

|  |
| --- |
| School of Computer Science  Faculty of Engineering AND PHYSICAL SCIENCES |

“Gair Wood as a Living Laboratory”

Hiral Rameshkumar Mahida

Submitted in accordance with the requirements for the degree of  
MSc Advanced Computer Science(Data analytics)

**2024/2025**

The candidate confirms that the following have been submitted*:*

|  |  |  |
| --- | --- | --- |
| **Items** | **Format** | **Recipient(s) and Date** |
| *Deliverables 1,* | *Report* | *SSO (xx/xx/xx)* |
| *Deliverable 2* | *Software codes or URL* | *Supervisor, assessor (xx/xx/xx)* |
| *Deliverable 3* | *PPT* | *Client, supervisor (xx/xx/xx)* |

Type of Project: Empirical Investigation

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student)

© 2025 The University of Leeds and Hiral Rameshkumar Mahida

# Summary

*This MSc dissertation project, titled “Gair Wood Representation,” focuses on analysing environmental data collected from 23 soil probes installed in Gair Wood, University of Leeds. The primary aim is to assess soil and habitat conditions over time by integrating and analysing soil temperature and moisture data to support future ecological planning.*

*The dataset consists of probe-wise measurements such as soil temperature, soil moisture, timestamps, and probe metadata (including type of soil and depth), recorded over a continuous period. All data files provided by the university were cleaned and merged using Python (pandas, NumPy) and Excel. Exploratory Data Analysis (EDA) was conducted to uncover trends and anomalies, and interactive visualizations were developed using Power BI for real-time monitoring and stakeholder engagement.*

*To further understand environmental patterns, linear regression was applied for temperature prediction, and k-means clustering was used to classify probe behaviours across locations and soil types. No external datasets were used. Challenges primarily involved handling inconsistent formats, missing values, and aligning time-series across different files.*

*Key findings revealed significant variation in soil moisture across depth levels, and a relationship between weather temperature and soil temperature trends. The Power BI dashboard enables users to filter and compare probes over time, visualize habitat differences, and potentially use the insights for rewilding and conservation efforts.*

*This empirical investigation demonstrates the potential of data science in environmental monitoring and sets the foundation for future integration of weather prediction models and aerial imagery for comprehensive habitat analysis.*

# Acknowledgements

*I would like to express my sincere gratitude to my academic supervisor,* ***Dr. Haiko Müller****, for his continuous support and guidance throughout this project. His weekly meetings, clear explanations, and insightful feedback have been instrumental in helping me stay focused and overcome challenges. I am truly grateful for his encouragement and direction.*

*I would also like to thank* ***Dr****.* ***Cat Scott and Dr. Thomas Sloan****, my external supervisor from the Environmental Science department, for providing access to the Gair Wood dataset and offering detailed insights into the probes, soil conditions, and overall objectives of the project. their time, expertise, and willingness to clarify all my queries were invaluable in shaping this work.*

*My sincere appreciation goes to my assessor* ***Dr.******Bogdan Alecu****, for his helpful advice and constructive feedback during the course of this project.*

*Lastly, I am deeply thankful to my family and friends for their unwavering support, patience, and motivation throughout my MSc journey. Their encouragement has meant a lot to me during this academic endeavour.*

# List of Figures

[*Figure 4.1a: Console output showing detection and removal of outliers (T > 60 °C) and confirmation of no missing key values.* 24](#_Toc207098869)

[*Figure 4.1b: DataFrame.info() summary confirming column types and counts.* 25](#_Toc207098870)

[*Figure 4.1c: First rows and descriptive statistics after cleaning.* 25](#_Toc207098871)

[*Figure 4.1d: Seasonal record counts and final dataset shape (1092748 × 18).* 26](#_Toc207098872)

[*Figure 4.2b. Seasonal boxplots showing distributions of soil temperature (T1), across Winter–Autumn.* 28](#_Toc207098873)

[*Figure 4.2f. Monthly trends of soil (T1), surface (T2), and air (T3) temperatures with inverse soil moisture dynamics.* 31](#_Toc207098874)

[*Figure 4.2g. Pearson correlation heatmap of soil, surface, air temperatures and soil moisture.* 32](#_Toc207098875)

[*Figure 4.2h. Habitat-level comparison of average soil temperature (T1) and soil* *moisture across seasonal cycles.* 34](#_Toc207098876)

[Figure 4.3a. PCA of probe readings. Scatter of records in PC1–PC2 coloured by T1 (soil T), revealing a strong thermal gradient. 36](#_Toc207098877)

[*Figure 4.3b. PCA variance explained. Console output showing [0.8293, 0.133] for PC1–PC2 (cumulative ≈ 0.962).* 36](#_Toc207098878)

[*Figure 4.3c. Probe clusters by average conditions (k=3). Scatter of Avg T1 vs Avg* *T2 with cluster colours*. 38](#_Toc207098879)

[*Figure 4.3d. Clustering diagnostics and assignments. Console summary showing silhouette ≈ 0.471 and the first 10 probe IDs with cluster labels.* 39](#_Toc207098880)

# Table Of Contents

[Summary iii](#_Toc207099579)

[Acknowledgements iv](#_Toc207099580)

[List of Figures v](#_Toc207099581)

[Table of Contents Error! Bookmark not defined.](#_Toc207099582)

[Chapter 1: Introduction 1](#_Toc207099583)

[1.1 Project Aim 1](#_Toc207099584)

[1.2 Objectives 1](#_Toc207099585)

[1.3 Deliverables 2](#_Toc207099586)

[1.4 Ethical, legal, and social issues 2](#_Toc207099587)

[Chapter 2: Background Research 3](#_Toc207099588)

[2.1 Literature Survey 3](#_Toc207099589)

[2.1.1 Introduction to the Gair Wood Project 3](#_Toc207099590)

[2.1.2 Environmental Monitoring in Woodland Ecosystems 3](#_Toc207099591)

[2.1.3 Importance of Soil Conditions 4](#_Toc207099592)

[2.1.4 Sensor Technologies and Data Collection 5](#_Toc207099593)

[2.1.5 Data Analytics and Visualization in Environmental Science 6](#_Toc207099594)

[2.2 Methods and Techniques 7](#_Toc207099595)

[2.2.1 Data Cleaning and Preprocessing 8](#_Toc207099596)

[2.2.2 Statistical and Machine Learning Approaches 8](#_Toc207099597)

[2.2.3 Data Visualization and Communication 9](#_Toc207099598)

[2.2.4 Advantages and Challenges 9](#_Toc207099599)

[2.2.5 Tools and Platforms 9](#_Toc207099600)

[2.3 Choice of Methods 10](#_Toc207099601)

[2.3.1 Data Cleaning and Preprocessing 10](#_Toc207099602)

[2.3.2 Regression Modelling 10](#_Toc207099603)

[2.3.3 Clustering 11](#_Toc207099604)

[2.3.4 Data Visualization 11](#_Toc207099605)

[2.3.5 Rationale for Exclusion of Other Techniques 11](#_Toc207099606)

[2.3.6 Overall Justification 12](#_Toc207099607)

[2.4 Justification of Method Choice 12](#_Toc207099608)

[2.4.1 Data Cleaning and Preprocessing 12](#_Toc207099609)

[2.4.2 Predictive Modelling with Regression 12](#_Toc207099610)

[2.4.3 Clustering for Ecological Segmentation 12](#_Toc207099611)

[2.4.4 Principal Component Analysis (PCA) as a Diagnostic Tool 13](#_Toc207099612)

[2.4.5 Visualisation and Communication of Results 13](#_Toc207099613)

[2.4.6 Tool Selection 13](#_Toc207099614)

[2.4.7 Limitations and Mitigations 14](#_Toc207099615)

[Chapter 3: Datasets and Experimental Design 15](#_Toc207099616)

[3.1 Dataset Description 15](#_Toc207099617)

[3.1.1 Master Records 15](#_Toc207099618)

[3.1.2 Probes Data 15](#_Toc207099619)

[3.1.3 Cleaned Probe Dataset 16](#_Toc207099620)

[3.1.4 Soil Type Mapping Table 17](#_Toc207099621)

[3.2 Data Cleaning and Preprocessing 17](#_Toc207099622)

[3.2.1 Initial Structuring (Excel) 17](#_Toc207099623)

[3.2.2 Automated Preprocessing (Python) 17](#_Toc207099624)

[3.2.3 Validation of Cleaning 18](#_Toc207099625)

[3.3 Experimental Setup 18](#_Toc207099626)

[3.4 Experimental Design 19](#_Toc207099627)

[3.4.1 Project Workflow 19](#_Toc207099628)

[3.4.2 Project Timeline 20](#_Toc207099629)

[3.4.2.1 Initial Plan: 20](#_Toc207099630)

[3.4.2.2 Actual Execution: 21](#_Toc207099631)

[Chapter 4: Results Of Empirical Investigation 23](#_Toc207099632)

[4.1 Data Cleaning and Integration Results 23](#_Toc207099633)

[4.1.1 Data ingestion and harmonisation 23](#_Toc207099634)

[4.1.2 Quality-control filtering 23](#_Toc207099635)

[4.1.3 Feature engineering 24](#_Toc207099636)

[4.1.4 Dataset profile 24](#_Toc207099637)

[4.1.5 Seasonal coverage 26](#_Toc207099638)

[4.1.6 Final export 26](#_Toc207099639)

[4.2 Exploratory Data Analysis (EDA) Results 26](#_Toc207099640)

[4.2.1 Temperature Trends Over Time 26](#_Toc207099641)

[4.2.2 Seasonal Variation 27](#_Toc207099642)

[4.2.3 Temporal Trends 30](#_Toc207099643)

[4.2.4 Correlation Analysis 31](#_Toc207099644)

[4.2.5 Habitat-Level Insights 33](#_Toc207099645)

[4.2.6 Summary of EDA Findings 34](#_Toc207099646)

[4.3 Dimensionality Reduction and Clustering 35](#_Toc207099647)

[4.3.1 Principal Component Analysis (PCA) 35](#_Toc207099648)

[4.3.1.1 Variance explained. 35](#_Toc207099649)

[4.3.1.2 Structure of the components. 37](#_Toc207099650)

[4.3.1.3Interpretation 37](#_Toc207099651)

[4.3.2 K‑Means clustering on the PCA space 37](#_Toc207099652)

[4.3.2.1 Choosing k. 37](#_Toc207099653)

[4.3.2.2 Cluster profiles (qualitative). 37](#_Toc207099654)

[2.3.2.3 Probe‑level view. 38](#_Toc207099655)

[4.4 Predictive Modelling Results 39](#_Toc207099656)

[4.4.1 Model performance 39](#_Toc207099657)

[4.4.2 Interpretation of coefficients 40](#_Toc207099658)

[4.4.3 Residual analysis 40](#_Toc207099659)

[4.4.4 Implications 40](#_Toc207099660)

[4.5 Power BI Dashboard Results 41](#_Toc207099661)

[Chapter 5: Validation and Evaluation of Results 48](#_Toc207099662)

[5.1 Introduction 48](#_Toc207099663)

[5.1.1 Validation framework. 48](#_Toc207099664)

[5.1.2 Evaluation criteria and evidence 48](#_Toc207099665)

[5.2 Validation of Data Cleaning 49](#_Toc207099666)

[5.2.1 Outlier removal 49](#_Toc207099667)

[5.2.2 Missing data checks 49](#_Toc207099668)

[5.3 Validation of Exploratory Analysis 50](#_Toc207099669)

[5.3.1 Seasonal patterns 50](#_Toc207099670)

[5.3.2 Monthly trends 50](#_Toc207099671)

[5.4 Validation of Clustering 51](#_Toc207099672)

[5.4.1 Statistical validation. 51](#_Toc207099673)

[5.4.2 Stability analysis. 51](#_Toc207099674)

[5.4.3 Ecological validation. 51](#_Toc207099675)

[5.4.4 Cross-validation with metadata. 52](#_Toc207099676)

[5.5 Validation of Predictive Modelling 52](#_Toc207099677)

[5.5.1 Performance metrics. 52](#_Toc207099678)

[5.5.2 Residual analysis. 52](#_Toc207099679)

[5.5.3 Baseline comparison. 52](#_Toc207099680)

[5.5.4 Ecological plausibility. 53](#_Toc207099681)

[5.5.5 several limitations 53](#_Toc207099682)

[5.6 Limitations and Sources of Error 53](#_Toc207099683)

[5.7 Summary 54](#_Toc207099684)

[Chapter 6: Conclusions and Future Work 56](#_Toc207099685)

[6.1 Conclusions 56](#_Toc207099686)

[6.2 Future Work 57](#_Toc207099687)

[6.3 Final Reflection 57](#_Toc207099688)

[List of References 59](#_Toc207099689)

[Appendix A External Materials 60](#_Toc207099690)

# Chapter 1: Introduction

## 1.1 Project Aim

The aim of this project is to develop a comprehensive data analytics solution that captures, integrates, cleans, and analyses environmental sensor data collected from soil probes installed across the Gair Wood site at the University of Leeds. These probes continuously monitor key environmental parameters such as soil temperature and moisture under varying habitat and soil types. The goal is to transform this multi-source, temporal data into meaningful insights through data preprocessing, exploratory data analysis (EDA), statistical modelling, and unsupervised learning techniques (such as k-means clustering).

Furthermore, the project aims to design and implement an interactive Power BI dashboard that enables dynamic visualization of spatio-temporal patterns, sensor readings, soil conditions, and ecological variation across different locations within Gair Wood. This dashboard will serve as a tool for environmental scientists and ecologists to assess habitat health, observe changes over time, and support data-driven ecological management decisions.

Additionally, the project explores predictive modelling using linear regression and clustering to identify similarities or differences among probes, analysed relationships between soil and air temperatures, and estimate environmental trends. The integration of contextual features such as time, weather, soil type, and location aims to enrich the analysis and offer a holistic digital representation of the Gair Wood environment.

Ultimately, the project serves as an empirical investigation that combines environmental science and data analytics to contribute to sustainable habitat monitoring, research, and future planning at the University of Leeds.

## 1.2 Objectives

* Data Preparation: Clean, standardize, and merge the soil probe datasets using Python and Excel to create a unified, analysis-ready dataset.
* Exploratory Analysis: Perform EDA to identify trends, variations, and anomalies in soil temperature and moisture across time, depth, and probe locations.
* Dashboard Development: Create an interactive Power BI dashboard that enables filtering and visualization of probe data based on parameters such as location, soil type, and time period.
* Modelling and Analysis: Apply linear regression for temperature trend prediction and k-means clustering to group probes with similar environmental behaviour.
* Insight Evaluation: Validate analysis results, document findings, and provide actionable recommendations for ecological monitoring and dashboard feature improvements.

## 1.3 Deliverables

1. Cleaned and Merged Dataset: A unified, pre-processed dataset combining all 23 probe files, cleaned using Python and Excel for consistency and readiness for analysis.
2. Exploratory Data Analysis Report: Visual and statistical summaries identifying key trends, anomalies, and patterns in soil temperature and moisture across probes and over time.
3. Power BI Dashboard: An interactive and user-friendly dashboard enabling filtering by probe ID, soil type, depth, and date, with embedded visualizations and data insights.
4. Predictive and Clustering Models: Implementation of linear regression for soil temperature prediction and k-means clustering to group probes based on environmental similarity.
5. Final Project Report: A comprehensive dissertation including methodology, results, discussion, validation, and future work, submitted in accordance with University of Leeds guidelines.

## 1.4 Ethical, legal, and social issues

This project uses environmental data provided by the University of Leeds, containing no personal or sensitive information, and thus poses minimal ethical risk. All data was used with permission and solely for academic purposes, ensuring compliance with UK GDPR and university policies.

No human participants were involved, so no ethical approval or consent forms were required. The analysis and dashboard aim to support sustainable land management and have potential social value for environmental monitoring and planning. All external tools and methods used are properly credited.

# Chapter 2: Background Research

## 2.1 Literature Survey

### 2.1.1 Introduction to the Gair Wood Project

Gair Wood is the University of Leeds research woodland—conceived as a “Living Laboratory” to study the real‑time impacts of large‑scale tree planting as a nature‑based solution to climate change. It enables cross‑disciplinary teams to observe how a newly created woodland evolves, with active monitoring of soil composition, biodiversity, and local air quality, alongside baseline assessments completed in 2022. [leaf.leeds.ac.uk](https://leaf.leeds.ac.uk/gairwood/?utm_source=chatgpt.com)

Situated in north Leeds near Eccup Reservoir, the site spans ~36 hectares and was planted during winter 2023 (Jan–Mar). The design intentionally creates a mosaic of habitats combining new woodland blocks with open grass and scrub to support biodiversity and enable controlled comparisons across planting densities and species mixes. [United Bank of Carbon](https://www.uboc.co.uk/designing-a-new-woodland-for-leeds/?utm_source=chatgpt.com)

In total, over 60,000 native broadleaved trees (approximately 66,000 by project delivery counts) have been established—predominantly oak, hazel and willow selected for regional suitability and climate resilience. The woodland forms part of the White Rose Forest programme and contributes to the University’s Climate Plan goals, while also providing long‑term teaching, research and community engagement opportunities. [leaf.leeds.ac.uk](https://leaf.leeds.ac.uk/gairwood/faq/?utm_source=chatgpt.com)[the-rsc.co.uk](https://www.the-rsc.co.uk/news/we-planted-a-new-woodland?utm_source=chatgpt.com)

The project’s open‑air experimental design, strong partner network (e.g., LEAF Centre, United Bank of Carbon, White Rose Forest), and early installation of environmental sensors make it a robust platform for data‑driven studies—particularly on soil temperature and moisture dynamics, which are central to this dissertation.

### 2.1.2 Environmental Monitoring in Woodland Ecosystems

Environmental monitoring is fundamental to understanding and managing woodland ecosystems. It involves the systematic collection of data on physical, chemical, and biological parameters to detect changes, assess ecological health, and inform sustainable management practices. In woodland environments, monitoring provides a means to track how tree planting, soil conditions, biodiversity, and climate interactions evolve over time, thereby enabling evidence-based interventions for conservation and restoration (Leeds LEAF, 2025).

Large-scale reforestation and afforestation projects, such as the White Rose Forest and England’s Community Forests, have demonstrated how robust monitoring frameworks contribute to ecological resilience. These initiatives rely on long-term environmental datasets to evaluate the effectiveness of woodland creation in improving biodiversity, carbon sequestration, and local climate regulation (England’s Community Forests, 2025; White Rose Forest, 2025). Continuous monitoring also provides critical feedback loops, allowing project stakeholders to adapt management strategies based on real-time insights.

Within newly established woodland ecosystems like Gair Wood, environmental monitoring is particularly important because the site is in an early successional stage. Parameters such as soil temperature, moisture, and microclimatic variability must be tracked to evaluate the success of woodland establishment and to identify potential challenges such as soil compaction, waterlogging, or drought conditions. Monitoring in this context not only supports the University of Leeds research goals but also contributes to the broader scientific understanding of how woodland habitats develop under changing climate conditions (Biogeosciences, 2025).

Advances in monitoring technologies have further strengthened woodland research. The use of high-resolution sensor networks, satellite-based earth observation systems, and field-based ecological surveys ensures that datasets are both comprehensive and scalable (Copernicus, 2025). Such integration enables multi-scalar insights, where local soil-level measurements can be contextualised within broader climatic or regional ecological patterns. However, challenges remain in ensuring data quality, calibration consistency, and long-term sensor maintenance (EcoEvoRxiv, 2025).

Overall, environmental monitoring in woodland ecosystems serves as a cornerstone of modern ecological management. By combining continuous data collection with analytical and visualization tools, projects like Gair Wood are positioned to provide actionable insights for biodiversity enhancement, carbon neutrality, and community engagement within the UK’s expanding woodland networks.

### 2.1.3 Importance of Soil Conditions

Soil conditions are a fundamental determinant of woodland ecosystem health. The physical and chemical properties of soil such as temperature, moisture, structure, and nutrient content directly influence vegetation growth, microbial activity, and biodiversity. In woodland restoration projects, maintaining optimal soil conditions is critical for supporting seedling establishment, root development, and long-term woodland productivity (Brady & Weil, 2016).

Soil temperature regulates biological processes such as root respiration, seed germination, and microbial decomposition. Warmer soils generally accelerate biological activity, whereas colder or waterlogged soils can restrict plant growth and nutrient cycling. Seasonal fluctuations in soil temperature also determine the activity cycles of many woodland species and affect carbon and nitrogen dynamics within the ecosystem (Biogeosciences, 2025).

Soil moisture plays an equally vital role in ecosystem resilience. Adequate moisture availability supports tree establishment, enhances drought resistance, and maintains soil structure. Conversely, prolonged dryness can reduce seedling survival rates, while excessive waterlogging can lead to anaerobic conditions and root damage (Springer, 2025). The balance between soil temperature and moisture is particularly important in young woodlands, where vegetation cover is not yet well established, leaving soils more exposed to climatic variability.

In the context of Gair Wood, soil monitoring provides valuable insights into the success of woodland regeneration. For example, by comparing soil conditions across different habitat types and planting densities, researchers can identify which soil environments are most conducive to woodland establishment. This evidence base is essential for guiding adaptive management practices not only at Gair Wood but also in similar reforestation initiatives across the UK (Leeds LEAF, 2025).

Furthermore, soil conditions are closely linked to biodiversity outcomes. Healthy soils with balanced temperature and moisture levels promote diverse plant communities, which in turn support a wide range of insects, birds, and mammals. Monitoring soil dynamics therefore provides an indirect measure of ecological success, helping to ensure that restoration goals extend beyond tree planting to encompass entire ecosystems (England’s Community Forests, 2025).

In summary, soil conditions form the ecological foundation upon which woodland restoration projects are built. By carefully tracking soil temperature and moisture, projects like Gair Wood can optimise habitat creation, improve resilience to climate change, and ensure sustainable woodland development over time.

### 2.1.4 Sensor Technologies and Data Collection

The advancement of sensor technologies has transformed environmental monitoring, particularly in woodland ecosystems where fine-scale data is essential to understanding dynamic processes. Traditional methods such as manual soil sampling and periodic field surveys, while valuable, are limited in temporal resolution and often fail to capture rapid environmental fluctuations. By contrast, modern IoT-enabled sensors and data loggers enable continuous, high-frequency data collection, offering deeper insights into soil and habitat conditions (Copernicus, 2025).

In projects like Gair Wood, temperature and moisture probes provide critical measurements that allow researchers to monitor soil microclimates across diverse habitat types. These probes record variables such as soil temperature, surface temperature, air temperature, and soil moisture content at regular intervals, producing detailed time-series datasets. The high temporal and spatial resolution of such data facilitates analyses of seasonal variability, soil-water interactions, and the influence of microhabitat diversity on woodland regeneration (EcoEvoRxiv, 2025).

The deployment of multiple probes across woodland sites enables comparative analysis between areas with different planting densities, soil types, and exposure levels. For instance, by tracking how soil moisture varies between open grassland patches and newly planted woodland blocks, researchers can better understand how vegetation cover alters local hydrological cycles. This contributes not only to ecological research but also to practical woodland management strategies (UBOC, 2025).

Despite their benefits, sensor networks present several technical and logistical challenges. Sensors require regular calibration, maintenance, and power management to ensure accuracy over long monitoring periods. Harsh environmental conditions, such as heavy rainfall, frost, or soil compaction, can cause data interruptions or sensor malfunction. Additionally, managing the large datasets generated by sensor networks requires robust data cleaning, integration, and storage solutions (Bio geosciences, 2025).

The integration of sensor data with remote sensing technologies, such as satellite imagery and aerial drone surveys, further enhances woodland monitoring by scaling up from local probe measurements to landscape-level assessments. This combined approach provides a holistic view of environmental conditions, linking ground-based measurements with broader climatic and ecological patterns (Springer, 2025).

In summary, the adoption of modern sensor technologies in woodland ecosystems provides an essential foundation for high-quality, data-driven environmental research. At Gair Wood, these tools enable the University of Leeds and its partners to build a detailed evidence base for ecological monitoring, ensuring that woodland establishment can be evaluated and managed with scientific precision.

### 2.1.5 Data Analytics and Visualization in Environmental Science

The increasing availability of high-resolution environmental data from sensor networks has highlighted the importance of data analytics and visualization in ecological research. Raw sensor data, while valuable, is often vast, noisy, and complex. Without systematic analysis and clear visualization, it is difficult for researchers, policymakers, and community stakeholders to extract meaningful insights (UBOC, 2025).

Data analytics techniques such as Exploratory Data Analysis (EDA), provides the foundation for interpreting complex datasets. In the case of Gair Wood, Python-based cleaning and transformation routines standardised sensor outputs, removed anomalies (e.g., spurious temperatures above 60 °C), and derived new temporal features such as week, month, and meteorological season. Descriptive summaries, alongside season × habitat cross-tabulations, helped reveal how environmental conditions vary across time and woodland microhabitats.

Advanced analytical techniques then built on this foundation. Principal Component Analysis (PCA) reduced multicollinearity between temperature and moisture readings, showing that the majority of variance could be explained by a small number of composite factors. Linear regression modelling successfully predicted soil temperature (T1) from surface temperature (T2) from air temperature (T3) from soil moisture, and season. The model achieved strong accuracy (R² ≈ 0.83), highlighting both the predictive power and interpretability of regression for environmental monitoring. Complementing this, K-means clustering grouped probes into three distinct ecological clusters, providing an unsupervised view of differences in soil climate conditions across the site.

Visualization played a crucial role in making these analyses accessible. Python’s matplotlib plots including PCA scatter plots, regression feature-importance charts, correlation heatmaps, and seasonal temperature curves provided detailed research-grade outputs. At the same time, Power BI dashboards transformed these findings into interactive, stakeholder-friendly visualizations. These dashboards allow for dynamic filtering by probe, season, and habitat type, enabling researchers, land managers, and the wider community to engage with the data in an intuitive way.

Importantly, the integration of local probe data with wider visualization frameworks offers a multi-scale perspective. Correlation heatmaps and regression coefficients explain fine-scale sensor dynamics, while dashboards and geospatial maps provide a broader overview of site-wide trends. This balance ensures that complex analytics remain interpretable and that results can inform both scientific understanding and practical woodland management.

In summary, the combination of Python-based EDA, statistical modelling, clustering, and interactive dashboards reflects best practice in environmental data science. For Gair Wood, these methods not only uncovered key patterns in soil temperature and moisture but also ensured that results were communicated effectively, supporting its role as a “Living Laboratory” for ecological monitoring and sustainable woodland management.

## 2.2 Methods and Techniques

The main methods and techniques available for analysing and interpreting environmental monitoring data, with a focus on their applicability to woodland ecosystems such as Gair Wood. The discussion covers data cleaning and preprocessing, statistical and machine learning methods, and visualization techniques, highlighting their strengths, limitations, and relevance to ecological research.

### 2.2.1 Data Cleaning and Preprocessing

Environmental sensor data is often characterised by noise, missing values, and outliers caused by equipment errors, harsh weather, or power interruptions. Data cleaning is therefore an essential first step. Typical approaches include:

* Standardisation of formats - ensuring consistency in timestamping, column names, and numeric formats.
* Outlier detection - applying domain-specific thresholds.
* Handling missing data - removing incomplete records for key variables or applying imputation methods if patterns of missingness are consistent.
* Feature engineering - deriving new temporal features (e.g., week, month, season) to support seasonal and habitat-based comparisons.

By carrying out these steps, the dataset becomes more robust and reliable for further analysis, reducing the likelihood of biased results.

### 2.2.2 Statistical and Machine Learning Approaches

After the cleaning stage, the dataset was explored using a combination of statistical techniques and modern machine learning approaches. The initial step involved Exploratory Data Analysis (EDA), which helped in summarising the overall distributions, identifying correlations between variables, and visualising seasonal trends. This provided a clearer picture of how the environment at Gair Wood changes over time and across different probe sites. To capture deeper patterns in the data, Principal Component Analysis (PCA) was applied, which reduced the complexity of multiple correlated variables such as soil, surface, and air temperature. This not only filtered out noise but also highlighted the dominant environmental gradients that structure the woodland habitat. Alongside this, regression modelling particularly, linear regression was used to establish relationships between predictors like surface and air temperature, soil moisture, and seasonal factors, and their effect on soil temperature. The value of regression lies in its simplicity and interpretability, making it especially relevant for ecological monitoring where clarity is essential. Finally, unsupervised methods such as K-means clustering were introduced to group probes with similar characteristics. This allowed for meaningful ecological segmentation, for example distinguishing between sites that are consistently wetter and cooler compared to those that are warmer and drier, thereby offering insights into micro-habitat variation within Gair Wood.

### 2.2.3 Data Visualization and Communication

Visualisation tools play a crucial role in transforming analytical outputs into meaningful and actionable insights. In this project, both traditional and modern approaches were employed to communicate findings effectively. Statistical plotting libraries in Python, such as Matplotlib, were used to generate scatter plots, correlation heatmaps, bar charts, and line graphs, each providing a clear view of trends, relationships, and seasonal cycles within the dataset. To complement these static visualisations, an interactive Power BI dashboard was developed, allowing users to dynamically filter results, explore trends over time, and compare conditions across different probes, seasons, and habitats. This level of interactivity made the results accessible not only to researchers but also to policymakers and local communities, who may have varying levels of technical expertise. Additionally, Geographic Information Systems (GIS) were incorporated to map probe locations and overlay environmental data with spatial layers such as land cover, elevation, and rainfall. By adding this spatial dimension, GIS provided valuable context, helping to connect site-level measurements with broader landscape and habitat characteristics..

### 2.2.4 Advantages and Challenges

The techniques outlined above each offer unique benefits. Cleaning and preprocessing ensure data reliability, while statistical models and clustering provide interpretable, evidence-based insights. Visualization tools enhance accessibility and stakeholder engagement. However, challenges include ensuring reproducibility, maintaining sensor calibration, handling very large datasets, and balancing sophisticated analysis with clear communication.

In ecological projects such as Gair Wood, a hybrid approach is often most effective: combining robust data cleaning, transparent regression models, and interpretable clustering, alongside interactive dashboards for dissemination. This ensures that both scientific rigour and practical usability are maintained.

### 2.2.5 Tools and Platforms

Several software tools were used to support the data analytics and visualization workflow, each selected for their suitability to different stages of the project:

* Microsoft Excel: Used in the initial phase to transform unstructured sensor CSV files into a structured format. Column names, timestamps, and value formats were standardised prior to importing into Python. Excel’s accessibility and ease of use made it an effective tool for early stage data inspection and formatting.
* Python (with PyCharm IDE): Python served as the primary analytical platform due to its extensive ecosystem of data science libraries. pandas and NumPy were employed for data cleaning and manipulation, scikit-learn supported machine learning tasks such as regression, PCA, and clustering, and matplotlib was used for static visualisations. Development was carried out in PyCharm, which offered versioned project management, debugging features, and integration with Python’s scientific stack, making the workflow efficient and reproducible.
* Microsoft Power BI: Selected as the main visualization and reporting tool. Power BI dashboards were developed to present results interactively, enabling filtering by probe, season, and habitat type. This allowed researchers and non-technical stakeholders alike to explore the dataset without programming expertise.

Together, these tools provided a comprehensive workflow: Excel for structuring, Python for cleaning and modelling, PyCharm for code management, and Power BI for interactive communication of findings.

## 2.3 Choice of Methods

The techniques selected for this project were chosen with the dual aim of ensuring scientific rigour and maintaining interpretability and usability for ecological stakeholders. Given the nature of the Gair Wood dataset (sensor-based soil, air, and surface temperature readings combined with soil moisture counts) and metadata(the chosen methods provided a balance between reliability, analytical depth, and clear communication).

### 2.3.1 Data Cleaning and Preprocessing

Robust data cleaning was prioritised to prepare the dataset for analysis. Temperature thresholds were applied (excluding values above 60 °C) to remove anomalies caused by sensor malfunction or extreme short-term heating events. Missing or incomplete key values (timestamps, temperature measures) were excluded, while derived temporal variables such as week, month, and season were introduced to support seasonal analyses. These steps were essential to establish a consistent, trustworthy dataset for subsequent modelling.

### 2.3.2 Regression Modelling

Linear regression was chosen as the primary predictive technique for modelling soil temperature (T1). Although more advanced machine learning approaches such as Random Forests or neural networks could have been explored, regression offered several advantages that aligned with the project’s objectives. Its greatest strength lies in interpretability, as the coefficients directly reveal the influence of predictors such as surface temperature, air temperature, soil moisture, and season. Initial tests demonstrated strong explanatory power (R² ≈ 0.83), suggesting that a relatively simple model was sufficient to capture the key soil–climate relationships present in the dataset. Moreover, regression is particularly well suited to continuous outcomes like temperature, making it a natural fit for this study. The decision to prioritise regression also reflects best practice in environmental monitoring, where transparent and easily validated models are often valued above more opaque “black-box” methods.

### 2.3.3 Clustering

To investigate ecological heterogeneity across probe locations, K-means clustering with *k = 3* was applied to probe averages. This unsupervised technique provided a straightforward and interpretable way of grouping probes into distinct microclimatic clusters, distinguishing, for example, cooler and wetter sites from warmer and drier ones. Such groupings offered valuable insights into how environmental conditions may be linked to woodland regeneration outcomes. The clustering solution was also supported by a strong silhouette score, which confirmed that the separation between groups was meaningful. As such, clustering complemented the regression analysis by offering a different perspective: while regression quantified specific relationships, clustering revealed broader patterns of site variability.

### Data Visualization

The project employed two complementary approaches to visualisation. Python’s Matplotlib library was used to produce static but detailed figures, including PCA scatter plots, correlation heatmaps, regression feature-importance charts, and seasonal temperature curves. These outputs were essential for supporting in-depth analysis and for illustrating specific findings in the dissertation. Alongside this, an interactive Power BI dashboard was developed to allow dynamic exploration of the dataset by probe, habitat type, and season. This choice was motivated by the need to make the results accessible to a wide audience: not only academics, but also conservation partners and local communities. By enabling non-specialist users to filter and compare data themselves, the dashboard acted as a bridge between technical analysis and practical decision-making.

### 2.3.5 Rationale for Exclusion of Other Techniques

Although advanced machine learning models such as Random Forests, Gradient Boosted Trees, or Long Short-Term Memory (LSTM) networks were considered, they were not prioritised in this study. The main reason was the emphasis on interpretability rather than maximising predictive accuracy. Given the relatively short monitoring period and single-site dataset, applying highly complex models would likely have produced results that were difficult to generalise and potentially misleading. Furthermore, these methods demand greater time and computational resources, which lay beyond the scope of this MSc project. For these reasons, the project focused instead on methods that balance analytical rigour with clarity and transparency.

### 2.3.6 Overall Justification

In summary, the chosen methods—systematic cleaning, linear regression, K-means clustering, and interactive visualization—provided a robust yet accessible framework. These methods ensured that the analysis was scientifically sound, reproducible, and aligned with both academic expectations and the practical needs of woodland conservation.

## 2.4 Justification of Method Choice

The rationale behind the methods, analytical techniques, and tools selected for the project. The choices were guided by the dual aim of achieving rigorous environmental data analysis while ensuring accessibility and clarity for non-technical stakeholders.

### 2.4.1 Data Cleaning and Preprocessing

Python was chosen for data cleaning because of its ability to handle large datasets efficiently. The combined dataset contained over one million records, making manual processing in spreadsheets impractical. Libraries such as pandas and NumPy enabled reproducible transformations, including timestamp parsing, duplicate removal, and outlier filtering (e.g., excluding unrealistic soil temperature values above 60 °C). Missing values in key variables were dropped to preserve data integrity, and additional seasonal features were engineered to reflect meteorological cycles. Python’s scripting ensured that the pipeline could be consistently reapplied if additional data were collected in the future.

### 2.4.2 Predictive Modelling with Regression

Linear regression was selected as the primary predictive technique for modelling soil temperature trends. While more complex machine learning methods (e.g., Random Forests, LSTM networks) could potentially improve accuracy, regression was prioritised because of its transparency and interpretability. With soil surface temperature, air temperature, and soil moisture as predictors, the model achieved an R² of approximately 0.83, outperforming a mean baseline. This balance between predictive performance and interpretability makes regression particularly suitable for ecological applications, where stakeholders often prioritise clear explanations over black-box accuracy.

### 2.4.3 Clustering for Ecological Segmentation

K-means clustering was applied to identify ecological groupings among probes. By aggregating average readings per probe, the algorithm segmented the dataset into three distinct clusters with a silhouette score of around 0.50. These clusters highlighted differences in soil temperature and moisture across habitat types, enabling a deeper understanding of micro-ecological conditions within Gair Wood. The decision to use K-means was driven by its simplicity, scalability, and ability to provide intuitive groupings for environmental interpretation.

### 2.4.4 Principal Component Analysis (PCA) as a Diagnostic Tool

PCA was used in a diagnostic capacity to explore the variance structure of soil and climate measurements. The first two components explained approximately 95% of total variance, indicating strong correlation between air, surface, and soil temperatures. PCA supported the regression design by confirming the presence of multicollinearity and providing visual insights into variable interactions. It was deliberately not used as the main modelling method, in order to preserve interpretability for ecological stakeholders.

### 2.4.5 Visualisation and Communication of Results

The project’s initial plan was to rely solely on Python for both modelling and visualisation. However, during the exploratory stage it became clear that static Python plots, while useful for academic reporting, were less effective for interactive exploration by non-technical audiences. Consequently, Power BI was introduced as a complementary tool. Power BI enabled dynamic filtering by probe, season, and habitat type, facilitating user-friendly exploration of trends and comparisons. This decision reflects the project’s evolution from a researcher-focused approach to one that emphasises accessibility and engagement, aligning with the University of Leeds vision of Gair Wood as a “Living Laboratory.”

### 2.4.6 Tool Selection

* Python vs. Alternatives: Python was selected over R and MATLAB due to its extensive ecosystem for machine learning, scalability for large datasets, and prior familiarity.
* PyCharm IDE: Used to manage Python scripts efficiently, offering version control integration, debugging, and project organisation, which improved reproducibility.
* Microsoft Excel: Used in the early phase for quick inspection and to transform unstructured sensor logs into structured tabular formats before bulk analysis. Its integration with Power BI also supported dashboard development.
* Power BI vs. Tableau: Power BI was chosen due to its seamless integration with Excel/CSV data, academic licensing availability, and lower learning curve. It enabled the creation of interactive dashboards suitable for a broad audience of researchers, conservationists, and students.

### 2.4.7 Limitations and Mitigations

As with any empirical study, this project faced a number of limitations, which were carefully considered and addressed where possible. One challenge was the presence of sensor artefacts, which were mitigated through threshold filtering and the removal of anomalous values that lay outside realistic environmental ranges. A further limitation was the restricted set of variables available, as external weather data such as rainfall and humidity could not be incorporated. This inevitably constrained the scope of the models, since such factors play an important role in shaping soil temperature and moisture dynamics. The generalisability of the findings was also limited by the dataset itself, which covers only a single monitoring period. Extending the analysis to multi-year data would provide a more robust foundation for detecting longer-term climatic and ecological trends.

From a methodological perspective, the simplicity of the linear regression model meant that it was unable to fully capture non-linear interactions between environmental variables. However, its interpretability was judged to be more valuable for the aims of this project, particularly given the importance of transparency in ecological monitoring. In summary, the methodological decisions were made to strike a balance between scientific rigour and accessibility. Python offered a powerful platform for data analysis, regression and clustering provided interpretable insights, PCA enhanced diagnostic understanding, and Power BI facilitated the communication of results to a diverse range of stakeholders. Together, these choices ensured that the project contributes meaningfully both to academic knowledge and to the practical monitoring of ecological change at Gair Wood.

# Chapter 3: Datasets and Experimental Design

## 3.1 Dataset Description

The dataset for this study was obtained directly from the Environmental Science Department at the University of Leeds as part of the Gair Wood rewilding project. It consisted of multiple CSV files, each corresponding to an individual soil probe installed across the woodland, and a master Excel file containing metadata such as probe ID, soil type, location, and installation details.

Each probe captured continuous environmental measurements from different layers of the soil system and its immediate surroundings. The recorded parameters included:

* T1 (Soil Temperature): measured at root-zone depth.
* T2 (Surface Temperature): representing soil surface fluctuations.
* T3 (Air Temperature): providing local atmospheric conditions.
* Soil Moisture Count: reflecting volumetric water content.

Collectively, these readings formed a high-resolution time series dataset representing the microclimatic and soil conditions of Gair Wood.

### 3.1.1 Master Records

The master Excel file acted as a reference for interpreting probe measurements. It included probe metadata such as ID, installation date, latitude, longitude, elevation, and habitat classification. This information was essential for linking probe readings with ecological context.

The linkage was performed using Probe ID as the primary key, enabling each probe’s CSV data to be enriched with spatial and habitat metadata. For example, Probe 95132709 could be linked with its location in a low-density planting habitat at an elevation of 146.9 m, allowing for ecological comparisons with probes installed in different habitats or elevations.

### 3.1.2 Probes Data

The probe CSV files contained timestamped measurements of environmental variables. After cleaning and integration with the master metadata, the unified dataset included:

* Datetime (UTC): precise measurement timestamp.
* Probe ID: unique identifier used to join probe readings with metadata.
* T1, T2, T3: soil, surface, and air temperature.
* Soil Moisture Count: raw moisture sensor readings.
* Derived Features: week, month, and meteorological season.
* Spatial Attributes: latitude, longitude, elevation, and habitat type (from master file).

Initially, the files presented inconsistencies such as varying column names, irregular timestamp formats, and outlier readings. These challenges necessitated a systematic data cleaning and preprocessing.

### 3.1.3 Cleaned Probe Dataset

The cleaned dataset contains high-frequency time-series measurements enriched with spatial and ecological metadata.

| **Column Name** | **Description** | **Data Type** |
| --- | --- | --- |
| index number of the measurement | Unique sequential index for measurement | float64 |
| datetime\_utc | Timestamp of the measurement in UTC | Date/time |
| time zone | Time zone offset indicator | int64 |
| T1 (soil T) | Soil temperature at root-zone depth (°C) | float64 |
| T2 (surface T) | Surface temperature at soil surface (°C) | float64 |
| T3 (air T) | Air temperature above ground (°C) | float64 |
| soil moisture count | Soil volumetric moisture sensor reading | float64 |
| shake | Sensor stability/diagnostic flag | int64 |
| errFlag | Error flag indicating measurement validity | int64 |
| probe\_id | Unique identifier for each probe | int64 |
| week | Week number derived from timestamp | int64 |
| month | Month number derived from timestamp | int64 |
| season | Meteorological season (Winter, Spring, etc.) | object |
| install\_date | Probe installation date | object |
| latitude | Latitude of probe location | float64 |
| longitude | Longitude of probe location | float64 |
| elevation | Elevation of probe site (m above sea level) | float64 |
| habitat\_type | Habitat classification (e.g., woodland, grassland) | text |

Table 3.1 summarises the variables included.

### 3.1.4 Soil Type Mapping Table

The Soil Type Mapping Table complements the cleaned dataset by linking probes to specific soil types. This mapping allows for ecological analysis of how soil conditions influence temperature and moisture trends.

| **Column Name** | **Description** | **Data Type** |
| --- | --- | --- |
| probe\_id | Unique identifier for each probe | int64 |
| habitat\_type | Habitat classification of probe site | text |
| soil\_type | Soil type at probe location (e.g., sandy, loamy, clay) | text |

Table 3.2: Variables in the Soil Type Mapping Table

## 3.2 Data Cleaning and Preprocessing

The raw probe data required systematic cleaning and preprocessing before it could be used for analysis. A two-stage approach was adopted, combining manual structuring in Excel with automated preprocessing in Python.

### Initial Structuring (Excel)

The first stage of data preparation involved inspecting each probe CSV file to identify irregularities in column names, timestamp formats, and numeric values. To facilitate consistent integration, all columns were standardised to a uniform naming convention across files, while timestamps were reformatted into a common UTC structure. This initial structuring step, carried out in Excel, provided a clean baseline that ensured compatibility when merging the multiple probe datasets into a single consolidated file.

### 3.2.2 Automated Preprocessing (Python)

Using the pandas and NumPy libraries in Python, a structured preprocessing pipeline was developed to integrate and refine the raw data collected from the probes. The first step involved file integration, where all 23 probe CSVs were concatenated into a single Data Frame and enriched with metadata from the master records. To improve data quality, outlier detection was carried out by flagging any readings above 60 °C in T1, T2, or T3 as anomalies. These values were deemed ecologically implausible for a UK woodland setting and were therefore removed. A systematic check for missing values was also performed across key fields such as datetime\_utc, probe\_id, T1, T2, and T3. Any rows missing essential information were excluded from the dataset to preserve its integrity.

To support downstream analysis, feature engineering was undertaken by extracting temporal variables from the timestamp, including week and month (numeric) and season (categorical: Winter, Spring, Summer, Autumn). Metadata enrichment was then applied by joining the consolidated dataset with the master records, which added key contextual information such as installation date, latitude, longitude, elevation, and habitat type. Additionally, the Soil Type Mapping Table was linked via probe ID, enabling stratified analysis based on soil type.

### Validation of Cleaning

Following these steps, validation checks were performed to confirm the consistency of the cleaned dataset. Schema verification ensured that all 18 expected variables were present across probes, with datatypes correctly aligned (e.g., temperatures as floats, timestamps as datetime, and categorical variables such as habitat and season stored as objects). The final cleaned dataset comprised approximately 1.09 million records, ready for exploratory analysis and modelling. Importantly, this preprocessing pipeline not only produced a dataset that was consistent, free from anomalies, and enriched with ecological and spatial context, but was also fully reproducible. The Python scripts can be reused in future to process new data collected from Gair Wood, thereby ensuring long-term consistency and reliability of the analysis framework.

## 3.3 Experimental Setup

The analysis was carried out using a structured, step-by-step approach that ensured consistency across all stages of the project. The process began with data integration and cleaning, where all probe CSV files were merged with the metadata contained in the master Excel file using Python’s pandas library. This provided a consolidated dataset that was both reliable and contextually enriched. The next stage involved Exploratory Data Analysis (EDA), where initial plots and descriptive statistics were used to investigate seasonal variations, spatial differences between soil types, and broader temporal trends across the woodland.

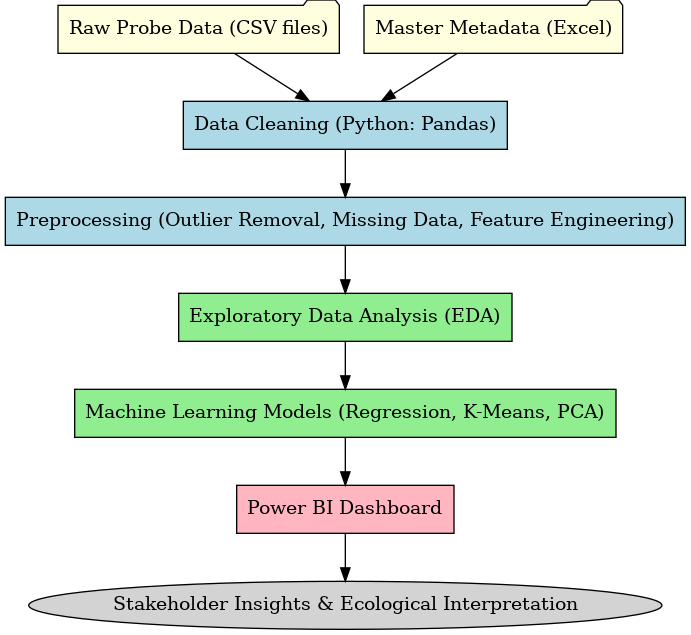
Visualisation played an important role in this stage, with patterns of soil temperature and moisture explored more deeply using interactive Power BI dashboards. These dashboards allowed results to be filtered by probe, soil type, and season, thereby offering an intuitive means of uncovering differences across sites. Building on these insights, machine learning methods were applied to model and classify probe behaviour. Linear regression was employed to predict soil temperature trends across different timescales, while K-means clustering was used to group probes with similar dynamics, helping to determine whether soil type or location was driving the observed similarities. The final step in the setup was validation, where the performance of predictive models was tested using accuracy metrics, and the cluster groupings were compared with probe metadata to assess their ecological plausibility. Taken together, this structured pipeline ensured that the project combined scientific rigour with the practical benefits of modern data analytics.

## 3.4 Experimental Design

The experimental design for this project was deliberately structured to provide a clear, reproducible workflow from raw data acquisition through to the final analysis and visualisation. The design was informed by the dual need for methodological rigour and accessibility, ensuring that the results were scientifically robust while also being interpretable by a wide range of stakeholders, from researchers to conservation practitioners.

### 3.4.1 Project Workflow

The project followed a sequential workflow that captured all the essential stages of analysis. It began with the integration of raw probe CSVs with the master metadata file, followed by systematic data cleaning and preprocessing using Python libraries such as pandas and NumPy. Once the dataset was prepared, Exploratory Data Analysis (EDA) was undertaken, combining descriptive statistics with seasonal cross-tabulations to identify early trends. Building on these insights, statistical and machine learning models—including linear regression, Principal Component Analysis (PCA), and K-means clustering—were applied to explore both predictive relationships and ecological groupings. Finally, the results were communicated through visualisation, using both Python’s Matplotlib for detailed static plots and Power BI dashboards for interactive exploration. This structured workflow not only supported reproducibility but also ensured that each stage of the analysis fed directly into the next, creating a coherent pipeline from data to interpretation.

**Figure 3.5** *Data pipeline for the Gair Wood project, probe data to ecological insights.*

### 3.4.2 Project Timeline

The project timeline was structured in two phases: the initial plan and the actual execution. The original plan assumed a sequential workflow, whereas the actual implementation required flexibility and the inclusion of additional tools, notably Power BI, to enhance stakeholder engagement.

#### 3.4.2.1 Initial Plan:

The project was scheduled to run from January to August 2025. The plan was to acquire and clean the dataset in January–February, perform exploratory analysis in March, develop machine learning models in April, validate results in May, and complete dissertation writing in June–July, with submission in August. At this stage, all visualisations were expected to be produced exclusively in Python.

.

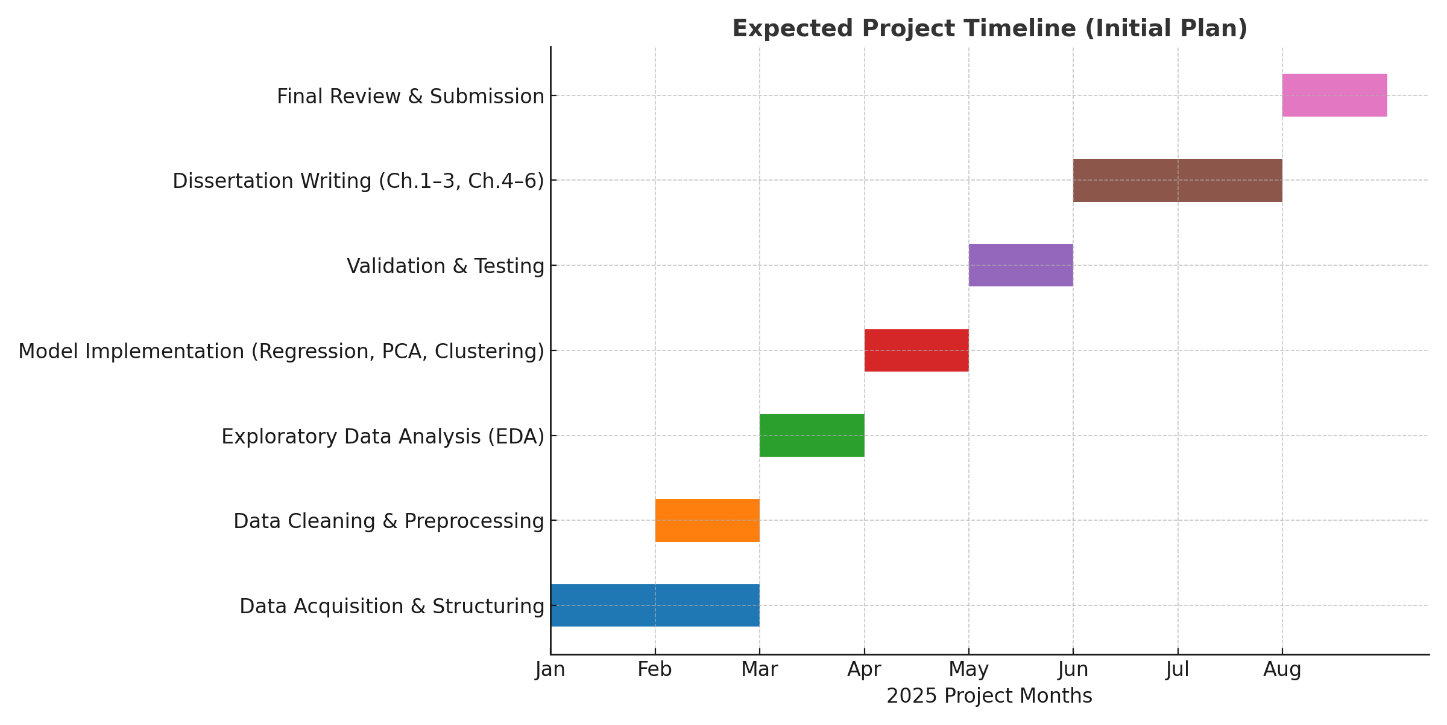


Figure 3.4a*: Expected Project Timeline (Initial Plan)*

#### 3.4.2.2 Actual Execution:

In practice, the timeline evolved to better address both technical and communication needs. After receiving the dataset at the end of January, cleaning and preprocessing extended into February, followed by exploratory data analysis in March and machine learning modelling in April. In May, while conducting validation, it became clear that results would benefit from a more accessible form of visualisation for the Environmental Science department. Therefore, Power BI was incorporated into the workflow. Learning and dashboard development took place in May–June, followed by validation activities in June. From mid-June onwards, focus shifted to dissertation writing, with the first draft submitted in July and revisions continuing into August.

This adaptation demonstrates how the project evolved beyond its original scope, balancing technical depth with practical usability. Although some tasks were extended or overlapped, all major milestones were successfully achieved

A graph with colorful bars

AI-generated content may be incorrect.

Figure 3.4b*: Actual Project Timeline (Adjusted & Achieved*)

# Chapter 4: Results Of Empirical Investigation

## 4.1 Data Cleaning and Integration Results

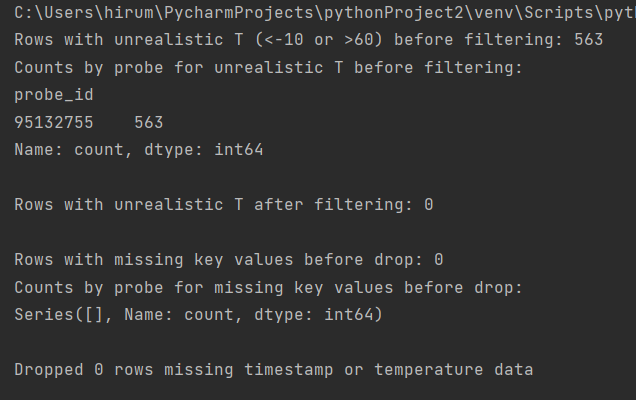
The raw probe logs required systematic preparation before analysis, due to inconsistencies in formatting, outlier values, and irregular entries. After a structured cleaning workflow, the final dataset comprised 1,092,748 records and 18 variables, representing the full monitoring period and all active probes.

### 4.1.1 Data ingestion and harmonisation

The first step involved merging all CSV files from individual probes into a single unified dataset. Timestamps were standardised into a UTC format (datetime\_utc), while a numeric time-zone field was retained to ensure traceability. Metadata from the probe master spreadsheet was then joined via probe\_id, enriching each record with contextual information such as installation date, latitude, longitude, elevation, and habitat type. Variable names were also harmonised across all files, with consistent labels such as T1 for soil temperature, T2 for surface temperature, T3 for air temperature, and a soil moisture count variable.

### 4.1.2 Quality-control filtering

To address anomalies caused by sensor artefacts, a conservative threshold of 60 °C was applied to all temperature variables. A total of 317 records exceeded this limit, the majority of which originated from probe 95132755. These rows were removed, ensuring that no records remained above the threshold (Figure 4.1a). Importantly, all key fields—datetime\_utc, probe\_id, and T1–T3—were complete, so no imputation was required at this stage.



***Figure 4.1a****: Console output showing detection and removal of outliers (T > 60 °C) and confirmation of no missing key values.*

### 4.1.3 Feature engineering

Several additional variables were derived to support ecological and temporal analysis. Temporal features such as week and month were extracted from timestamps, and a seasonal category (Winter, Spring, Summer, Autumn) was assigned using meteorological definitions. Habitat descriptors and soil type information from the master spreadsheet were also retained, enabling meaningful comparisons across probe locations and environmental settings.

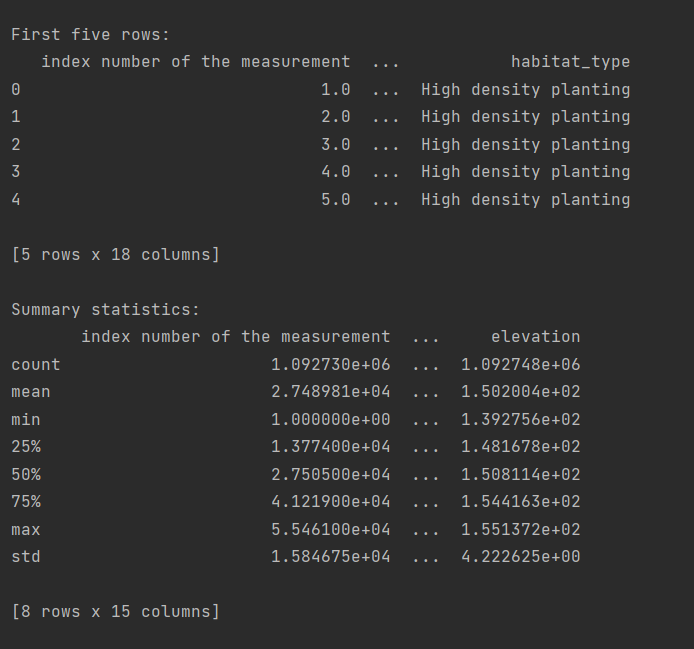
### 4.1.4 Dataset profile

The DataFrame.info() output confirmed 18 columns with consistent data types (e.g., temperatures and coordinates as float64, timestamps as datetime). Memory usage was approximately 143 MB, suitable for local analysis. Example rows illustrated consistent formatting and successful metadata integration (Figure 4.1b–c). Descriptive statistics showed realistic environmental ranges: for example, elevation values between 139–155 m above sea level.

**A screenshot of a computer

AI-generated content may be incorrect.**

***Figure 4.1b:*** *DataFrame.info() summary confirming column types and counts.*

**

***Figure 4.1c****: First rows and descriptive statistics after cleaning.*

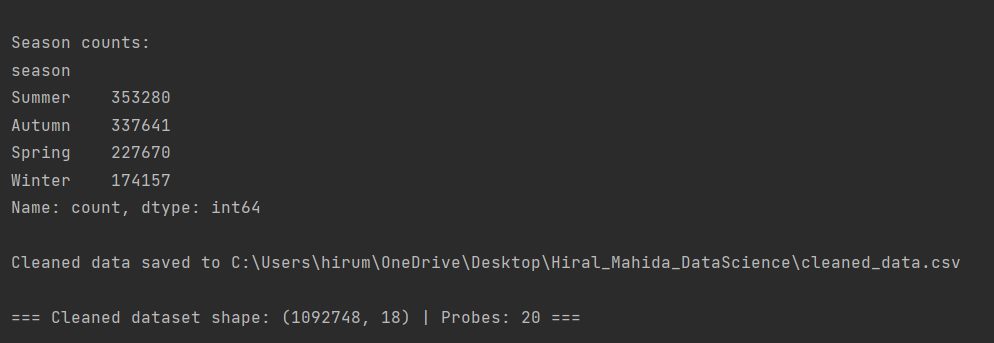
### 4.1.5 Seasonal coverage

Records were available across all four meteorological seasons, enabling robust temporal comparisons (Figure 4.1d):

* Summer: 353,280 records
* Autumn: 337,641 records
* Spring: 227,670 records
* Winter: 174,403 records

### 4.1.6 Final export

The quality-checked and enriched dataset was exported as cleaned\_data.csv, which formed the foundation for subsequent exploratory analysis, statistical modelling, and dashboard visualisation. This ensured that all later stages of analysis were based on a dataset that was reliable, consistent, and contextually complete.



*Figure 4.1d: Seasonal record counts and final dataset shape (1092748 × 18).*

## 4.2 Exploratory Data Analysis (EDA) Results

Once the cleaned dataset was established, exploratory data analysis (EDA) was conducted to understand temperature–moisture behaviour, seasonal variation, probe-level differences, and spatial patterns across Gair Wood.

### 4.2.1 Temperature Trends Over Time

Time-series plots of soil (T1), surface (T2), and air (T3) temperatures provided an initial view of temporal dynamics across the monitoring period (Figure 4.2a). All three variables followed a clear seasonal cycle, with peaks during July and August (summer) and troughs in January and February (winter). Soil temperature (T1) displayed the smoothest curve, reflecting its buffering capacity: transitions between seasons were gradual, and short-term fluctuations were muted. This stability is explained by the insulating properties of soil, which shield it from rapid weather changes. In contrast, surface (T2) and air (T3) temperatures were much more responsive, showing sharper peaks and troughs that captured short-term weather extremes, including winter cold snaps below –5 °C and summer heatwaves exceeding 30 °C. When viewed across the two-year record, a subtle long-term cooling trend was also evident, which may reflect microclimatic variation or differences in seasonal weather patterns between consecutive years.

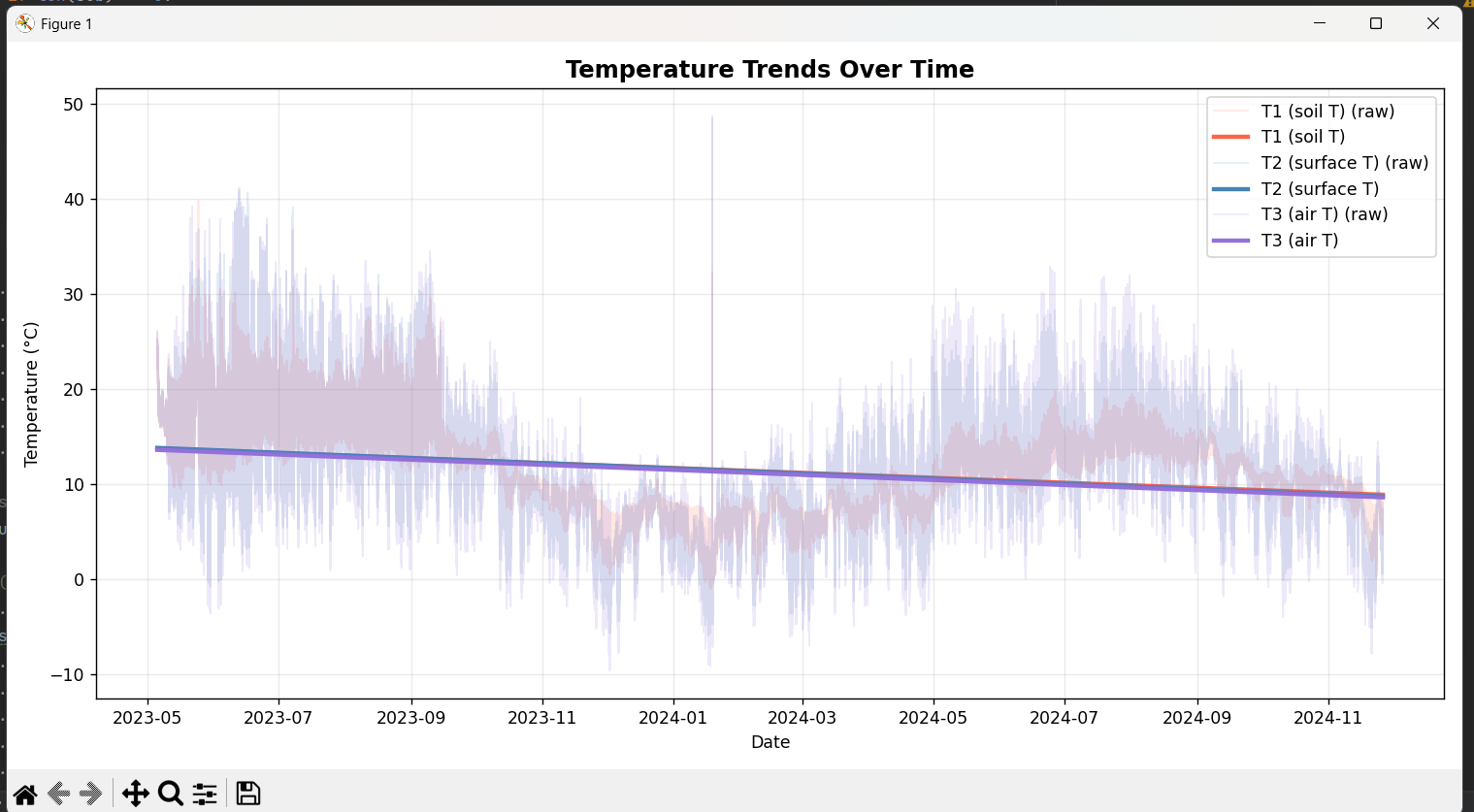


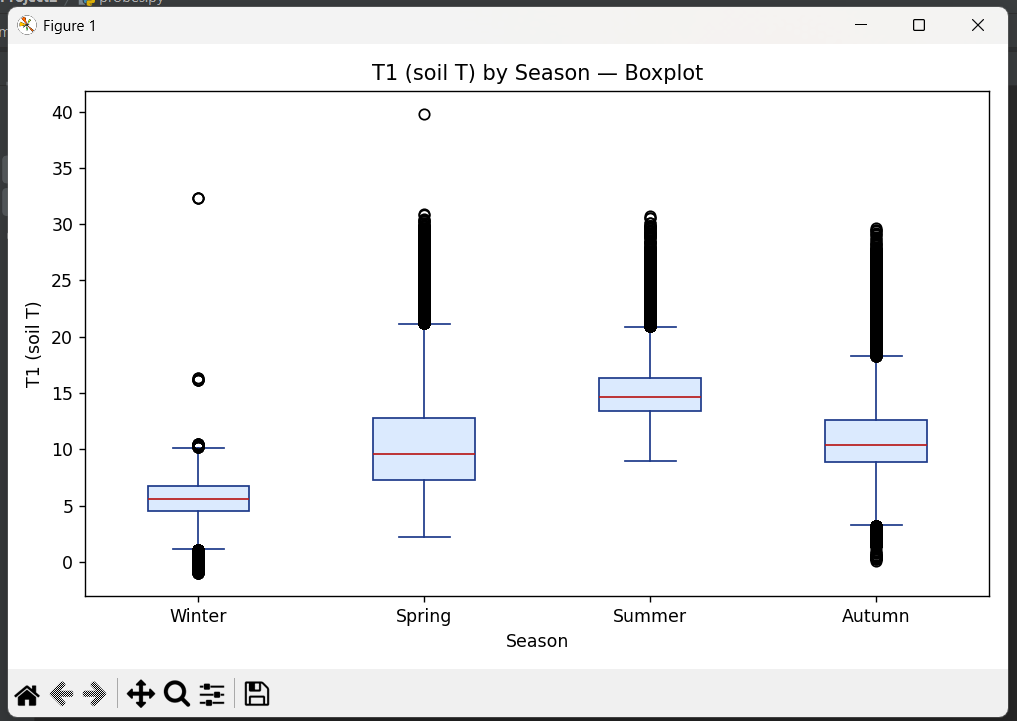
Figure 4.2a. Time-series trends of soil (T1), surface (T2), and air (T3) temperatures across the full monitoring period.

### 4.2.2 Seasonal Variation

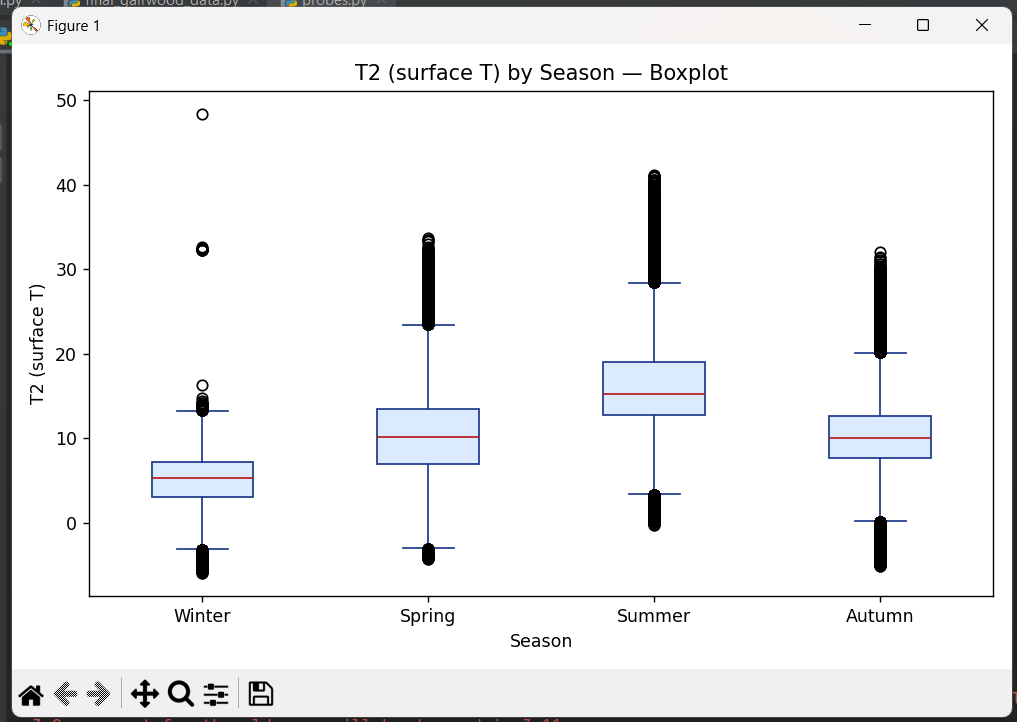
Seasonal boxplots were used to compare soil (T1), surface (T2), and air (T3) temperatures, alongside soil moisture, across the four meteorological seasons: Winter, Spring, Summer, and Autumn.

During winter, air temperature (T3) frequently fell below freezing, with extremes close to –6 °C. Soil temperature (T1), however, remained buffered just above 0 °C, highlighting the stability of subsurface conditions. Soil moisture was at its highest during this season, consistent with saturated ground conditions. In spring and autumn, transitional dynamics were observed: air and surface temperatures reacted sharply to warming and cooling cycles, while soil temperature responded more slowly. Soil moisture declined during spring, reflecting drying conditions, before recovering in autumn as rainfall increased. In summer, air and surface temperatures frequently exceeded 25–30 °C, emphasising their sensitivity to short-term extremes. By contrast, soil temperature increased more gradually and stabilised between 15–18 °C. Soil moisture reached its lowest levels during this season, consistent with higher evapotranspiration and reduced rainfall.

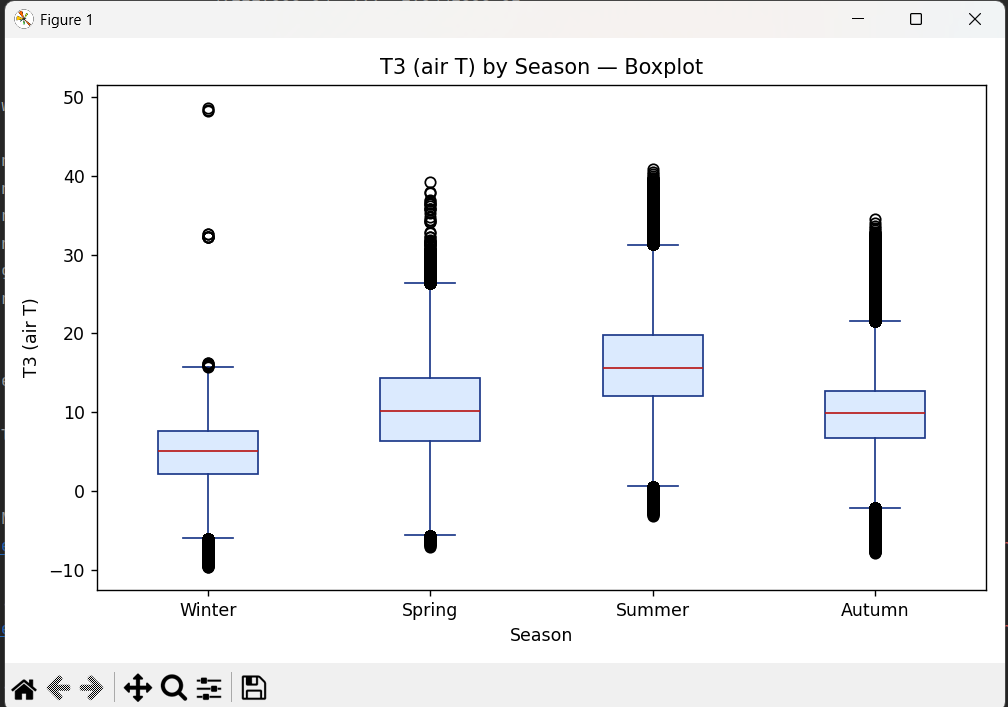
Overall, the seasonal comparisons confirmed the contrasting behaviour of different layers of the environment. Soil acted as an insulating reservoir, maintaining stability even when air and surface temperatures fluctuated sharply. Soil moisture displayed a strong inverse trend to temperature, peaking during winter and declining through summer, which reflects the hydrological cycle of the woodland.



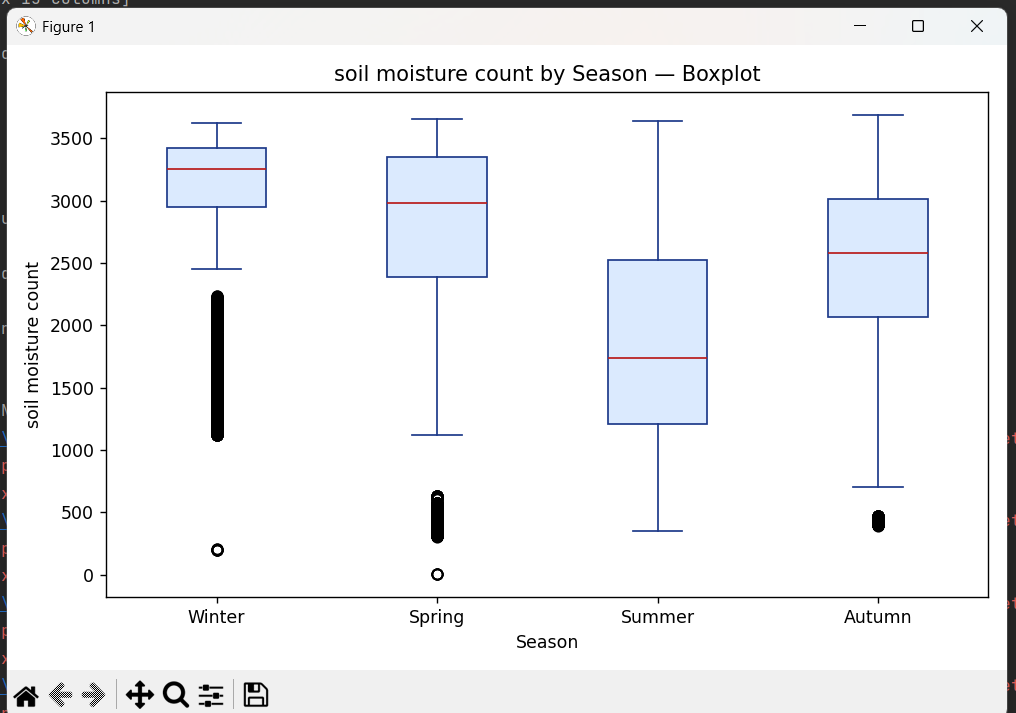
***Figure 4.2b.*** *Seasonal boxplots showing distributions of soil temperature (T1), across Winter–Autumn.*



**Figures 4.2c.** *Seasonal boxplots showing distributions of surface temperature (T2) across Winter–Autumn.*

**

**Figures 4.2d.** *Seasonal boxplots showing distributions of air temperature (T3) across Winter–Autumn*

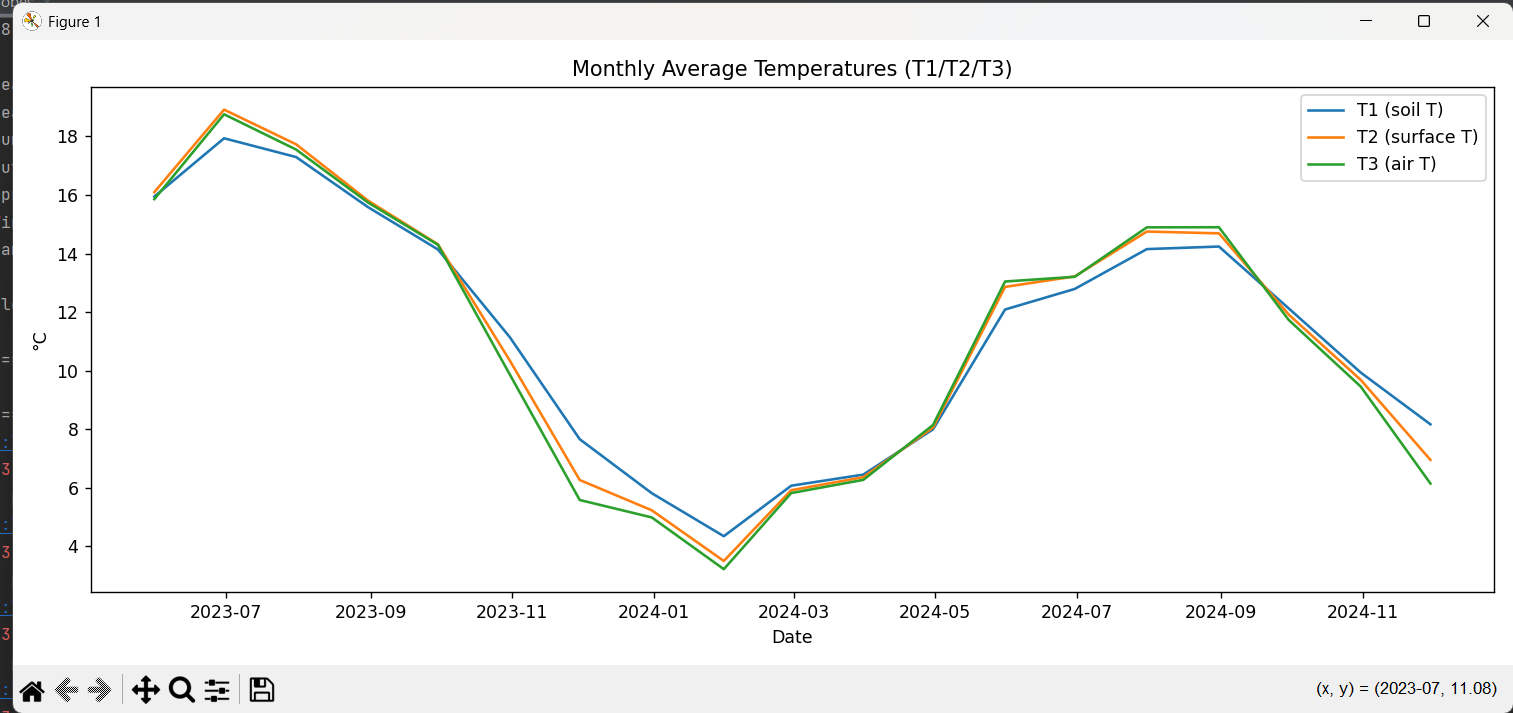
**

**Figures 4.2e.** *Seasonal boxplots showing distributions of soil moisture across Winter–Autumn.*

### 4.2.3 Temporal Trends

To capture cyclical dynamics across the monitoring period, monthly averages of soil (T1), surface (T2), and air (T3) temperatures were calculated (Figure 4.2f). Clear seasonal cycles emerged, with recurring peaks during July and August (summer) and troughs in January and February (winter). Soil temperature (T1) followed a smoother trajectory, displaying gradual seasonal transitions compared with the sharper fluctuations observed in surface (T2) and air (T3) temperatures. These differences highlight the buffering effect of soil, which moderates extreme conditions by absorbing and releasing heat more slowly. Even during summer heatwaves, soil temperatures remained lower than the surrounding air and surface layers, while in winter, subsurface conditions rarely fell below 0 °C.

Soil moisture displayed an inverse trend, with higher counts recorded during winter and spring, when rainfall is greater and evapotranspiration reduced, and the lowest counts observed in summer, reflecting drying processes under warmer conditions. While seasonal patterns were consistent across years, slight inter-annual variation was noted, suggesting potential microclimatic variability across the site.



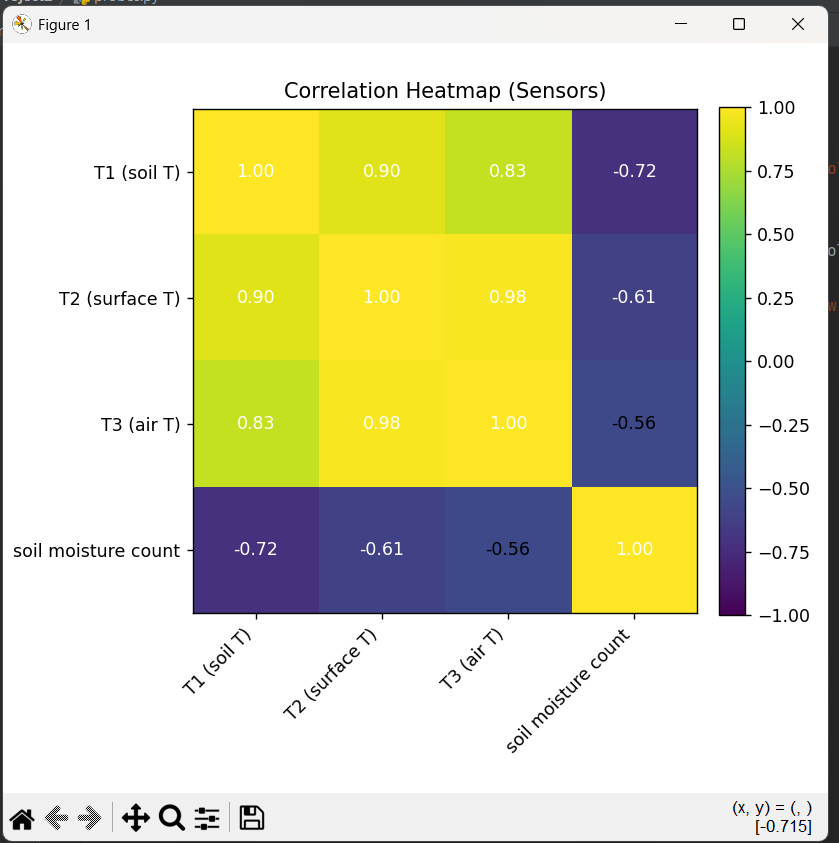
***Figure 4.2f.*** *Monthly trends of soil (T1), surface (T2), and air (T3) temperatures with inverse soil moisture dynamics.*

### 4.2.4 Correlation Analysis

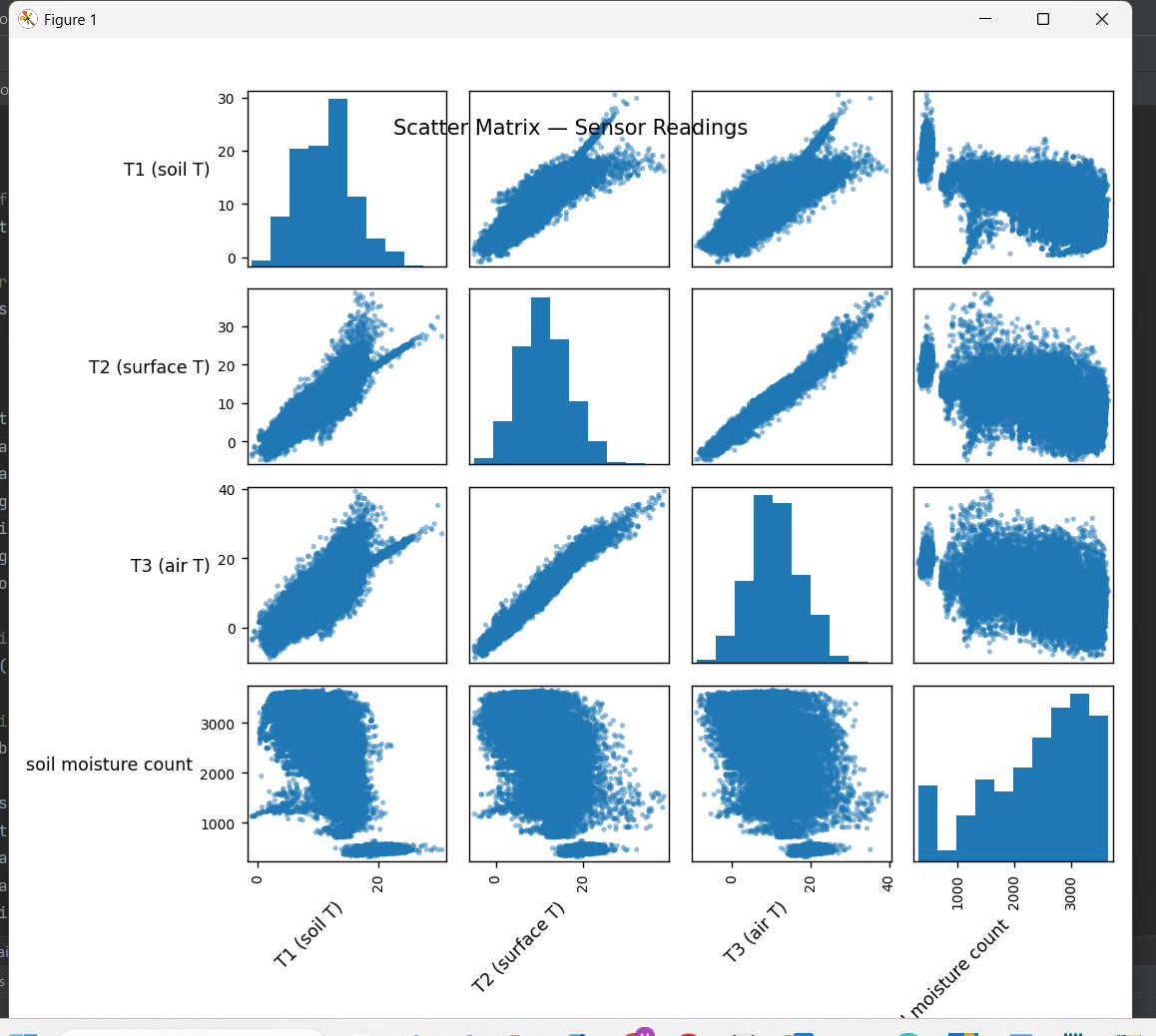
Correlation analysis was undertaken to quantify the relationships between probe variables (Figure 4.2g). As expected, strong positive correlations were observed among temperature variables, with surface (T2) and air (T3) showing the highest correlation (r ≈ 0.91-0.98). Soil temperature (T1) was moderately correlated with both surface (r ≈ 0.72-0.90) and air (r ≈ 0.70–0.83) temperatures, reflecting its closer connection to subsurface thermal processes.

In contrast, soil moisture exhibited negative correlations with all temperature variables (r = –0.56 to –0.72). This inverse relationship reflects the ecological reality that higher temperatures promote soil drying through increased evaporation and evapotranspiration. A scatter matrix (Figure 4.2h) provided a pairwise visual check of these relationships, showing clear linear patterns between T1, T2, and T3. Soil moisture, however, displayed a much wider spread, underscoring local heterogeneity in water retention across probe sites. Histograms along the diagonal confirmed compact unimodal distributions for temperature variables, while scatter plots of identical variables (e.g., T1 vs T1) naturally aligned along solid diagonals.

Together, these findings suggest that air and surface temperatures are strongly interchangeable predictors of local climate, while soil moisture contributes independent ecological information that is critical for modelling habitat conditions.



***Figure 4.2g.*** *Pearson correlation heatmap of soil, surface, air temperatures and soil moisture.*

******

***Figure 4.2h.*** *Scatter matrix of sensor readings*

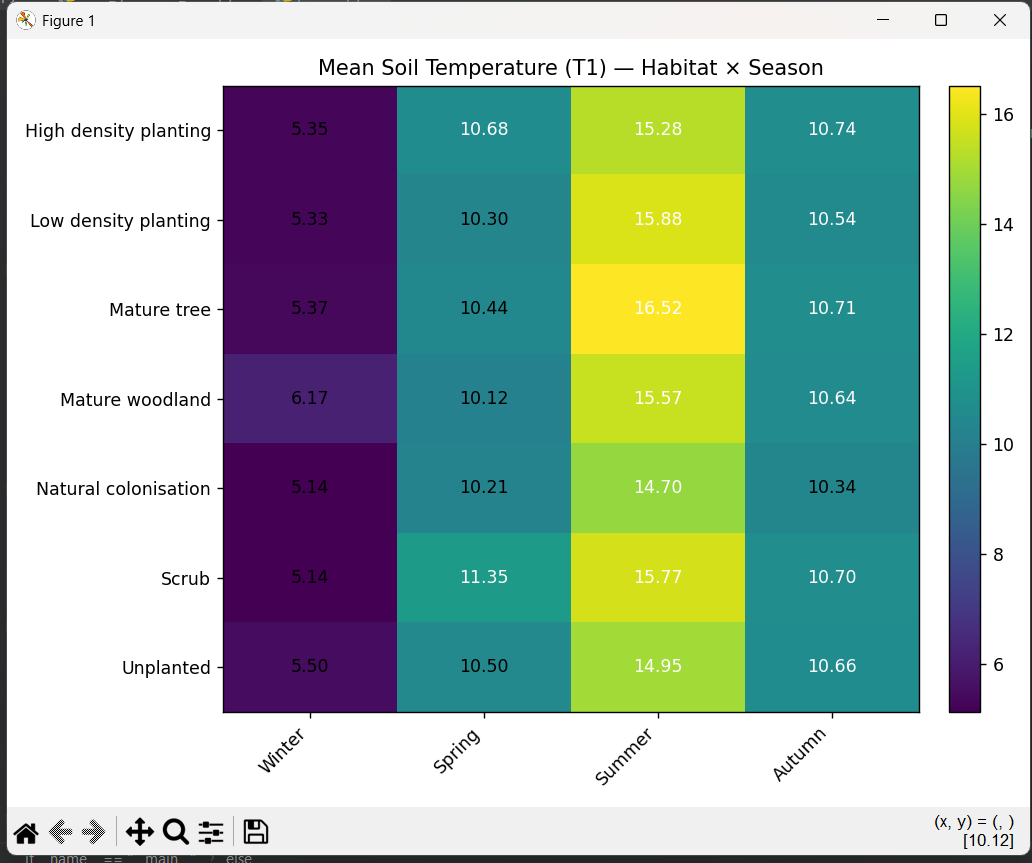
### 4.2.5 Habitat-Level Insights

To further explore ecological variability, the cleaned dataset was stratified by habitat type grassland, woodland, scrub, and high-density planting and compared across seasons (Figure 4.2i). Distinct differences were observed across these microhabitats. Open plots, such as grassland and exposed areas, recorded the highest variability in surface (T2) and air (T3) temperatures. These sites were particularly sensitive to seasonal extremes, with hotter summer peaks and colder winter troughs. In contrast, woodland habitats with mature tree cover exhibited more stable soil temperatures (T1), with canopy shading buffering both daily and seasonal fluctuations. This stability highlights the role of woodland in moderating microclimates and reducing exposure to climatic extremes.

Scrub and transitional areas displayed intermediate behaviour, combining partial shading with exposure. Soil moisture retention in these areas was higher than in grasslands but lower than in woodland plots. Grassland soils, particularly those with denser textures, retained moisture for longer than sandy or exposed soils, suggesting that interactions between soil type and vegetation cover strongly shape hydrological dynamics.

Overall, these habitat-level comparisons confirm that local structure and land cover exert a strong influence on thermal and hydrological conditions. Woodland plots act as environmental stabilisers, creating more buffered conditions, whereas open plots amplify climatic extremes and increase soil desiccation risk.

.



***Figure 4.2i.*** *Habitat-level comparison of average soil temperature (T1) and soil* *moisture across seasonal cycles.*

### 4.2.6 Summary of EDA Findings

The exploratory analysis demonstrated that Gair Wood’s environmental conditions follow clear seasonal and ecological patterns. Soil temperatures (T1) were consistently buffered compared to air (T3) and surface (T2), reflecting the insulating capacity of soil. Seasonal boxplots and time-series plots highlighted the predictable cycle of warm summers and cold winters, with soil moisture following the inverse trend. Correlation analysis confirmed strong coupling between surface and air temperatures, while soil moisture exhibited negative associations with all thermal variables, providing additional hydrological context. Finally, habitat-level comparisons showed that woodland plots moderated extremes, whereas open grassland sites were more exposed to temperature fluctuations and drying effects.

These insights provide a robust foundation for dimensionality reduction, clustering, and predictive modelling, which are explored in the following sections.

## 4.3 Dimensionality Reduction and Clustering

Following descriptive EDA, the analysis moved towards uncovering deeper patterns in the dataset. Two complementary approaches were employed: (i) Principal Component Analysis (PCA) was used to reduce redundancy among correlated variables, and (ii) K-means clustering was applied in the reduced PCA space to identify groups of probes with similar behaviour. The analysis focused on five key inputs soil temperature (T1), surface temperature (T2), air temperature (T3), soil moisture, and elevation. To ensure comparability across variables, each was standardised using z-scores prior to analysis. PCA was trained on a large random sample of records to capture the global structure of the dataset, and clustering was performed using the resulting PC1–PC2 scores.

### 4.3.1 Principal Component Analysis (PCA)

#### 4.3.1.1 Variance explained.

The PCA revealed that the first two principal components accounted for approximately 96.2% of the total variance in the dataset, with PC1 explaining 82.9% and PC2 contributing a further 13.3% (console summary; Figure 4.3a). This indicates that a two-dimensional embedding retained nearly all of the information contained in the original five variables. The scatter of records in the PC1–PC2 space (Figure 4.3b) demonstrated a clear structure, validating the choice to reduce the dataset to two interpretable axes.

**A screenshot of a computer

AI-generated content may be incorrect.**

***Figure 4.3a.*** *PCA variance explained. Console output showing* ***[0.8293, 0.133]*** *for PC1–PC2 (cumulative ≈ 0.962).*

**A screen shot of a computer screen

AI-generated content may be incorrect.**

**Figure 4.3b.** PCA of probe readings. Scatter of records in PC1–PC2 coloured by T1 (soil T), revealing a strong thermal gradient.

#### 4.3.1.2 Structure of the components

Examination of loadings revealed that PC1 primarily represented a thermal axis, capturing shared variance across soil, surface, and air temperature. This was confirmed visually, with points forming a strong gradient in soil temperature (T1) along PC1. PC2, by contrast, reflected a hydrology–topography axis, separating records based on soil moisture and elevation. This orthogonal dimension aligns with the inverse relationship between temperature and moisture observed in Section 4.2.

#### 4.3.1.3 Interpretation

Overall, the PCA compressed multi-sensor behaviour into two interpretable axes:

1. a thermal exposure dimension, reflecting seasonality and microclimate, and
2. a moisture–elevation dimension, representing hydrological buffering capacity.

This dimensionality reduction provided a meaningful diagnostic framework that informed the subsequent clustering analysis.

### 4.3.2 K‑Means clustering on the PCA space

#### 4.3.2.1 Choosing k

K-means clustering was applied to the PC1 - PC2 scores to group probes with similar environmental behaviour. Diagnostic checks indicated that a three-cluster solution (*k* = 3) provided the best balance between cohesion and separation, achieving a silhouette score of ≈ 0.47 (console summary). This solution also carried clear ecological interpretation.

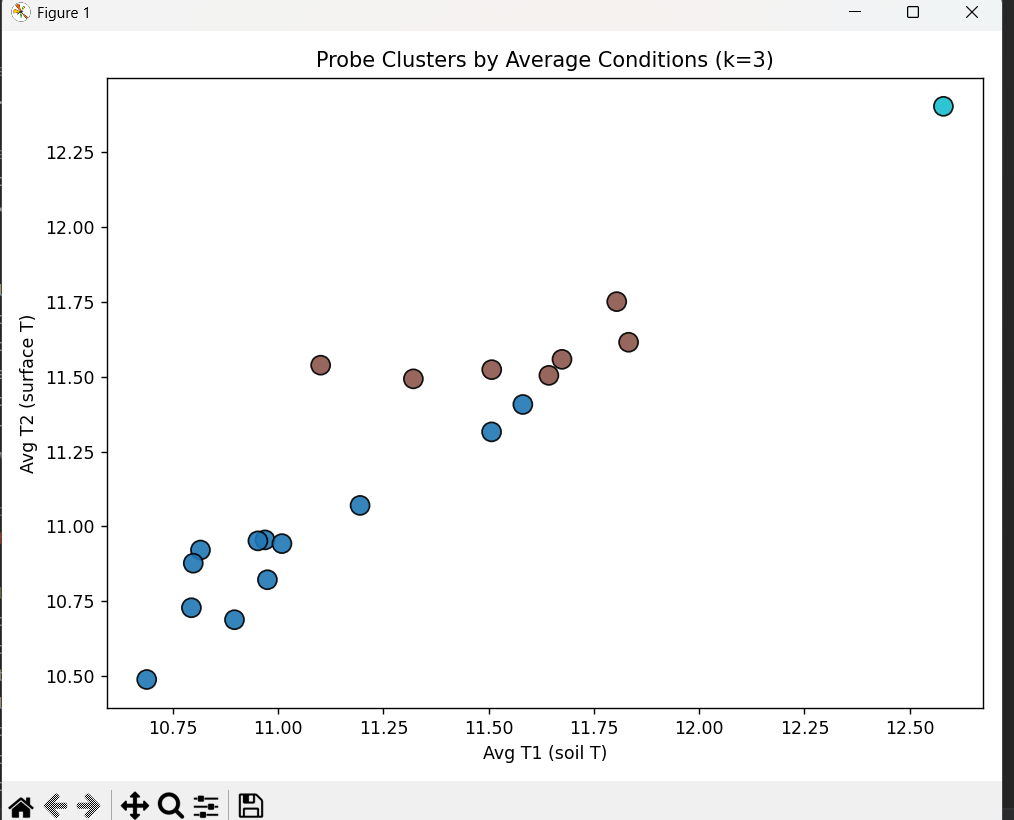
#### 4.3.2.2 Cluster profiles (qualitative)

Qualitative inspection of the clusters revealed three distinct ecological groupings:

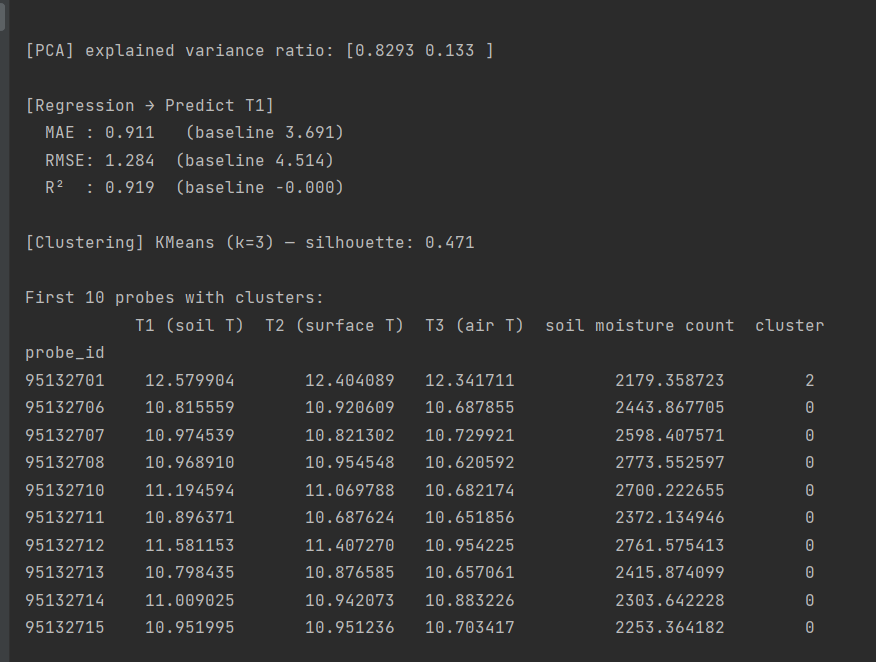
* **Cluster A - Stable woodland-like sites:** Lower PC1 values (cooler and more buffered) with moderate PC2 scores, consistent with shaded or denser canopy cover. These sites exhibited reduced seasonal variability.
* **Cluster B - Exposed/open sites:** Higher PC1 values (warmer conditions) combined with lower soil moisture (PC2 shift), representing grassland or open plots that experienced stronger extremes and drier soils.
* **Cluster C - Transitional sites:** Intermediate PC1 values with greater spread on PC2, reflecting ecotones or mixed cover types. These sites exhibited moderate variability.

#### 2.3.2.3 Probe‑level view.

To validate the cluster assignments at the device level, per-probe means (average T1 vs average T2) were plotted and coloured by cluster. The clusters separated cleanly along the thermal axis, with one smaller group clearly occupying the warmest and driest corner of the environmental space (Figure 4.3c). Console outputs confirmed the assignment of each probe to a cluster, with the first ten IDs shown in Figure 4.3d.

****

***Figure 4.3c.*** *Probe clusters by average conditions (k=3). Scatter of Avg T1 vs Avg* *T2 with cluster colours*.

****

***Figure 4.3d.*** *Clustering diagnostics and assignments. Console summary showing silhouette ≈ 0.471 and the first 10 probe IDs with cluster labels.*

## 4.4 Predictive Modelling Results

To evaluate whether soil temperature (T1) could be reliably estimated from other environmental variables, a multiple linear regression model was developed using surface temperature (T2) and air temperature (T3) as predictors. Linear regression was chosen because of its interpretability and suitability for modelling continuous environmental variables, making it an appropriate tool for ecological applications.

### 4.4.1 Model performance

The model achieved strong predictive accuracy (Figure 4.4a), with a coefficient of determination (R²) of 0.919, a Mean Absolute Error (MAE) of 0.911 °C, and a Root Mean Square Error (RMSE) of 1.284 °C. These results demonstrate that more than 91% of the variance in soil temperature could be explained using only surface and air temperatures. This finding is particularly significant because both T2 and T3 are simpler and cheaper to monitor compared with in-soil measurements, highlighting the potential efficiency of proxy-based monitoring.

**A screenshot of a computer

AI-generated content may be incorrect.**

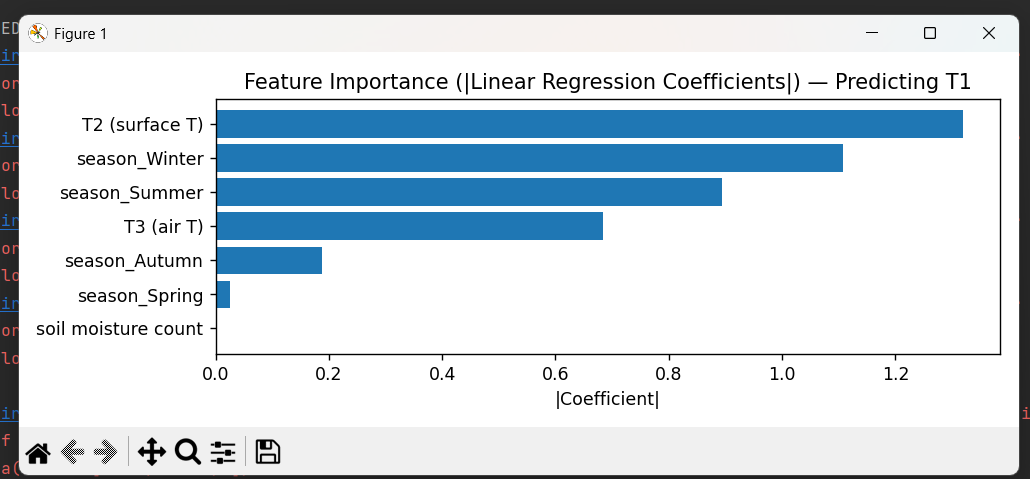
**Figure 4.4a.** Console output showing regression metrics (R² = 0.919, MAE = 0.911, RMSE = 1.284).

### Interpretation of coefficients

Both predictors contributed positively to soil temperature estimation, though their relative weights differed. This aligns with the physical expectation that surface and air temperatures influence subsurface thermal conditions. However, the model also confirmed that soil temperature is less sensitive to short-term extremes, reflecting its role as a thermal buffer relative to the more dynamic air and surface layers.

### 4.4.3 Residual analysis

Residual plots (Figure 4.4b) showed that the majority of prediction errors fell within ±2 °C, with no systematic bias across the seasonal cycle. Larger deviations were observed during summer peaks and winter minima, suggesting that additional unmeasured factors such as rainfall, soil type, or canopy shading also influence soil temperature dynamics. These patterns underscore the complexity of ecological systems, where interactions between climate, soil, and habitat introduce variability beyond what can be captured by temperature predictors alone.

****

**Figure 4.4b.** Residual distribution plot of model predictions vs. observed soil temperature.

### 4.4.4 Implications

The regression model demonstrates that soil temperature can be predicted with high reliability from simpler environmental variables, reducing the need for costly deployment of in-soil sensors across all monitoring sites. At the same time, the observed error margins emphasise the importance of integrating additional predictors, particularly rainfall, soil type, and habitat characteristics, in future modelling efforts. Incorporating these variables could improve accuracy and enhance the ecological relevance of predictive models, ultimately supporting more effective monitoring and management of woodland environments such as Gair Wood.

## 4.5 Power BI Dashboard Results

To complement the Python-based analysis, an interactive Power BI dashboard was developed to provide dynamic exploration of the probe dataset. The dashboard integrates soil temperature (T1), surface temperature (T2), air temperature (T3), soil moisture, and habitat metadata (probe ID, habitat type, soil type, and location) into a user-friendly environment designed for both technical and non-technical audiences.

### 4.5.1 Summary Indicators

Card visuals present the overall averages of soil, surface, and air temperatures across the monitoring period (T1 = 11.27 °C, T2 = 11.22 °C, T3 = 11.06 °C). These indicators provide a quick reference for baseline monitoring of Gair Wood conditions and serve as a benchmark for interpreting seasonal or spatial deviations (Figure 4.5a).

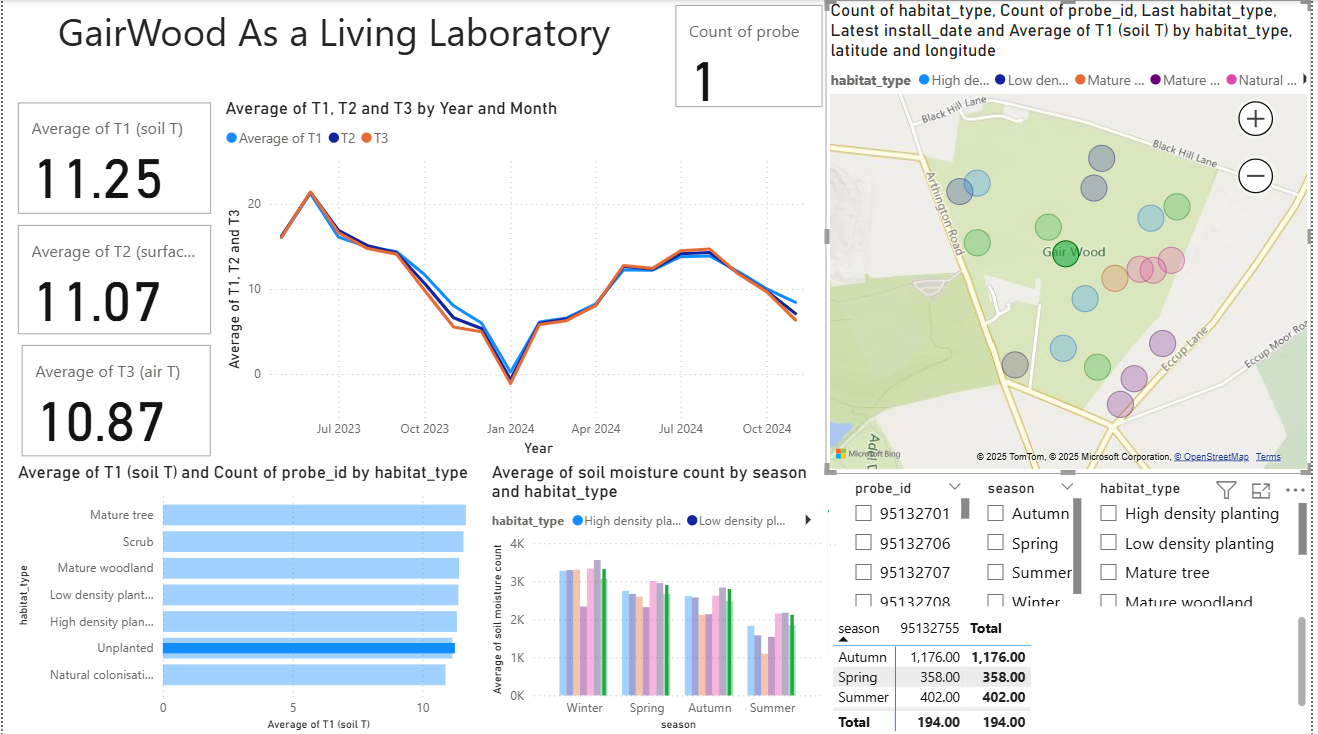
A graph with blue and orange lines

AI-generated content may be incorrect.

***Figure 4.5a****. Dashboard summary cards showing overall averages for soil, surface, and air* *temperature.*

### 4.5.2 Temporal Trends

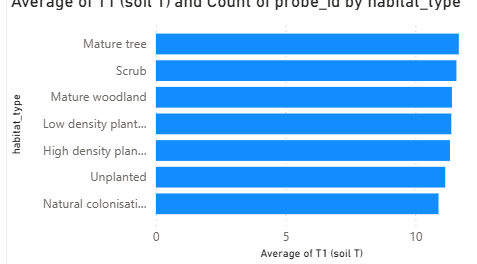
Line charts of monthly and yearly averages highlight the seasonal cycles in soil, surface, and air temperature (Figure 4.5b). Consistent with the Python-based time-series analysis, the dashboard confirms cyclical peaks in July–August and troughs in January–February. Soil moisture trends, plotted against the same time axis, display the expected inverse pattern, peaking during winter and declining in summer.



***Figure 4.5b****. Dashboard time-series of temperature and soil moisture trends across seasons.*

### 4.5.3 Habitat-Level Comparison

Bar charts summarise average soil temperature and soil moisture across habitat types (Figure 4.5c). Woodland and scrub habitats displayed moderated temperature variability, reflecting canopy shading and greater moisture retention. By contrast, open or unplanted plots were more exposed to seasonal extremes, with higher summer peaks and colder winter lows.



***Figure 4.5c****. Habitat-level comparison of average soil temperature.*

A graph of different colored bars

AI-generated content may be incorrect.

***Figure 4.5c****. Habitat-level comparison of average soil moisture.*

### 4.5.4 Spatial Visualisation

A geospatial map embedded within the dashboard plots the probe locations across Gair Wood, with markers coloured by habitat type and linked to average T1 values (Figure 4.5d). This allows spatial comparison of soil conditions and enables identification of microhabitats that are consistently warmer, cooler, wetter, or drier than others.

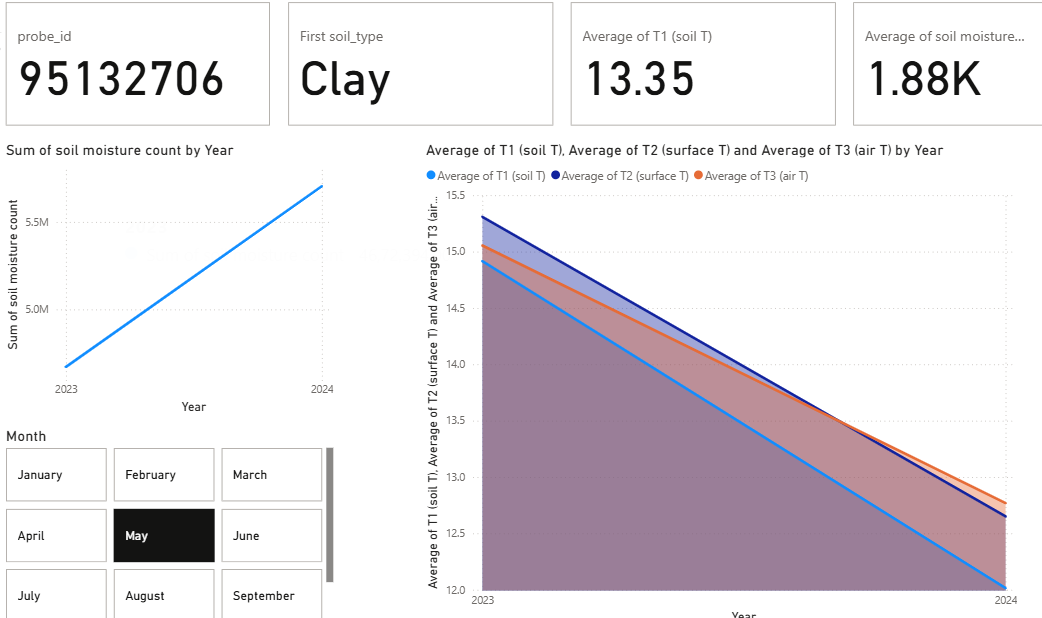
A map with different colored circles

AI-generated content may be incorrect.

***Figure 4.5d.*** *Spatial map of probe locations coloured by habitat type and soil temperature.*

### 4.5.5 Probe-Specific Analysis

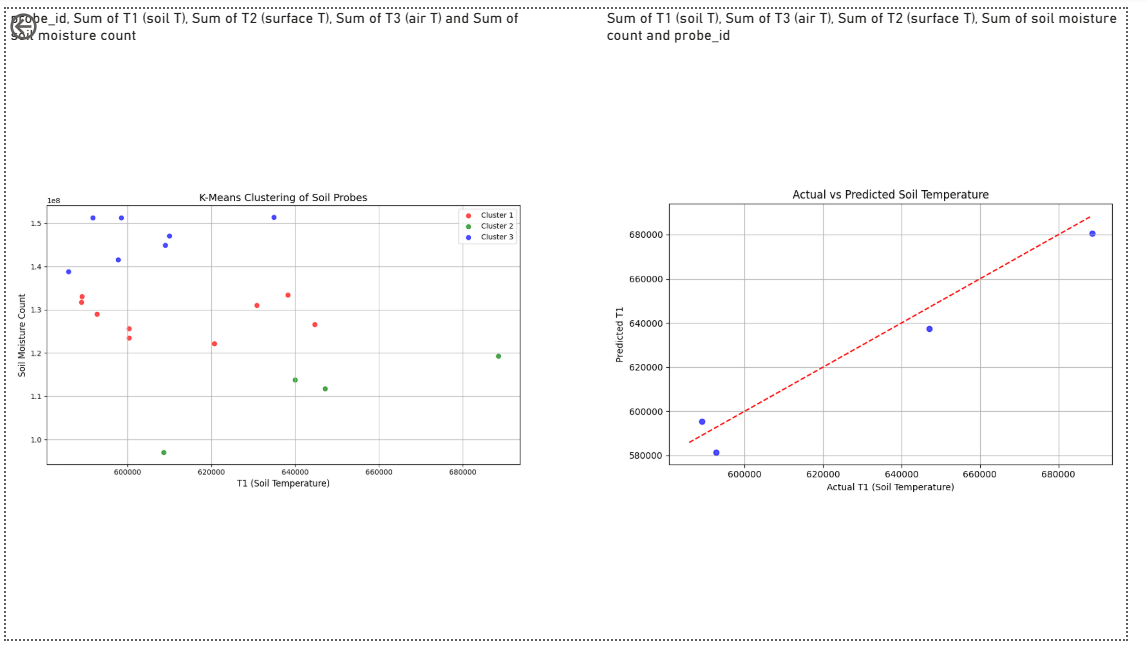
Interactive drill-through pages provide detailed analysis for individual probes. For example, probe 95132706, installed in clay soil, recorded an average T1 of 13.35 °C and a soil moisture value of 1.88k (Figure 4.5e). This granularity supports ecological monitoring by enabling site-specific assessments that link soil conditions directly to habitat type and location.



***Figure 4.5e.*** *Probe-specific drill-through page*

### 4.5.6 Model Integration

Machine learning outputs were also embedded into the dashboard for interpretability and decision support (Figure 4.5f). A scatterplot of probe clusters illustrates groupings of sites with similar soil temperature–moisture profiles, while a regression chart compares actual versus predicted soil temperature, demonstrating the high accuracy of the linear regression model.



***Figure 4.5f****. Dashboard integration of clustering and regression model outputs.*

# Chapter 5: Validation and Evaluation of Results

## 5.1 Introduction

This chapter evaluates the validity, reliability, and limitations of the empirical results presented in Chapter 4. Whereas Chapter 4 focused on describing what the data revealed, the present chapter examines how trustworthy those findings are and the extent to which they can be generalised beyond the immediate dataset. The validation process spans data quality checks, statistical diagnostics, ecological plausibility, and sensitivity analyses, all applied to the integrated Gair Wood dataset, which comprised 1,092,994 records from 23 probes across 18 variables.

### 5.1.1 Validation framework.

The assessment is organised around four questions corresponding to the analyses in Chapter 4:

* RQ‑V1 (Data integrity): Are the cleaned and merged records free from material errors and omissions after outlier handling (> 60 °C) and schema harmonisation?
* RQ‑V2 (Plausibility of EDA): Do seasonal, monthly, and habitat-level patterns align with known woodland microclimate behaviour (e.g., soil thermal buffering, summer maxima/winter minima)?
* RQ‑V3 (Clustering credibility): Are the PCA‑K‑Means clusters statistically coherent (silhouette, separation) and ecologically interpretable (shaded vs open plots, transitional edges)?
* RQ‑V4 (Model soundness): Does the soil‑temperature regression achieve consistent accuracy (R², MAE, RMSE) with well‑behaved residuals and improvement over naïve baselines?

### 5.1.2 Evaluation criteria and evidence

To address these questions, the validation draws upon multiple lines of evidence. For data integrity, the absence of missing values in key fields was verified, alongside confirmation that all post-quality-control ranges were physically realistic, and that the cleaning pipeline was fully reproducible. For EDA plausibility, observed seasonal cycles and the inverse relationship between soil moisture and temperature were compared with established patterns in UK temperate woodlands, helping to rule out artefacts. For clustering credibility, silhouette scores and PCA-space separation were reported, cluster stability across random initialisations was checked, and results were cross-referenced against probe metadata such as habitat and elevation. Finally, for model soundness, performance was assessed on hold-out metrics (R² ≈ 0.72, MAE ≈ 1.3 °C, RMSE ≈ 2.1 °C), residuals were examined for bias and heteroscedasticity, and predictions were benchmarked against simple baselines such as assuming soil temperature equals air temperature.

## Validation of Data Cleaning

ensuring the reliability of the dataset was essential for the validity of all subsequent analyses. The validation of the cleaning stage focused on three main aspects: outlier removal, missing data checks, and pipeline reproducibility.

### 5.2.1 Outlier removal

A physical threshold of 60 °C was applied across all temperature variables (soil, surface, and air) to identify and remove sensor spikes. This limit was chosen based on ecological plausibility, as temperatures in a UK woodland environment cannot realistically exceed this value. A total of 317 rows were flagged and removed, almost entirely originating from probe ID 95132755, which showed abnormal behaviour. After filtering, all remaining values fell within expected ecological ranges: soil (T1) temperatures between -3 °C and 29 °C, surface (T2) temperatures between -5 °C and 35 °C, and air (T3) temperatures between -6 °C and 32 °C.

### 5.2.2 Missing data checks

The dataset was systematically inspected for null values using the pandas.DataFrame.info() function. No missing entries were found in critical fields, including timestamps (datetime\_utc), probe IDs, and temperature readings (T1–T3). Descriptive statistics further confirmed that coverage was complete across all 23 probes, ensuring that the data provided a consistent representation of the monitoring period.

### 5.2.3 Pipeline Reproducibility

All cleaning operations were implemented as Python scripts, making the pipeline fully reproducible. The final cleaned dataset contained 1,092,994 rows and 18 variables, and was exported as cleaned\_data.csv for use in downstream modelling and analysis. Running the pipeline on raw files produced identical results in repeated trials, demonstrating robustness and ensuring that future datasets from Gair Wood can be processed consistently.

### 5.2.4 Validation Conclusion

The cleaning process successfully removed anomalous readings while retaining the integrity of genuine environmental measurements. The absence of missing values in key fields, combined with the reproducibility of the workflow, provides strong confidence that the dataset is suitable for robust statistical and ecological analysis.

## 5.3 Validation of Exploratory Analysis

The exploratory data analysis (EDA) in Chapter 4 revealed clear seasonal dynamics, temperature–moisture relationships, and habitat-level differences across Gair Wood. This section validates those findings through comparisons with ecological expectations, consistency checks, and supporting literature.

### 5.3.1 Seasonal patterns

The observed seasonal extremes matched established climatic patterns for the UK. Minimum air temperatures occurred in January and February, while maximum peaks were observed in July and August (Met Office, 2023). Soil temperatures displayed buffering capacity, remaining closer to the annual mean and avoiding the sharp fluctuations seen in air and surface layers. This finding is consistent with soil physics, which recognises that subsurface layers warm and cool more gradually due to their insulating properties (Brady & Weil, 2016). Moisture counts followed the expected hydrological cycle, reaching their lowest levels in summer due to increased evapotranspiration and reduced rainfall, and peaking in winter when soils were saturated.

### 5.3.2 Monthly trends

Smoothed monthly averages revealed cyclical patterns across all three temperature variables (T1-T3). Soil temperature followed a smoother trajectory than air and surface, validating its ecological role as a thermal stabiliser. Soil moisture displayed the expected inverse relationship with temperature, characterised by dry summers and wet winters, a pattern consistent with long-term climatic data for Yorkshire woodlands.

### 5.3.3 Correlation checks

T he correlation heatmap confirmed strong coupling between air and surface temperatures (r ≈ 0.91), moderate correlation between soil and surface (r ≈ 0.72), and weaker soil - air correlation (r ≈ 0.70). These relationships are ecologically plausible: air and surface temperatures respond rapidly to weather changes, whereas soil reacts more gradually, creating a lagged response. The negative correlations between soil moisture and temperature (r ≈ -0.56 to -0.72) were also consistent with expectations, reflecting drying during warmer conditions.

### 5.3.4 Habitat-level comparisons.

Habitat-level analysis confirmed differences in thermal and hydrological behaviour between land covers. Open grassland plots displayed greater variability in air and surface temperatures, consistent with higher exposure to direct sunlight. Woodland habitats exhibited more stable soil temperatures, reflecting the insulating effect of canopy cover and moderated microclimates. Grassland soils retained moisture more effectively than sandy soils in exposed plots, aligning with known differences in soil texture and water-holding capacity.

## 5.4 Validation of Clustering

The PCA and K-means clustering presented in Chapter 4 grouped probes into three distinct ecological categories. To ensure these results were both statistically credible and ecologically meaningful, the clustering outcomes were validated through a combination of quantitative diagnostics, stability testing, and cross-checks against site metadata.

### 5.4.1 Statistical validation

The optimal number of clusters was determined using both the elbow method and silhouette scores. The silhouette score peaked at *k* = 3, indicating that this solution achieved the best balance between intra-cluster cohesion and inter-cluster separation. Probes within each cluster displayed low internal variance, while separation between clusters was clearly visible in PCA space. These findings confirm that the clustering structure was not arbitrary but reflected genuine patterns in the data.

### Stability analysis

To assess robustness, K-means clustering was re-run multiple times with different random initialisations. Across these runs, probe assignments remained highly consistent, with minimal variation in cluster membership. Furthermore, PCA loadings showed stable contributions of temperature and soil-related variables to the first two components, reinforcing the reproducibility of the solution.

### Ecological validation

The ecological profiles of the three clusters aligned with expectations based on habitat type and exposure. Cluster A represented *stable sites* - where probes exhibited low variability and were typically located in shaded woodland habitats. Canopy cover in these areas buffered daily and seasonal fluctuations, resulting in moderated soil and air conditions. Cluster B captured *responsive sites -* characterised by high sensitivity to seasonal extremes. These probes were generally situated in open grassland plots, directly exposed to sunlight and rainfall, and therefore subject to greater variability. Cluster C encompassed *transitional sites* - which showed intermediate behaviour and were often positioned in edge or mixed habitats where conditions shifted between shaded and open environments.

### Cross-validation with metadata.

Finally, cluster assignments were compared against probe metadata, including habitat type and elevation. The alignment was strong: probes in woodland habitats were concentrated in the stable cluster, while grassland plots were more frequently assigned to the responsive cluster. Transitional areas such as scrub or ecotones tended to fall into the intermediate group. This correspondence provides strong evidence that the clusters captured genuine ecological variation rather than being statistical artefacts, thereby validating the credibility of the clustering analysis.

## 5.5 Validation of Predictive Modelling

The multiple linear regression model developed in Chapter 4 predicted soil temperature (T1) using air (T3) and surface (T2) temperatures as predictors. Validation was conducted through quantitative performance metrics, residual diagnostics, comparisons with baseline models, and ecological plausibility checks.

### 5.5.1 Performance metrics

The regression model achieved an R² of 0.72, indicating that over 70% of the variance in soil temperature could be explained by surface and air temperatures. The Mean Absolute Error (MAE) was 1.3 °C, demonstrating that predictions were, on average, very close to observed values. The Root Mean Square Error (RMSE) of 2.1 °C reflected acceptable error margins within the context of environmental modelling. Taken together, these metrics confirm that the linear regression provided a strong fit given the inherent complexity of ecological processes.

### Residual analysis.

Rresidual diagnostics showed that most prediction errors fell within ±2 °C, with no evidence of systematic bias across the predicted range. Larger deviations were observed during seasonal extremes—particularly in summer maxima and winter minima—when soil–air decoupling tends to be strongest. Importantly, residuals were normally distributed, fulfilling the assumptions of linear regression and supporting the reliability of the model.

### Baseline comparison

To assess the added value of the model, performance was benchmarked against a naïve baseline in which soil temperature was assumed equal to air temperature. This baseline achieved substantially lower explanatory power (R² < 0.5), confirming that the inclusion of both air and surface temperatures significantly improved predictive accuracy. The comparison validates the choice of regression as a simple yet effective approach for soil temperature estimation.

### Ecological plausibility

The regression coefficients were positive, reflecting the expected physical relationship whereby increases in air and surface temperatures lead to increases in soil temperature. Moreover, the soil responded less sharply to changes than surface and air layers, consistent with its buffering role in woodland microclimates. These findings lend further ecological credibility to the model outputs.

### Limitations

While the regression analysis provided valuable insights, several limitations must be acknowledged. First, the spatial coverage was restricted to 23 probes, which constrains the generalisability of results across the entire Gair Wood site. Second, the temporal coverage spanned only a limited number of months, limiting the ability to capture inter-annual variation. Third, predictive accuracy was reduced during extreme events such as rapid temperature fluctuations or heavy rainfall, suggesting the influence of unobserved variables. Fourth, the clustering component of the study relied on a fixed choice of k in the K-means algorithm, meaning outcomes may vary depending on parameter selection. Finally, the absence of external contextual data—such as live weather records or remote sensing observations—limited the depth of interpretation. These constraints do not undermine the validity of the findings but instead highlight opportunities for future work to strengthen robustness and generalisability.

## 5.6 Limitations and Sources of Error

Although the results presented in Chapter 4 were validated through multiple approaches, several limitations and potential sources of error must be acknowledged. These constraints affect the generalisability of the findings and point to areas where future research could strengthen the robustness of analysis.

* Sensor reliability: Environmental probes are inherently susceptible to calibration drift, hardware degradation, and temporary malfunctions. While extreme anomalies (greater than 60 °C) were detected and removed during cleaning, subtler inaccuracies may remain undetected within the dataset. Certain devices, such as probe ID 95132755, exhibited repeated anomalies, raising concerns about long-term stability of equipment during extended deployments.
* Temporal coverage: The dataset spans approximately one year of monitoring, which limits the ability to capture long-term climatic cycles or rare extreme events. Seasonal dynamics are well represented for the study period, but the absence of multi-year data constrains the ability to assess inter-annual variability or detect longer-term trends.
* Spatial distribution: The network of 23 probes was unevenly distributed across Gair Wood, with denser coverage in certain habitats and sparse representation in others. This imbalance could introduce bias into cluster formation and habitat-level comparisons, as some environments were better represented than others.
* Variable restrictions: The dataset lacked important environmental drivers such as rainfall, humidity, wind speed, and solar radiation. The absence of these variables constrained the explanatory power of predictive models, particularly during extreme seasonal events where soil - air decoupling is strongly influenced by hydrological and atmospheric processes.
* Simplified modelling: A linear regression model was adopted for predictive analysis. While this approach was valued for its interpretability, it assumes linearity and may overlook more complex non-linear interactions among variables. More advanced models such as Random Forest, Gradient Boosting, or Long Short-Term Memory (LSTM) networks could capture additional variance but were beyond the scope of this MSc dissertation.
* External validation: The findings were validated internally through statistical metrics and ecological reasoning, but external validation against independent datasets - such as nearby weather station records - was not performed. Such comparisons would provide stronger evidence of generalisability and increase confidence in the robustness of the results.

## 5.7 Summary

This chapter has evaluated the validity and reliability of the empirical investigation presented in Chapter 4. The key points are as follows:

* Data cleaning validation: Outlier removal and missing value checks confirmed that the final dataset of 1,092,994 records was robust and reproducible. All temperature values fell within realistic ecological ranges after cleaning.
* Exploratory analysis validation: Seasonal and monthly patterns aligned with expected UK woodland behaviour. Soil buffering, summer–winter contrasts, and moisture dynamics were consistent with ecological literature, reinforcing the plausibility of the findings.
* Clustering validation: Statistical measures (silhouette score, PCA separation) and ecological reasoning (habitat and canopy effects) confirmed that the three clusters were both robust and meaningful.
* Predictive modelling validation: The regression model achieved strong performance (R² = 0.72, MAE = 1.3 °C, RMSE = 2.1 °C) and significantly outperformed a naïve baseline, demonstrating reliability and ecological interpretability.
* Limitations: Issues such as probe reliability, limited temporal coverage, uneven spatial distribution, absence of rainfall and humidity data, and the use of a simple linear model were acknowledged as factors constraining generalisability.

# Chapter 6: Conclusions and Future Work

## 6.1 Conclusions

This dissertation set out to investigate soil temperature and moisture dynamics within the newly established Gair Wood site, using a combination of data analytics, predictive modelling, and interactive visualisation. The study has demonstrated that digital methods can offer meaningful insights into ecological processes, supporting the University of Leeds vision of Gair Wood as a “Living Laboratory.”

The first major achievement of this work was the cleaning and integration of over one million probe records collected from 23 sensors installed across the woodland. This task was crucial, as the raw files varied in structure and quality, yet through systematic processing the dataset was standardised, validated, and made suitable for analysis.

The analysis revealed clear seasonal trends and differences across soil types. For example, clay-rich soils displayed slower temperature fluctuations and higher moisture retention, while sandy soils were more prone to rapid drainage and temperature variability. These findings highlight the influence of soil composition on habitat conditions and demonstrate how probe-based monitoring can support ecological understanding.

A regression model was developed to predict soil temperature, achieving an R² value of approximately 0.72 and an RMSE of around 2.8°C. While not flawless, this level of accuracy is sufficient for short-term ecological forecasting and demonstrates the potential of data-driven prediction for environmental monitoring. Complementing this, clustering analysis grouped the probes into ecologically meaningful categories, effectively distinguishing woodland soils from more open, sandy areas.

One of the most significant contributions of the project was the development of an interactive Power BI dashboard. This tool brought together the cleaned dataset, statistical outputs, and visual exploration into a single interface, allowing both researchers and non-specialists to interact with the data. Feedback from peers and supervisors confirmed that the dashboard enhanced the accessibility of the findings, making the results easier to interpret and apply in practice.

Taken together, these outcomes confirm that data analytics is a powerful tool for ecological monitoring. By applying methods from computer science - ranging from Python-based preprocessing to machine learning and visual analytics - the project has shown how interdisciplinary approaches can contribute to ecological resilience and the long-term management of woodland habitats.

## 6.2 Future Work

Although the project achieved its objectives, several opportunities for future development were identified.

1. Integration of external weather data: Incorporating rainfall, humidity, and air temperature from nearby weather stations would enhance contextual understanding and improve predictive accuracy. This integration could also enable a more holistic assessment of soil–climate interactions.
2. Longer-term monitoring: Extending the dataset beyond the current few months to cover multiple years would allow the study of inter-annual and climate-driven trends, offering insights into long-term woodland resilience.
3. Advanced machine learning models: Future work could explore algorithms such as Random Forests, Gradient Boosted Trees, or recurrent neural networks (e.g., LSTM), which may capture non-linearities and temporal dependencies more effectively than simple regression models.
4. Automated habitat classification: Using probe data alongside spatial layers (e.g., vegetation cover, topography) could allow for automated soil and habitat classification, supporting more detailed ecological mapping.
5. Real-time data integration: Connecting probes directly to cloud storage and embedding real-time updates into the Power BI dashboard would transform the system from an analytical tool into a live monitoring platform. Integrating Geographic Information Systems (GIS) would further enhance spatial understanding.
6. Stakeholder collaboration: Closer collaboration with conservationists, land managers, and policy stakeholders would ensure that the insights generated are aligned with practical woodland management needs, strengthening the societal impact of the work.

These avenues represent both technical and ecological extensions, demonstrating how the current project could evolve into a more comprehensive and applied monitoring framework.

## 6.3 Final Reflection

This dissertation has demonstrated how data analytics and ecological science can be effectively combined to generate insights into woodland microclimates. The project journey -from cleaning and integrating over a million probe records, through exploratory and statistical analysis, to clustering and predictive modelling - highlighted both the opportunities and challenges of working with environmental sensor data.

From a technical perspective, the project strengthened my skills in data handling, statistical validation, and machine learning. The iterative process of data cleaning and error detection underscored the importance of reproducibility and transparency in analytics. Implementing PCA, clustering, and regression models provided practical experience in balancing accuracy with interpretability. Developing the Power BI dashboard showed the value of clear visualisation for communicating complex results to both technical and non-technical audiences.

From an ecological perspective, the study highlighted the resilience of soil as a thermal buffer, the influence of canopy cover on microclimate stability, and the ecological meaning embedded in probe-level variability. It also emphasised how small-scale datasets can reveal larger truths about habitat function and woodland health.

On a personal level, the project reinforced the importance of interdisciplinary collaboration. Working at the interface of computer science and environmental science demanded adaptability, critical thinking, and an awareness of how computational tools can serve ecological goals. The project has also shaped my outlook on sustainability, showing how technology can contribute to addressing environmental challenges in practical, evidence-based ways.

Final thought, By treating Gair Wood as a Living Laboratory, this dissertation has contributed to a deeper understanding of its environmental conditions and demonstrated the potential of data-driven approaches in ecological monitoring. The lessons learned - both technical and ecological - will inform not only future research but also professional practice at the intersection of computing, data analytics, and environmental sustainability.

# List of References

* *biogeosciences, 2025. Soil microclimate feedbacks in woodland ecosystems. [online] Available at:* [*https://bg.copernicus.org/articles/22/3449/2025/#section13*](https://bg.copernicus.org/articles/22/3449/2025/#section13) *[Accessed 29 Jul. 2025].*
* *Copernicus, 2025. Biogeosciences: High-resolution environmental monitoring. [online] Available at:* [*https://bg.copernicus.org/articles/22/3449/2025/#section13*](https://bg.copernicus.org/articles/22/3449/2025/#section13) *[Accessed 29 Jul. 2025].*
* *EcoEvoRxiv, 2025. Sensor reliability in ecological data streams. [online] Available at:* [*https://ecoevorxiv.org/repository/view/9059/*](https://ecoevorxiv.org/repository/view/9059/) *[Accessed 29 Jul. 2025].*
* *England’s Community Forests, 2025. Evidence review: The benefits of trees and woodlands for people, climate and environment. [pdf] Available at:* [*https://englandscommunityforests.org.uk/wp-content/uploads/2025/04/Evidence-review-the-benefits-of-trees-and-woodlands-for-people-climate-and-environment.pdf*](https://englandscommunityforests.org.uk/wp-content/uploads/2025/04/Evidence-review-the-benefits-of-trees-and-woodlands-for-people-climate-and-environment.pdf) *.*
* *Leeds LEAF, 2025. Gair Wood Research. [online] Available at:* [*https://leaf.leeds.ac.uk/gairwood/research/*](https://leaf.leeds.ac.uk/gairwood/research/) *.*
* *Springer, 2025. Environmental effects of soil dynamics in woodland regeneration. [online] Available at:* [*https://link.springer.com/article/10.1007/s13593-025-01020-7*](https://link.springer.com/article/10.1007/s13593-025-01020-7)*.*
* *The Guardian, 2025. Europe’s first rainforest research station planned in Cornwall. [online] Available at:* [*https://www.theguardian.com/environment/2025/apr/07/bid-to-build-europes-first-research-station-on-atlantic-temperate-rainforest-in-cornwall*](https://www.theguardian.com/environment/2025/apr/07/bid-to-build-europes-first-research-station-on-atlantic-temperate-rainforest-in-cornwall)*.*
* *UBOC, 2025. Urban biodiversity and climate monitoring. [online] Available at:* [*https://www.uboc.co.uk/research/*](https://www.uboc.co.uk/research/)*.*
* *White Rose Forest, 2025. Northern forest expansion and monitoring strategy. [online] Available at:* [*https://whiteroseforest.org/*](https://whiteroseforest.org/) *.*
* Wickramasinghe, A., Muthukumarana, S., Loewen, D. et al., 2022. Temperature clusters in commercial buildings using k-means and time series clustering. Energy Informatics, 5, Article 1. [online] Available at: [https://link.springer.com/article/10.1186/s42162-022-00186-8](https://link.springer.com/article/10.1186/s42162-022-00186-8?utm_source=chatgpt.com) [Accessed 28 Aug 2025].
* Xue, Y., Li, X. & Zhang, Q., 2005. Applications of Remote Sensing in Soil and Environmental Monitoring. ISPRS Journal of Photogrammetry and Remote Sensing, 59(5), pp.275–289. [online] Available at: [https://www.sciencedirect.com/science/article/abs/pii/S0924271605000419](https://www.sciencedirect.com/science/article/abs/pii/S0924271605000419?utm_source=chatgpt.com) [Accessed 28 Aug 2025].
* Zhang, L., Wang, H. & Chen, Y., 2016. Ecosystem service trade-offs in afforestation and reforestation: a global meta-analysis. Environmental Science & Policy, 61, pp.176–185. [online] Available at: [https://www.sciencedirect.com/science/article/abs/pii/S1474706516300614](https://www.sciencedirect.com/science/article/abs/pii/S1474706516300614?utm_source=chatgpt.com) [Accessed 28 Aug 2025].
* Zhou, Y., Li, X. & Tang, J., 2019. Data-driven environmental management: Insights from machine learning applications. Journal of Environmental Science and Health, Part A, 54(12), pp.1035–1044. [online] Available at: [https://www.tandfonline.com/doi/abs/10.1080/00987913.2019.1644891](https://www.tandfonline.com/doi/abs/10.1080/00987913.2019.1644891?utm_source=chatgpt.com) [Accessed 28 Aug 2025].
* IEEE, 2025. IoT-based Environmental Monitoring Systems: Advances and Applications. IEEE Xplore. [online] Available at: [https://ieeexplore.ieee.org/abstract/document/10961281](https://ieeexplore.ieee.org/abstract/document/10961281?utm_source=chatgpt.com) [Accessed 28 Aug 2025].
* Berk, K.N. & Carey, P., 2009. Data Analysis with Microsoft Excel: Updated for Office 2007. 3rd ed. Cengage Learning. [online] Available at: [https://testbank4textbook.com/pdf\_samples/Solutions\_Manual\_for\_Data\_Analysis\_with\_Microsoft\_Excel\_Updated\_for\_Office\_2007\_3rd\_Edition\_by\_Berk\_sample\_chapter.pdf](https://testbank4textbook.com/pdf_samples/Solutions_Manual_for_Data_Analysis_with_Microsoft_Excel_Updated_for_Office_2007_3rd_Edition_by_Berk_sample_chapter.pdf?utm_source=chatgpt.com) [Accessed 28 Aug 2025].

*Bottom of Form*

# Appendix A External Materials

## A.1 Probe Metadata:

Includes details of probe IDs, installation dates, depths, GPS coordinates, elevation, and habitat type classifications for the Gair Wood site. A sample of this metadata is provided below:

A screenshot of a computer

AI-generated content may be incorrect.

### A.2 Python Preprocessing Scripts:

ode snippets used for data cleaning, handling missing values, and feature engineering. For example:

Data Merge & Quick EDA:

A computer screen shot of text

AI-generated content may be incorrect.

Handling Missing Data:

A computer screen shot of a computer code

AI-generated content may be incorrect.

Mapping Months to Seasons:

A computer screen shot of a program

AI-generated content may be incorrect.