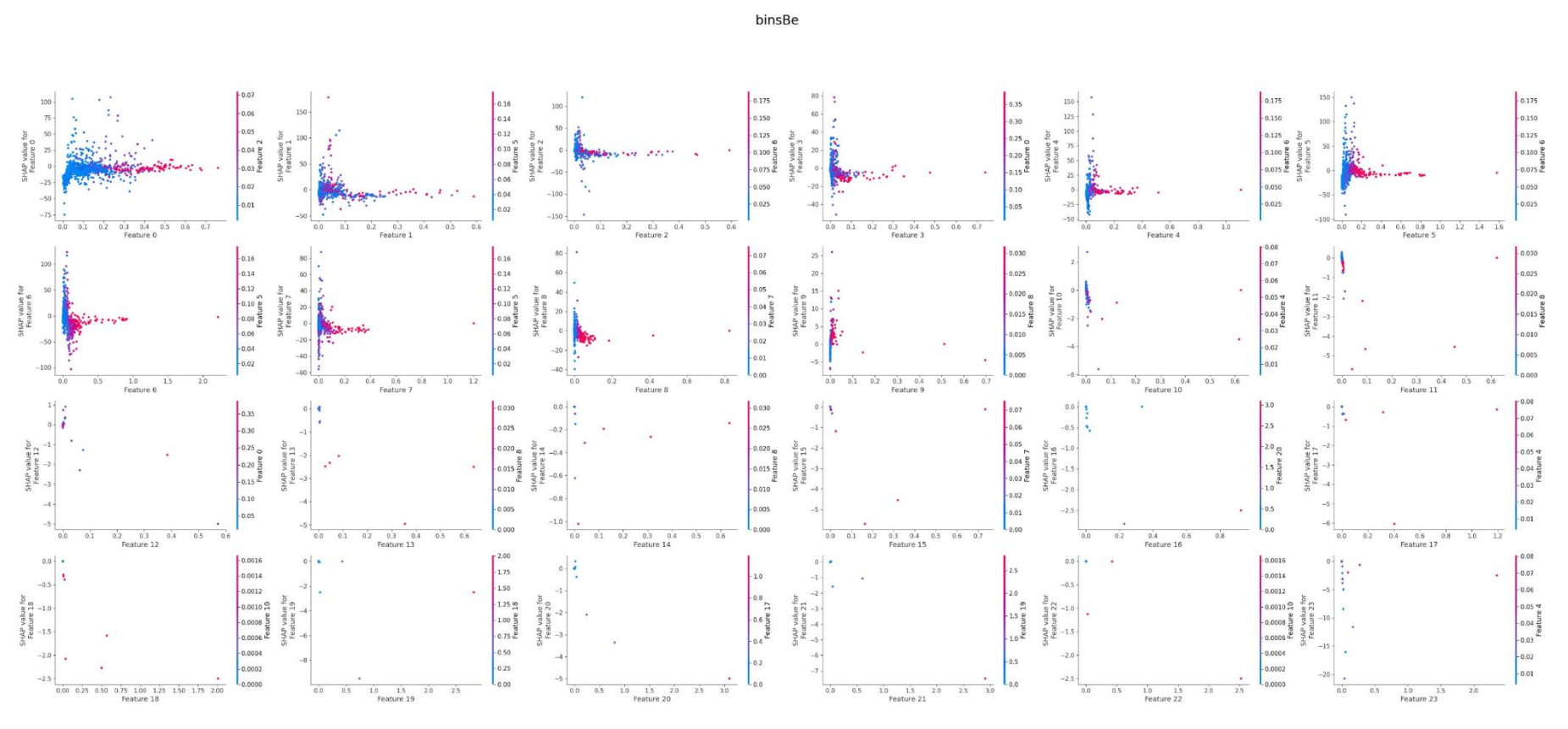


Figure 8.5*: Partial dependence plots for* Betula *pollen model based on OPC bins only. Plot for each input feature/bin (e.g. Bin0, Bin1, …, Bin23) listed across in rows then columns. Y axis: SHAP value (impact on model output), X axis: normalised feature value, colour: value of feature with greatest interaction.*

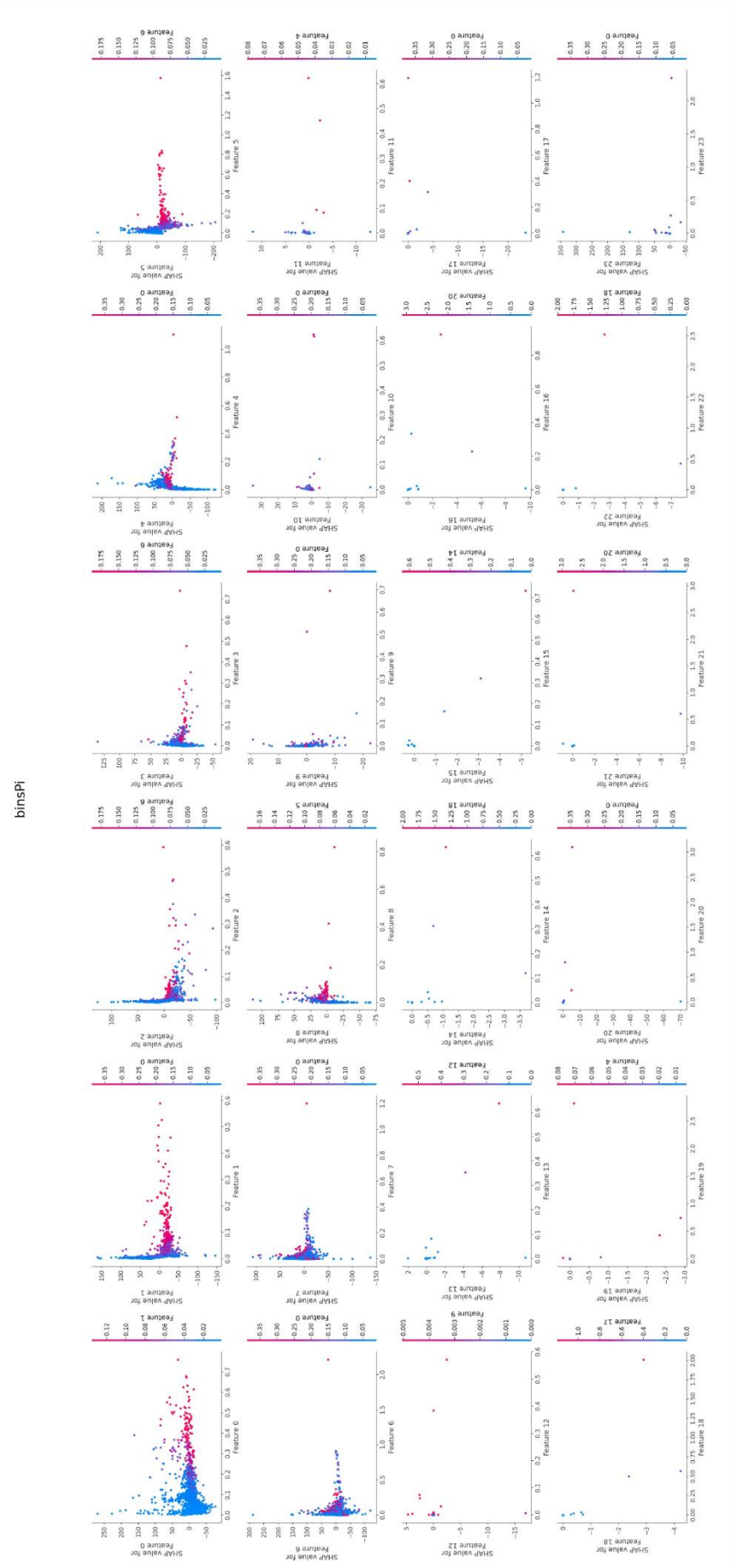


***Figure 8.6**: Partial dependence plots for* Pinus *pollen model based on OPC bins only. Plot for each input feature/bin (e.g. Bin0, Bin1, …, Bin23) listed across in rows then columns. Y axis: SHAP value (impact on model output), X axis: normalised feature value, colour: value of feature with greatest interaction.*

***Table 8.5**: Table summarising the general correlation each feature had on the model output, based visually on the SHAP bee swarm plots in in Figure 3.4 (and Figure 8.1 in the Appendices for models including meteorological features). The matrix plot in Figure 3.4 in the paper is a condenses version of this, where the OPC bin rows are from the models trained without meteorological variables and the RH and temperature rows are taken from the those trained with meteorological variables. Positive/red means that feature tended to increase the model output value with greater values of that feature, whereas for negative signs/blue this means greater values of that feature tended to decrease the model output value.*

Feature	Correlation with TP		Correlation with Po		Correlation with Qu		Correlation with Be		Correlation with Pi	
	bins	binsRHT	bins	binsRHT	bins	binsRHT	bins	binsRHT	Bins	binsRHT
Bin0	(+)	(+)	+	+	(-)	(-)	~	(+)	(+)	(+)
Bin1	-	-	(-)	-	-	-	(-)	~	(-)	-
Bin2	-	-	-	-	-	~	-	~	-	-
Bin3	(-)	(-)	~	(-)	-	~	(-)	-	~	(-)
Bin4	+	~	+	~	-	(-)	+	(-)	+	~
Bin5	-	~	-	(+)	+	(+)	(+)	(+)	-	(-)
Bin6	(-)	(-)	+	(-)	-	(-)	-	-	(-)	~
Bin7	+	~	(+)	(+)	~	(-)	~	(-)	(-)	~
Bin8	(+)	~	(+)	~	-	(-)	~	-	+	(+)
Bin9	~	(+)	(-)	+	-	~	+	+	(-)	~
Bin10	~	~	~	~	-	~	-	~	~	~
Bin11	~	~	~	~	-	~	-	~	~	~
RH		-		(-)		~		(-)		-
Temp.		+		+		~		(-)		+

Section 5: Effect of removing certain bin input features from models

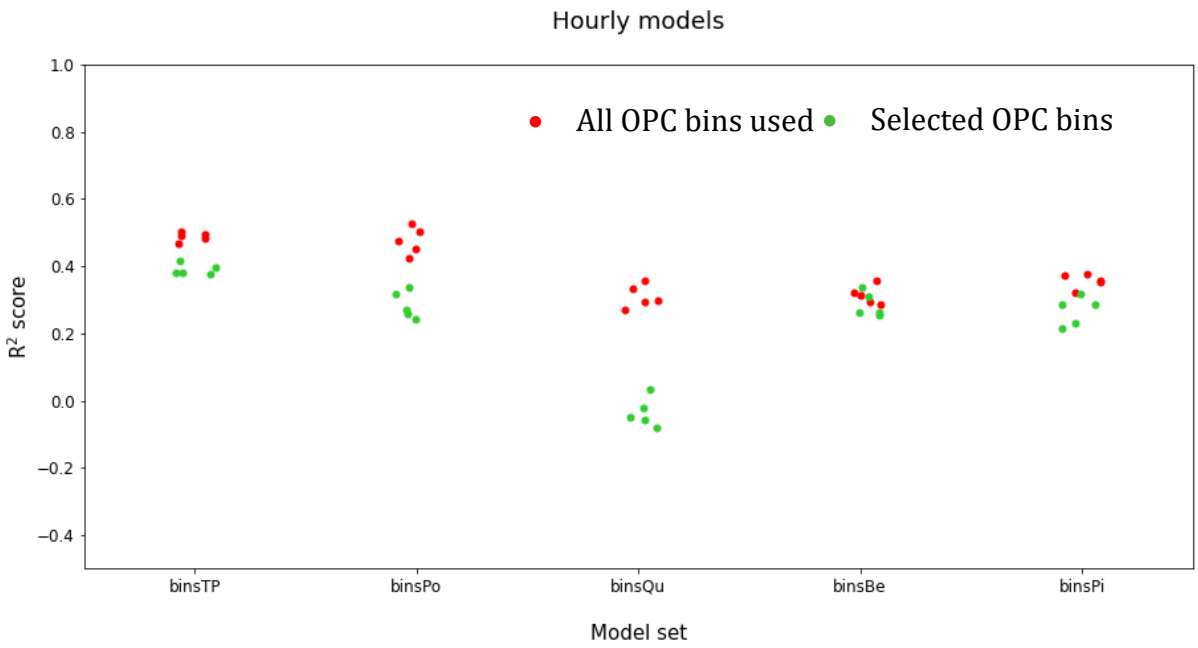


***Figure 8.7**: R^2 scores for models with selected OPC bin input features (green) compared with original models that used all 24 OPC bins as input features. For total pollen (TP) the bins used are 0,2-4,6-8; for* Poaceae *(Po) the bins used are 0,2-4,6-8; for* Quercus *(Qu) the bins used are 4-8; for* Betula *(Be) the bins used are 0-8; for* Pinus *(Pi) the bins used are 0-4, 6-8.*

9. Supporting information for Chapter 4

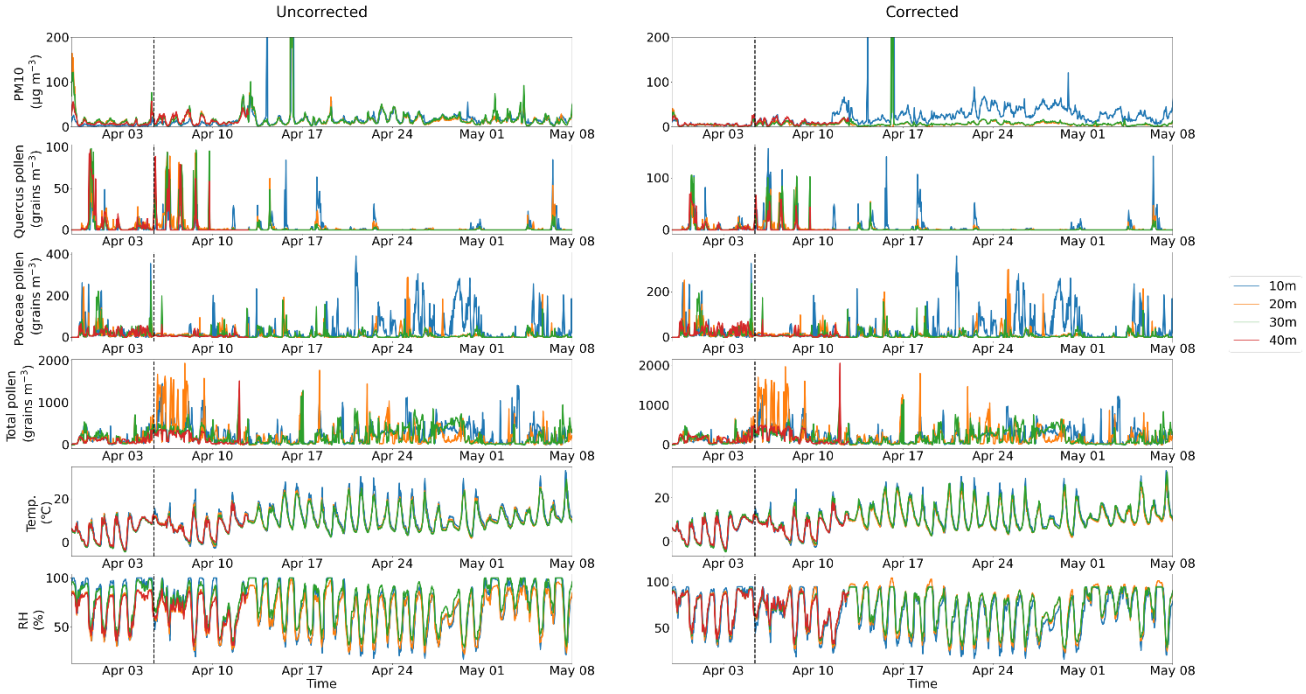


Figure 9.1: *Time series of PM_{10},* Quercus *pollen,* Poaceae *pollen, total pollen number concentration, temperature, and RH from top to bottom. On the left are uncorrected time series, while those on the right are corrected as described in Chapter 4.*

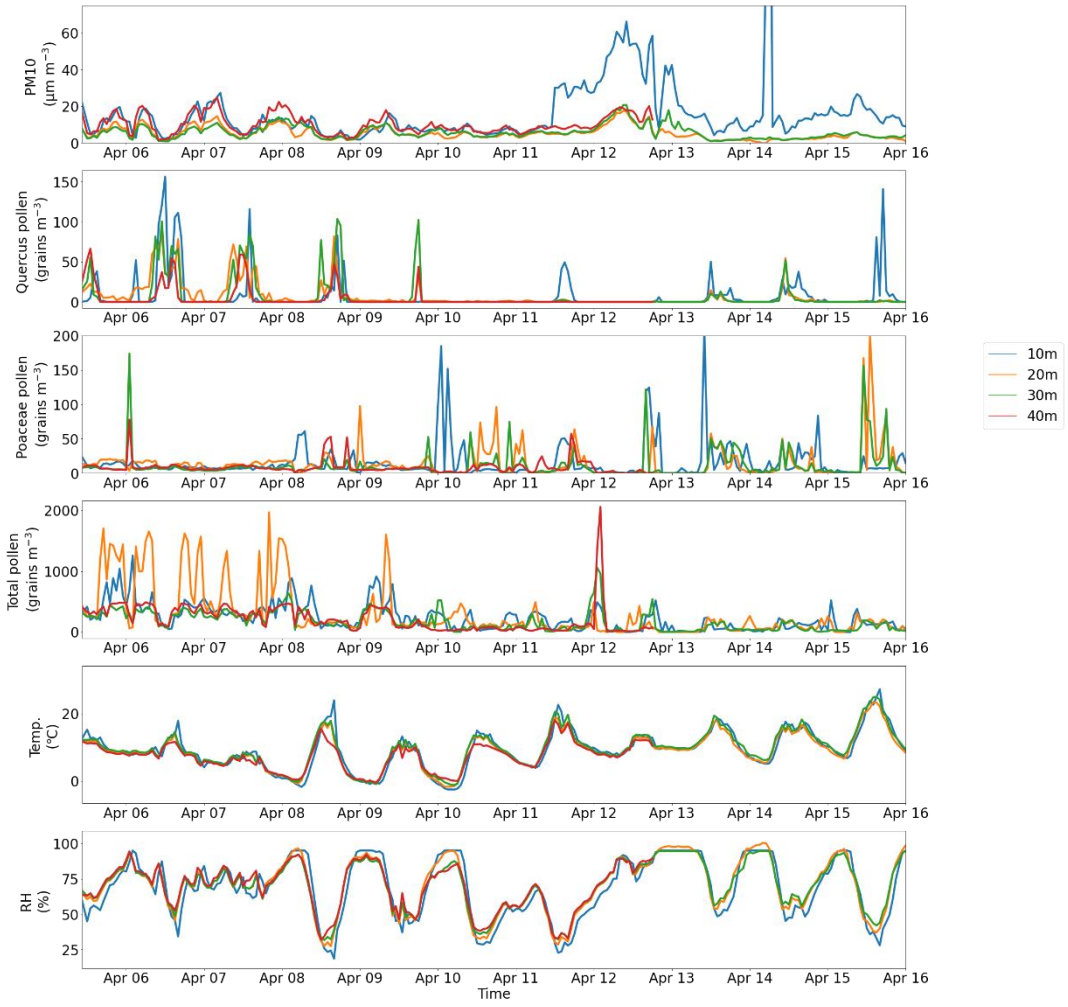


Figure 9.2: *Zoomed in time series from Figure 9.1 with only corrected data and from the start of the 'tower' time period (5th April) up to 16th April 2022.*

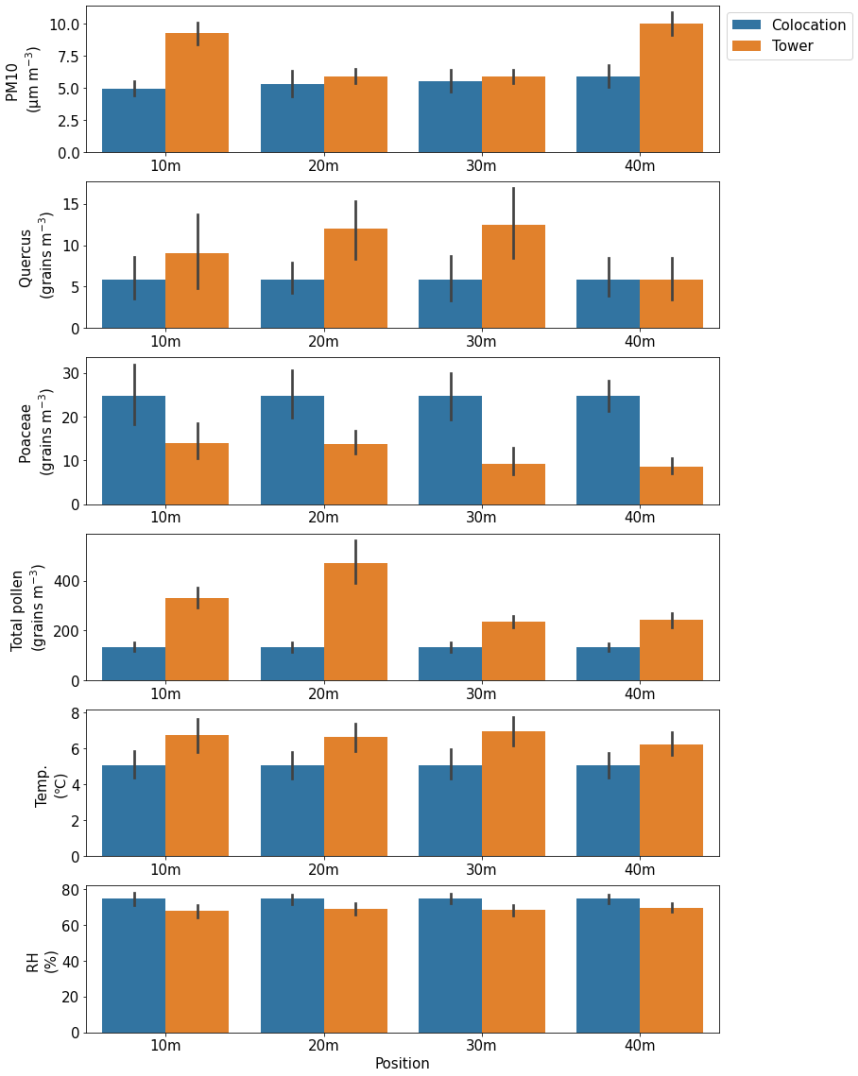


Figure 9.3: *Bar plots of mean values for each variable with 95% confidence intervals for each sensor (x-axis) and colocation/tower period (colour) respectively. Blue = colocation period; orange = tower period. Values for the tower period were only included up until 11th April to exclude the later period when the 40 m sensor stopped logging and the 10 m sensor PM_{10} values became anomalous from the other sensors.*

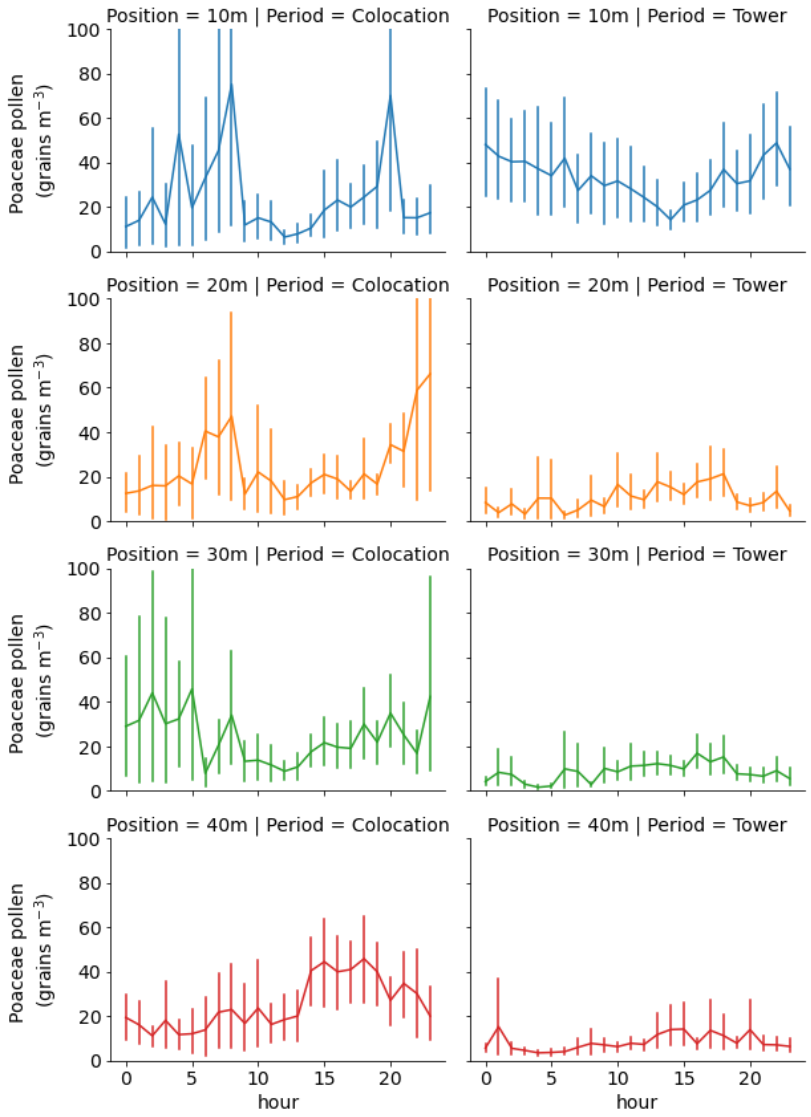


Figure 9.4: *Diurnal trend for* Poaceae *pollen.* Poaceae *pollen number concentrations averaged over each hour of the day, for each sensor (row), for both colocation (left column) and tower (right column) periods. The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May).*

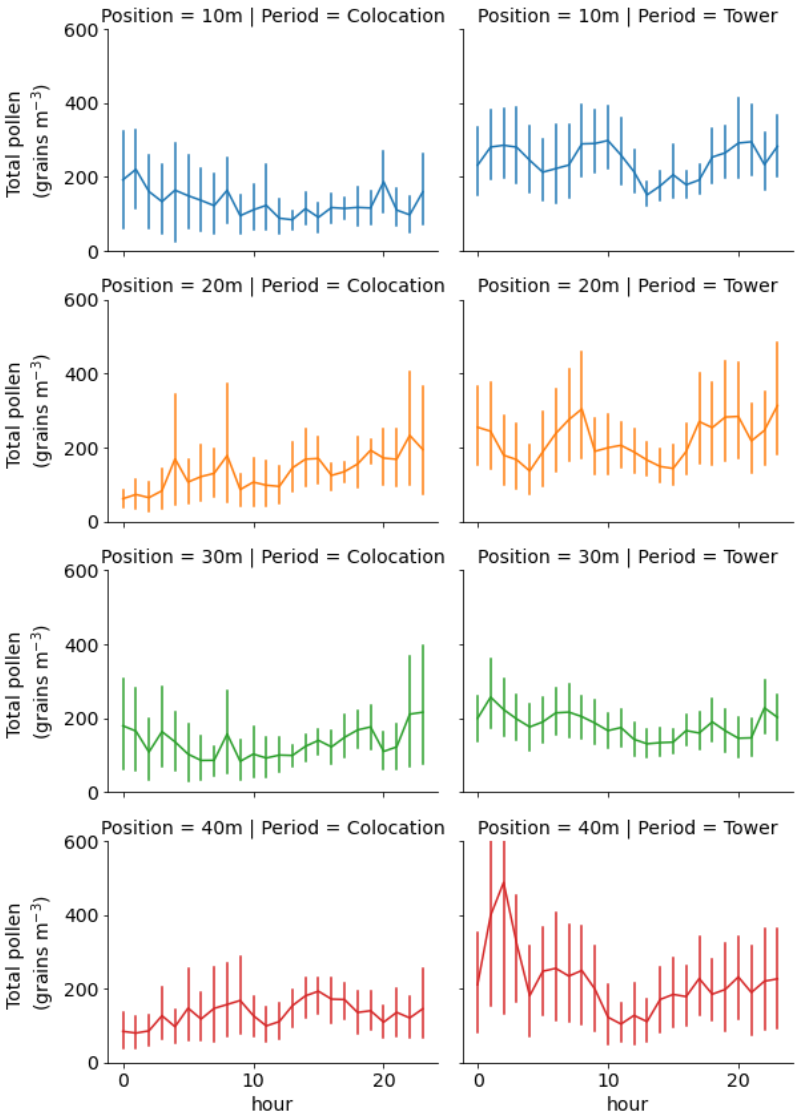


Figure 9.5: *Diurnal trend for total pollen. Total pollen number concentrations averaged over each hour of the day, for each sensor (row), for both colocation (left column) and tower (right column) periods. The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May). (The y-axis has been capped at 600 grains m^{-3}.)*

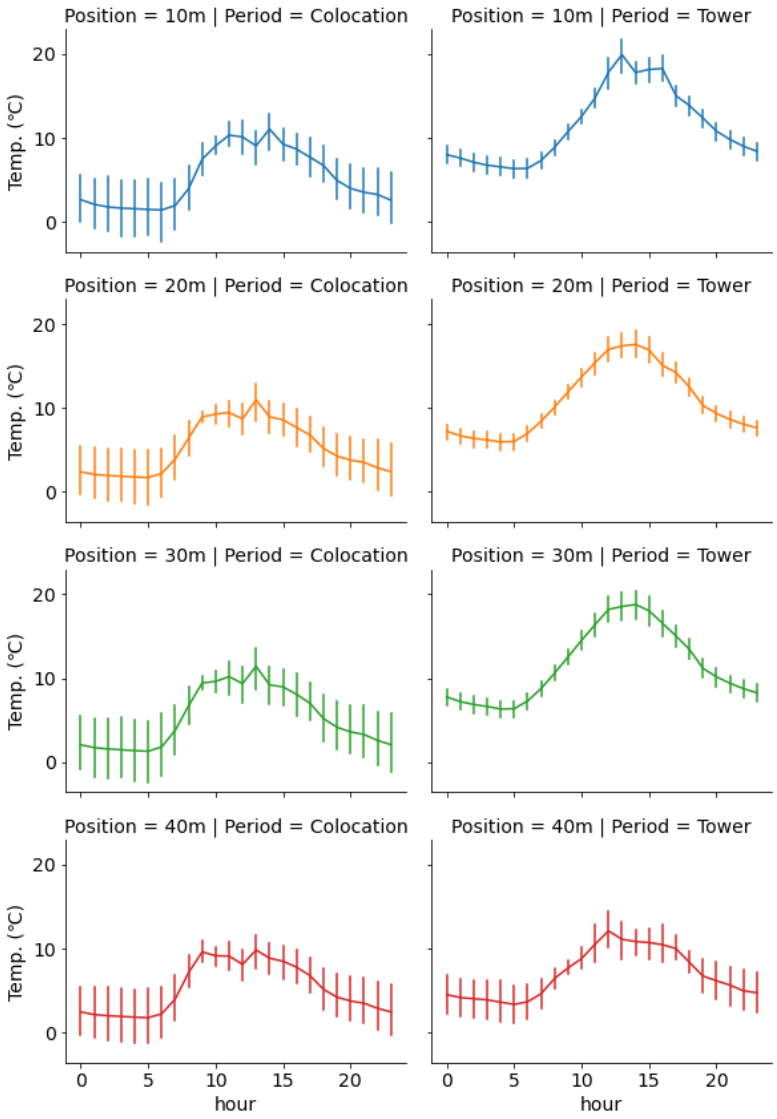


Figure 9.6: *Diurnal trend for temperature. Temperature averaged over each hour of the day, for each sensor (row), for both colocation (left column) and tower (right column) periods. The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May).*

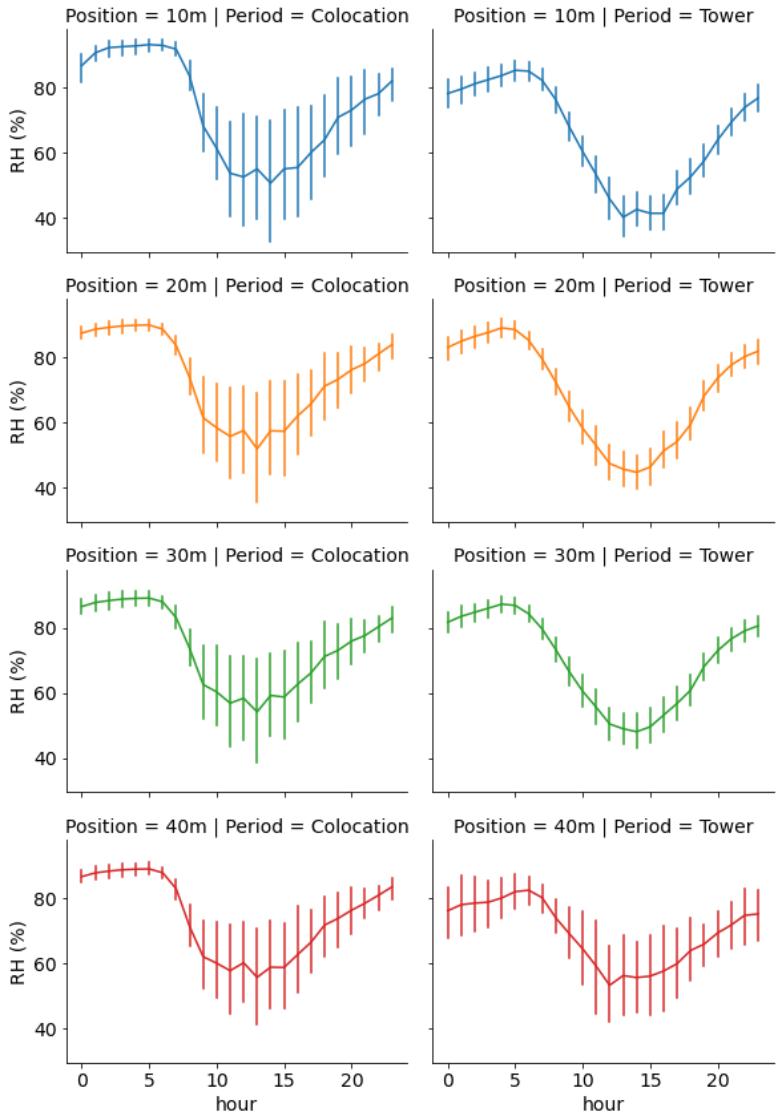


Figure 9.7: *Diurnal trend for relative humidity (RH). RH averaged over each hour of the day, for each sensor (row), for both colocation (left column) and tower (right column) periods. The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May).*

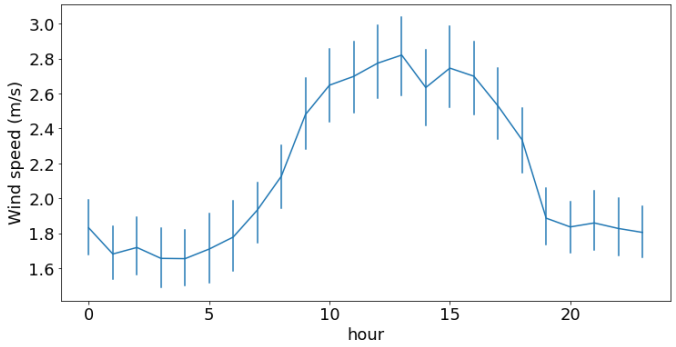


Figure 9.8: *Diurnal trend for wind speed. Wind speeds averaged over each hour of the day for whole period (colocation and tower). The line represents the average (mean) while the bars represent the 95% confidence interval across the period of data collection (until 8th May). The positioning of the instrument monitoring wind was consistent through both colocation and tower periods and irrespective of the height of the low-cost sensors, hence only one summary plot has been constructed for the whole period.*

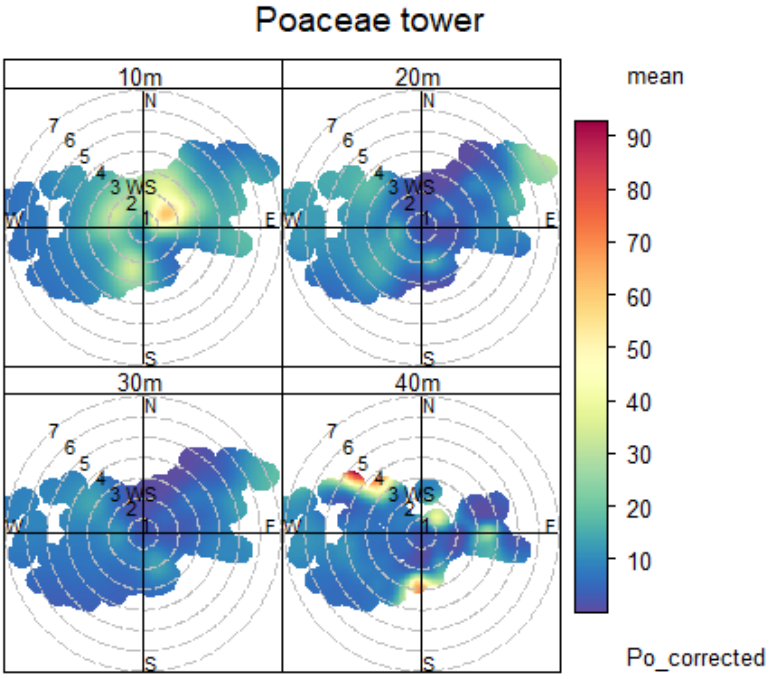


Figure 9.9: *Polar plots showing corrected* Poaceae *pollen proxy number concentration (colour) as a function of wind speed and direction for each sensor height during the tower period.*

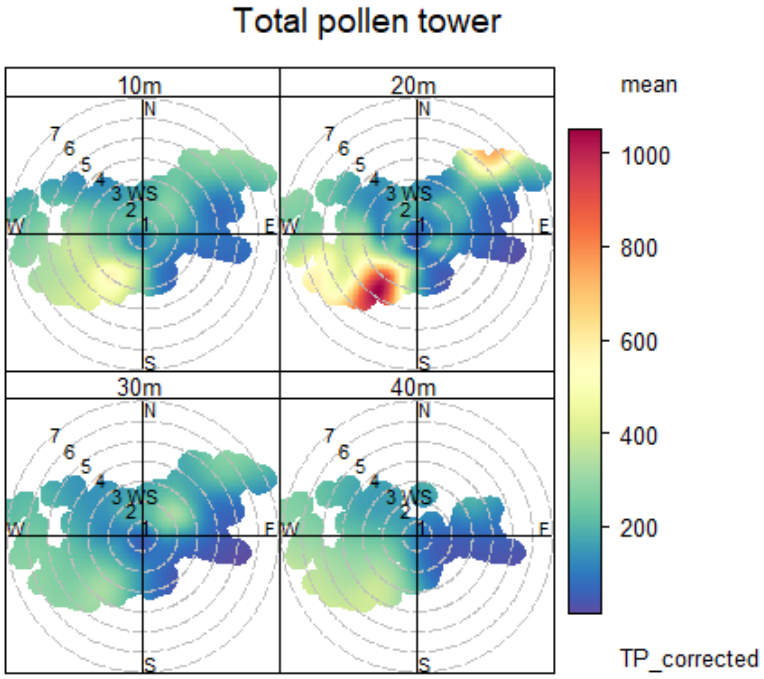


Figure 9.10: *Polar plots showing corrected total pollen proxy number concentration (colour) as a function of wind speed and direction for each sensor height during the tower period.*

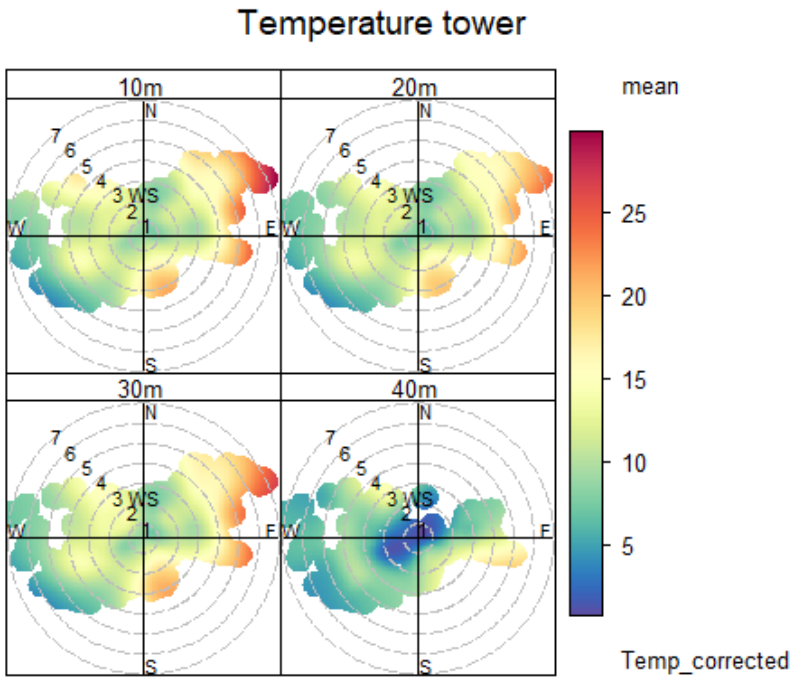


Figure 9.11: *Polar plots showing corrected temperature in °C (colour) as a function of wind speed and direction for each sensor height during the tower period.*

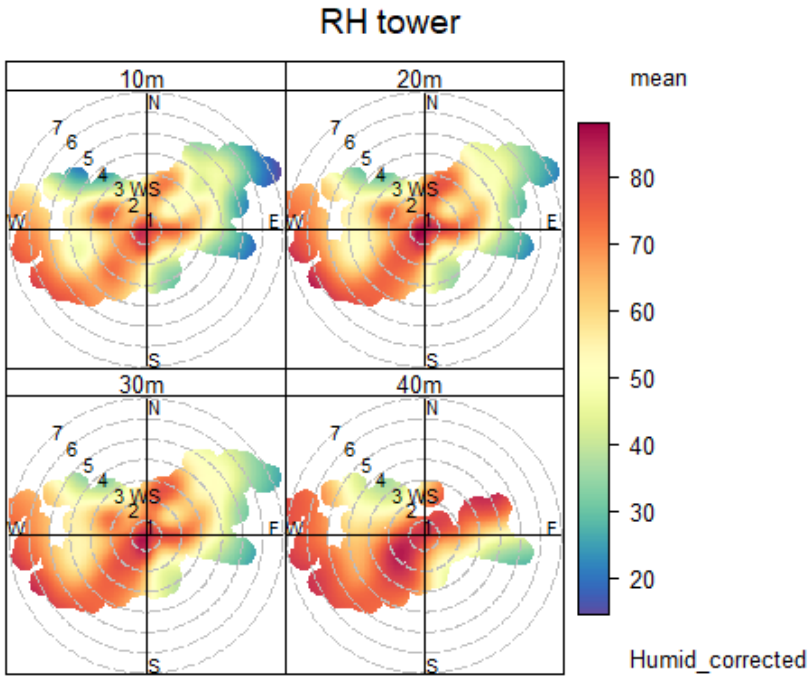


Figure 9.12: *Polar plots showing corrected relative humidity in % (colour scale) as a function of wind speed and direction for each sensor height during the tower period.*

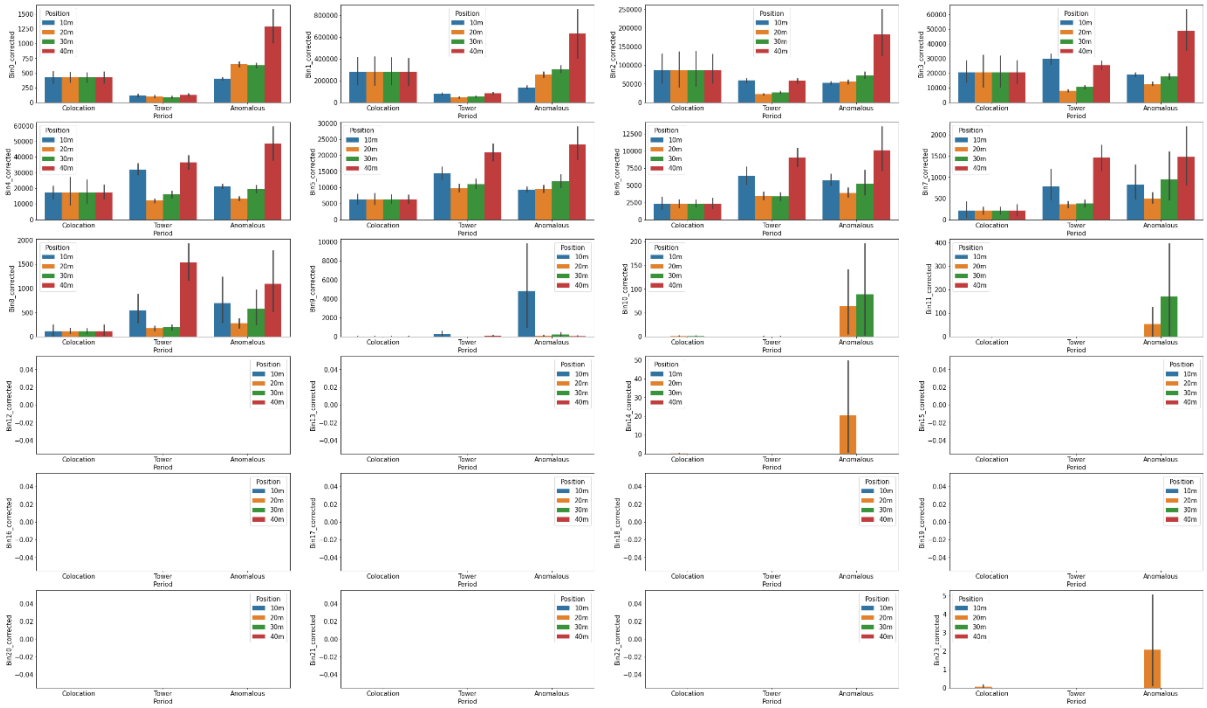
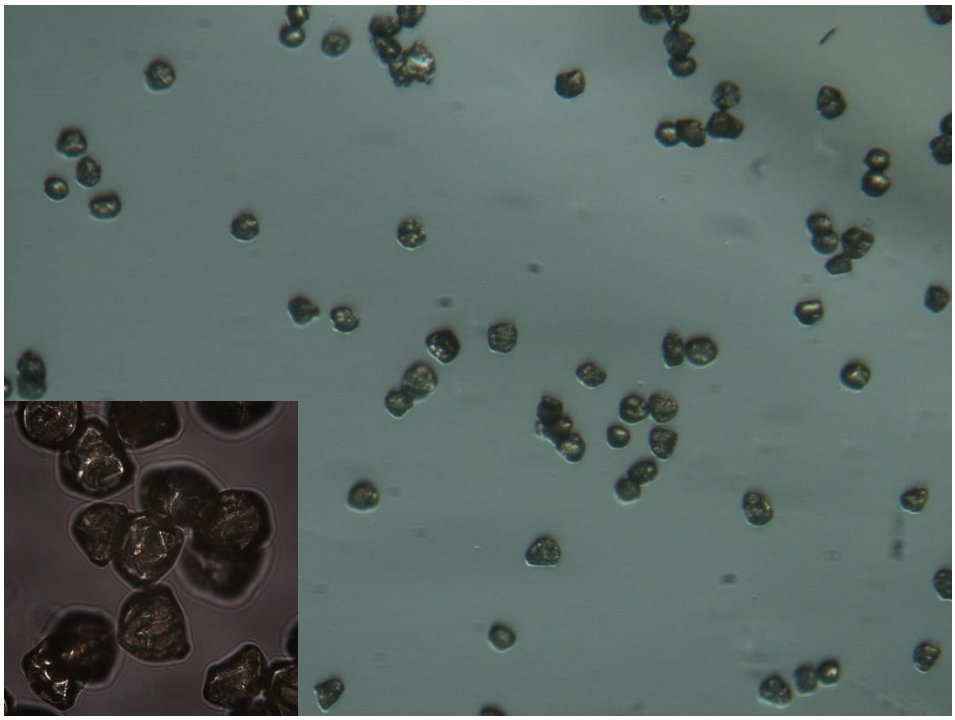


Figure 9.13: *Bar plots of mean values for each OPC bin number concentrations (corrected) with 95% confidence intervals for each time period (x-axis) and sensor (colour) respectively. Blue = 10m; orange = 20m; green = 30m; red = 40m. The 'colocation' and 'tower' periods are as described previously, while the 'anomalous' period included specifically dates from 11th April when the 10m sensor started reporting distinctively higher PM_{10} concentrations than other sensors. This figure demonstrates the anomalously high concentrations (with confidence intervals between 1000-10,000 m^3) observed specifically for bin 9 during this period compared to adjacent bins of similar particle size.*

10. Supporting information for Chapter 5



s

Figure 10.1*: Microscope images of pollen grains (above:* Lilium orientalis*; below:* Populus deltoides*) from the samples used in this work.*



Figure 10.2: *Photograph of the acoustic levitator chamber and macroscope used in this study.*

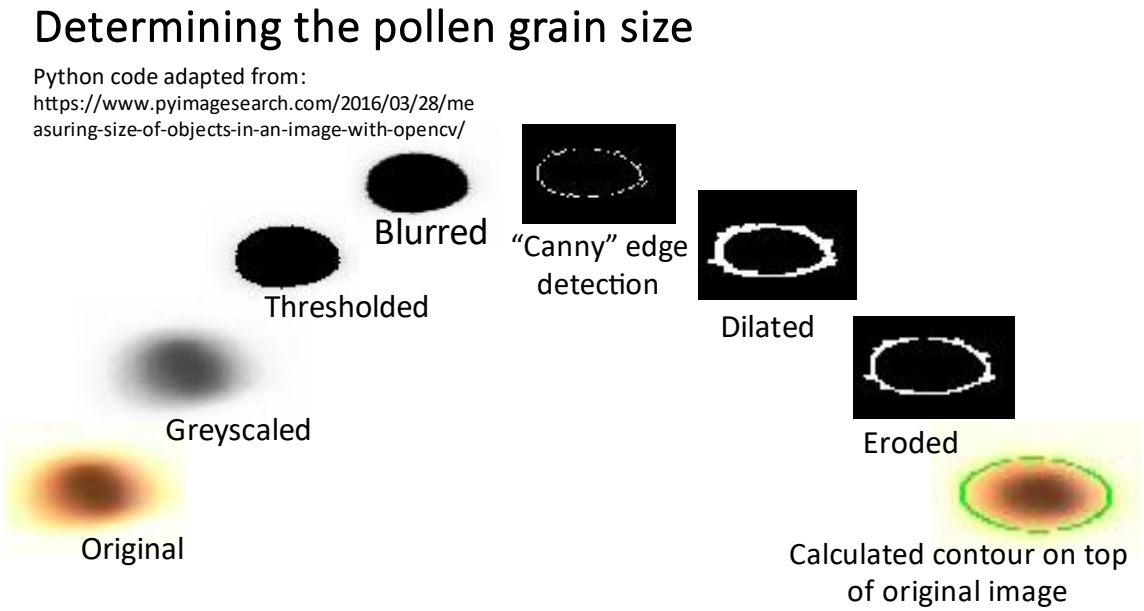


Figure 10.3: *Methodology of the python program designed to draw contours around the pollen grains in each image.*

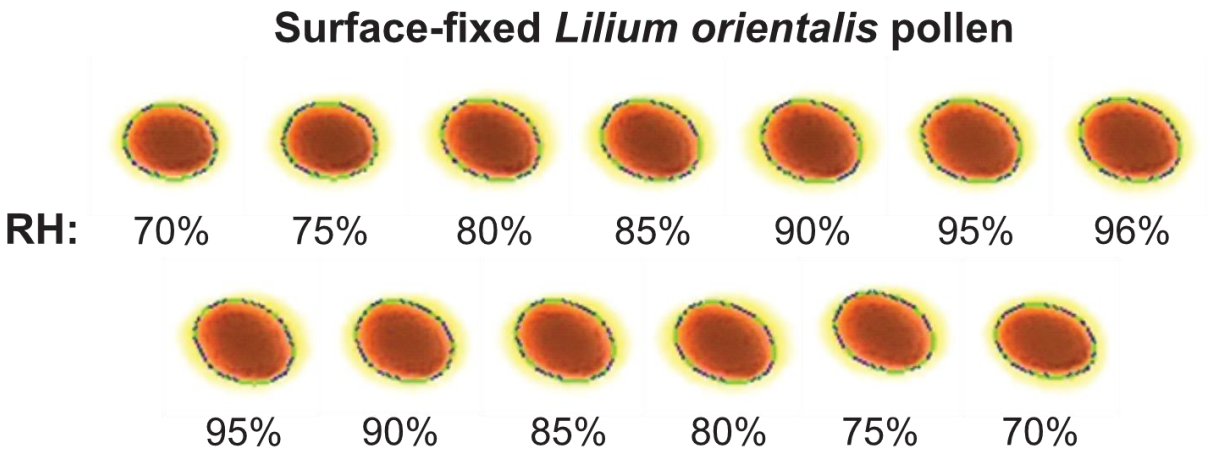


Figure 10.4: *Example snapshots of surface-fixed* Lilium orientalis *pollen grains undergoing increasing and decreasing relative humidity changes between 70 and 96%. Blue pixels denote the contour which the program outlined and the area was determined from while the green visualises the ellipse fitted in each case and from which 'polar' and 'equatorial' diameters were determined. This demonstrates the consistency of the pollen grain silhouettes in the surface-fixed state.*

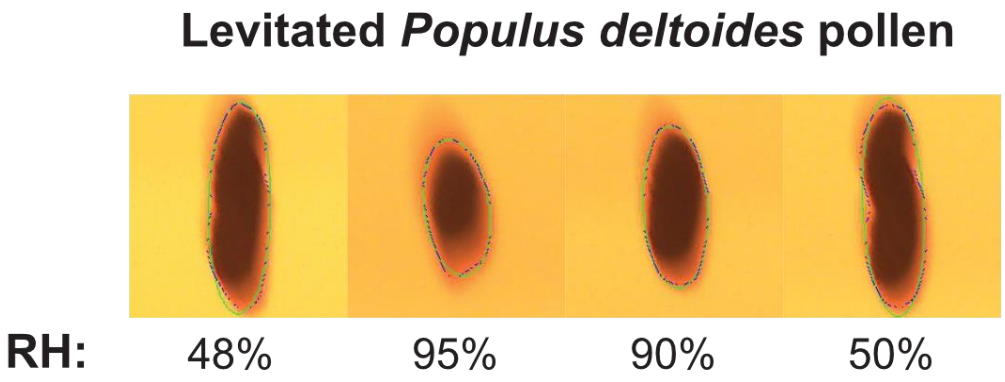


Figure 10.5: *Example snapshots of levitated* Populus deltoides *pollen grains undergoing increasing and decreasing relative humidity changes between 48 and 95%. Blue pixels denote the contour which the program outlined and the area was determined from while the green visualises the ellipse fitted in each case and from which 'polar' and 'equatorial' diameters were determined. This demonstrates the large variance in silhouette observed in the levitated state due to the instability and frequent movement of the pollen grains.*