Dublin City University

School of Computing

Explainable Machine Learning for Knowledge Discovery in Environmental Science

by

Adam Stapleton, B.Sc.

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Supervisor: Prof. Mark Roantree
Co-Supervisor: Prof. Elke Eichelmann (School of Environmental Science, University College Dublin)

November 2025

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Adam Stapleton
ID No.: 20214892
Date: January 9, 2026

Generative AI Statement

Statement on the Use of Generative AI in this Thesis

In accordance with DCU's Position Statement on the Use of Artificial Intelligence Tools and the guidelines for responsible use of generative AI in doctoral research, I declare the following use of generative AI tools in the production of this thesis:

Tools Used

Throughout the development of this thesis, I utilized the following generative AI tools:

- **Claude (Anthropic)**
  - October 2024 - December 2024: Claude 3.5 Sonnet
  - January 2025 - October 2025: Claude 4 Sonnet 4.5 and Claude 4 Sonnet 4

Applications and Use Cases

1. LaTeX Document Preparation and Formatting

- **Template Development**: AI assisted in creating and modifying LaTeX templates for consistent thesis formatting
- **Troubleshooting**: Resolution of compilation and formatting errors, particularly with bibliography management and table formatting.

2. Writing Support

- **Grammar, Spelling and Consistency Checking**: Identification and correction of grammatical errors and typos throughout the manuscript. Ensuring consistent use of terminology (e.g., XAI vs. IML) and notation throughout the thesis.
- **Writing Feedback**: Identification of sections that required clearer phrasing and improved readability while maintaining academic tone. Feedback on overall structure and flow of the draft thesis leading to identification of areas for improvement.

- **Dictaphone**: Conversion of spoken word to text for the fast addition of draft sections of the thesis.

3. Code Development and Debugging

- **Debugging**: Identification and resolution of errors in Python code including but not limited to data loading issues, memory bottlenecks, typos.
- **Code Enhancement**:
  - Refactoring code for improved readability or to make code more pythonic (i.e. using better code structure that aligns with python best practices).
  - Implementation of additional functionality to existing (e.g. adding the option to change the colormap on a mapping function, extending a function to iterate over multiple instances with the addition of error logging).
- **Documentation**: Generation of inline comments for better code maintainability

4. Literature Search and Synthesis

- **Article Identification**: Identifying and locating relevant academic papers across a range of topics for subsequent review and either synthesis or rejection for inclusion in the thesis and publication.
- **Citation Verification**: Checking the accuracy and appropriateness of references.

Prompt Examples

- **LaTeX debugging:** `I have this text at the start of my chapter but the bibentry isn't working in order to produce the full citation. Please identify a solution based on the attached main.tex file.`
- **Code Support:** `I keep getting this message to say that a port is already open but it shouldn't be. How to I make sure there's no client open before I run my code?`
- **Paper Identification:** `My supervisor suggested I get a more recent reference than the salati paper for the below section. Please suggest relevant papers with direct links to access.`

> The Amazon has also been shown to be weakening in its resilience (Forzieri et al., 2022; Chen et al., 2024a) and ability to self-support via evapotranpiration driven water recycling across the basin (Salati et al., 1979), exacerbated by anthropogenic disturbances (Wang et al., 2024).

- **Dictaphone:** I want to use this chat as a dictaphone for my discussion and conclusions section. So, essentially, what I'm gonna do is I'm going to speak want you to transcribe what I'm saying and add it into one big document. That I can then copy and paste into my Overleaf latex. Starting now. The tools I have applied in this thesis Well established well tested, scientifically grounded. And packaged into utilities that are readily available. For variety of research. The novelty comes in the application of these specifically... (Errors in spoken transcription are then corrected by the model and compiled in LaTeX format).

Iterations

- **Code Support:** After running that code I get this error:

  > Found existing client at http://127.0.0.1:40853/status Closing existing client and cluster... GPU setup failed with error: name 'LocalCUDACluster' is not defined

- **Paper Identification:** Now reviewing this document suggest the most relevant papers missing from the discussion that would be relevant to giving an introduction to the field of environmental science.

- **Dictaphone:** Add a section header for this section that will talk about how these are accessible tools that can provide useful insights that many climate researchers could and should be using.

Critical Reflection

Benefits

The use of generative AI tools provided several advantages:

- **Efficiency**: Significantly reduced time spent on formatting, debugging, and routine coding tasks.

- **Quality Enhancement**: Improved code readability and document consistency.

- **Learning**: Exposure to best practices in LaTeX formatting and Python programming.

Limitations and Challenges

Several limitations were observed:

- **Accuracy**: AI occasionally provided incorrect or outdated information, inefficient or incorrect code, requiring verification or rejection.
- **Over-reliance Risk**: Conscious effort was required to maintain independent critical thinking, independent problem solving, primary code authorship, detailed review of academic literature and text authorship.
- **Hallucinations**: The propensity for hallucination of academic articles that did not exist was avoided by prompting for direct links to access academic articles. References produced without links were ignored.

Ethical Considerations

Throughout the use of AI tools, I maintained the following ethical standards:

- All substantive intellectual contributions, research design, data analysis, and interpretations remain my own work
- AI was used as a supportive tool rather than a replacement for critical thinking and analysis
- No sensitive research data or unpublished results were shared with AI systems
- All AI-suggested content was thoroughly reviewed and adapted to ensure accuracy and originality
- The core research questions, methodology development, and scientific insights were developed independently

Declaration

I declare that while generative AI tools were used to enhance the technical quality, efficiency of production of research and presentation of this thesis, all research conception, experimental design, data collection and analysis, result interpretation, and scientific conclusions are my own original work or the work of my collaborators, where my specific contributions have been detailed in the associated Declarations of Authorship (App. E). The AI tools served primarily as sophisticated assistants

for technical improvements to code and writing alongside literature discovery, but did not contribute to the fundamental intellectual content or scientific discoveries presented in this thesis.

Signed: Adam Stapleton
Date: January 9, 2026

Acknowledgments

I would like to first and foremost express my sincere gratitude to my supervisors, Prof. Mark Roantree and Prof. Elke Eichelmann, for their continuous guidance and encouragement throughout this research. Their expertise, support and hard work have been invaluable in shaping this thesis and aiding me in opening up new opportunities for academic collaboration and new research directions.

My thanks to that end go to Dr. Mauricio Cruz Mantoani who first introduced me to his colleagues in Brazil and connected me with the vibrant research community there. I am deeply grateful to Prof. Celso Von Randow and Prof. Cléo Quaresma Dias Junior for hosting me during my placement in Brazil. The experiences I had visiting INPE and ATTO will stay with me for my whole life and I am deeply honoured to have had these opportunities and collaborations.

To Prof. Noel O'Connor and Prof. Brian MacNamee, without whom I never would have begun this research journey I am extremely grateful for all the work you have put into ML-Labs and providing these opportunities. Deepest thanks go also to Angela Lally, whose seemingly magical ability to solve any practical issues and endless knowledge of byways of DCU's internal systems were invaluable in navigating this Ph.D. Thanks also to everyone in ML-Labs, particularly Vicky Flanagan, Carla Naltchayan and Antonella Ferrecchia, without whom we would all be sitting twiddling our thumbs rather than successfully completing our doctoral studies.

And finally to my family and friends, thanks to all of you for getting me through. In particular my Mum, for not only proof-reading my thesis but in supporting me at every stage of this journey and always encouraging me to pursue higher education. I think I probably have enough degrees now.

Funding

This work was supported by Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183) and through the Insight SFI Research Centre for Data Analytics at Dublin City University (Grant Number 12/RC/2289_P2, co-funded by the European Regional Development Fund). Additional support was provided by the Ministry of Science, Technological Development and Innovations of Republic of Serbia (451-03-1524/2023-04/16) and by COST Action CA20108, supported by COST (European Cooperation in Science and Technology).

Data Availability and Acknowledgements

My sincere thanks go to all researchers involved in the collection of these data, without whom none of this reseach would have been possible.

AmeriFlux Eddy Covariance Data. Eddy covariance data from wetland sites were obtained from the AmeriFlux network, made available under the AmeriFlux CC-BY-4.0 License through https://ameriflux.lbl.gov/. Specific DOIs for each site dataset: Twitchell Wetland West Pond AmeriFlux ID US-TW1 (https://doi.org/10.17190/AMF/1246147), Twitchell East End Wetland AmeriFlux ID US-TW4 (https://doi.org/10.17190/AMF/1246151), Mayberry Wetland AmeriFlux ID US-MYB (https://doi.org/10.17190/AMF/1246139), and Sherman Island Restored Wetland AmeriFlux ID US-Sne (https://doi.org/10.17190/AMF/1418684). Funding for the AmeriFlux data portal was provided by the U.S. Department of Energy Office of Science. I would like to thank the site PI Dennis D. Baldocchi for enabling access to this extensive and high-quality dataset, and the technicians Joseph Verfaillie and Daphne Szutu, along with the many Berkeley Biometeorology Lab postdocs, PhD students, and field assistants for their dedicated work in maintaining the field sites and collecting and processing the data.

ERA5 Reanalysis. Atmospheric, land surface, and energy balance variables were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis dataset (Hersbach et al., 2020). Monthly data (1950–2022) at 0.1° spatial resolution are freely available through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/) following user registration.

MERGE Precipitation. High-resolution precipitation estimates combining gauge observations with satellite data were obtained from the MERGE dataset (Rozante et al., 2010). Data are available through the Center for Weather Prediction and Climate Studies (CPTEC/INPE) at http://ftp.cptec.inpe.br/modelos/tempo/MERGE/.

MapBiomas Land Cover. Land use and land cover classifications (Collection 6.0 and Collection 8) were derived from the MapBiomas project (Souza et al., 2020; MapBiomas Project, 2022). Annual maps (1985–2020) at 30 m resolution are freely available at https://mapbiomas.org/ for non-commercial research use.

SRTM Topography. Terrain characteristics were derived from the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global digital elevation model, distributed by the U.S. Geological Survey (**?**NASA, 2013). Data are freely available through https://earthexplorer.usgs.gov/.

PRODES Deforestation. Annual deforestation data (2000–2017) were obtained from Brazil's National Institute for Space Research (INPE) Program for Calculating Deforestation in the Amazon (PRODES). Data are publicly available at

http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/amazon/.

ATTO Boundary Layer Height Data. Boundary layer height data for the ATTO site (5-minute averaged and upsampled z_i data) are available through the ATTO data repository (ATTO Data Repository, 2024). I acknowledge the Max Planck Society and the Instituto Nacional de Pesquisas da Amazônia (INPA) for continuous support. For the operation of the ATTO site, I acknowledge the support by the German Federal Ministry of Education and Research (BMBF contract nos. 01LB1001A, 01LK1602B and 01LK2101B) and the Brazilian Ministério da Ciência, Tecnologia e Inovação (MCTI/FINEP contract 01.11.01248.00) as well as the Reserva de Desenvolvimento Sustentável Uatumã (SDS/CEUC/RDS-Uatumã). CQDJ acknowledges the support from CNPq (Processes: 307530/2022-1; 406884/2022-6; 406307/2023-7). I would like to especially thank all the people involved in the technical, logistical, and scientific support of the ATTO project and the field staff.

GoAmazon Campaign Data. The 16-second ceilometer-based z_i data for the T3 site are available through the GoAmazon campaign (ARM Research Facility, 2014a,b). The T3 Data were obtained from the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research Program.

CloudRoots, LBA, and CAFE-Brazil Campaign Data. Data from the CloudRoots campaign are available through (CloudRoots Campaign, 2024). The LBA data archive can be accessed via the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) (Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC), 2024). CAFE Brazil campaign data are available through the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory Data Repository (NCAR Earth Observing Laboratory, 2024). I acknowledge the important contributions of these major research initiatives in the Amazon region: CloudRoots, the Large Biosphere Atmosphere Experiment in Brazil (LBA), and CAFE-Brazil (Chemistry of the Atmosphere: Field Experiment in Brazil).

Forest Inventory Data. Tree density, species diversity, and richness estimates were derived from the Amazon forest inventory database compiled by ter Steege et al. (ter Steege et al., 2023). Access to plot-level data requires approval from the Amazon Tree Diversity Network, and data requests should be submitted through http://www.biodiversityresearch.org/research.php with a detailed research proposal. I would like to thank the Amazon Tree Diversity Network and the numerous field researchers, botanists, and institutions across the Amazon basin who contributed to this invaluable forest inventory database spanning decades of fieldwork.

Soil Phosphorus Data. Gridded estimates of available, organic, and total soil phosphorus were obtained from Random Forest model predictions developed for

tropical South America by Darela-Filho et al. (Darela-Filho et al., 2024). Access to these data requires permission from the original modeling team. I thank the researchers who developed and made available these critical soil nutrient datasets.

Wood Density Data. Community-weighted mean wood density estimates were derived from the global wood density database compiled by Mo et al. (Mo et al., 2024). Processed gridded data are available upon request from the corresponding author. I acknowledge the extensive work involved in compiling wood density measurements from forest plots worldwide.

Environmental Disturbance Data. Edge effects, fire occurrence, drought stress, and logging intensity data were obtained from the environmental disturbance dataset compiled by Lapola et al. (Lapola et al., 2023). I thank the research team for developing this comprehensive assessment of anthropogenic and climatic disturbances across the Amazon basin.

Gross Primary Productivity Data. GPP estimates used as the target variable were obtained from the GOSIF product developed by Li and Xiao (Li and Xiao, 2019) and can be accessed with permission from the authors through the Global Ecology Data Repository (https://globalecology.unh.edu/data/GOSIF-GPP.html, accessed January 2025). I thank the developers of the GOSIF product for making this high-quality global GPP dataset available to the research community.

I gratefully acknowledge all field researchers, data collectors, and institutions that contributed to the datasets used in this study.

Software and Tools

The algorithms used in this study were implemented in Python (Python Software Foundation, 2024), using the scikit-learn package for general machine learning functionality (Pedregosa et al., 2011a), as well as standalone packages such as LightGBM (Ke et al., 2017) and XGBoost (Chen and Guestrin, 2016).

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AWS	Automated Weather Station
BLH	Boundary Layer Height
CNN	Convolutional Neural Network
DP	Dew Point
ERA-5	European Reanalysis, Fifth Generation
ESS	Earth System Science
ET	Evapotranspiration
GPP	Gross Primary Productivity
LGB	Light Gradient Boosting
LSM	Land Surface Model
MAE	Mean Absolute Error
MBE	Mean Bias Error
ML	Machine Learning
NN	Neural Network
nRMSE	Normalised Root Mean Square Error
R^2	Coefficient of Determination
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Square Error
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence
XGB	eXtreme Gradient Boosting
XML	eXplainable Machine Learning

List of Publications

This thesis is based on the following publications and manuscripts:

Publications as First Author

1. **Stapleton, A.**, Eichelmann, E., & Roantree, M. (2022). A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning. *Applied Computing and Geosciences*, 16, 100105. **[Chapter 3]**

2. **Stapleton, A.**, Dias-Junior, C. Q., Von Randow, C., D'Oliveira, F. A. F., Pöhlker, C., de Araújo, A. C., Roantree, M., & Eichelmann, E. (2025). Intercomparison of machine learning models to determine the planetary boundary layer height over Central Amazonia. *Journal of Geophysical Research: Atmospheres*, 130(6), e2024JD042488. **[Chapter 4]**

3. **Stapleton, A**. Aline Anderson de Castro, Celso Von Randow, Mark Roantree, and Elke Eichelmann (2026). Discovering Regional Vulnerability Patterns of Gross Primary Productivity in Amazon Rainforests with XAI.*[Journal of Geophysical Research: Biogeosciences]*. [Status: In preparation]˙[Chapter 5]

Publications as Contributing Author

1. Lalic, B., **Stapleton, A.**, Vergauwen, T., Caluwaerts, S., Eichelmann, E., & Roantree, M. (2024). A comparative analysis of machine learning approaches to gap filling meteorological datasets. Environmental Earth Sciences, 83(24), 679. **[Chapter 2]**

Conference Presentations

1. Cuong, D. V., Le-Khac, P. H., **Stapleton, A.**, Eichlemann, E., Roantree, M., & Smeaton, A. F. (2022). Managing Large Dataset Gaps in Urban Air Quality Prediction: DCU-Insight-AQ at MediaEval 2022. arXiv preprint arXiv:2212.10273.

2. Eichelmann, E., **Stapleton, A.**, Foley, P., & Roantree, M. (2023, May). Evapotranspiration Partitioning Using Machine Learning. In *35th Conference on*

Agricultural and Forest Meteorology/14th Fire and Forest Meteorology Symposium/Sixth Conference on Atmospheric Biogeosciences. AMS.

3. **Stapleton, A.**, Dias, C. Q., von Randow, C., Roantree, M., Eichelmann, E., & D'Oliveira, F. A. F. (2025, January). An Intercomparison of Machine Learning Models to Determine the Planetary Boundary Layer Height Using Surface Level Measurements. In *105th Annual AMS Meeting 2025* (Vol. 105, pp. 448779).

Author Contributions: The candidate led the conceptualization, experimental design, implementation of all machine learning models, data preprocessing and analysis, and manuscript preparation for each first-author publication. Specifically, this included: developing code, machine learning frameworks for comparing multiple algorithms and feature selection methodologies; implementing recursive feature elimination with feature importance heuristics; conducting clustering analysis with SHAP explanations for regional pattern identification; discovering novel relationships between environmental variables and target outcomes; creating figures and statistical analyses; and drafting complete manuscripts across all sections. The candidate's contributions ranged from 70-80% across the three first-author publications. For the contributing author publication, the candidate developed the novel gap creation methodology and experimental design (including feature set comparisons), implemented all machine learning models (Random Forest, LightGBM, Linear Regression) and reference spatial interpolation, conducted data preprocessing and primary data analysis including feature importance assessment, and drafted the manuscript including introduction, related research, methodology, results and initial discussion sections (60% contribution).

Abstract

Title:*Explainable Machine Learning for Knowledge Discovery in Environmental Science* **Author:***Adam Stapleton*

With climate change increasingly impacting communities and around the globe, understanding the Earth system has never been more vital. The many petabytes of data that are generated from remote sensing efforts such as satellite observations and ground-based measurement networks have made the fields of environmental science and Earth system science ripe for the application of machine learning. The fundamental question that drives the research presented in this thesis is how ML models can be meaningfully integrated into environmental science workflows to enhance our understanding and modelling of complex Earth system processes. This thesis demonstrates through three applications the utility of machine learning for the modelling of complex systems in environmental science. In addition these works demonstrate that explanations of the models enable the discovery new relationships within these systems purely from data with little or no prior knowledge. The first application is to the partitioning of evapotranspiration into its components, evaporation and transpiration by predicting the evaporation component from data where only the total evapotranspiration is measured. The data used are from four wetlands in California and served as a preliminary study that introduces the framework for the methodology that is used for the other studies. A key finding uncovered by simple model explanations is that methane flux, a feature whose relationship with evapotranspiration is not generally examined, may contribute to further biophysical process understanding. The second application is to the local-scale modeling of the atmospheric boundary layer height in central Amazonia. This study found gradient boosted ensemble models using all available features to perform best. A modified recursive feature elimination algorithm identified minimal sets of 5–7 surface measurements sufficient for accurate boundary layer height prediction, demonstrating potential for wider spatial monitoring using cost-effective sensors. The study revealed previously unrecognized variables that strongly contributed to boundary layer height predictions, such as deeper soil temperature measurements (40 cm). The final application is to the large-scale modelling of gross primary productivity in primary forest the Amazon basin utilising clustering methods to separate similar regions and understand the complex relationship between climatic variables with gross primary productivity in those regions. Shapely explanations enabled deeper understanding of the relationships discovered by the machine learning models, in particular identifying the vulnerable peripheral regions with increased variability

in gross primary productivity under the influence of deforestation and degradation pressures.

Chapter 1

Introduction

Climate change is one of the defining challenges of the 21st century and one of humanity's greatest existential threats. The extent of our influence on the makeup of our planet, its atmosphere and ecosystems has led to the informal emergence of a new geological epoch defined by the impacts of human activities: the Anthropocene (Crutzen, 2002).

Simultaneously, humanity is witnessing unprecedented advances in artificial intelligence (AI) that for the first time match and exceed human capabilities on numerous cognitive tasks (Bubeck et al., 2023). These advances have largely been driven by progress in the field of Machine Learning (ML), particularly the development of deep learning, enabling the modeling of highly complex and non-linear functions (LeCun et al., 2015). Recent years have witnessed transformative developments including transformer architectures that have revolutionized natural language processing (Vaswani et al., 2017), diffusion models that generate photorealistic images and videos from text descriptions (Rombach et al., 2022; Xing et al., 2024), and foundation models for a diverse range of tasks such language, vision, robotics, reasoning and human interaction (Bommasani et al., 2021). Large language models have demonstrated human-level performance on a variety of language based tasks, including chained reasoning on multidisciplinary graduate level tasks, coding and multi-modal medical reasoning (OpenAI, 2023; Zhao et al., 2023; Guo et al., 2025; Wang et al., 2025). In the medical sciences, AI systems have achieved remarkable breakthroughs: the previously open problem of accurately predicting a protein's 3D structure (relevant to research on genetics, drug discovery and the understanding of diseases) was solved by AlphaFold (Jumper et al., 2021) and AI models now exceed human radiologists in medical image analysis (McKinney et al., 2020). While human strategic games like chess were previously solved by traditional computing systems, AI systems have recently demonstrated superhuman performance in complex strategic games like Go and StarCraftII (Silver et al., 2016; Vinyals et al., 2019).

It is the nature of any transformative technology such as AI that it presents as much as an opportunity for humanity as it does a potential risk (Kokotajlo et al., 2025). The opacity of modern AI systems—their "black box" nature poses significant challenges for scientific understanding and high-stakes decision-making. This is particularly concerning in environmental science, where model predictions inform

critical policy decisions affecting billions of people and planetary-scale systems.

The pressing need for understanding the complex and interwoven systems that constitute our natural environment has never been more urgent. Climate tipping points, biodiversity loss, and ecosystem collapse require not only accurate predictions but also deep mechanistic understanding of the underlying processes (Steffen et al., 2015). The massive datasets generated by decades of scientific research, satellite observations, and sensor networks have created unprecedented opportunities for ML applications, but only if we can interpret and trust the insights these models provide.

Despite marked advances in environmental ML applications, significant gaps remain in low-data scenarios, model interpretability, and integration of ML models into larger Earth system modelling frameworks. Understanding the inner workings of these "black boxes" is essential not only to improve trust in ML models but to ensure their generalizability to future climate scenarios that may not be represented in training data (Reichstein et al., 2019).

Ultimately, the utility of any model lies in how humans can use the information and understanding it provides. This ranges from uncovering new understanding of the mechanistic functioning of complex systems, to making predictions about future climate and ecosystem states, to informing policy decisions for climate change adaptation and mitigation. The propensity of ML systems to process data volumes that are practically infeasible for humans enables discovery of relationships and processes in systems that may be missing in models based on top-down theoretical and mathematical approaches. In contrast, the modelling of these relationships, however accurate in the context that the models are trained and tested in, do not automatically confer mechanistic understanding of these complex systems or the processes therein. That is to say that there is no guarantee the models will hold for a variation of the same system, such as in a different location or under climate change (Beucler et al., 2024). This change in the data used for model inference with respect to the training data is referred to as a distribution shift (Koh et al., 2021).

It is therefore the aim of this thesis is to bridge the gap between bottom-up, data-driven approaches to modeling and top-down theoretical frameworks that have historically dominated Earth system science. The applications presented will demonstrate the broad applicability of ML algorithms paired with explainability techniques to not only accurately model complex environmental systems but to discover new hypotheses about the fundamental processes governing these systems.

1.1 Background on Machine Learning

Throughout the thesis the term ML will be used to indicate a computer program that improves via some process that changes its internal composition, referred to as

learning (Mitchell, 1997).
The field of ML has its roots in statistics and computer science, with stastistical models such as linear regression included in the parthenon of ML algorithms despite having been first introduced in the 19th Century (Legendre, 1806; Gauss, 1877). Most foundational work in ML was undertaken in the middle of the 20th Century with the advent of modern computing. Artificial Neural Networks (NNs), logical models that mimic the structure of neurons in the brain, were first introduced by McCulloch and Pitts (1943) and the extension of this architecture to include many hidden layers of neurons would later enable the deep learning revolution.

The field of ML and AI entered what is called the AI winter in the 1970s and 1980s after failing to deliver on the promise of human-level intelligence expected by early pioneers such as Turing (Russell and Norvig, 2020). The advent of the internet and advanced computing technologies in the late 20th Century introduced vast volumes of data that enabled the training of much more complex models (Halevy et al., 2009). Deep learning (i.e. the application of NNs with many hidden layers), by virtue of their ability to approximate any continuous function (Hornik et al., 1989) led to the rapid advancement of the sub-disciplines of natural language processing, image recognition and generative AI (Schmidhuber, 2015). Though deep learning heralded a new era in advancement in the capabilities of ML-based systems, it also introduced new problems. The complex, layered nature of the operations occurring in a deep NN make them opaque to human understanding. This lack of understanding has led to many cases of "successful" AI systems finding shortcuts by way of spurious correlations in the data, such as image classifiers that used watermarks, text overlays or background elements instead of the visual features that should have been learned (Lapuschkin et al., 2019).

The tension between model accuracy and inherent model interpretability has been a persistent challenge in ML development since the early days of statistical models that were the precursor to modern ML. Ensemble methods emerged as a compromise to this tension as well as a response to the computationally intensive NNs, combining multiple simple models to produce complex functions. Ensembles have some degree of intrinsic interpretability by virtue of the ability to take statistics across sub-models. Random forests, introduced by Breiman (2001), provided built-in feature importance measures by examining statistics across constituent decision trees. For the scale of most modern problems, particularly within the field of environmental science, complex models appear to be unavoidable in order to accurately model systems with both the generalisability and specificity required by the intrinsic complexity of these systems (Bonan and Doney, 2018; Reichstein et al., 2019). The computational expense, energy requirements, and extensive hyperparameter tuning needed for deep learning architectures have motivated continued interest in efficient

ensemble methods (Strubell et al., 2019; Schwartz et al., 2020; Fan et al., 2023). The advancement of ensembles via XGBoost (Chen and Guestrin, 2016) led to their wide proliferation in the 2010s due to its exceptional performance in a wide number of ML competitions (Nielsen, 2016), subsequently followed by the LightGBM model with increases in computational efficiency (Ke et al., 2017). With the great promise of ML models in mind it is next most important to understand what the current state of the art is in interpreting the inner workings of these models in order to improve their accuracy and reliability and to ensure that models learn meaningful patterns rather than spurious correlations. This is particularly important in environmental science where mechanistic understanding is often as important as or more important than accuracy, both for its intrinsic values as well as the propensity for the models to generalise.

1.2 Interpretability and Explainability of ML Models

Can a human understand the outputs of a ML model and the internal processes that produce them? This question belies the heart of interpretability and explainability.

The nascency of this field has led to broad disagreement as to some basic definitions of what constitutes interpretability or explainability. There is overlapping use of the terms eXplainable Artificial Intelligence (XAI), Interpretable Machine Learning (IML) and other variations that are used to define the field itself with no absolute consensus as to which should be preferred (Saeed and Omlin, 2023; Graziani et al., 2023; Schwalbe and Finzel, 2024; Saarela and Podgorelec, 2024).

Multiple systematic literature reviews have been identified including those covering XAI applications (Saarela and Podgorelec, 2024), taxonomies (Graziani et al., 2023; Schwalbe and Finzel, 2024) and meta-analyses of challenges and future research directions (Saeed and Omlin, 2023). One of the most commonly identified needs for XAI and IML research across multiple studies is that of increased formalism, including definitions of terms (Saeed and Omlin, 2023).

The field is heterogeneous in terms of its terminology and still evolving, and certain terms are more applicable in certain contexts. “Interpretable AI/ML” seems to be more prominent in the scientific community whereas “Explainable AI/ML” is used more in a public setting (Adadi and Berrada, 2018). Notwithstanding, the term that will be used to refer to ML models with explanations throughout this thesis will be XAI. The term has gained general usage in both academia and industry as a catch-all term for the use of explanations with intelligent computing systems. Though the more accurate term in this context may be "Explainable ML", deference

to the popular nomenclature gives this research better alignment and discoverability within the context of the existing research.

XAI techniques can be categorised by three defining features: *when* the explanation happens (Timing), whether the explanations are *global or local* (Scope) and whether the explanation can be applied to *any* model or whether it is *model specific* (Applicability) (Saarela and Podgorelec, 2024). XAI techniques can also be classified by their explanation mechanism which may rely on feature importance (Fisher et al., 2019; Saarela and Jauhiainen, 2021), attribution or saliency (for images, with pixels as features) (Simonyan et al., 2013; Fong and Vedaldi, 2017; Zintgraf et al., 2017), counterfactuals (Wachter et al., 2017), examples and prototypes (Koh and Liang, 2017; Li et al., 2018b; Yeh et al., 2018), rules (Towell and Shavlik, 1993; Mitra and Hayashi, 2000; Castro et al., 2002), surrogate models (Ribeiro et al., 2016; Lundberg and Lee, 2017a), and visualizations, including of outputs of hidden layers of neural networks (Erhan et al., 2009; Zeiler and Fergus, 2014).

Of these techniques SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017b) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) are the most widely used across applications in all domains (Saarela and Podgorelec, 2024). Though LIME has the advantage of being model-agnostic it has been noted that LIME lacks the stability of SHAP as it fails to produce the same explanations for the same inputs consistently (Saarela and Geogieva, 2022). SHAP has the advantages of having both model-agnostic and model-specific explainers with local explanations used to create global explanations in a way that is mathematically guaranteed (Lundberg and Lee, 2017b).

Applications of XAI have been mostly concentrated in the health domain due the high-requirement for trust, safety and accountability in the use of ML models in this domain (Saarela and Podgorelec, 2024). The second largest domain for XAI applications is environmental science, with over 15% of applications surveyed carried out in this field (Saarela and Podgorelec, 2024). An outline of the applications of XAI in environmental science will be given in Section 1.5.1.

1.3 An Overview of the Research Application Area: Environmental and Earth System Science

Environmental science is the study of the natural environment including its physical, chemical, geological and biological components and how they interact. It is a multidisciplinary field, drawing on multiple areas of scientific investigation in order to understand the Earth system and how it produces weather and climate on local and global scales. It is heavily interrelated with the field of Earth Systems Science

(ESS) with the distinction that ESS is concerned more with Earth as entire system in a holistic, global sense, including the coupling and interactions of phenomena across scales. In contrast environmental science as a discipline is concerned with phenomena in the natural environment on all scales and less focused on the system as a whole. These two fields of study are deeply interrelated and overlapping. The research of phenomena at local and synoptic scales feeds into process-based modelling at regional, continental and climatic scales which in turn enables the modelling of the Earth System as a whole. The global and decadal evolution and interactions between these larger scale phenomena requires its own treatment and pertains more to ESS. This modelling, both computational and theoretical enables all predictions of future states of the Earth's climate and ecosystems and our understanding of their development and implications for humanity and life on Earth.

Two of the most intensively studied components of the Earth system are the carbon and water cycles. The study of both ranges from scales of leaf-level photosynthesis involving carbon respiration (and therefore capture of CO_2 and storage as C) and transpiration (and therefore water release) to ecosystem-wide budgets of energy, water and carbon (Bonan, 2008; Beer et al., 2010; Reichstein et al., 2013). Carbon budgets and anthropogenic influences on these have been widely studied due to the role of CO_2 as a greenhouse gas, with the Global Carbon Project accounting for human influences on sinks and sources of carbon in the biosphere (including plant matter and soils), lithosphere, cryosphere, atmosphere and oceans (Friedlingstein et al., 2023).

Understanding of water dynamics is important not only for the management of water resources but for ecosystem health, cloud processes and the exchange of energy between the land and the atmosphere (Oki and Kanae, 2006; Huntington, 2006; Trenberth et al., 2007). The modeling of vegetation dynamics is critical to both carbon and water cycles (Poulter et al., 2015; Bonan and Doney, 2018).

Other areas of research that are critical to the understanding of the Earth system and future climate include methane cycles (a greenhouse gas with a 100-year global warming potential 28-34 times more potent than CO_2) (Myhre et al., 2013; Saunois et al., 2020), land use and land cover change (Lawrence et al., 2012), modeling of the cryosphere (Bamber et al., 2018), ocean circulation dynamics (Caesar et al., 2018), cloud formation and dynamics (Bony et al., 2015), and ecological processes and human influence on biodiversity (Chapin III et al., 2000; Chen et al., 2011; Jenkins et al., 2013).

1.3.1 Process-based Models

The majority of Earth system modelling has historically been carried out using process-based models and therefore our best estimates of the current and future state of the Earth system depends on these models alongside observations. Process-based models attempt to capture the physical, chemical, biological, geological and, increasingly anthropogenic processes that make up the Earth system from first-principles. To this extent they explicitly encode understanding of these systems rather than learning them from data, though the understanding of these relationships was discovered empirically before being encoded in theoretical, mathematical and numerical models.

Land Surface Models (LSMs) simulate energy, water and carbon exchange between the land surface and the atmosphere by modelling vegetation dynamics, soil processes and hydrology. Examples include the Community Land Model (CLM) (Lawrence et al., 2019) and the Joint UK Land Environment Simulator (JULES) (Best et al., 2011). Earth System Models (ESMs) simulate environmental processes and their interactions on global scales over multi-decadal to centennial timescales by coupling atmosphere, ocean, sea ice, and land components into integrated modeling systems (Eyring et al., 2016). Examples include the Community Earth System Model (CESM) (Danabasoglu et al., 2020) and the Hadley Centre Global Environmental Model (HadGEM) (Kuhlbrodt et al., 2018; Williams et al., 2018).

These models and many others are ensembled into the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) to give best estimates of the state of the Earth system under future climate change. This in turn informs the Intergovernmental Panel on Climate Change (IPCC) reports on future climate scenarios (IPCC, 2023), the most comprehensive and globally integrated collection of the scientific community's best efforts to peer into the future of our planet and attempt to influence its course for the better.

Limitations of Process Based Models

One of the greatest challenges in Earth system modelling and environmental science is the divergence of LSMs and ESMs in their predictions of future climate states and the timelines to reach them (Zhang and Chen, 2021). This occurs due to the accumulation of errors in estimations when considering many complex components all interacting in a highly coupled and interdependent way (Hawkins and Sutton, 2009; Deser et al., 2012).

Structural uncertainties in these models arise from incomplete understanding of the underlying processes and the need to simplify or approximate models for computational tractability (Palomas et al., 2024; Peatier et al., 2024). Scale mismatches

between the development of models and their application create various problems. For example, models of forests (whose fundamental processes are studied at a leaf, tree or local level) are scaled to entire ecosystems without guarantees that these models will represent the system well at that scale (Bonan et al., 2014; Fisher et al., 2018). In addition, large-scale models must be parameterised on a regular grid for numerical simulation. This introduces uncertainty due to processes that occur at scales lower than the grid resolution, such as turbulence or vegetation dynamics (Pitman, 2003; Avissar and Pielke Sr, 2006; Qin et al., 2023).

In addition, there are many model parameters that cannot be derived theoretically but have to be estimated or calibrated from limited data (Hourdin et al., 2017). Computational constraints limit both the complexity and resolution of models (Lawrence et al., 2019; Balaji, 2021), resulting in trade-offs between process detail, scale and computational capacity.

These limitations have motivated hybrid approaches (See and Adie, 2021) that combine process-driven and data-driven modelling, setting the stage for the ML applications discussed in this thesis which will be further motivated and outlined in Section 1.7.

1.4 Machine Learning in Environmental Science

ML in its simplest form (linear regression) has been used in meteorology since as early as the mid-20th Century (Malone, 1955). The revival of NNs brought a new wave of ML applications in ESS that matched or exceeded the state of the art in traditional modelling approaches in cloud classification (Lee et al., 1990), rainfall prediction (French et al., 1992) and the land-cover classification of multispectral remote sensing images (Benediktsson et al., 1990) among many others. These methods demonstrated that NNs could capture complex, non-linear relationships where traditional statistical methods were inferior and where traditional numerical techniques were more computationally expensive. With the introduction of random forests (RF) (Breiman, 2001), many areas of environmental science, climate science and ESS saw breakthroughs that became integrated into regular operations in a broad array of applications. For example, Gislason et al. (2006) provided an influential early demonstration of RF for land cover classification, contributing to RF becoming a widely adopted method in operational remote sensing applications.

The deep learning revolution led to many advances that ushered in a new era of ML applications in climate and environmental science. Liu et al. (2016) applied deep convolutional NNs (CNNs) to the detection of extreme weather events in climate datasets.This paper was significant as it demonstrated that CNNs could remove the need for manual feature-engineering, a highly complex and time-consuming task

by replacing subjective expert-defined thresholds with data-driven pattern recognition. The pace of progress due to ML has been staggering. In 2021, the question was asked: "Can deep learning beat numerical weather prediction?" (Schultz et al., 2021). Just two years later a team of researchers at Google DeepMind unveiled GraphCast (Lam et al., 2023), a spherical-graph-based NN architecture that matched the predictive accuracy of the European Centre for Medium Range Forecasts (ECMWF) best numerical models with orders of magnitude in decreases in computational resource requirements. However, it cannot be understated that these breakthroughs were made possible by the remarkable efforts of the large team of researchers at the ECMWF to unify the incredibly diverse and voluminous data pertaining to weather and Earth System observations into ERA-5 (Hersbach et al., 2020) and ERA-5 Land (Muñoz-Sabater et al., 2021) which were used as the training data for GraphCast. ERA-5 is the fifth generation of atmospheric reanalysis of the global climate, produced by the assimilation of several observational datasets from satellites, meteorological and eddy covariance stations and radiosondes. It covers the entire global atmosphere and provides spatial and temporal data products for a range of meteorological and climatological variables (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). The full breadth of the advances in ESS by way of application of ML has been the subject of multiple review papers, including the role of ML and AI in process understanding in ESS (Reichstein et al., 2019), the role of ML and AI in Earth System modelling (Reichstein et al., 2019; See and Adie, 2021; Sun et al., 2022), in combatting climate change (Rolnick et al., 2022), the prediction of climate extremes (Materia et al., 2024) and for biodiversity protection and conservation (Silvestro et al., 2022; Reynolds et al., 2025).

1.5 The Need for Explainability in Environmental Science

The remarkable progress in ML models and the advances that have been made possible by innovations in this field are evident. ML models have been shown to be more accurate with lower computational resource needs in many applications in ESS and are able to capture complex non-linear behaviors with remarkable fidelity. With these models now taking over the roles previously fulfilled by process-driven numerical models, several issues come to the fore. With the propensity of ML models to learn spurious correlations, how do modellers know that the algorithms being deployed are genuinely modelling the underlying physical, chemical, biological and geological processes that they should be, *as they should be*? How can policymakers trust the outputs of these models and their validity, especially considering that they

are being used to inform decisions that affect millions of people and their livelihoods? Do we know that these models will generalise to scenarios outside of the training data, such as future climate change?

This extensibility and generalisability has historically been one of the advantages of complex dynamical systems models. If a system has been well-modelled from first principles, the same relationships should hold true under any domain shift. The same is not true for ML models due to their stochastic nature and dependency on the distribution of the data they were trained on.

Huang et al. (2025) identified four main stakeholders for XAI in ESS and their needs:

- *Policymakers* require trust in the AI models to make informed decisions that affect large populations and ecosystems.
- *Forecasters* need to be able to understand the predictions the model is making, both to develop trust in the model outputs and to assess accuracy and uncertainty in operational settings.
- *AI modellers* require tools to diagnose and improve models, identifying biases, errors, and areas for enhancement in model performance and reliability.
- *Geoscientists* need to identify and understand the underlying processes that are identified in data in order to improve understanding in their domain and enhance both theoretical and numerical process-based models.

ML approaches have the potential to revolutionise the fields of environmental science and ESS, but integrating ML into global climate and environmental models requires the ability to diagnose and assess functioning of the models and provide assurance as to their accuracy and integrity.

1.5.1 Existing Work on Interpretability in Earth System Science

XAI, IML and XML methods have seen an increase in use in environmental science applications in recent years with a 90% growth rate from 2015 to 2023 (Schiller et al., 2025). Of these 575 articles, 135 used Shapely values or SHAP as the method of explanation with the next most popular being feature importance or permutation feature importance (27 and 18 articles respectively). Other popular methods included Partial Dependence Plots (22), LIME (21), Saliency Maps (15) and accumulated local effect plots (10) (Schiller et al., 2025). Schiller et al. (2025) raise the question as to whether the popularity of SHAP is due to its versatility, accessability

and familiarity rather than it being the best suited XAI tool for applications within ESS.

The majority of XAI applications in ESS are focused on communicating model decisions, diagnosing and improving models and providing scientific insights (Huang et al., 2025) with only a small number of applications specifically intended to improve trust in models (Schiller et al., 2025).

Huang et al. (2025) found that the majority of applications in ESS utilizing XAI have been in the areas of hydrology (focusing on water cycle processes, streamflow prediction, groundwater modeling, and hydrological partitioning), land (encompassing soil science, vegetation dynamics, land-atmosphere interactions, and terrestrial carbon cycling), atmosphere (including weather forecasting, atmospheric chemistry, precipitation processes, and climate pattern recognition), and hazards (such as droughts, wildfires, floods, landslides and earthquakes, where XAI helps identify risk factors and improve early warning systems).

1.6 Research Questions and Objectives

The specific problem this thesis addresses is to demonstrate that XAI can model complex processes and uncover new scientific hypotheses from data in diverse domains within environmental science. Furthermore to do so in a context that would be difficult or impossible for a human analyst to do so (without the introduction of simplification in the models) by modelling complex, real-world systems. The general approach will be to use the simplest, most computationally efficient tools first and expand on complexity as required. The ethos here is that suitable results can be obtained without the need for excessive energy expenditure or the use of models that are less accessible to environmental science researchers (Schwartz et al., 2020).

This thesis aims to answer the following research questions:

1. **RQ1:** Can ensemble models produce accurate predictions with respect to other ML and process-based models, given their known reduction in computational cost particularly in comparison with deep learning approaches?

2. **RQ2:** Do ML methods for gap-filling environmental science data improve downstream prediction tasks and do explanations assist in diagnosing and improving these models?

3. **RQ3:** Can algorithmic feature selection methods identify previously unknown relationships between environmental processes, and what insights do these discoveries provide about complex system functioning?

4. **RQ4:** Does the systematic application of ML with explanations produce reliable scientific knowledge discovery or improvements in systemic understanding when applied to diverse environmental processes across different ecosystems?

To address these research questions, the following objectives are defined:

1. **Objective 1:** Develop and validate a methodology for model development including algorithmic feature selection and model explanations.

2. **Objective 2:** Implement and evaluate the proposed approach on three distinct datasets for diverse applications.

3. **Objective 3:** Analyze and assess the results, show that they advance understanding in their specific domains and provide recommendations for future work based on these results.

1.7 Applications in this Thesis

The applications considered in this thesis pertain more to the field of Environmental Science but as the scale of the applications increases (both spatially and temporally) the relevance to ESS becomes clearer. All three applications have a specific focus on biosphere-atmosphere interactions.

The three processes that are examined in this thesis are Evapotranspiration (ET), atmospheric boundary layer dynamics (specifically Boundary Layer Height (BLH)) and Gross Primary Productivity (GPP). This progressive approach is taken to build in complexity through the three studies, tackling some of the open areas that are either not fully understood or the modelling of which could be improved by the utilisation of ML approaches. Before any of these studies are presented, a preliminary study on the application of ML to gap-filling of meteorological data is presented in Chapter 2. This work is critical due to the myriad of sources of gaps in environmental data, including those used in the other studies, and the need to fill these gaps before ML model training. The study serves to validate the methodology for gap-filling that is applied in the studies in Chapters 3 and 4.

ET is one of the most critical processes in understanding water recycling and its role on both local and large-scale weather patterns. The modelling and understanding of the components of ET (referred to as partitioning), evaporation and transpiration, is critical for process-based Earth System models as well as for applications in agriculture and the management of water resources (Oki and Kanae, 2006; Lawrence and Chase, 2007). In order to examine the methodological benefits of testing a range of ML algorithms and using simple explanation methods to extract new hypotheses about the underlying processes driving ET within this system, a set

of data for which there exists a baseline set of ML-based partitioning was selected (Eichelmann et al., 2021b). This work will be presented in Chapter 3.

Boundary layer height dynamics are extremely important for understanding the transport of volatile organic compounds (nucleating particles for cloud formation) as well as the exchange of energy, water and trace gases between the biosphere and the atmosphere. Despite the existence of sophisticated numerical models for the BLH that capture deep theoretical understanding of the BLH in many regimes (Vilà-Guerau de Arellano et al., 2015), there are discrepancies between ERA-5 and remote sensing measurements of the BLH in the Amazon region. It is proposed that data driven approaches may be able to identify previously unidentified drivers or processes and that understanding these relationships may lead to the improvement of theoretical and numerical models. This work serves as a proof of the applicability of ML methods and basic explanations to process understanding in a relatively well understood system with a clear theoretical framework. This work will be presented in Chapter 4. The final study examines the most complex application, modelling Gross Primary Productivity across the entire Amazon basin. This work retains the ecosystem focus of the Amazon rainforest from the previous study while moving to a larger spatial extent with a lower temporal resolution (though similar coverage). Shapely explanations are introduced to give a more robust and mathematically grounded method of examining the discovered relationships between drivers and GPP. This work will be presented in Chapter 5. In Chapter 6 the implications of the research presented in this thesis on these fields of study will be discussed as well as future research directions.

1.8 Contributions and Significance

The contributions of this thesis can be summarised as follows

- **Discovery of novel environmental relationships:** methane flux as connected to ET in wetlands, deep soil temperature influencing BLH dynamics, regional vulnerability patterns in Amazon GPP.
- **Novel gap creation scheme:** A novel scheme for the creation of gaps for the testing of gap-filling methods was introduced that ensures full coverage of the dataset while avoiding existing gaps in the data.
- **Empirical evidence on XAI reliability:** systematically tested ML methods with explanations, demonstrating the utility of gain-based feature importance as a diagnostic tool while demonstrating its advantages and disadvantages as an explanatory tool alongside the testing of more robust methods such as SHAP.

- **Practical guidance for environmental ML:** minimal sensor requirements for BLH prediction (5-7 variables), evidence that no universal algorithm exists even across similar sites for ET partitioning, demonstrated that simple, parsimonious and computationally efficient ensembles outperform or match other models.
- **Methodological insights:** systematic comparison of explanation methods across data scenarios, documentation of advantages and disadvantages for scientific discovery.

1.9 Thesis Structure

This thesis is organized into seven chapters:

Chapter 1 introduces the research problem and objectives.

Chapter 2 presents a preliminary study on gap-filling for meteorological datasets, applicable across other time-series data in environmental science.

Chapter 3 presents the first study and introduces the methodological framework applied to the partitioning of evapotraspiration using a dataset for which there are a baseline set of results with which to compare.

Chapter 4 presents work modelling the atmospheric boundary layer height in the Amazon region. This represents an application of equal complexity for a completely distinct application, adapting the methodology and validating its generalisability.

Chapter 5 presents the final study on gross-primary productivity modelling in the Amazon, representing the most complex and largest scale application. This chapter also introduces SHAP as the method of explanation, enhancing the integrity of the methodology.

Chapter 6 discusses the implications and limitations of the findings as well as outlining future research directions.

Chapter 7 concludes the thesis and summarises the main findings.

Chapter 2

A comparative analysis of machine learning approaches to gap filling meteorological datasets

The following chapter has been published in Environmental Earth Sciences and the accepted version is reproduced here under the copyright agreement of the publisher Springer Nature
Lalic, B., Stapleton, A., Vergauwen, T., Caluwaerts, S., Eichelmann, E., and Roantree, M. (2024). A comparative analysis of machine learning approaches to gap filling meteorological datasets. Environmental Earth Sciences, 83(24):679.

This chapter serves as validation for the ML gap-filling methods that will be applied to input data in Chapters 3 and 4. Most ML models require continuous data for training and therefore any rows with missing data must either be removed or the gaps filled with appropriate replacement. The data in the first two applications are not extensive in volume and, therefore, it is desirable to retain as much of the data as possible as it has been shown that ML performance increases directly proportionally to data volume (Halevy et al., 2009). Gaps are paramount in environmental science data due to the variety of sources of missing data and therefore any applications in this field are likely to benefit from appropriate gap-filling methods. In addition, it is also critical to avoid the biasing of the data away from certain periods that pertain to the sources of the gaps. These periods may contain vital information about certain system behaviours that may be critical for model training and generalisation.

2.1 Abstract

Observational data of the Earth's weather and climate at the level of ground-based weather stations are prone to gaps due to a variety of causes. These gaps can inhibit scientific research as they impede the use of numerical models for agricultural, meteorological and climatological applications as well as introducing analytic biases. In this research, different machine learning techniques are evaluated together with traditional approaches to gap filling automated weather station data. When filling

gaps for a specific data stream, data from neighbouring weather stations are used in addition to reanalysis data from the European Centre for Medium-Range Weather Forecasts atmospheric reanalyses of the global climate, ERA-5 Land. A novel gap creation method is introduced that provides 100% coverage in sampling the dataset while ensuring that the sampled data are randomly distributed. Gap filling across a range of different gap lengths and target variables are compared using a range of error functions. The variables selected for modelling are mean air temperature, dew point, mean relative humidity and leaf wetness. Our results show that models perform best on gap-filling temperature and dew point with worst performance on leaf wetness. As expected, model performance decreases with increasing gap length. Comparison between machine learning and reanalysis approaches show very promising results from a number of the machine learning models.

2.2 Introduction

Ground-based weather stations play a critical role in the monitoring of the Earth's weather systems and local meteorology as well as providing reference data for climatological and Earth system models (Peterson and Vose, 1997). Before the advent of satellite measurements, these data were the only record of the evolution of the Earth system for climatologists. Increasing the spatial coverage of ground-based weather stations has been a focus of the meteorology and climatology communities over the course of recent decades. These data act as a key source of validation for satellite data as well as numerical models for meteorological and climatological applications such as those used in the climate model inter-comparison projects of the United Nations (Eyring et al., 2016). Automated weather stations (AWS) have become more common as they require less manual labour once deployed. Given their automated nature and the diverse environments in which these stations are deployed, there are many sources of gaps in these data. Sensor malfunction, exceptional system maintenance, errors in data logging or transfer and the removal of erroneous measurements are among some of the common causes of gaps in these data.

Gap-filling is the process of calculation, estimation or prediction of data values that are missing from a dataset. Interpolation is the task of estimating the value of a variable at a new point in space or time that has not been sampled in the original data (Miller, 1997). In this work, machine learning (ML) techniques for spatial interpolation are used for the task of gap-filling. It is possible that the methods described may also be applicable (with adaptation) to other forms of interpolation. This research fills gaps in data series of 4 meteorological variables (hourly values of air temperature (TA), dewpoint pressure (DP), relative humidity (RH) and leaf wetness (LW)) measured by 10 automated weather stations (AWS) in the northern

region of Serbia. Data was taken from the Forecasting and Warning Service in Plant Protection (PIS) Serbian Network freely available on the FAIR Micromet Portal (Roantree et al., 2023). The FAIRNESS Cost Action (EU Cost Action, 2021) established a knowledge sharing platform to provide FAIR micrometeorological data to the research community. A further reliable source of data that is consistently used for gap filling meteorological data comes from reanalysis data such as ERA-5 (Bessenbacher et al., 2022; Dumitrescu et al., 2020; Lompar et al., 2019; Lipson et al., 2022). ERA-5 is the fifth generation of global reanalysis products developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). Reanalysis assimilates observations from disparate, irregularly placed sources using numerical models to compute the state of the atmosphere-surface system on a regularly sampled spatial grid.

The variables selected for gap-filling are commonly measured meteorological variables, though not all of these variables are equally treated in the literature in terms of gap-filling studies. In addition, some of these variables are standard ERA-5 Land output (Muñoz-Sabater et al., 2021) (TA as 2m temperature, DP as 2m dewpoint pressure) while others are not included in ERA5: LW is leaf wetness duration and RH is relative humidity. Those variables that are not included in ERA-5 may be gap-filled only by methods that exploit some indirect correlation with the variables that are included, rather than having a direct counterpart. However, machine learning techniques can use these correlations to more accurately predict the missing values where no direct counterpart is available in the ERA-5 data. The drawbacks associated with using reanalysis data are: a) most probably, there are no corresponding grid points associated with the location of the weather stations and therefore, require debiasing or spatial interpolation before reanalysis data can be used for gap-filling; and b) there is a 5-7 day delay (depending on the source) before this data becomes available. For gaps contained in the intervals between minimum and maximum temperatures (i.e. a few hours) and outside periods of extreme temperatures, linear regression can successfully fill gaps in hourly data. However, machine learning techniques have the potential to exploit more complex underlying correlations within the data to more accurately predict missing values in near real-time.

2.2.1 Existing Approaches

Due to the overlap between the methodologies of gap-filling and interpolation studies and the varied nature of the form and treatment of meteorological data (meteorological data may be treated as time series, "flat" spatial data/images or spatiotemporal series), the research discussed in this section will focus primarily on works that specifically discuss gap-filling. In addition, spatiotemporal methods often focus on

data that are sampled regularly in space such as those from satellites, rather than the irregularly sampled data obtained by meteorological measurements. These methods will not be examined in this work but provide an avenue for further research.

An examination of the literature highlights that there are many studies of air temperature gap filling (Aslan, 2010; Boomgard-Zagrodnik and Brown, 2022; Cerlini et al., 2020; Daly et al., 2000; Hartkamp et al., 1999; Gad et al., 2017; Morales-Moraga et al., 2019; Lompar et al., 2019; Padial-Iglesias et al., 2022; Pape et al., 2009; Razavi et al., 2018; Saleem and Ahmed, 2016; Stahl et al., 2006; Tobin et al., 2011). Only two studies were identified that filled gaps in relative humidity (Kørner et al., 2018; Saleem et al., 2018) and none that filled gaps in leaf wetness or dewpoint pressure. Statistical techniques for spatial interpolation such as multiple regressions (Stahl et al., 2006), kriging (Garen et al., 1994; Hartkamp et al., 1999; Tobin et al., 2011), inverse-distance weighting (Daly et al., 2000) and thin-plate splines (Pape et al., 2009) have also been used for the gap-filling of meteorological variables. Methods for filling gaps in air temperature, vapour pressure deficit, wind speed, global radiation, and long-wave incoming radiation at 153 FLUXNET sites using ERA-5 Interim debiased using linear regression have been outlined in Vuichard et Papale (Vuichard and Papale, 2015). Recently, methods that used ERA5 reanalysis data were proposed for the filling of gaps in near-surface air temperature using linear methods for the debiasing of these data (Lompar et al., 2019).

More recently, machine learning (ML) methods have been applied to fill gaps in a range of eddy covariance variables at FLUXNET sites. The baseline for most papers is that of Moffat et al. (Moffat et al., 2007) which outlines a comprehensive overview of 15 methods for the gap-filling of net carbon fluxes. The authors compare machine learning, numerical, time-series and statistical models for gap-filling. They also outline a methodology for designing the artificial gap-creation to cover a range of different gap scenarios across very short, short, medium and long-term gaps that is subsequently used in other gap-filling studies (Dengel et al., 2013; Kim et al., 2020; Jiang et al., 2022) and also in this study. A wider range of gap lengths and more recent advancements in ML methods for time series have also been compared in Mahabatti et al. (Mahabbati et al., 2021). The Facebook Prophet model for time series was compared with ANNs, random forest (RF), and eXtreme Gradient Boost (XGB) as well as linear and traditional methods such as marginal distribution sampling (MDS) (Reichstein et al., 2005). It was found that random forests slightly outperformed the more complex machine learning models that showed comparable performance and the MDS method also performed well.

ML methods have also been used for gap filling methane flux data at peatland, wetland and rice paddy sites (Kim et al., 2020) and at a range of FLUXNET sites (et. al., 2021). In the former, principal component analysis was used to transform

features before input into three ML algorithms (Artificial Neural Networks (ANN), RF and Support Vector Machines) as well as one traditional method (Marginal Distribution Sampling (MDS). In the latter, Lasso Regression, RF and ANNs were tested against the MDS method with 10 uniquely sampled training and test sets generated for each site. A novel gap creation method was also introduced (et. al., 2021) to artificially introduce gaps of varying length so that they replicate the distribution of both duration and location of gaps observed in the dataset.

Ensembles of multiple methods of gap filling have also been compared, finding that this methodology minimises the deficiencies of any one particular model (Lucas-Moffat et al., 2022). Lucas et al. (Lucas-Moffat et al., 2022) also extend the gap creation methodology of Mofatt et al. (Moffat et al., 2007) to sample the entire dataset by imposing only a single gap length of either 30 minutes or 24 hours at a time. This technique is problematic in the training and evaluation of the performance of a machine learning model as it may lead to overestimation of the predictive performance of the model. In effect, a low percentage of the data is isolated for testing and as a result, the model has *"seen"* most of the data in training.

2.2.2 Contribution

In this work, we evaluate methodologies for filling gaps in meteorological data. These variables were carefully chosen to represent a standard set of AWS measurements that are used in applications in urban and rural areas, including forests. These include air temperature (TA), relative humidity (RH), dewpoint pressure (DP) and duration of leaf wetness (LW). This research directly addresses the problem of the shortage of gap-filling methodologies for wider set of meteorological variables than is normally tackled, and provides methods to achieve this in near real-time, i.e. within one day after the last measurement is taken. This is particularly important for leaf wetness, a complex variable that is difficult to measure or calculate, but is of vital importance in assessing the development of plant fungi. An additional source of novelty arises from those experiments where the ML model is trained using reanalysis data for meteorological variables (top soil layer temperature, surface atmosphere pressure, precipitation volume, evaporation and potential evaporation). While these variables are not the target of the gap filling methods, they affect energy and water balance at the Earth's surface. Thus, they are important in determining air temperature, humidity and leaf wetness duration.

2.3 Materials and Methods

2.3.1 Automated Weather Station Data

Phenological and meteorological data used in the study were obtained from the Forecasting and Reporting Service for Plant Protection (PIS) observational network in Serbia. Hourly values of air temperature and humidity, soil temperature, precipitation, and leaf wetness have been archived since 2011 using 166 AWSs located in orchards, vineyards, and crop fields across Serbia. More details about PIS meteorological and biological observational network and information system can be found in (Lalic et al., 2020). For this study, only locations with at least five years of continuous daily measurements were selected, amounting to 10 sites in total. The distribution of sites is shown in Figure 2.1 with site metadata presented in Table A.1.

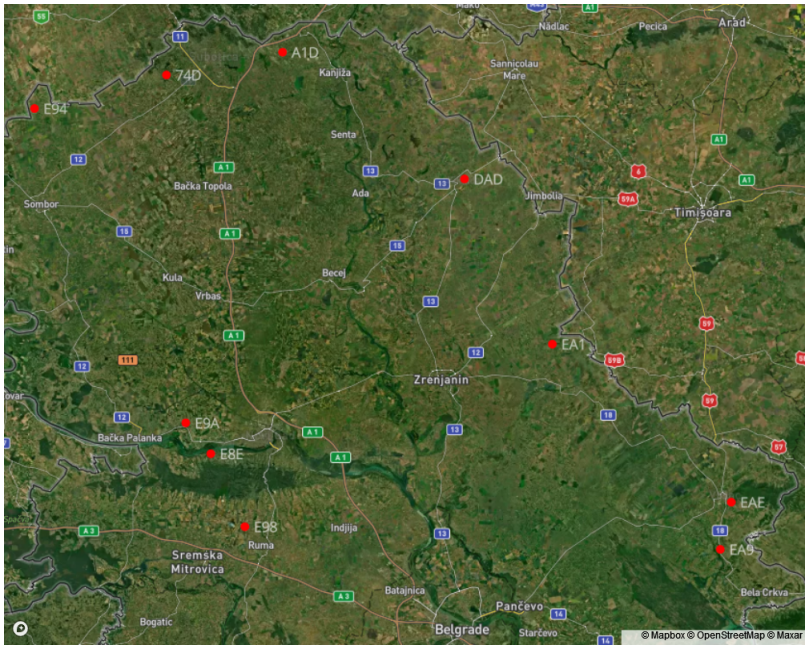


Figure 2.1: Satellite Map illustrating the distribution of AWS sites. OpenStreetMap contributors, distributed under the Open Data Commons Open Database License (ODbL) v1.0. Red circles are used to indicate the AWS sites.

The meteorological data comprises hourly values of: mean, maximum and minimum air temperatures, relative humidity and calculated specific humidity. Air temperature and humidity sensors are radiation-shielded and mounted at 1.5 m height (approximately mid-crown), a resistance soil thermometer is positioned at 20 cm depth, an electronic rain gauge is set between the rows of trees, and the leaf wetness duration sensor is attached to a leaf surface within the crown. The data period is between 2013 and 2021 with the exception that three sites, E98, 74D, and A1D, commencing in 2014. The data is largely complete with less than 5% gaps at most

ID	Region	Location	Plant	Years	Long.	Lat.	Elevation
DAD	Kikinda	Kikinda	Plum	2013-2021	20.47	45.83	73
E9A	Novi Sad	Cenej	Apple	2013-2021	19.58	45.28	78.12
E8E	Novi Sad	Cerevic	Peach	2013-2021	19.66	45.21	151
E98	Ruma	JazakIrig-Kudos	Apple	2014-2021	19.76	45.05	140.5
E94	Sombor	Ridjica	Apple	2013-2021	19.09	45.98	83
74D	Subotica	Ljutovo	Apple	2014-2021	19.51	46.06	121.41
A1D	Subotica	Backi Vinogradi	Peach	2014-2021	19.88	46.11	90.49
EAE	Vrsac	Vrsac	Grape	2013-2021	21.32	45.11	128
EA9	Vrsac	Crvena Crkva	Plum	2013-2021	21.28	45.00	83
EA1	Zrenjanin	Sutjeska	Plum	2013-2021	20.75	45.46	84

Table 2.1: Weather Station Metadata sorted by Region.

sites. At EAE, LW is not measured and at E98, approximately 35% of the data is missing for *TA*, *DP* and *RH*. Orchards and vineyards were selected due to their continuous time series, relatively common canopy architecture (for this region), and management practices. The elevation of the sites ranges from between 78.12 to 140.5 metres above sea level.

CODE	DAD	E9A	E8E	E98	E94	74D	A1D	EAE	EA9	EA1
DAD	0.0									
E9A	92.4	0.0								
E8E	93.4	10.0	0.0							
E98	102.9	29.2	19.4	0.0						
E94	108.1	86.7	96.4	115.8	0.0					
74D	**78.5**	86.9	95.2	114	**3.6**	0.0				
A1D	55.2	95.2	101.5	118	62.7	**29**	0.0			
EAE	104	137.6	130.6	122.7	198.8	176	157.8	0.0		
EA9	111.9	136.9	129.3	119.6	202.5	181.4	164.7	12.6	0.0	
EA1	46.5	93.6	89.6	89.9	141.2	117.1	98.9	59.2	65.9	0.0

Table 2.2: Weather Station Distance (km) Matrix

Table 2.2 displays the distance in kms between each pair of sites shown in Figures 2.1 and 2.2. Calculated using the haversine function (https://pypi.org/project/haversine/, 2024), this information feeds into methods that use nearby stations to help filling gaps. For example, gap filling for station 74*D* will include data from *DAD*, *E*94 and *A*1*D* as highlighted in Table 2.2.

2.3.2 ERA-5 Reanalysis

Another important source of information regarding the meteorological landscape can be acquired from the European ReAnalysis (ERA) repository produced by European

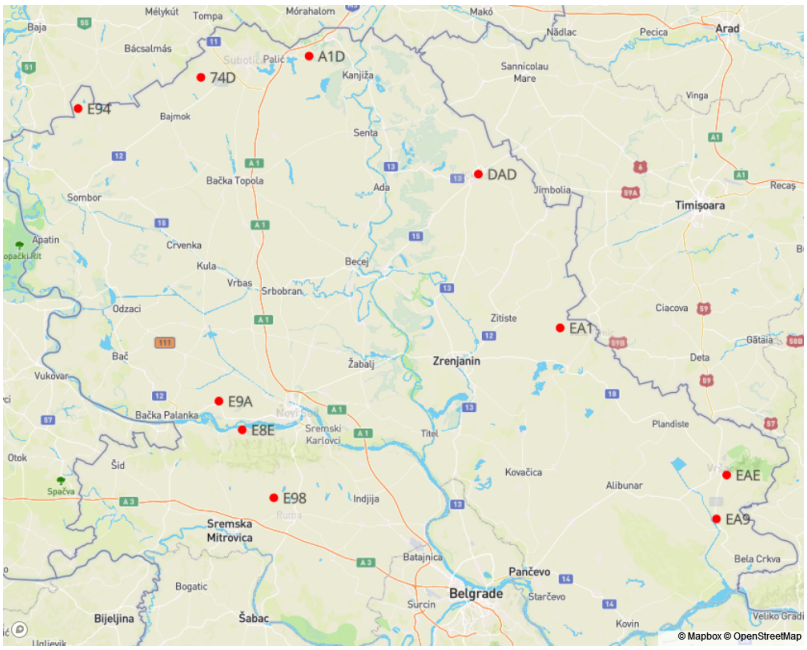


Figure 2.2: Geographical Maps illustrating AWS site distribution with distance calculations presented in Table 2.2. OpenStreetMap contributors, distributed under the Open Data Commons Open Database License (ODbL) v1.0. Red circles are used to indicate the AWS sites.

Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis is a scientific method that combines measurements of aspects of the Earth's surface and atmosphere with numerical models of the processes governing these systems. Data from many disparate and irregularly placed sources are combined using ensemble data assimilation methods (Hersbach et al., 2020; Muñoz-Sabater et al., 2021) to provide initial and boundary conditions for the numerical models simulating the state of the surface-atmosphere system. The fifth generation of reanalyses, ERA-5, consisted of three phases: ERA-5 Interim, the initial release of these reanalyses with a spatial resolution of 80km; ERA-5 (Hersbach et al., 2020) with a spatial resolution of 30km; and ERA-5 Land (Muñoz-Sabater et al., 2021) with an increased spatial resolution of 9km. All ERA-5 data are at a temporal resolution of 1 hour per sample.

Of the wide range of variables available in ERA-5 Land, 7 are typically selected as input variables to various statistical and physical models, or in our study, machine learning models. These include: total precipitation (tp), 2 metre temperature ($t2m$), 2 metre dewpoint pressure ($d2m$), evaporation (e), surface pressure (sp), potential evaporation (pev) and soil temperature at level 1 ($stl1$). These variables were selected due to their direct relationship with or influence on the physical processes of the variables being gap-filled.

Feature Set	Feature		Description
Time Features	Hour		Hour of the day (0-23)
	Month		Month of the year (1-12)
	Year		Calendar year (2013-2021)
AWS	TA	(×3)	Mean air temperature [°C]
	DP	(×3)	Mean dew Point temperature [°C]
	RH	(×3)	Mean relative Humidity [%]
	P	(×3)	Precipitation [mm]
	LW	(×3)	Leaf wetness [min]
ERA5	t2m	(×3)	2 metre temperature [K]
	stl1	(×3)	Soil temperature (Level 1) [K]
	d2m	(×3)	2 metre dew point temperature [K]
	tp	(×3)	Total precipitation [m]
	e	(×3)	Total evaporation [m of water equivalent]
	pev	(×3)	Potential evaporation [m]
	sp	(×3)	Surface pressure [Pa]
AWS+ERA5	All above	features	

Table 2.3: Feature set description. All feature sets contain the three *Time Features* variables. Each of the AWS feature sets contains hourly data from the nearest three closest Stations (AWS). ERA5 uses different methods to create the feature sets (see section 2.3.3) and also includes three nearby grid points. The Spatial approach uses only the same variable as the target variable from the 3 closest AWS to the target site.

2.3.3 Methods

Gap Filling Methods

In all, five methods were used to gap-fill missing data. As a baseline for analysing the comparative performance of the three ML models, two non-ML approaches were used. A brief description of the methods is presented here with further detail provided in tables 2.4 A.2 and A.3 in the discussion on section 2.4.

- **Debias.** A popular approach using ERA5 data which is debiased to compensate for the fact that a single input value represents a broad area which may have different local environmental conditions, elevation etc.
- **Spatial.** This approach is based on simple statistical methods where *mean* spatial interpolation is used, taking data for the same variable at the nearest three AWS as input.
- **LR.** The simplest ML model used is that of the linear regression which acts as a baseline for ML models.
- **RF.** Random forests are selected as they have been found to be effective for other gap-filling tasks (et. al., 2021), are relatively interpretable models and are well optimised in the scikit-learn package (Pedregosa et al., 2011a).
- **LGB.** A gradient boosted ensemble is also tested as this class of algorithm have been noted performance in terms of reduced training time and increased accuracy in gap-filling meteorological time series (Kørner et al., 2018). Specifically the Light Gradient Boosted Machines (LGB) model has not been utilised for gap filling in other studies to the best of the authors knowledge due to its relative recency in development. LGB models are noteworthy due to its increased speed performance in training over other gradient boosted ensembles and its notable performance on a number of machine learning tasks (Ke et al., 2017).

Feature Sets

Different feature sets were employed either because they best suited a particular method or to evaluate and understand the different performance levels of individual machine learning models. In addition, ERA5 data is used in different ways: for the Debias method,it is used to directly interpolate missing data while for machine learning approaches,it is used purely as input to machine learning methods.

Debias. A simple debiasing for ERA5-land was developed where a period of 10 days before (leading period) and 10 days after (trailing period) the gap is selected.

For both periods a bias is calculated. To fill gaps, the ERA5-land values are used and the average bias (leading and trailing period) is subtracted from the ERA values.

Machine Learning Models. For all ML models, every feature set contains the $TimeFeatures$ variables. For the AWS feature set, data are taken from the three nearest AWS to that being gap filled. For the ERA5 feature set, data are taken from the nearest ERA-5 grid-point to the AWS as measured by the Haversine function. Data from the grid point 0.2 degrees north and east and 0.4 degrees south and west of this point are also used. These are chosen to give a fairer comparison of the effect of distance on correlation with the target variable and to examine the utility of the input features in improving model accuracy as it is hypothesised these data may contain information about future meteorological conditions for the location in question.

Spatial. The Spatial approach uses the same three closest AWS to the target site.

2.3.4 Gap Creation

In order to fully assess the performance of each of the gap-filling methods, a range of scenarios for gaps of different lengths similar to (Moffat et al., 2007) was used. Gaps of very short (1 hour), short (4 hour) medium (36 hour) and long (288 hour) gap durations were introduced to the data set at random points. As discussed in Section 2.2.1, the concern that previous sampling methods did not cover the entire dataset was raised and addressed in (Lucas-Moffat et al., 2022). However, the method used requires that each gap be sampled individually, allowing a disproportional amount of data for training of models for each gap, limiting the evaluation of the model's ability to generalise across the entire distribution of the data. A novel gap creation approach is proposed that involves the following steps. First, split the dataset into consecutive blocks for each of the gap lengths being tested whilst avoiding existing gaps in the dataset. Enumerate each of these blocks and store these indices. Randomly select the locations of the gaps by sampling the index of the blocks (for a total of 10% of the data) and removing these indices from the set to be sampled from. Repeat until all indices have been sampled (i.e. the entire dataset has been covered). A total of 10% of the data is allocated for testing at each iteration, leaving 90% of the data for model training. This results in 10 sets of artificial gaps for each site and for each gap length, effectively creating a form of 10-fold temporal cross validation. This process is undertaken for each of the 4 target variables at each of the 10 sites. In this way, gaps are randomly distributed but also cover the entire dataset.

The Debias method used a different bespoke gap creation method. For each station, a random timestamp is selected; this timestamp is the start of an artificial

gap if the leading, trailing and gap period does not contain a gap in the observations. Gaps are created with sizes 1, 4, 36 and 288 hours (the shorthand 1h, 4h, 36h and 288h will be used throughout the rest of the paper) for consistency with the ML experiments. For each station and gap size, a total of 150 artificial gaps are created. For the ERA5-land data, hourly data of the nearest gridpoint was used at the native grid. To calculate the relative humidity,the Magnus approximation (Alduchov and Eskridge, 1996), was used.

2.3.5 Validation

The metrics used during model validation are the coefficient of determination (R^2), normalised Root Mean Squared Error (nRMSE), Mean Bias Error (MBE) and Mean Absolute Error (MAE), calculated as shown in equations 4.1, 5.7, 5.9 and 5.8 respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.1}$$

$$nRMSE = \frac{RMSE}{\sigma} = \frac{1}{\sigma}\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}} \tag{2.2}$$

$$MBE = \frac{\sum_{i=1}^{N} y_i - \hat{y}_i}{N} \tag{2.3}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{2.4}$$

Across all equations: $\hat{y}_i$ denotes the predicted value of the i-th sample; y_i is the corresponding true value for N total samples in the test set; and σ is the standard deviation of y_i in the test set.

2.4 Results and Discussion

There are five dimensions by which we measure and validate each of the gap filling models across a large set of experimental configurations: meteorological variables (*TA, DP, RH, LW*); feature sets (combinations of input variables including temporal information, data from other AWS stations, ERA-5); three types of machine learning algorithms (linear regression, random forests, LightGBM) combined with two non-ML algorithms; gap sizes of 1, 4, 36 and 288 hours; and site location. In this section, the results for each model are presented across different target parameters and gap sizes. In total, 1,720 experiments were conducted as per Table 2.4 with 4 separate measures used to validate predictive accuracy: R^2, RMSE, nRMSE and MAE. Of

these, RMSE and nRMSE are presented in the main text, while the others are included in the Supplementary Material. In addition, we will focus our discussion around the larger sizes and the higher performing models. However, the entire set of results is available online (Roantree, 2024).

Target ↓	**Model →**	**RF**	**LGB**	**LR**	**Spatial**	**Debias**	Total
dewpoint pressure	*DP*	120	120	120	40	40	440
humidity	*RH*	120	120	120	40	40	440
leaf wetness	*LW*	120	120	120	40	0	400
temp	*TA*	120	120	120	40	40	440
Total *(by Model)*		480	480	480	160	120	1,720

Table 2.4: A total of 1,720 experiments were conducted: 1,440 machine learning experiments; 160 using a spatial algorithm and 120 experiments using ERA-5 debiasing. The average result was selected for each of 10 sites for four gap sizes (40 results) with the three ML models using three different feature sets (120 results).

The discussion begins with a summary of results across models, gap lengths and feature sets for two contrasting target variables before a more detailed focus on the performance of predictive models. The rationale behind this approach is practical. Typically, the presence of gaps in the data series of interest for this study ranges from 14% (LW) to 18% (TA). In other words, if one variable is missing, most likely all variables are missing for that time step, suggesting that the origin of the problem often lies not in the sensor itself but in the entire AWS. Therefore, a methodology that offers good results across multiple variables may often be a superior option than one that provides excellent results for a single variable while performing poorly for other(s). This is important from a user's perspective, as in agriculture application studies for example, commonly two or more variables are used simultaneously.

2.4.1 Best Performing Models

Before we begin, we recall that our gap creation for ML methods effectively creates a form of 10-fold temporal cross validation. These 10 sets of results were averaged before being presented here. We now proceed with a look at how models perform in general across the different target variables. As *temperature* and $leaf\ wetness$ offer contrasting target types, we focus on their results although all result breakdowns are in Supplementary Materials (Roantree, 2024) (as HighLevel-TargetVar-Only). Results for the TA and LW variables are presented in Table 2.5 (note that LW was not predicted in the Debias experiments).

Despite the contrasting variables under analysis, it is evident that two of the machine learning functions, Random Forest (RF) and Light GBM (LGB), perform best with highest predicting accuracy for all 4 gap sizes, with close to identical

TA	**GapSize**	**nRMSE**	**RMSE**	**MAE**	**R2**
RF	1	0.1424	1.348	0.9501	0.9792
RF	4	0.1546	1.4629	1.0492	0.9754
LGB	4	0.1567	1.4825	1.0859	0.9747
LGB	1	0.1569	1.4842	1.0806	0.9734
LGB	36	0.1626	1.534	1.1189	0.9726
RF	36	0.1657	1.5628	1.1248	0.9714
LGB	288	0.1741	1.6133	1.1624	0.9677
RF	288	0.1786	1.6553	1.1773	0.9657
Debias	4	0.2304	2.226	1.7157	0.9467
LR	1	0.2319	2.1904	1.5668	0.9386
LR	4	0.2321	2.1912	1.568	0.9385
LR	36	0.2328	2.1926	1.5715	0.9378
LR	288	0.2369	2.1925	1.5834	0.9341
Debias	36	0.2379	2.2296	1.7064	0.9432
Debias	288	0.2446	2.3242	1.7814	0.94
Debias	1	0.2464	2.3035	1.7498	0.9384
Spatial	288	0.2856	2.6487	1.8499	0.9056
Spatial	36	0.2874	2.7073	1.8557	0.9069
Spatial	4	0.2874	2.7137	1.8556	0.9074
Spatial	1	0.2874	2.7148	1.8557	0.9075
LW	**GapSize**	**nRMSE**	**RMSE**	**MAE**	**R2**
RF	1	0.6120	13.1560	7.5323	0.6024
RF	4	0.7014	15.1076	8.8305	0.4941
LGB	1	0.7102	15.3242	9.3529	0.4843
LGB	4	0.7373	15.9084	9.7039	0.4466
LGB	36	0.7723	16.6499	10.1791	0.3956
RF	36	0.7906	17.0455	10.2336	0.3671
LGB	288	0.7914	16.9772	10.4372	0.3659
RF	288	0.8197	17.5895	10.7592	0.3208
LR	1	0.8547	18.5284	13.0907	0.2658
LR	4	0.8554	18.5379	13.0993	0.2645
LR	36	0.8580	18.5663	13.1300	0.2600
LR	288	0.8631	18.5759	13.1557	0.2509
Spatial	1	1.0309	22.1262	14.9373	-0.1290
Spatial	4	1.0311	22.1238	14.9373	-0.1297
Spatial	36	1.0324	22.1046	14.9286	-0.1328
Spatial	288	1.0343	22.0759	14.9333	-0.1375

Table 2.5: Predictive performance for **Temperature** and **Leaf Wetness**, aggregated across feature sets and sites. Scores are ranked by normalised root mean square error score (in column **nRMSE**), with columns root mean square error (RMSE), Mean Absolute Error (MAE) and R^2 provided for comparison purposes.

ranking across both target variables. As expected, the performance across all 5 methods decreases as the gap size increases.

2.4.2 Analysis By Feature Set

For a better understanding of the performance across different feature sets, the next analysis uses a 36h gap size, for the best performing ML models (LGB, RF) and for all selected variables. In order to focus on the ML models this means that both Debias and Spatial feature sets are removed from this presentation of results.

Feature Set	Variable	Model	nRMSE	RMSE	MAE	R2
AWS-ERA5	**temp (TA)**	**LGB**	0.1169	1.0945	0.7663	0.9863
AWS-ERA5	**temp (TA)**	**RF**	0.1205	1.1287	0.7867	0.9854
AWS	temp (TA)	LGB	0.1206	1.1294	0.789	0.9854
AWS	temp (TA)	RF	0.1244	1.1652	0.8193	0.9845
AWS-ERA5	**dewpoint (DP)**	**LGB**	0.1604	1.2274	0.9146	0.9742
ERA5	temp (TA)	LGB	0.1625	1.5214	1.1379	0.9736
AWS-ERA5	**dewpoint (DP)**	**RF**	0.1655	1.2669	0.9321	0.9725
ERA5	temp (TA)	RF	0.1655	1.5497	1.1467	0.9726
AWS	dewpoint(DP)	LGB	0.1728	1.3218	0.98	0.97
AWS	dewpoint (DP)	RF	0.1766	1.351	0.9959	0.9687
ERA5	dewpoint (DP)	LGB	0.2109	1.6152	1.2301	0.9554
ERA5	dewpoint (DP)	RF	0.2173	1.6635	1.2533	0.9527
AWS-ERA5	**humidity (RH)**	**LGB**	0.3162	7.0494	4.6286	0.8997
AWS-ERA5	**humidity (RH)**	**RF**	0.3254	7.2555	4.6653	0.8938
AWS	humidity (RH)	LGB	0.3319	7.3985	4.8897	0.8896
AWS	humidity (RH)	RF	0.3402	7.5837	4.8816	0.884
ERA5	humidity (RH)	LGB	0.4149	9.2512	6.4168	0.8277
ERA5	humidity (RH)	RF	0.4197	9.3601	6.3497	0.8236
AWS-ERA5	**leaf wetness (LF)**	**LGB**	0.7379	16.2725	9.4174	0.4547
AWS-ERA5	**leaf wetness (LF)**	**RF**	0.7587	16.7272	9.6429	0.4235
AWS	leaf wetness (LF)	LGB	0.7588	16.7333	9.8349	0.4234
ERA5	leaf wetness (LF)	LGB	0.7808	17.2195	10.4383	0.3894
AWS	leaf wetness (LF)	RF	0.792	17.4629	9.884	0.372
ERA5	leaf wetness (LF)	RF	0.795	17.5241	10.5399	0.3671

Table 2.6: Feature set performance by variable and ML model, ranked by nRMSE. Results were for the (randomly chosen) 74D site with a fixed gap length of 36. Each variable has 6 result combinations from two models × 3 feature sets. The best performing Model/Variable combinations are in boldface.

In Table 2.6, we can see how the three feature sets performed for each target variable using the two best performing models. There is a clear discrimination by target variable with TA easiest to predict and LW most difficult to predict. Within each target variable subsection, there is clear evidence that the AWS-ERA5 feature set performs best, followed closely by AWS. It also highlights the poor performance of ERA5 which requires far more complex debiasing to make the feature set competitive. This analysis is also performed for all sites in a single combined report as part of supplementary materials (Roantree, 2024) (as Drilldown-FeatureSet).

The relatively small difference in performance between AWS-ERA5 and AWS is noteworthy as one may decide the overhead of colecting using ERA5 data (debiasing effort and lag times) is unnecessary. For TA the difference for LGB is 0.0037 and for RF is 0.039 (or approx. 4%). Even as performance degrades across variables, the difference between AWS-ERA5 and AWS feature sets remains relatively small (DP is 8%, DP is 5%, LF is 4%).

2.4.3 Analysis By Site for Gap Sizes 36 and 288

Now using the best performing (LGB) model and (AWS-ERA5) feature set, we examine across sites where we would expect a correlation between performance and gap distribution and volume. In Figure 2.3, the results of the gap analysis are shown. Here, it is useful to understand the (missing data) thresholds for each model before a decline in performance occurs.

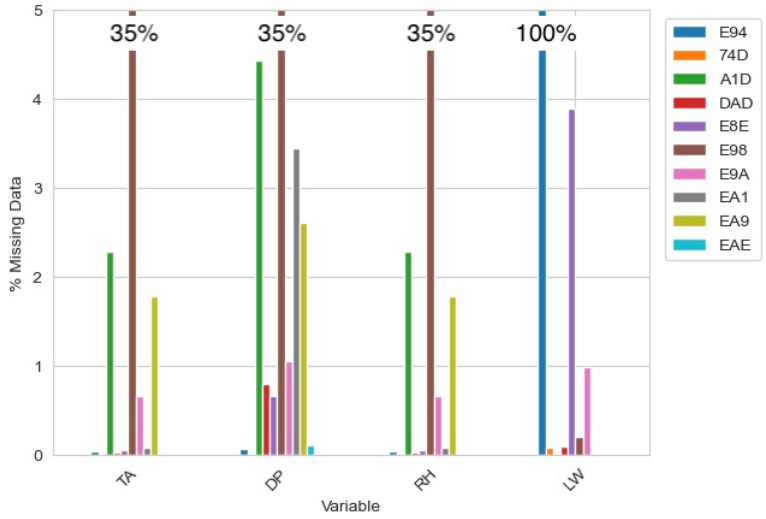


Figure 2.3: Percentage of missing data for each variable at each site. At EAE, LW is not measured and and at E98, approximately 35% of the data is missing for TA,DP and RH.

In general terms, models at site 74D performed best for three of the target variables as illustrated in Table 2.7, while site E9A performed best for LW. Models at site E94 also performed well for variables other that LW. More specifically, the performance of the LGB model for TA, in comparison to other variables, is still best with low (16%) variation in nRMSE across sites. Significant differences should be noticed between RH and DP variables in both, magnitude (RH=49%, DP=30%) and variability (RH=24%, DP=38%) of nRMSE (Tab. 2.7). Inventory of sites indicates high nRMSE and RMSE for the same locations for both RH and DP which is to be expected since DP is not a measured value but the temperature taken from (temperature) sensors at $RH = 100\%$. This implies that DP valus are

	TA	*TA*	*RH*	*RH*	*DP*	*DP*	*LW*	*LW*
Site	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE
74D	**0.1169**	1.0945	**0.3162**	7.0494	**0.1604**	1.2274	0.7379	16.2725
A1D	0.131	1.2602	0.6138	19.1393	0.3122	2.7552	0.7337	17.4434
DAD	<u>0.1219</u>	1.1745	0.4907	21.153	0.4576	3.4685	<u>0.677</u>	15.8504
E8E	0.1983	1.7748	0.6127	18.4281	0.3493	2.5502	0.7028	13.4459
E94	0.1334	1.2694	<u>0.3611</u>	6.6491	<u>0.1689</u>	1.3337	1.0029	18.7857
E98	0.1445	1.3637	0.3818	8.3889	0.1785	1.2849	0.7037	17.7952
E9A	0.1489	1.401	0.4704	13.0336	0.3074	2.2804	**0.6662**	15.9189
EA1	0.1479	1.4216	0.6724	26.1903	0.476	3.4809	0.7741	14.1796
EA9	0.1601	1.5371	0.4623	14.135	0.3232	2.3509	0.7568	15.8371
EAE	0.1532	1.4387	0.4991	12.5654	0.2501	1.7806	0.7195*	15.5511*

Table 2.7: Results for all 4 target variables at each site location for LGB model, AWS-ERA5 feature set and 36h gap length. Results have top sites in boldface and 2nd best underlined. *Results for *LW* for site EAE can be ignored as data was synthetic.

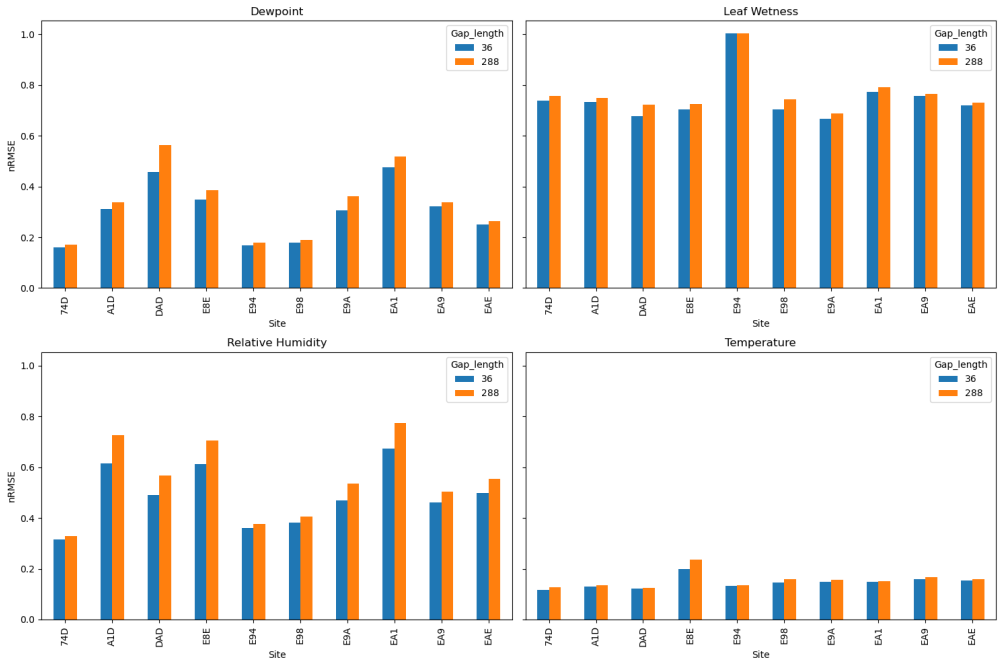


Figure 2.4: Tables 2.7 and 2.8 in graph format showing nRMSE results. *Results for LW for site EAE can be ignored as data was synthetic

affected by operability of two sensors: thermometer and hygrometer. Since *RH* can be calculated using *DP* and *TA*, for locations with high nRMSE for *RH*, these gaps can be filled using synthesized values for both *DP* and *TA*.

Figure 2.4 is useful when comparing gap sizes of 36h and 288h, as results in Table 2.8 are very similar to those achieved for gap size 36h. Sites DAD and 74D swap first and second positions but the difference is very slight. Otherwise, the results for the larger gap size are identical. Figure 2.4 graphs also illustrate clearly

	TA	TA	RH	RH	DP	DP	LW	LW
Site	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE
74D	<u>0.1271</u>	1.1551	**0.3286**	7.2819	**0.1723**	1.2799	0.7558	16.4859
A1D	0.1342	1.2863	0.7266	22.1916	0.3388	2.9623	0.75	17.7932
DAD	**0.1256**	1.1977	0.5677	24.2922	0.5635	4.1967	<u>0.7238</u>	16.7572
E8E	0.235	2.0309	0.7058	21.0747	0.3868	2.7059	0.7259	13.8351
E94	0.1363	1.2922	<u>0.3754</u>	6.8474	<u>0.1798</u>	1.3864	1.0027	18.7817
E98	0.1583	1.4491	0.4045	8.8794	0.189	1.3369	0.7432	18.6724
E9A	0.1562	1.4421	0.5345	14.6158	0.3627	2.6126	**0.6867**	16.0932
EA1	0.1524	1.4502	0.7735	29.995	0.5186	3.7476	0.7918	14.4455
EA9	0.1671	1.5884	0.5032	14.8382	0.3368	2.4451	0.7636	16.0138
EAE	0.159	1.4796	0.5533	13.9236	0.2646	1.8526	0.7292	15.665

Table 2.8: Results for all 4 target variables at each site location for LGB model, AWS-ERA5 feature set and 288h gap length. Results have top sites in boldface and 2nd best underlined. *Results for LW for site EAE can be ignored as data was synthetic.

the predictive performance of the machine learning algorithm for TA and DP when compared with RH and LW.

2.4.4 Analysis of ERA5-Debias for Gap Sizes 36 and 288

In this section, we repeat the site-based analysis for the same two gap sizes but on this occasion, we examine results for the ERA5 gap filling experiment which used a debiasing method with bespoke feature set. Below we present the results for gap size 36h with broadly similar results available in supplementary materials (Roantree, 2024) (as Drilldown-Sites). Interestingly, while the LGB model favoured the small gap size in terms of predictive performance, a similar difference in performance is not evident in figure 2.5, with predictions for smaller gap sizes sometimes performing worse.

When comparing the performance of LGB and Debias models using nRMSE, the results for the LGB model are significantly better for RH (0.49 vs. 0.63), slightly better results for TA (0.15 vs. 0.24), and broadly similar results for DP (0.3). However, one interesting feature of the Debias technique is the significantly lower variability of nRMSE among locations (RH - 19% vs. 24%; TA - 6% vs. 16%; DP - 15% vs. 38%). The results achieved by ERA5 using Debias for TA are of the same rank obtained by researchers in (Lompar et al., 2019) for lowland stations using the same gap scale.

For longer gaps (288h), the ranking achieved by ERA5 using Debias shows a slight improvement over the best performing (LGB) ML model. Here, we observe the same average nRMSE with very similar nRMSE variability across variables.

	TA	*TA*	*RH*	*RH*	*DP*	*DP*
Site	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE
74D	**0.2193**	1.9951	0.5603	12.2803	<u>0.2588</u>	1.963
A1D	0.2318	2.3516	0.8287	27.9684	0.3538	3.1939
DAD	<u>0.2265</u>	2.1055	0.5357	22.9723	0.3942	2.7406
E8E	0.2733	2.3281	0.7234	22.1005	0.2981	2.2615
E94	0.2351	2.1487	0.5977	11.0261	0.2903	2.3704
E98	0.2359	2.1415	<u>0.5188</u>	11.3809	0.2871	1.7935
E9A	0.2396	2.3117	0.6104	14.8365	<u>0.2583</u>	1.9283
EA1	0.2355	2.1733	0.7842	30.5984	**0.2529**	1.7905
EA9	0.2563	2.549	**0.4848**	15.5768	0.3001	2.0484
EAE	<u>0.2260</u>	2.1912	0.5728	14.4042	0.2809	2.1539

Table 2.9: Results for 3 target variables at each site location for Debias model with ERA5 feature set for 36h gap length. Results have top sites in boldface and 2nd best underlined.

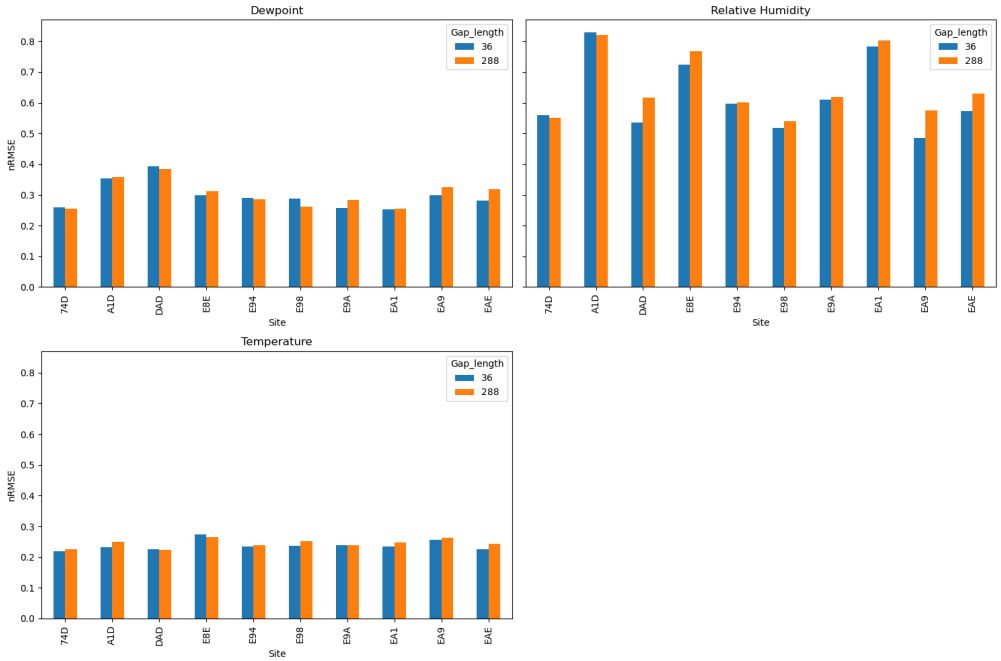


Figure 2.5: Tables 2.9 and 2.10 in graph format showing nRMSE results.

However, the LGB model does not perform as well for *RH*.

2.4.5 Summary and Limitations

Based on this analysis of gap filling methods together with an array of results, we can conclude that the selection of a method for gap filling in meteorological time series data depends not only on the inherent characteristics of the data but is also heavily influenced by the timing constraints specified by the data owner and/or user (Tab. 2.11). Specifically, the urgency with which gap filling needs to be performed

	TA	TA	RH	RH	DP	DP
Site	nRMSE	RMSE	nRMSE	RMSE	nRMSE	RMSE
74D	0.226	2.1571	0.5508	12.2385	0.2545	2.0337
A1D	0.2492	2.4647	0.8211	26.4904	0.3577	3.3994
DAD	0.2238	2.2657	0.6171	26.2077	0.3841	2.7031
E8E	0.2641	2.3473	0.7682	21.9664	0.3118	2.4023
E94	0.239	2.2026	0.6004	11.0708	0.2854	2.3516
E98	0.2511	2.2365	0.5397	11.6851	0.2607	1.7815
E9A	0.2378	2.2774	0.6185	17.2978	0.2834	2.0064
EA1	0.2481	2.4359	0.8024	31.5856	0.255	1.8315
EA9	0.2634	2.5886	0.5752	18.0249	0.3251	2.31
EAE	0.2431	2.2663	0.6297	15.1979	0.3197	2.2819

Table 2.10: Results for 3 target variables at each site location for Debias model with ERA5 feature set for 288h gap length. Results have top sites in boldface and 2nd best underlined.

- whether within one day or one week (or longer) following a measurement failure - plays a critical role in selecting the appropriate methodology. This temporal constraint influences the choice of techniques, as different methods may vary in their processing times and the immediacy with which they can address data discontinuities, thereby affecting the overall responsiveness and applicability of the gap-filling process. Finally, one must take into account the computational skills of data owners and users which is also an important factor in methodology selection.

Gap (h)	Ability	Period after Measurement Failure	
		Day(s)	Week(s) or more
< 4	Low, high	Linear interpolation	
< 24	Low	Sinusoidal method	
< 24	High	LGB	LGB or ERA5-Land Debias
Day(s)	Low	Data from nearest representative weather station	
Day(s)	High	LGB	LGB or ERA5-Land Debias

Table 2.11: Gap filling methodology selection according to user ability (computational skills) and timing constraints.

2.5 Conclusions

In this paper, we compared different techniques for filling gaps in typical AWS measurements. Two ML models (LGB and RF) and one dynamical-statistical hybrid (ERA 5 Land debias) showed best performances with slightly better results for ML methods and significant advantage that it can be applied in near-real time (on a daily basis). Both ML models are trained using not only the variables of interest but also other relevant data eg. precipitation and from reanalysis variables. We believe that this makes these experiments more trustworthy. Results obtained for TA and DP are promising, while our results would indicate that RH might be better calculated using TA and DP. With respect to LW, there remains no sufficiently robust method for gap filling, instead relying on a calculation where precipitation is available, and evaporation can be calculated from available data. Indeed, for all gap lengths, LW is confirmed as a biggest challenge for gap filling.

One potential weakness of ML methods is lack of physical explainability of results. Therefore, our current research focus is on the introduction of ERA5 Land trained, physics-guided neural networks (Hao et al., 2023) in filling gaps in AWS data series what we believe could see a paradigm change in addressing gap filling in meteorological data series.

Chapter 3

Evapotranspiration Partitioning

The following chapter has been published in Applied Computing and Geosciences and the accepted version is reproduced here under the copyright agreement of the publisher Elsevier.
Stapleton, A., Eichelmann, E., and Roantree, M. (2022). A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning. Applied Computing and Geosciences, 16:100105.

This chapter sets the foundations of the three applications presented in this study by introducing a methodological framework for the comparison of multiple machine learning models alongside data driven methods for feature selection. A key component of this study is the availability of existing ML results with which the outputs of this study can be compared. Eichelmann et al. (2021b) first introduced this methodology using neural networks and this study expands on the range of machine learning algorithms tested so as to systematically cover different classes of algorithm and establish which should be used in further studies. The recursive feature selection with feature importance as a heuristic is shown to be effective in both identifying optimal feature sets as well as discovering new scientific hypotheses. These are key elements of the methodology that will be carried forward in later chapters.

3.1 Abstract

A deeper understanding of the drivers of evapotranspiration and the modelling of its constituent parts (evaporation and transpiration) may be of significant importance to the monitoring and management of water resources globally over the coming decades. In this work a framework was developed to identify the best performing machine learning algorithm from a candidate set, select optimal predictive features and rank features in terms of their importance to predictive accuracy. The experiments conducted in this work used 3 separate feature sets across 4 wetland sites as input into 8 candidate machine learning algorithms, providing 96 sets of experimental configurations. Given this high number of parameters, our results show strong evidence that there is no singularly optimal machine learning algorithm or feature set across all of the wetland sites studied despite their similarities. At each of the

sites at least one model was identified that improved on the predictive performance of our baseline. A key finding discovered when examining feature importance is that methane flux, a feature whose relationship with evapotranspiration is not generally examined, may contribute to further biophysical process understanding. This work demonstrates the applicability of a machine learning framework for evapotranspiration partitioning that is independent of domain knowledge, producing improved models for partitioning and identifying new and useful predictive features.

3.2 Introduction

Evapotranspiration (ET) is the process by which water is exchanged between the biosphere and the atmosphere. Better understanding of ET processes and their drivers in various environments is important for the entire terrestrial hydrological cycle that governs the transport and recycling of the water that supports, for example, our fresh water supplies (Oki and Kanae, 2006) (Zeng et al., 2018). Observations of the Earth's atmosphere and biosphere over the last number of decades have indicated an intensifying hydrological cycle (Brutsaert and Parlange, 1998) (Pascolini-Campbell et al., 2021) and an increase in the number of people living in water stressed areas (Oki and Kanae, 2006). Modelling efforts over this period have shown disagreements, with evidence indicating a decline in global terrestrial ET caused by a reduction in available moisture supply (Jung et al., 2010) and more recently, indication of an increase in global terrestrial ET due to increasing land temperature (Pascolini-Campbell et al., 2021). ET is a process composed of two main parts: Evaporation (E), the physical process, and Transpiration (T), a biologically modulated process that occurs through the stomata of plants. A better understanding of the drivers of ET and the modelling of each of its constituent parts may be of significant importance to the monitoring and management of water resources globally over the coming decades. ET research contributes to many important components of global climate modelling including cloud formation (of relevance due to their role in the absorption and reflection of solar radiation and the transfer of energy between environments) and moisture availability (Gerken et al., 2018; Green et al., 2017; Pielke et al., 1998; Schlesinger and Jasechko, 2014; Trenberth et al., 2009). The partitioning of ET into its constituents is vital in reducing the associated uncertainty in climate land surface models and satellite remote sensing projects such as ECOSTRESS (Fisher et al., 2020) as current models are validated on combined ET data only (Stoy et al., 2019). The usage of machine learning (ML) in the domain of biosphere-atmosphere exchange has seen an increase in recent years with the availability of large, open source Eddy Covariance (EC) data sets such as FLUXNET (Baldocchi et al., 2001) and AmeriFlux (Novick et al., 2018) enabling more data

intensive approaches. Applications of ML in the domain of biosphere-atmosphere exchange have mostly focused on gap-filling of EC data (et. al., 2021) but some success has been achieved in the application of ML techniques to the partitioning of gas fluxes (Tramontana et al., 2020), prediction of fluxes (Tramontana et al., 2016), spatial interpolation (Lin et al., 2002), and upscaling of EC data (Bodesheim et al., 2018; Jung et al., 2009). As the EC method measures total water flux, the goal of partitioning in this work is to determine the individual contributions of E and T to the net flux.

Our previous work (Eichelmann et al., 2021b) introduced a novel, data-driven ET partitioning method and applied neural networks on micro-meteorological data collected from four wetland sites in California (Eichelmann et al., 2021b). Artificial neural networks (NN) were used to partition ET into E and T by training these networks to predict E during periods where T can assumed to be negligible. From this, T can be estimated by subtracting the predicted E from total ET. In this paper a broader range of ML algorithms are compared alongside the NN tested in the previous work, expanding on the complexity of the models via a novel feature selection process. The previous work utilised predictive features via domain expertise only and in this work additional features are selected via their correlation with the target and their effect on increasing predictive performance. This work seeks to address three research questions. Firstly, is there a ML algorithm that performs as well or better than those tested in our previous work on the task of predicting E? Secondly, can ML be utilised to identify an optimal set of predictive features that improves predictive performance? Thirdly, do the features identified contribute to our understanding of the processes mediating ET in the wetland sites in this study? Identical datasets to the previous work are used in this work (as described in Section 3.3.1) and the results from this previous work are used as a partial baseline for comparison in Section 3.5.

3.3 Background

3.3.1 Data

The data utilised in this work are obtained using the Eddy Covariance (EC) method (Aubinet et al., 2012) from measurement towers across four wetland sites in the Sacramento-San Joaquin river delta in Northern California: Twitchell Wetland West Pond (AmeriFlux ID: US-TW1) (Valach et al., 2021a)(WP), Twitchell East End Wetland (AmeriFlux ID: US-TW4) (Eichelmann et al., 2021a)(EE), Mayberry Wetland (AmeriFlux ID: US-MYB) (Matthes et al., 2021)(MB), and Sherman Island Restored Wetland (AmeriFlux ID: US-Sne)(Shortt et al., 2021)(SW). The locations

of the sites are displayed graphically in Figure 3.1.

Figure 3.1: Satellite view of wetland sites included in this study.

This method ascertains the flux of trace gases by measuring the covariance between fluctuations in vertical wind velocity and the mixing ratio of the gas in question. The data from all sites are available under an open-source license as part of the AmeriFlux network and can be accessed through the AmeriFlux data sharing platform (**?**Novick et al., 2018). The sites have been described in detail elsewhere ((Detto et al., 2010; Eichelmann et al., 2018; Hatala et al., 2012; Knox et al., 2015)) and the reader is referred to these works for a more complete description. The four sites are all freshwater marsh wetlands that have been constructed by the Department of Water Resources to manage soil subsidence in the area. The observation period for each site differs in length with approximately 10 years of data for MB (October 2010 to October 2020), 8 for WP (July 2012 to September 2020), 7 for EE (November 2013 to September 2020) and 4 for SW (May 2016 to April 2020). All sites, with the exception of WP, underwent flooding within the measurement period. The longest standing of the four sites is WP having been established in 1998. The initial flooding period is of note as it provides a period in which vegetation has not yet been established and thus, it can be assumed that T is negligible during this period. The vegetation cover (within the EC footprint at the latest measurement in 2018 (Valach et al., 2021b)) varies between the sites with 97% cover at WP, 64% at MB, 96% at EE and 45% at SW. The lower vegetation cover at SW can be explained by the fact that it is the newest wetland to be established, constructed in 2016. The dominant vegetation species at all sites are tule (*Schoenoplectus acutus*) and cattail (*Typha* spp.) (O'Connell et al., 2015). Continuous fluxes of water vapour and other trace gases were measured using the EC method. In addition to the EC data, micro-meteorological and environmental data were also obtained for each of the sites including the following variables with a known relationship with ET; air temperature (TA); water temperature (TW); soil temperature (TS); relative humidity (RH); atmospheric pressure (AP); net radiation ($RNET$); water table depth

(WT); vapor pressure deficit (VPD); sensible heat exchange (H); friction velocity ($u*$); vegetation greenness index from camera data (GCC) and the target variable water flux (labelled total ET or wq). The data frequency is at 30 minute intervals and where the data were recorded at higher frequencies, the mean was computed for that 30 minute period. Pre-processing of the data to remove spikes, filter for instrument malfunctioning and gap-filling procedure for certain data has been described in detail in (Eichelmann et al., 2018). For EC flux features with missing data, a NN procedure was used for imputation (Baldocchi et al., 2015; Knox et al., 2015) and is detailed in our previous work (Eichelmann et al., 2021b). Meteorological variables were imputed using data from nearby weather stations where data were available. For any remaining features with missing data, linear interpolation was used.

3.3.2 Machine Learning Algorithms

In this paper, a variety of supervised ML algorithms were utilised and the resulting models compared for performance on the prediction task described in Section 3.4.1. The algorithms tested can be broadly grouped into 3 categories: parametric regressors, non-parametric regressors and ensembles (Géron, 2019). The scikit-learn library (Pedregosa et al., 2011b) was used for model building in addition to the XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) libraries.

Parametric models, such as linear and ridge regression models, produce a predictive function by assuming a model with a fixed functional form and a finite number of parameters learned from data (i.e. optimised relative to the loss function obtained by comparing model predictions with ground truth values). Due to their simplicity, linear parametric models are extremely fast in training and prediction but can suffer from underfitting if the true distribution of the data is more complex.

Non-parametric models, such as K-nearest neighbours (KNN), decision trees (DT) and Support Vector Machines (SVM) make little or no assumptions about the functional form of the model seek to learn both the functional form and the function's parameter values from the data. Non-parametric models produce a more flexible predictive function, thereby allowing them to better model more complex distributions. This increase in complexity can lead to overfitting and an increase in both training time and the volume of data required to fit more complex models.

Ensemble methods such as Gradient Boosting, XGBoost and LightGBM, combine the predictions of many simple models, referred to as weak learners or base learners, to produce a predictive model. Base learners can be trained in parallel and combined using methods such as bagging or stacking, or sequentially using methods such as boosting (Géron, 2019). For all ensemble methods tested, the base learners were DT. Ensemble methods are generally less prone to overfitting while still retain-

ing sufficient complexity to arrive at a reasonable approximation of the underlying distribution.

3.4 Framework Methodology

This section begins with a description of the approach used for ET Flux Partitioning. The proposed framework is then described as the following series of methodological steps: data preparation, feature selection, construction of baseline models, final model training, evaluation and comparison. These steps are represented as a flow in Figure 3.2.

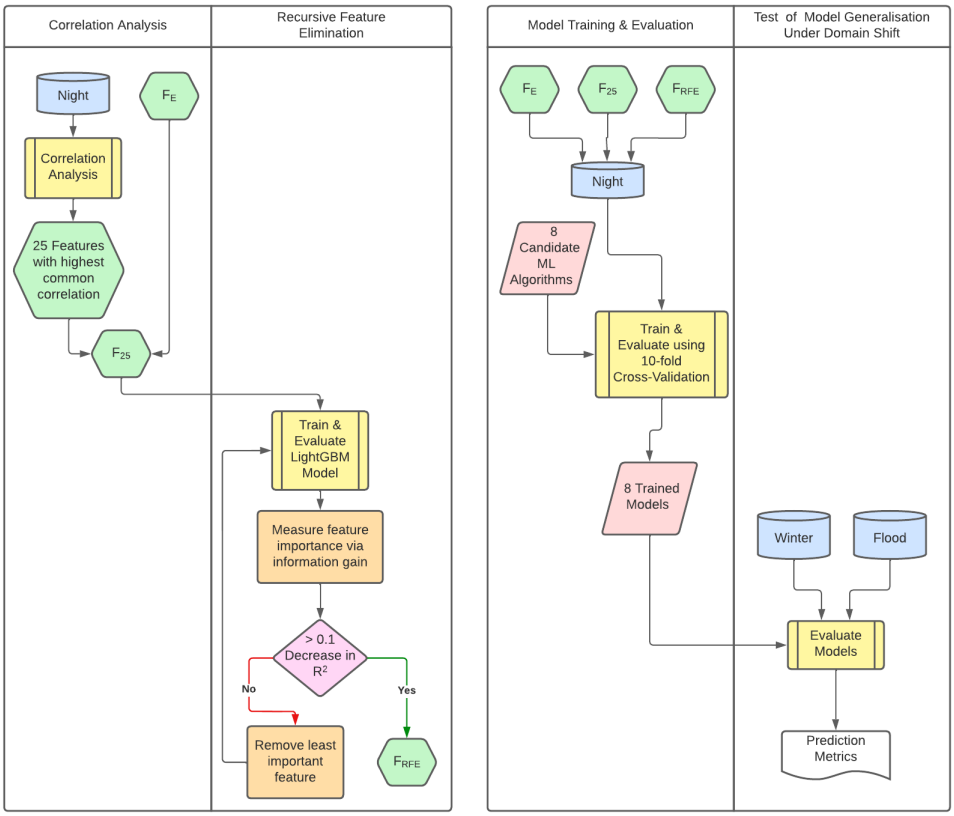


Figure 3.2: The process flow for the entire framework is split into two for the sake of legibility. On the left hand side the processes for obtaining the two additional feature sets, F_{25} and F_{RFE} are described. On the right hand side the processes for training and evaluating the models are described. Each orange rectangle represents a standalone process, each yellow rectangle represents a process with multiple components (the details of which are included in the text). Each green hexagon represents a feature set, each blue cylinder represents a data set, each red parallelogram represents a set of ML models and a pink diamond represents a decision. Some processes are carried out across all 4 sites, such as the Correlation Analysis. Other processes are carried out on each site individually, such as the Recursive Feature Elimination.

Each set of experiments were repeated across the four wetland sites for each set

of algorithms.

3.4.1 Evapotranspiration Partitioning

A novel, data-driven method to ET partitioning (drawing from previous work on carbon dioxide flux partitioning (Tramontana et al., 2020)) was presented previously in (Eichelmann et al., 2021b), with a brief outline here. Given the difficulty in establishing ground-truth data for the component contributions of E and T to overall ET, a number of assumptions are used to establish periods during which T can be assumed to be negligible and therefore, taken to be 0 in calculations. During the night, plant stomata are assumed to be closed and therefore not transpiring (confirmed with leaf level measurements at these sites). Utilising this assumption, the night-time data (Night) are used to train models to predict E, which can then be subtracted from total ET to give predicted the values for T. Explicitly the relationship between E and T can be expressed using Equation 3.1.

$$\begin{aligned} ET &= T + E \\ T_{Night} &\simeq 0 \\ ET_{Night} &= E \end{aligned} \tag{3.1}$$

As there are no measured ground truth data for the individual components E and T, assumptions about T during other periods of the year are used to determine two further test sets to evaluate the methodology, namely the day-time data from the initial flooding period (Flood) and day-time data from the winter senescent months (Winter). The principal purpose of evaluating with these test sets is to examine the performance of the models under a domain shift. Namely, this is the ability of the models to generalise to situations where the underlying distribution of the predictive features is different to that observed in training, which is performed on Night data only. Further information on this domain shift can be found in the supplementary material of our previous work (Eichelmann et al., 2021b). During the initial flooding period of each of the wetland sites, vegetation had not yet been established and therefore, T was not occurring. During the winter months the vegetation are observed to be senescent and here again, T is negligible. An additional set of core assumptions are used in determining the timing of the onset and duration of these periods. The zenith angle of the sun being greater than 90Doctor of Philosophy is used to determine the night-time periods. Visual determination of the level of vegetation from camera observation of the sites is used to determine the onset of vegetation after the initial flooding period, also referred to as "greenup". Lastly, the months of December, January and February are taken as the senescent periods.

Limitations of these assumptions are discussed in Section 3.5.5.

3.4.2 Data Preparation

In order to reduce the dimensionality (number of features) of the data to be computationally tractable for model building, the following approach to feature selection was undertaken. First reduce the candidate number of features using correlation and completeness analyses. Secondly, explore possible feature sets of different sizes and in different combinations using Recursive Feature Elimination. This process is described fully in Section 3.4.4.

As in (Eichelmann et al., 2021b), domain knowledge was used to inform the selection of features that have a known relationship with water flux. These were *VPD, GCC, u*, TA, RNET, WT, H* and ecosystem respiration estimated from an exponential relationship between night-time carbon flux and temperature as performed in (Reichstein et al., 2005) ($ER_{Reichstein}$). In addition, three time features were added: *year*, *month* and *the day of the year (DOY)*. This forms the first feature set for testing, denoted by the identifier F_E. All features with a completeness less than 80% (i.e. missing greater than 20% of the data) for the measurement period were discarded. Soil and water temperature measurements taken at various depths showed low levels of completeness and the depths at which measurements were taken was not consistent across sites. In order to obtain a usable feature for soil and water temperature that is comparable between sites, the measurements were consolidated by computing the mean of the sensor values across all depths to create two new features, *TS (mean)* and *TW (mean)*. For the remaining features, linear interpolation was used to replace the remaining missing data and a correlation analysis was undertaken to extract the most likely useful features. Of the 50 most highly correlated features at each site, the 25 features that were common across all sites in that subset of 50 were selected and added to the F_E feature set. The resultant feature set was labelled F_{25}. This approach was taken as the hypothesis is that features that do not have a correlation with the target feature that is common across multiple sites will be less likely to have an underlying physical causal relationship (i.e. the correlation is more likely to be spurious) and thus can be removed. It is noted that selecting the most highly correlated features that are common across the 4 sites is equivalent to removing the features that have no common correlation and the features that have a lower average correlation across the sites.

3.4.3 Model Comparison

In order to test and compare a suitably diverse set of algorithms for model building, initial testing examined 38 algorithms from the scikit-learn library (Pedregosa et al.,

2011b) alongside two additional ensemble algorithms; LightGBM (Ke et al., 2017) and XGBoost (Chen and Guestrin, 2016). The models are compared in order to ascertain which algorithm will be most suited to the model building task, including but not limited to an improvement in model performance in terms of fitting well to the training data, generalising well to unseen data and computational cost. Any models whose predictions had a negative coefficient of determination (R^2) with the target at any site were immediately discarded. From the remaining models, a subset of the best performing models (or simplest model in the case of equal model performance) were selected across 3 different categories of models; parametric, non-parametric and ensemble. The inclusion of different categories of ML algorithms is undertaken to prevent a loss in diversity from the initial set of algorithms tested. The models selected for final testing and comparison were linear regression, ridge regression, KNN, DT, Gradient Boosting Decision Trees, LightGBM and XGBoost. Default hyper-parameters were used for all algorithms.

To reduce the effect of sampling bias in training and testing the models, 10-fold cross-validation was applied. The Night data are split into 10 randomly sampled subsets (folds) for cross-validation. The models are trained and evaluated 10 times wherein at each iteration, one of the folds is removed and the models are trained on the remaining 9 folds. At each iteration, the models are evaluated on the held-out fold of the Night data as well as the entire Winter and Flood data. The process is then repeated with the previous fold replaced for training and the next fold removed for testing. At the end of the procedure, the mean value for each of the metrics is obtained.

3.4.4 Recursive Feature Elimination

In order to identify a feature set that contains maximal information with the minimum number of features a recursive feature elimination (RFE) method is used. A lower number of features is desired to reduce model complexity and subsequently reduce the chance of overfitting and to combat the so-called "curse of dimensionality" (Han et al., 2011) whereby an increase in the number of features leads to a lower number of samples per unit volume of the feature space. In this method, a LightGBM model is trained on Night data with the features from the F_{25} feature set for each of the four sites. Each model is then used to obtain a metric for the relative importance of each feature at that site. The metric that is used to measure feature importance is the sum of the gains in model performance, as measured by reduction in Root Mean Squared Error (RMSE), of all branches of base learner DT using that feature. The least important feature is then removed from the feature set. A new model is trained on the resulting, smaller feature set and the process is

repeated until no features remain. Cross-validation is applied at each iteration and the mean of the model performance metrics are recorded. The optimal features for each site are determined to be the feature set that preceded a 0.1% decrease in R^2 for the hold-out Night data as the number of features is iteratively decreased. The feature sets obtained for each site are then compared for commonalities and those features that were of low significance and in the optimal feature set for only one site are discarded and the remaining feature set is labelled F_{RFE}. It is hypothesised that this feature set approaches the minimum number of features needed to capture all information needed to model E and T from the available data.

3.4.5 Evaluation

In order to select the most appropriate set of metrics for evaluating model performance, the nature of the data must be taken into consideration. In contrast to a conventional supervised learning problem where the ground-truth data were obtained under known conditions, the ground-truth data for the experiments in this study are based on an assumption about approximate levels of T occurring under different conditions. During the night-time, winter and initial flooding periods, the assumptions governing negligible T are slightly different. Therefore, we expect that some common metrics for the evaluation of a regressor (e.g. RMSE and mean average error) may lead to difficulty in comparing model performance across test sets as the level of actual T occurring may vary and be non-negligible in some cases. This may lead to increases in measures of predictive error that are not attributable to poor predictive performance but rather to deviations in the data caused by a confounding variable that is not present in the training data (namely T arising in total measured ET where the model assumes that the total measured ET should be measuring E only).

Each of the Night, Flood and Winter datasets have different data distributions (Eichelmann et al., 2021b) and it is the performance of the models on data whose values lie outside the range of the training data (referred to as unseen data) that must be evaluated.

Therefore, a metric that determines how closely the variations in predictions of E follow the variations in total measured ET across all test sets is required. For this reason, the metrics chosen for evaluation are R^2, Adjusted R^2 (R^2_{Adj}) and slope of line of best fit between ground truth and predictions (m). R^2_{Adj} enables comparison between feature sets as this metric adjusts for the number of features used in order to account for the often spurious increase in R^2 when additional features are added to a model.

$$R^2_{Adj} = 1 - \frac{(1 - R^2) \times (p - 1)}{p - q - 1} \tag{3.2}$$

Equation 3.2 describes Adjusted R^2 where R^2 is the R^2 of the model, p is the number of samples and q is the number of features.

Slope is chosen in order to validate one of the biophysical constraints of any partitioning model, namely that the slope of the line of best fit between the predicted E and the ground truth (total ET) never exceeds 1 for any of the data. A slope greater than 1 would indicate that E had exceeded net ET which would lead to negative T, violating the biophysical constraint that negative T cannot occur.

At each iteration of the cross-validation procedure, the metrics are obtained for the removed fold of the Night data in addition to the entire Winter and Flood data. The mean of the metrics for all 10 iterations is then reported as the metric for that model.

3.5 Results & Discussion

In this section experimental results are presented where, for all four wetland sites, identical feature sets and experimental configurations were used. The results are reported on a per site basis as the goal is to compare how each of the models generalise to unseen data for the same site they were trained on. All results are the mean values of the metric across 10 cross-validation folds.

3.5.1 Model Comparison Results

Figure 3.3 shows the R^2_{Adj} values for all sites, algorithms and feature sets tested for the Night, Winter and Flood data.

Figure 3.3 shows that an improvement in model performance was obtained on Night data as well as in generalising to Winter and Flood data over and above that of the baseline results (Eichelmann et al., 2021b), where the baseline results are those that utilised NN-based models and the F_E feature set, indicated by a grey circular icon.

In general, results show that addition of the extra 25 features from the correlation analysis gave some improvement in model performance across all model types when compared to the baseline feature set, F_E. In addition, it is seen that reduction in features from 36 to 19 in going from F_{25} to F_{RFE} either resulted in further incremental improvement for the best performing models, or did not drastically decrease model performance. All sites had more than one model which failed to generalise well to the Winter and Flood data. This observation is important as it indicates that a site-specific approach to model building may be more favourable. The sites

modelled in this paper are all biologically similar: all wetlands with the same species composition, same climate and similar management. As described in Section 3.3.1 there are some known differences between the sites, such as the ratio of open water to vegetation cover and the utility of this framework may be best realised when building models that contain not only the general features relevant for modelling a particular ecosystem but also those features that are relevant for modelling that particular site. Gradient Boosting and LightGBM based models performed well at all sites except WP with LightGBM notably performing best on the Night data at all sites. At WP linear parametric models such as Ridge or Linear Regression performed best with most other models failing to generalise to Winter data at this site. The low computational time and resource requirements and high predictive performance on Night, Flood and Winter data would suggest that LightGBM or Ridge regression would be ideal candidate models in most cases. Many of the features are known to exhibit a high-level of non-linearity in their relationship with ET, particularly when considering the shift from night to day. For example, *RNET* is approximately constant and negative at night while positive during the day. This may go towards explaining why simpler models (such as linear parametric models or models that used less features) performed worse in some cases - the generated hyper-plane may not have sufficient complexity to model the underlying relationship between the predictive features and the target feature. It is also noted that most of the models failed to generalise well for the Winter data at WP, indicating that there may be particularities about this site that were not captured in the features or in the learned predictive function. This may be due to the fact that WP has differences in its composition to the other sites being the oldest of the 4 sites, as previously discussed in Section 3.3.1.

3.5.2 RFE Results

Figure 3.4 displays the results of the RFE process wherein features are iteratively removed from the F_{25} feature set until only 1 feature remains. The feature set selected for testing (F_{RFE}) is that which precedes a 1% reduction in R^2_{Adj} on the Night data. F_{RFE}, the feature set generated by the RFE process, contains all features from F_E except TA and $ER_{Reichstein}$. It is also noted that the temperature information may already be captured sufficiently in the TW or TS variables. It is evident that the number of features needed for an optimally performing model varies from site to site, indicating the difficulties in determining a universally optimal feature set. For example, at SW a model with just 3 features generalises best on Winter data and generalises better than the feature set chosen by the RFE process for that site. In contrast, a model with 6 features for that site generalises best on Flood data with

Table 3.1: Feature importance ranked in order of importance for each site where the features obtained by the RFE process are denoted by F followed by the site label and the relative importance of that feature at that site is given by I followed by the site label. Features that were omitted from the final feature set (F_{RFE}) are indicated by a strike-through, highly important features indicated in bold, features of interest in italics and the threshold for significant feature importance indicated by a horizontal line for each site.

Rank	F (EE)	I (EE)	F (SW)	I (SW)	F (MB)	I (MB)	F (WP)	I (WP)
1	**TW (mean)**	0.282	**u (mean)**	0.546	**u***	0.273	**H**	0.413
2	**u (mean)**	0.218	VPD	0.179	*wm*	0.146	**u***	0.208
3	*c (mean)*	0.123	u*	0.053	year	0.113	RNET	0.098
4	RH	0.087	H	0.044	RH	0.099	VPD	0.053
5	year	0.060	*wm*	0.032	VPD	0.094	RH	0.048
6	VPD	0.042	TW (mean)	0.028	u (mean)	0.082	DOY	0.024
7	WT	0.037	RH	0.025	TW (mean)	0.052	year	0.021
8	H	0.031	t (mean)	0.020	DOY	0.050	TA	0.018
9	u*	0.030	DOY	0.017	GCC	0.030	uw	0.018
10	uw	0.027	WT	0.015	H	0.027	GCC	0.016
11	DOY	0.015	uw	0.009	WD	0.014	~~uu~~	0.015
12	RNET	0.012	time	0.008			*wm*	0.013
13	TS (mean)	0.011	ww	0.007			WT	0.010
14	ww	0.004	RNET	0.004			WD	0.009
15	*wm*	0.004	GCC	0.004			time	0.008
16	~~ze~~	0.004	TS (mean)	0.004			TW (mean)	0.006
17	GCC	0.004	year	0.003			TS (mean)	0.005
18	~~$stat_q$~~	0.003					~~ts (mean)~~	0.004
19	t (mean)	0.002					~~sos~~	0.004
20	WD	0.002					~~er_{linear}~~	0.003
21							~~vv~~	0.003

a reduction in performance in generalising to Winter data.

3.5.3 Feature Importance

Table 3.1 lists the features selected using the RFE process for *each* site. Feature (F) columns rank features by their importance while Importance (I) columns give the relative proportion of total gain in performance contributed by that feature, normalised to sum to 1. The features that were not included in F_{RFE} are denoted by a strike-through. A full list of the features tested and their descriptions can be found in the supplementary material. As noted in Section 3.5.2, there is an overlap with the features previously selected using domain knowledge and many of the new features selected relate to processes that are known mediators of E and T. Our previous work highlighted the importance of VPD and $u*$ as they both relate to energy transport (Eichelmann et al., 2021b) which affects E as it is a form of latent energy. VPD is a measure of dryness of the air which increases transport of water across this gradient from high to low moisture and $u*$ is a measure of turbulence which also increases the transport of energy away from the surface. Most of the features identified by this framework can be grouped into their relationships with

E as the energy available for evaporation (*TW*, *H*, *RNET*, *TS*, *tbar*), the moisture gradient driving *E* (*RH*, *VPD* and *WT* to a lesser degree), the turbulent processes transporting water vapor away from the surface (*u**, *u (mean)*, *uw*, *ww* and *WD*) and the temporal patterns of ET (*year*, *DOY* and *time*). For description of variable labels please refer to Table B.1 in the supplementary material.

If a threshold of 0.2 is set for highly important and 0.05 ($\pm$10%) for significantly important, an examination of table 3.1 indicates 4 features (highlighted in bold) as being of high importance in accurately predicting ET: *u(mean)* at the EE and SW sites; *u** at the MB and WP sites; *H* at the WP site and *TW (mean)* at the EE site. If the features that are deemed to be highly or significantly important are examined it is observed that EE has 6, SW has 4, MB has 8, and WP has 5 features. This indicates that the majority of the predictive performance is attributable to these features. Two variables (highlighted in italics) which were unexpectedly ranked as important were carbon dioxide concentration (*c (mean)*) at EE and methane flux (*wm*) at MB. It is hypothesised that the relevance of *c (mean)* may be due to its connection to microbial activity via soil respiration wherein carbon dioxide and water are transported in the same way. The connection with *wm* is not as clear as there are multiple pathways through which methane can be released; diffusion, ebullition, and plant mediated transport. The fact that *wm* appears as an important predictive feature for *E* could indicate that there is mostly diffusive transport occurring which would follow the same physical processes as evaporation.

Identifying new features may reveal previously unknown connections between components of the system for further study with the potential to improve understanding of the underlying biophysical processes. This process is significantly enabled by this objective and data-driven framework.

While this work focused on using half hourly flux data, recent research on the use of high-frequency (10 or 20 Hz) EC data in the partitioning of methane fluxes (Iwata et al., 2018; Taoka et al., 2020) and in the partitioning of water vapour and carbon dioxide fluxes (Klosterhalfen et al., 2019; Scanlon and Sahu, 2008; Scanlon and Kustas, 2010; Scanlon et al., 2019; Skaggs et al., 2018; Zahn et al., 2022) provide an avenue for further research. The latter utilises the similarity between non-stomatal (respiration and E) and stomatal (photosynthesis and T) components, a methodology that could possibly be amenable to ML techniques or that could serve as a comparison for the outputs of our methodology.

3.5.4 Additional Results

All other sites were tested for the slope of line of best fit between total ET and the predictions for the Night, Winter, Flood periods as well as for the daytime

data outside of these periods. The single site for which a slope greater than 1 was observed was SW during the Winter period and only for the Decision Tree and Linear models. It should be noted however that the slopes for other models at SW were quite close to 1, indicating that the modelling of the Winter data for this site requires further investigation as these predictions violate the biophysical constraints outlined in Section 3.4.5. These findings are in line with those of our previous study (Eichelmann et al., 2021b). The details of these results can be found in the supplementary materials in Tables B.3, B.5, B.7 and B.8.

3.5.5 Limitations

Combining features across sites as part of the RFE feature selection process may have led to the inclusion of features that were site specific i.e. relevant to the predictions at one site but not adding useful information at another site. Therefore, this methodological pipeline may be more useful on a site specific basis to identify useful features for that site only and reduce them to the optimal number of features.

A large percentage of the data for the target has been imputed for all sites and additionally a small percentage of features were imputed with a variety of methods being used for imputation. Building models that use this data carry the errors and limitations of the imputation methods and may introduce noise to the data, particularly where linear interpolation was used. Gap-filling of the target data as well as H and wm was carried out using NN-based methods as discussed in our previous work (Eichelmann et al., 2021b), as well as linear interpolation of the remaining missing data used in these experiments, which may have effects on the error of our models. It is noted however, that there is no clear relationship between the linear interpolation carried out in this study and the performance of the models at any particular site. This is a topic that requires further investigation and the inclusion of more comprehensive methods for gap-filling that are outside the scope of the current work. More information on gaps in the data can be found in the supplementary materials (Figure B.1). Assumptions around the onset of the different periods where T is considered to be negligible may also lead to the introduction of noise to the target feature where T could be low but non-negligible.

Further model improvements could be obtained through the tuning and optimisation of the hyper-parameters of the models implemented. This forms a potential direction for further experimentation along with the testing and optimisation of different NN architectures exploiting the feature sets obtained in this research to allow for better comparison with previous modelling efforts (Eichelmann et al., 2021b). An investigation into the performance of the SW models which used the two smaller feature sets obtained from RFE may also yield further model improvements. Fur-

ther research should focus on determining if these methods generalise to other sites, including other freshwater marsh sites and other sites in the FLUXNET network of different types. The challenge for generalising this methodology to sites of a different land, vegetation or climate class is that the underlying physical assumptions may be different, potentially rendering the method inapplicable or requiring modification of the core methodology. A definitive choice of feature set or algorithm across all sites was not possible from the results of our model comparison, indicating that while some features may be common and of relevance across similar ecosystems (i.e. modelling two different wetlands) some features may be specific to a particular site. It is the combination of these more general features with more specific features that may lead to more accurate data-driven modelling of more heterogenous systems and the potential identification of previously unknown drivers or mediators of E and T for further study.

3.6 Conclusions

In this work a new framework by which climate scientists can test the efficacy of multiple ML algorithms and identify suitable predictive features from a high-dimensional candidate set has been presented. The result is a ranking of the candidate algorithms, a generally optimal feature set and an understanding as to how features contribute to model performance (predictive accuracy). For validation, micro-meteorological datasets were used with this framework to produce a model with an optimal balance between complexity and model performance. The framework adopts an objective (i.e. without usage of domain knowledge) view of feature selection and demonstrated an improvement on the baseline (Eichelmann et al., 2021b) which used a subjective approach to feature selection.

Algorithm ranking identified that ensemble models (such as LightGBM or Gradient Boosting) or linear parametric models would likely perform well on this task at other sites, generalising well to unseen data. However this was not a universal result, with simpler linear parametric models performing best at WP, indicating that there are key differences between the sites that necessitate an approach that tailors the models to individual sites.

The RFE process identified new features from the data that improved model performance. The use of information gain as a metric to iteratively remove features also allows for a direct comparison as to which features were most important at each site, providing the basis for further work, either in transferring these learnings to new sites or refining the models for these sites.

The examination of feature importance highlighted an obscure biophysical link in the case of carbon dioxide concentration and methane flux which improves our

understanding of the physical and biological processes involved.

In conclusion, this method provides new evidence of the contribution of ML to ET partitioning. The independence of the framework from explicit domain knowledge indicates that this approach may be domain agnostic, meaning that this method may have applications on other datasets, either for different EC flux sites or on entirely unrelated data.

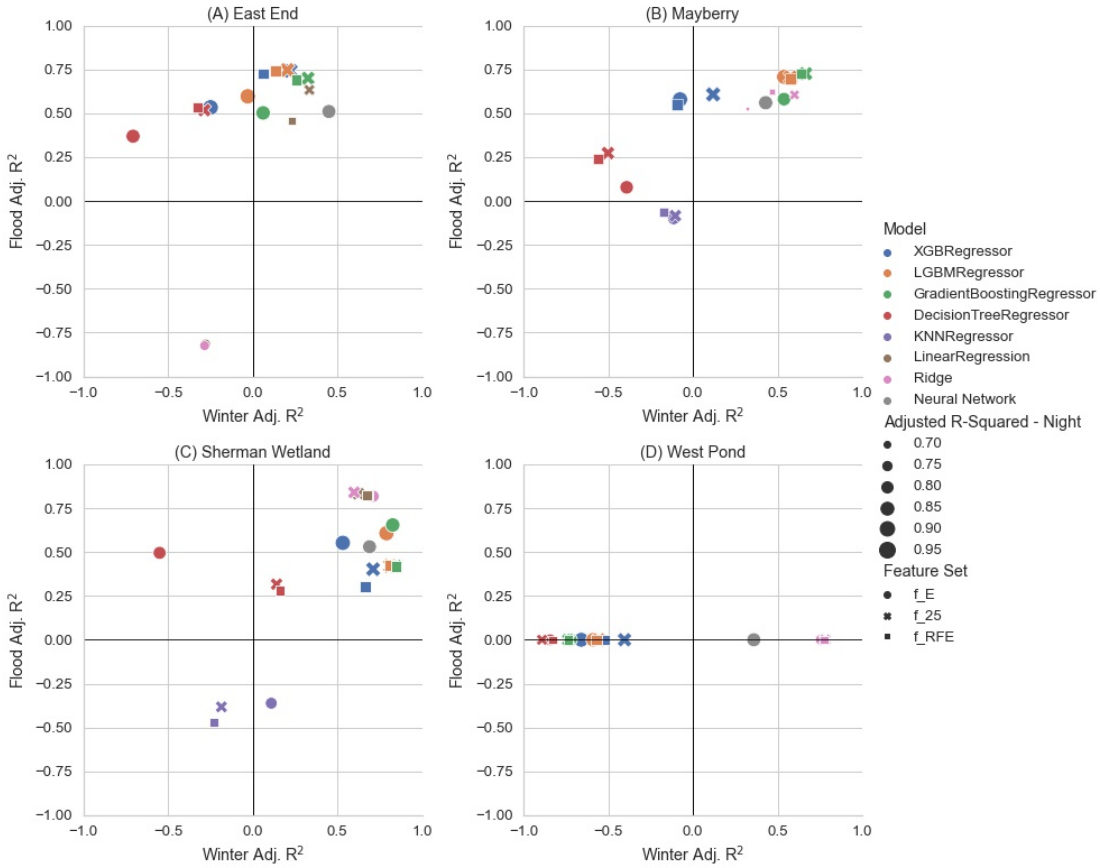


Figure 3.3: Results of model comparison for the four sites being studied. The x-axis plots the Adjusted R^2 (R^2_{Adj}) values for predictions on data from winter month and the y-axis plots the R^2_{Adj} values for predictions on data from the initial flooding period, testing the ability of the models to generalise to unseen data. The colour of the marker indicates the algorithm used in model building and the shape of the marker indicates the feature set being tested. The size of the marker indicates the R^2_{Adj} values for predictions on the hold-out Night data, demonstrating how well the models perform on data that is identically distributed to the training data. Therefore, the best performing models are those with the largest markers that are closest to the upper right corner of the graph. The x- and y-axis lines along the origin are displayed to allow for ease of identification of those models that fail to generalise well (i.e. models with $R^2_{Adj} < 0$). As WP does not have data from the initial flooding period, the results are displayed along the x-axis only.

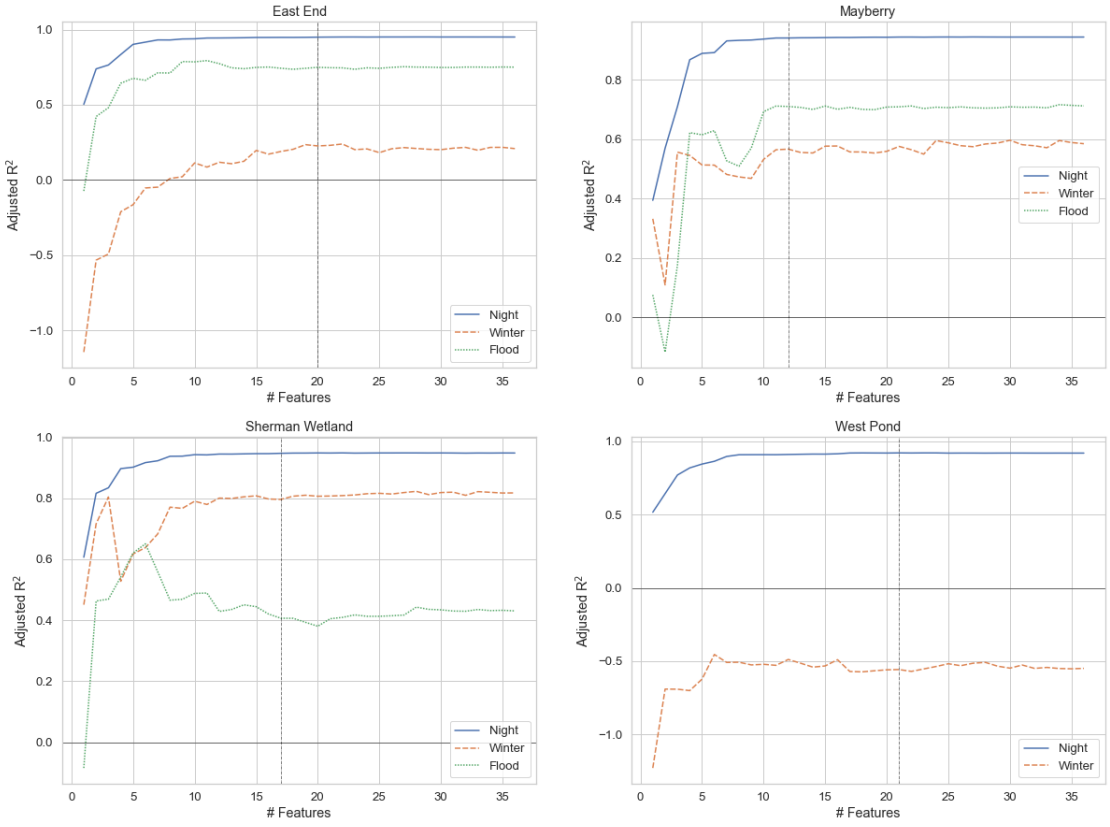


Figure 3.4: Results of the RFE process for each of the 4 sites tested with number of features on the x-axis and R^2_{Adj} results on the y-axis. The iterations start on the right and move towards 0 as RFE iteratively decreases the number of features until only 1 feature remains for each of the sites and each of the test sets; Night, Winter and (where available) Flood. A vertical line on each graph indicates the number of features selected, where the optimal feature set is determined to be the last feature set preceding a 0.1% reduction in R^2_{Adj}.

Chapter 4

Boundary Layer Height Modelling

The following chapter has been published in Journal of Geophysical Research: Atmospheres and the accepted version is reproduced here under the copyright agreement of the publisher John Wiley & Sons, Inc.
Stapleton, A., Dias-Junior, C. Q., Von Randow, C., Farias D'Oliveira, F. A., Pöhlker, C., de Araújo, A. C., Roantree, M., and Eichelmann, E. (2025). Intercomparison of machine learning models to determine the planetary boundary layer height over central amazonia. Journal of Geophysical Research: Atmospheres, 130(6):e2024JD042488

This chapter takes forward the key elements of the methodological framework presented in Chapter 3, refines and further validates the key components by applying them to a new dataset with different objectives for the ML prediction task. The aim of the study is to demonstrate that ensembles, identified in the previous studies to be efficient and accurate, can produce accurate models for these new data as well as produce new scientific hypotheses with the application of the feature selection methodology and simple gain-based feature importance as explanations. The study also introduces the Amazon rainforest environment as the biome being studied which is carried through to the final application in Chapter 5.

4.1 Abstract

The planetary boundary layer height (zi) is a key parameter in meteorology and climatology, influencing weather prediction, cloud formation, and the vertical transport of scalars and energy near Earth's surface. This study compares multiple Machine Learning (ML) models that predict zi from surface measurements at two sites in Central Amazonia - the Amazon Tall Tower Observatory (ATTO) and the Manacapuru site of the GoAmazon experiment (T3). Models were trained on ceilometer data with radiosonde measurements used for validation. We evaluated model performance by withholding approximately 10% of the data (as complete months) for testing, comparing predictions against ERA-5 reanalysis data using *RMSE*, *nRMSE* and R^2 metrics. Our results show that gradient boosted ensemble models using all available features perform best. A modified recursive feature elimination algo-

rithm identified minimal sets of 5-7 surface measurements sufficient for accurate *zi* prediction, demonstrating potential for wider spatial monitoring using cost-effective sensors. The study revealed previously unrecognized variables influential in determining *zi*, such as deep soil temperature measurements (40cm), suggesting new avenues for investigating land-atmosphere interactions. This study demonstrates the applicability of ML models to model *zi*.

Plain Language Summary

The height of the planetary boundary layer (PBL) is important for understanding weather, cloud formation, and the movement of pollutants and energy near the Earth's surface. This study compares different machine learning models to predict the height of the PBL using data collected from the surface, such as temperature, energy and meteorological measurements. The performance of these models is compared with radiosonde measurements and existing weather prediction data from the ERA-5 dataset at two locations in Central Amazon. This research highlights the potential of machine learning to improve our ability to model the PBL height accurately. This study found that you can predict the height of the PBL accurately using only 5-7 ground-based measurements. This is important because it suggests we could monitor boundary layer height over wider areas using simple, low-cost sensors combined with short-term measurement campaigns to obtain measurements of the PBL height for model training. The research also discovered some unexpected factors that influence boundary layer height such as soil temperature measured deep underground (40cm). This opens up new directions for study and helps us better understand what controls the height of the boundary layer.

4.2 Introduction

The planetary boundary layer (PBL) is the lowest portion of the atmosphere, directly influenced by the Earth's surface through exchanges of heat, moisture, and momentum (Garratt, 1994; Kaimal and Finnigan, 1994). Our understanding of the PBL has been greatly advanced over the last century including how it systematically varies both in terms of geography and its temporal structure as it relates to diurnal, seasonal and climatic scales (LeMone et al., 2019). The PBL therefore not only connects the surface and atmosphere of the Earth system but also bridges weather and climate as well as being the portion of the atmosphere in which we live and breathe.

The height of the PBL (*zi*) is a critical parameter in meteorological applications, affecting surface-atmosphere interactions, air quality, weather forecasting, the

atmospheric hydrological cycle and climate projections (Beamesderfer et al., 2022; Caughey, 1984; Helbig et al., 2021; Menut et al., 1999; Stull, 2012). Among these are the influence of the PBL on cloud formation, an active topic of research that is particularly relevant to the Amazon region (Andreae et al., 2004; Rosenfeld et al., 2016; Vilà-Guerau de Arellano et al., 2020). The importance of this research is underscored by the reiteration in consecutive reports from the Intergovernmental Panel on Climate Change that one of the most significant sources of uncertainty in current climate projections is that of clouds, their formation and evolution in climate-cloud feedback (IPCC, 2023).

Traditional methods for estimating *zi*, such as radiosonde measurements, offer high accuracy but are limited by their temporal resolution, preventing comprehensive diurnal cycle analysis (Guo et al., 2021; Seidel et al., 2010; Stull, 2012). Remote sensing techniques, including lidar systems, have addressed these limitations by providing high temporal and vertical resolution data (Hennemuth and Lammert, 2006; Sawyer and Li, 2013). Instruments such as Doppler lidar (Barlow et al., 2011; Tucker et al., 2009), ceilometer (Eresmaa et al., 2006; van der Kamp and McKendry, 2010), Raman lidar (Summa et al., 2013), micro-pulse lidar (Melfi et al., 1985) and aircraft sounding (Dai et al., 2014) have been employed to track the evolution of the PBL over a diurnal course, enhancing our understanding of its dynamics. Space-based methods have also been developed more recently such as passive infrared sounding, passive microwave sounding and Global Navigation Satellite System (GNSS) radio occultation (Teixeira et al., 2021) which provide greater spatial coverage.

The Amazon rainforest, with its unique terrain and climatic conditions, poses distinct challenges for observations of the PBL. The region's dense vegetation and high humidity complicate traditional techniques, requiring robust approaches. In the Amazon, field campaigns such as GoAmazon 2014/5 (Carneiro et al., 2016; Martin et al., 2016) and long-term sites such as the Amazon Tall Tower Observatory (ATTO) (Andreae et al., 2015; Souza et al., 2023) have provided extensive observational data. However, logistical and budgetary constraints have resulted in sparsity of data in these large remote regions, as observed with similar biosphere-atmosphere studies in the region such as the LBA experiment (Lahsen and Nobre, 2007).

Reanalysis data, such as those from the ERA-5 dataset (Hersbach et al., 2020), offer a promising alternative by combining observational and modeling data to provide spatio-temporally continuous information. ERA-5 is the fifth generation of atmospheric reanalysis of the global climate, produced by the assimilation of several observational datasets from satellites, meteorological and eddy covariance stations and radiosondes. It covers the entire global atmosphere and provides spatial and temporal data products for a range of meteorological and climatological variables (Hersbach et al., 2020; **?**). Studies have shown that ERA-5 provides a reasonable rep-

resentation of the atmosphere, though discrepancies with observational data, such as overestimation or underestimation of *zi*, are noted (Dias-Júnior et al., 2022; Zhang et al., 2020).For instance, Zhang et al. found that ERA-5 overestimated *zi* over the US by 18-41%. Comparisons with ERA-5 estimates over Amazonia show underestimates when compared with remote sensing observations, notably displaying discrepancies with the timing of the evolution of the PBL as it transitions between day and night (Dias-Júnior et al., 2022). One of the potential benefits of utilising machine learning (ML) techniques in this case would be to improve upon the accuracy and spatial and/or temporal coverage of estimates of *zi* in comparison to ERA-5.

Sleeman et al. have demonstrated the effectiveness of machine learning in denoising lidar/ceilometer backscattering profiles using denoising autoencoders to improve the detection of *zi* under cloudy conditions using convolutional networks. Other studies subsequently demonstrated the success of ML models in refining retrievals of *zi* for lidar backscattering in different environments using k-means (Liu et al., 2022; Rieutord et al., 2021), AdaBoost (Rieutord et al., 2021), support vector machines (Ye et al., 2021), and Gradient Boosted (de Arruda Moreira et al., 2022) algorithms.

Krishnamurthy et al. have predicted *zi* with radiosonde measurements as ground truth, using a random forest algorithm with surface meteorological data and parameters derived from Doppler lidar as input in the Great Southern Plains, USA. Subsequent works have used similar methodologies with surface meteorological data only as input. Molero et al. have tested four machine learning algorithms (linear regression, regression tree, support vector machine and Gaussian process regression) with ceilometer data as ground truth at a site in Madrid, Spain. Su and Zhang have applied a multi-structure deep neural network to a long-term (27 year) data set at the Great Southern Plains, radiosonde records as ground truth augmented with high-resolution micro-pulse lidar and Doppler lidar data. The authors also tested the model's ability to generalise to the GoAmazon (Martin et al., 2016) and CACTI (Cloud, Aerosol, and Complex Terrain Interactions; middle-latitude mountain) sites, showing good agreement with radiosonde ground truth with errors comparable with that of a ceilometer at those sites.

The most comprehensive study which adopted ML models to predict *zi* was developed by (Guo et al., 2024), assimilating data on a global scale covering the entire diurnal cycle of *zi* by integrating high-resolution radiosonde measurements, ERA-5 reanalysis, and the Global Land Data Assimilation System (GLDAS) product. However, this study did not include any observations in the Amazon or any other rainforest.

Contribution. This work derives estimates for *zi* using an intercomparison of

machine learning techniques using ground station data as input to the models. The novelty of this work is in the wider range of ML algorithms tested and the specific focus on the Amazon region. The paper utilises a novel methodology for feature optimisation, introduced in (Stapleton et al., 2022), wherein feature importance in LightGBM models is used as a heuristic to remove the least important features successively until the optimal feature set has been obtained. This optimal feature set demonstrates that highly accurate models for *zi* can be trained using few ground measurements (5-7), highlighting that wider modelling of *zi* may be possible using these inexpensive measurements alongside short-term *zi* measurement campaigns. The paper also includes a robust examination of feature importance, allowing further investigation into the functioning of these "black-box" algorithms and understanding the unique dynamics of the PBL and its drivers in the Amazon region. Important features were identified that were not known to have a significant influence on *zi* (e.g. soil temperature at 40cm), identifying new avenues for research and expanding understanding of the drivers of *zi*.

4.3 Data & Methods

4.3.1 Data

The data used in the study come from two experimental sites in the Central Amazonian tropical rainforest (Figure 4.1). The first site is the Amazon Tall Tower Observatory (ATTO, 02°08.752' S, 59°00.335' W, https://www.attoproject.org/, accessed on 31 Aug 2024) (Andreae et al., 2015), a permanent experimental site, established in 2012 and located in the Uatumã Sustainable Development Reserve approximately 150 km northeast of the city of Manaus in Brazil. The site's experimental compound includes the tall tower (325 m) and two 80m towers, along with a broad suite of meteorological sensors, trace gas and aerosol particle analyzers and a ceilometer. Radiosonde data are available only as part of fixed term campaigns and therefore have limited temporal coverage. The second site was established for a fixed duration as part of the Green Ocean Amazon (GoAmazon, https://www.arm.gov/research/campaigns/amf2014goamazon, accessed on 19 July 2024) campaign and is referred to as the T3 site (Martin et al., 2016). The GoAmazon campaign was conducted between 1 January 2014 through 31 December 2015. The T3 site was established approximately 70 km downwind towards the west-southwest of Manaus (03°12'36" S, 60°36'00" W), the furthest site from the pollution of the city out of the 9 research sites that made up the GoAmazon experiment. Meteorological towers, radiosondes, a ceilometer and surface flux instruments were among the experimental equipment at the site.

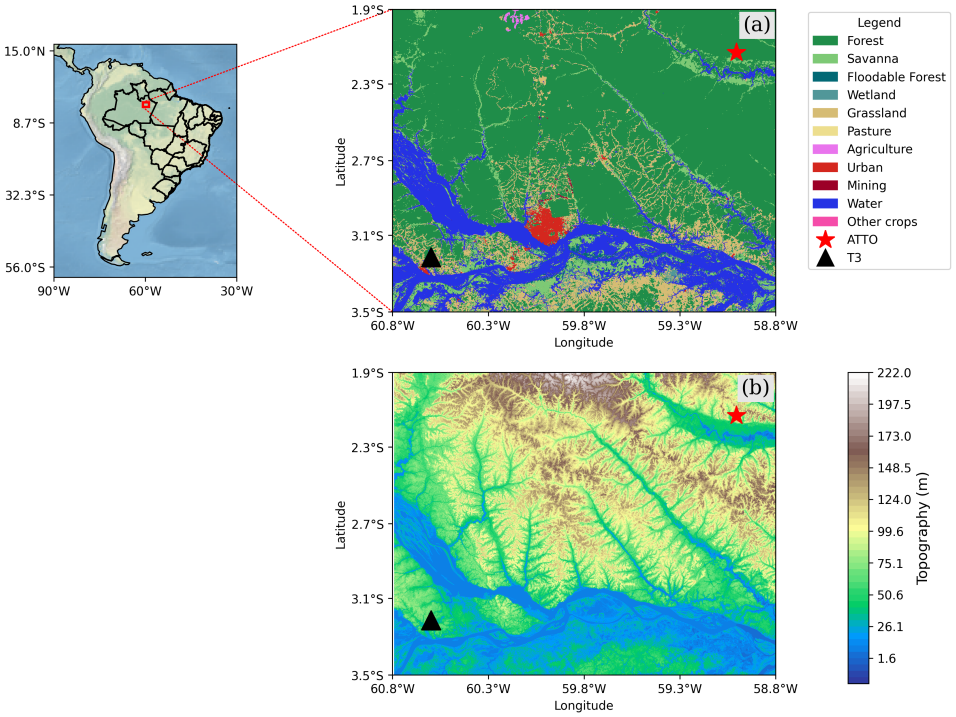


Figure 4.1: (Above) Geographic locations of the experimental sites T3 (black triangle) and ATTO (red star) in the Amazon basin with detail of the land use type; (Below) Topography of the study region. Data Sources: Land cover data: MapBiomas Project - Collection 8 of the Annual Land Use Land Cover Maps of Brazil; Topography: NASA Shuttle Radar Topography Mission (SRTM; 2013). Shuttle Radar Topography Mission (SRTM) Global. Distributed by OpenTopography. See Open Research Section for data access links.

Ceilometer

At the T3 site, a Vaisala CL31 ceilometer (Helsinki, Finland) was employed, while at the ATTO site, a CHM 15k-ceilometer (originally sold by Jenoptik AG, Jena, Germany, since 2014 sold by Lufft GmbH, Fellbach, Germany) was used. There are significant gaps in these data in 2016, 2018, 2019 and 2022 and no data recorded in 2017 and 2021. Both instruments use LIDAR-type remote sensing techniques, operating in the near-infrared wavelength range (900 to 1100 nm), to record the intensity of optical backscatter by emitting an autonomous vertical pulse. These LIDAR measurements, obtained every 16 seconds, depend on the aerosol concentration in the atmosphere, which is higher in the PBL compared to the free atmosphere above. The sharp drop-off of backscattering measurements in the entrainment zone (the zone wherein aerosols from the wet convective boundary layer and the dry atmosphere above mix) is used to determine *zi* (Cohn and Angevine, 2000). The variance V(z) at each height z is calculated as the squared difference between the backscatter signal R(z) and the mean backscatter (mR) computed from 15 m to 2500 m. The initial BLH is identified at the height of maximum variance. For

subsequent timesteps, the BLH is located at the maximum variance within a height range centered on the previous BLH value, with these ranges determined through visual analysis of daily backscatter profiles. During rain events, when precipitation is detected in the backscatter profile, BLH values are set to null.

The advantages of ceilometers include their ability to perform high-frequency measurements over long periods, providing a continuous three-dimensional mapping of aerosols with applications such as remote sensing of pollutants, and industrial and natural emissions. The primary drawbacks is their difficulty in *zi* determination which can lead to ambiguous results (Uzan et al., 2020). The ceilometer's high temporal resolution (16 seconds) and vertical resolution (∼100 meters) make it highly effective for tracking *zi* throughout its diurnal and nocturnal cycles, as demonstrated by previous studies (Geiß et al., 2017; Kotthaus et al., 2016; Morris, 2016). However it is noted that, at the ATTO site, the ceilometer may not accurately track the nocturnal course of the PBL due to difficulties in differentiating between the residual layer and the PBL using backscattering methods, leading to overestimation (de Souza et al., 2023).

As *zi* estimates depend on the aerosol concentration, measurement quality differs between the two sites. The T3 site, being closer to Manaus and receiving urban aerosols, provides more robust measurements of the PBL top compared to the pristine rainforest conditions at ATTO where aerosol concentrations are significantly lower (Cohn and Angevine, 2000). This is particularly relevant during nighttime measurements at ATTO, where the ceilometer may have difficulty differentiating between the residual layer and the PBL due to the very clean conditions in the nocturnal boundary layer. This can lead to overestimation of *zi* as the instrument may detect the residual layer height instead of the true nocturnal *zi* (Carneiro et al., 2021). Quality control comparisons with radiosonde measurements during the CloudRoots and CAFE-Brazil campaigns (Table 4.1) help validate the ceilometer measurements, showing good agreement with R^2 values of 0.85 and 0.80 respectively, though with some systematic underestimation as indicated by the negative MBE values. The seasonal and diurnal cycles for the ceilometer data at ATTO can be seen in de Souza et al.. Quality control flags are also used at the T3 site to remove erroneous data, the methodology for which has been described in Martin et al..

Here's the updated table with the number of soundings added:

It is acknowledged that, in general, the radiosonde may give a more accurate representation of *zi*, however ceilometer data are selected for training of ML models due to their greater temporal coverage and higher number of data samples for training. The reliability of ceilometer data in the Amazon region has been discussed in the literature and it has been shown to be more accurate than sodar, wind profiler, lidar

Table 4.1: Results of Quality Control (QC) of the ATTO ceilometer data compared with radiosonde measurements from two measurement campaigns.

Campaign	**# Soundings**	**MBE (m)**	**RMSE (m)**	R^2
CloudRoots	40	-41.49	209.46	0.85
CAFE-Brazil	61	-63.65	167.20	0.80

Note: MBE = mean bias error; RMSE = root mean square error; R^2 = coefficient of determination.

and microwave radiometer measurements when compared to radiosonde (Carneiro and Fisch, 2020).

Radiosonde

Radiosonde data are used as a comparative reference for validation at both sites, the results of which are discussed in Section 4.4.

At the ATTO site, radiosonde measurements were conducted using a Graw DFM-09 system (Germany). Launches occurred seven times daily at 02:00, 06:00, 08:00, 11:00, 14:00, 18:00, and 20:00 local time (LT). The radiosonde data were used to calculate vertical profiles of potential temperature (θ) and specific humidity (q), which were then used for *zi* determination. The radiosonde data at the ATTO site are limited to specific campaigns and include data from the Intensive Operating Period (IOP) in 2015 (Dias-Júnior et al., 2019), Cloudroots in August 2022 (Vilà-Guerau de Arellano et al., 2020; de Arellano et al., 2024) and CAFE Brazil (Curtius et al., 2024) in December 2022 and January 2023. The *zi* calculation method varied depending on the time of day. During the convective boundary layer (CBL) phase, the profile method was employed, where *zi* was identified as the vertical level exhibiting both an increase in potential temperature and reduction in specific humidity across three or more vertical bins. For the nocturnal boundary layer (NBL), *zi* was determined as the height at which the vertical θ gradient became either zero or less than 0.01 K km^{-}1 when measured from the surface. As the radiosonde data at the ATTO site are limited to specific campaigns, the feasibility of using these data in training the supervised ML models in this paper is limited.

The radiosonde instrumentation at the T3 site has been described in detail in Martin et al.,Carneiro et al. and Carneiro and Fisch, a summary is provided here. Radiosondes were launched using a Vaisala DigiCORA (MW12) system (Helsinki, Finland) coupled with RS92SVG radiosondes. The radiosondes were attached to meteorological balloons ascending at an average rate of 5 m s^{-1}. Regular measurements were conducted four times daily at 02:00, 08:00, 14:00, and 20:00 local time (LT), with additional launches at 11:00 LT during Intensive Operating Periods (IOPs)

to better characterize the convective phase. At the ATTO site the vertical profiles of the radiosonde launches are used to calculate the potential temperature (θ) and specific humidity (q), which then allowed to calculate the PBL height as follows: in its daytime phase, the heights were identified by the profile method (Heffter, 1983), in which *zi* is the vertical level with an increase in potential temperature and a reduction in specific humidity, for three or more layers (vertical bins). In the night phase, heights were determined where the θ vertical gradient was null or less than a defined number (0.01 km km^{-1}) from the surface (Carneiro et al., 2021). At the T3 site four methods are used to determine *zi* from the observed data; the Bulk Richarson method with critical thresholds of 0.25 and 0.5 (Richardson, 1920), the Heffter method (Heffter, 1980) and the Liu-Liang method (Liu and Liang, 2010). The Bulk Richardson method with a critical threshold of 0.25 is utilized throughout our analysis ceilometer measurements are found to agree more closely with this method.

4.3.2 Methods

This study employs a systematic approach to evaluate machine learning models for predicting planetary boundary layer height. The methodology consists of four main components: (1) data collection and pre-processing from two sites in the Amazon region, (2) implementation of multiple machine learning algorithms ranging from simple linear models to complex ensemble methods, (3) feature selection through recursive feature elimination, and (4) comprehensive model evaluation using multiple metrics and cross-validation. Each component is designed to address specific challenges in PBL height prediction while maintaining scientific rigor and reproducibility. The methodology of this work can be summarised as follows: High frequency (0.0625 Hz) ceilometer measurements of *zi* are pre-processed as described in Section 4.3.2 and upsampled to 30 minute frequency to match the input data. Gaps in input data are filled using a modified version of the Mean Diurnal Variation (MDV) gap-filling method (Falge et al., 2001) as described in Section 4.3.2. Approximately 10% of the data (as entire months) at each site are removed from the training set for each site. These data are used for model validation and evaluation. Individual machine learning models for each site are trained using three different feature sets as input; *Time* (*Year, Month, Day, Hour*), *All* (all available features) and *Optimal* (features obtained by RFE process). Model predictions are compared with ground truth and evaluation metrics (see Section 4.3.4) are obtained.

Data Pre-processing

At the ATTO site high frequency data of backscatter profiles were averaged across a 5 minute window which was then used to determine *zi*, resulting in 288 values daily. This is done so as to reduce the uncertainties associated with the determination of *zi* from the high-frequency data and to make the estimates more robust. The full procedure for the detection of the PBL from backscattering data has been described by de Souza et al.. At the T3 *zi* values are obtained for every backscatter profile resulting in values every 16 seconds. These high frequency data from the were cleaned before upsampling to remove any sudden jumps in values, outliers (detailed below) and other data that may be due to sensor error rather than true measurements. Any entries with a quality flag of 1 were ignored and all other original data were kept before pre-processing. The data are were then upsampled to 5 minute frequency. In order to match with the frequency of the input data, *zi* measurements from both sites were further upsampled to 30 minute frequency, improving their robustness by averaging across a 30 minute window. Some days were excluded from analyses due to abnormal PBL cycle behavior, such as *zi* maxima exceeding 3 km, and high percentages of missing data.

Outlier Detection. At the T3 site outliers in the high-frequency *zi* data are removed by standard deviation thresholds in a rolling 1 hour window. In order to obtain a continuous time series over which to conduct rolling window analyses, the high-frequency data is temporarily filled using the mean of the values at the same 16 second timestamp across that entire month. If a point y_t is outside of 1.5 standard deviations of the data in a 30 minute window either side of that point, the point is removed. Next filtering of spikes is conducted using a custom algorithm using any of the three following criteria to remove points y_t: (a) if the difference between y_t and its neighbour y_{t-1} is 200 metres or greater; (b) if the relative change in value is greater than a threshold (i.e. $|y_t - y_{t-1}|/y_t > 0.5$); (c) if the deviation from the mean is greater than 0.5 times the mean (i.e. $|y_t - \hat{y}|/\hat{y}$ where $\hat{y}$ is the mean of y at that time across that month). Here y is *zi*, measured in metres. These criteria are chosen as they indicate (a) non-physical changes in the boundary layer height over a 16 second interval (b & c) difficulties in determining nocturnal *zi* from backscatter profiles leading to low magnitude but high relative changes in the data. The thresholds and window sizes are chosen by testing a range of values and visually inspecting the data so that the spikes that are determinable by eye are removed while minimising the the introduction of gaps to the data.

Quality Control of input data Quality control checks were carried out on the input data at both sites ensure data integrity and reliability. Visual inspection of time series of all data revealed periods requiring targeted interventions, including:

the recalculation of longwave radiation values for the 2018-2019 period, the consolidation of Photosynthetically Active Radiation (PAR) measurements at multiple measurement heights after 2021, and selective removal of variables during periods of questionable instrument performance (e.g., friction velocity (u*) from 2020-2022, momentum flux from 2020). Features missing more than 50% of the data were removed.

Gap Filling.

Gaps in the input data at each of the sites are filled using a combination of two methods. Missing values in the target variable (zi) did not undergo any gap-filling to maintain the integrity of model validation and prevent artificial inflation of performance metrics. Where possible the data are filled using a modified version of the MDV procedure introduced by (Falge et al., 2001). In this method, the mean of data at the same time of day in the adjacent days in a 14 day window (7 days either side of the gap) are used to fill the gap. Where available, the data in a 14 hour window (7 hours either side of the gap to be filled) were also used due to observations that many of the variables have high autocorrelation within this time period. In this case the mean of the hourly values within both a 14 hour window and at the same time of day in a 7 day window is used. Where there were no data within a 14 day window to fill the gaps, the gaps were filled using the mean of that feature at that time of day across the entire month. The method is tested in terms of R^2 score in its skill at predicting the existing data in order to determine how it may extrapolate to the gaps. If the R^2 score was below 0.5 then a LightGBM model using all other available features was trained and the predictions used to fill gaps in the data. The LightGBM model was more accurate across all features but was not used to fill all gaps as this may introduce multicolinearity between features that may impair the performance of the final predictive models. It was also observed in testing that LightGBM model predictions had different range and standard deviation than the observed data for some features making the extrapolations less trustworthy. The modified MDV method was chosen due to its simplicity and robustness and does not suffer from these issues.

4.3.3 Machine Learning

Machine learning is a broad family of computational techniques that utilise data to build models. This paper compares different machine learning algorithms for the supervised multivariate regression task of predicting zi from surface-level meteorological, energy flux and soil data. The models are chosen for diversity and fall into 5 categories: linear models, neural networks, tree-based, clustering and ensem-

bles. The regression algorithms tested were linear (least-squares), k-means , decision tree, random forest (Breiman, 2001), Gradient Boosted Machines (Friedman, 2001), LightGBM (Ke et al., 2017) and a multi-layer perceptron Neural Network (NN). These algorithms, their basic working principles as well as their benefits and drawbacks are described in C.0.1 Table C.1. The algorithms are implemented in python using the scikit-learn package (apart from LightGBM (Ke et al., 2017)and XGBoost (Chen and Guestrin, 2016) which are standalone packages), a common and efficiently implemented scientific framework for machine learning (Pedregosa et al., 2011a). Hyperparameter and architecture optimisation experiments were not included in the results as trial experiments showed that this significantly increased computation time (and therefore energy expenditure) and resulted in only marginal gains in performance (2-3%) or decrease in performance as compared to the default hyperparameters of the models used.

In order to validate the models, a systematic sub-sample of the data was removed from the training data and used to compare the predictions of the model against this ground truth data which the model has never seen. This subset of the data is referred to as the holdout or test set. In testing the ML models 10-fold cross validation is employed - roughly 10% of the data is removed by selecting random months from the dataset, repeated 10 times so as to prevent the experiments from being biased by the month selection. For the purpose of running an individual LightGBM model to display predictions against ground truth, ERA-5 and radiosonde (see Figures 4.2, 4.3, 4.7, 4.8) the test set for ATTO consisted of all data for the months of October 2015, April 2016, October through December of 2019 and August 2022. For T3 the months of April 2014, October 2015 and November 2015 were used. These periods are chosen to satisfy (a) periods where radiosonde data were available at the ATTO site, (b) an overlap in time periods between the two sites, (c) a distribution of different periods of the year and (d) the end of each dataset where the models are extrapolating forward in time to an unseen period of data rather than interpolating between observation periods. This is significant for 2022 and 2023 data as there is a significant gap with the nearest previous year of data and the periods selected are the only measurements available for this year and therefore is a strong test of the models' ability to generalise to unseen data.

Recursive Feature Elimination

In order to obtain a Pareto optimal set of features (i.e. the lowest number of features resulting the highest performance on the out-of-sample test set relative to the model using all available features) a Recursive Feature Elimination (RFE) algorithm, informed by feature importance as a heuristic, is applied. The RFE process proceeds as follows: First, a LightGBM model is trained using all available features. Feature

importance scores are calculated based on the cumulative reduction in training error attributed to each feature across all decision trees in the ensemble. Features are then iteratively removed one at a time, starting with the lowest importance score. At each iteration, a new model is trained with the reduced feature set and its performance is evaluated. This process continues until the R_2 score decreases by more than 0.05 compared to the initial model. The feature set that achieves the best performance while minimizing the number of required measurements is selected as optimal.

4.3.4 Evaluation Metrics

In order to evaluate and compare model performance, the following metrics are used: in eq.4.1, coefficient of determination (R^2); in eq.5.7, Root Mean Squared Error ($RMSE$); and in eq.4.3, normalised RMSE ($nRMSE$), Mean Average Error (MAE) and Mean Bias Error (MBE). Across the 3 equations: $\hat{y}_i$ denotes the predicted value of the i-th sample; y_t is the corresponding true value for N total samples in the test set; and σ is the standard deviation of y_t in the test set.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_t - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_t - \bar{y}^2)} \tag{4.1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_t - \hat{y}_i)}{N}} \tag{4.2}$$

$$nRMSE = \frac{RMSE}{\sigma} \tag{4.3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_t - \hat{y}i| \tag{4.4}$$

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y}i) \tag{4.5}$$

These metrics are chosen for the following reasons: R^2 quantifies the proportion of variance in the dependent variable that is predictable from the independent variables, providing a measure of model performance relative to a naive mean-value predictor; $RMSE$ and MAE can heuristically be thought of as measures of the average error in metres that the model makes in predicting zi; $nRMSE$ allows for direct comparison of the relative error across sites where the magnitude and variation of zi (in terms of mean and standard deviation) are not equal; MAE and MBE are give estimates of how much the models are over- or under-predicting zi and, in the case of of MBE in which direction.

4.4 Results & Discussion

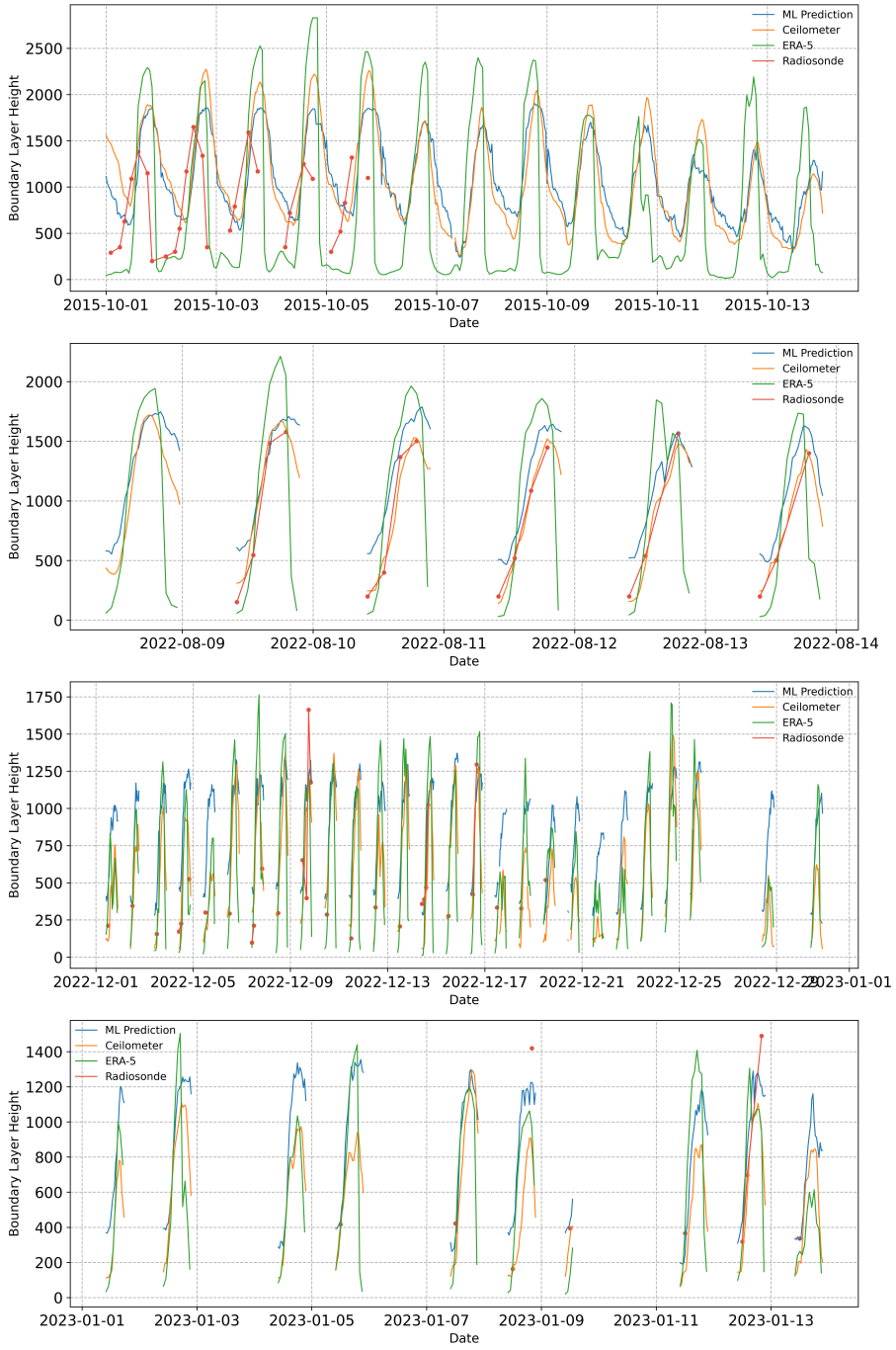


Figure 4.2: Comparison of boundary layer height (zi) measurements and predictions at the ATTO site during hold-out test periods. Blue line shows ceilometer measurements (ground truth), orange line shows predictions from the LightGBM model trained on all available features, green line shows ERA-5 reanalysis predictions, and red points indicate radiosonde measurements. Gaps in the time series indicate periods where ground-truth data were unavailable. The selected periods demonstrate the model's ability to generalize to unseen data across different seasons and years.

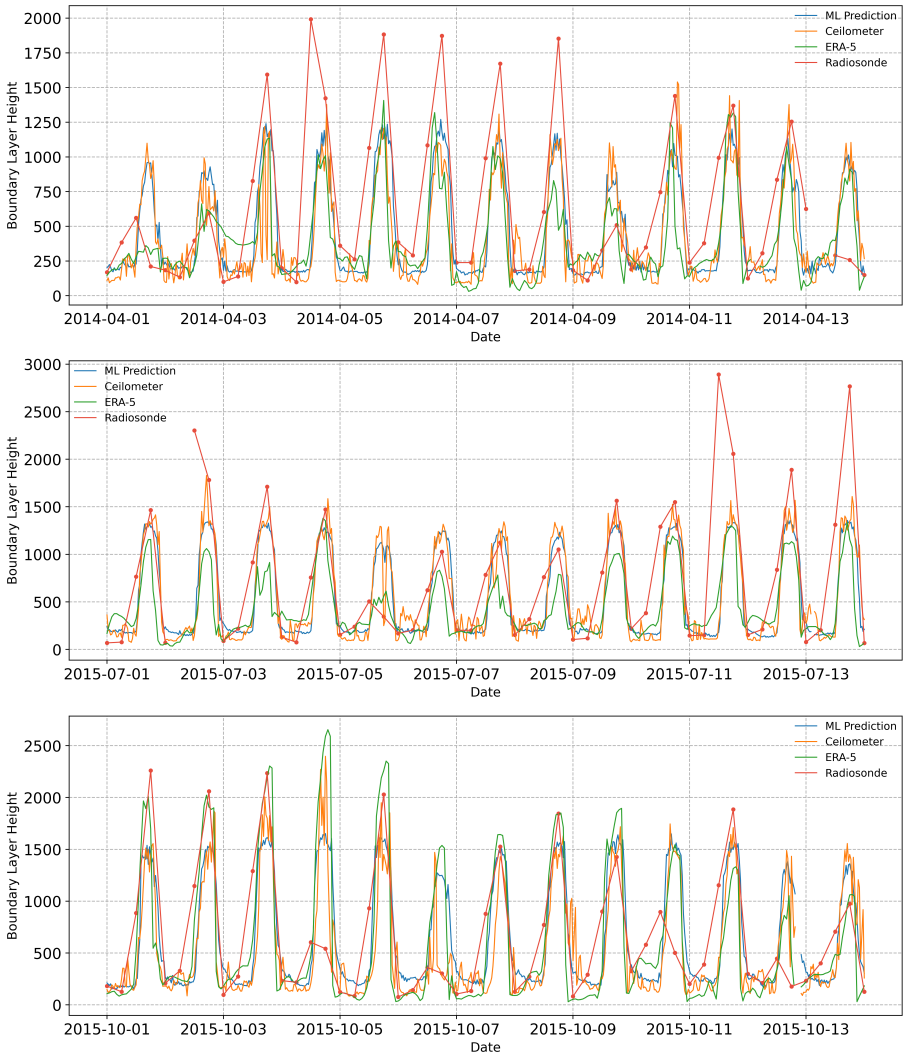


Figure 4.3: Ground truth for boundary layer height (zi) as measured by ceilometer (blue) at T3 compared with zi predicted by the Light Gradient Boosted Machines (LGBM) Regressor (orange), ERA-5 predictions (green) and radiosonde measurements (red). Where no ground-truth data were available, the data are omitted. The data are from hold-out test months, demonstrating the model's ability to generalise to unseen data.

Figures 4.2 and 4.3 present the ground truth ceilometer values for *zi* compared to the predictions of a LightGBM model trained on all available features and the ERA-5 predictions from the nearest available grid point for each of the two sites, ATTO and T3, respectively. The months selected are those data that were held-out for testing and not used for training the machine learning models and therefore provide a test of how well the model predicts *zi* for unseen periods. Visual inspection of both Figures reffig:predictions-ATTO and 4.3 indicates that the LightGBM predictions more closely follow the diurnal course of *zi* in general as compared to ERA-5. Radiosonde measurements of *zi*, available 4 times daily at the T3 site and at varying times at the ATTO site depending on the specific campaign, are also included in Figure 4.3 for comparison. It is important to note that the LightGBM model, having been trained on ceilometer measurements of *zi* may not accurately represent the *true* value of *zi* due to the shortcomings of the ceilometer measurements. Therefore, the evaluation metrics presented later in Figure 4.4 and Table 4.2 should be interpreted with the caveat that the model performance is with respect to the specific task of predicting ceilometer *zi*. This is particularly important when examining night-time predictions as it is known that both ceilometer and radiosonde measurements may overestimate nocturnal *zi*, as discussed in Section 4.2.

Figure 4.4 presents the out-of-sample test *RMSE* and R^2 results for all model configurations tested. The best machine learning models at both sites show noticeable improvements in predicting the ceilometer measured *zi* compared to ERA-5 predictions across all metrics where, in general, ensemble methods demonstrated superior predictive performance, with gradient boosting algorithms (LightGBM, XGBoost, Gradient Boosted Machines) benefiting from explicit regularization mechanisms, while Random Forests control model complexity through bagging and random feature selection at each split. These approaches effectively mitigate overfitting while capturing complex non-linear relationships. Multi-Layer perceptron NNs do not perform well and more complex models that utilise time series information (such as Transformers, Recurrent NNs and LSTMs) could potentially increase performance. These models come with the caveat that continuous data are needed and therefore the target data would have to be gap-filled first. We argue that the need for extensive tuning (which is expensive both in terms of time and computational cost) as well as the increases in prediction time (LightGBM models run orders of magnitude faster than complex NNs in training and prediction (Ke et al., 2017)) and the limitation of needing continuous data for sensors that lead to many missing values are sufficient reasons to advocate for the use of Ensemble methods over neural networks for this task.

Table 4.2 displays the results for the best performing model across all site and input feature set combinations. LightGBM is selected over Gradient Boosting for

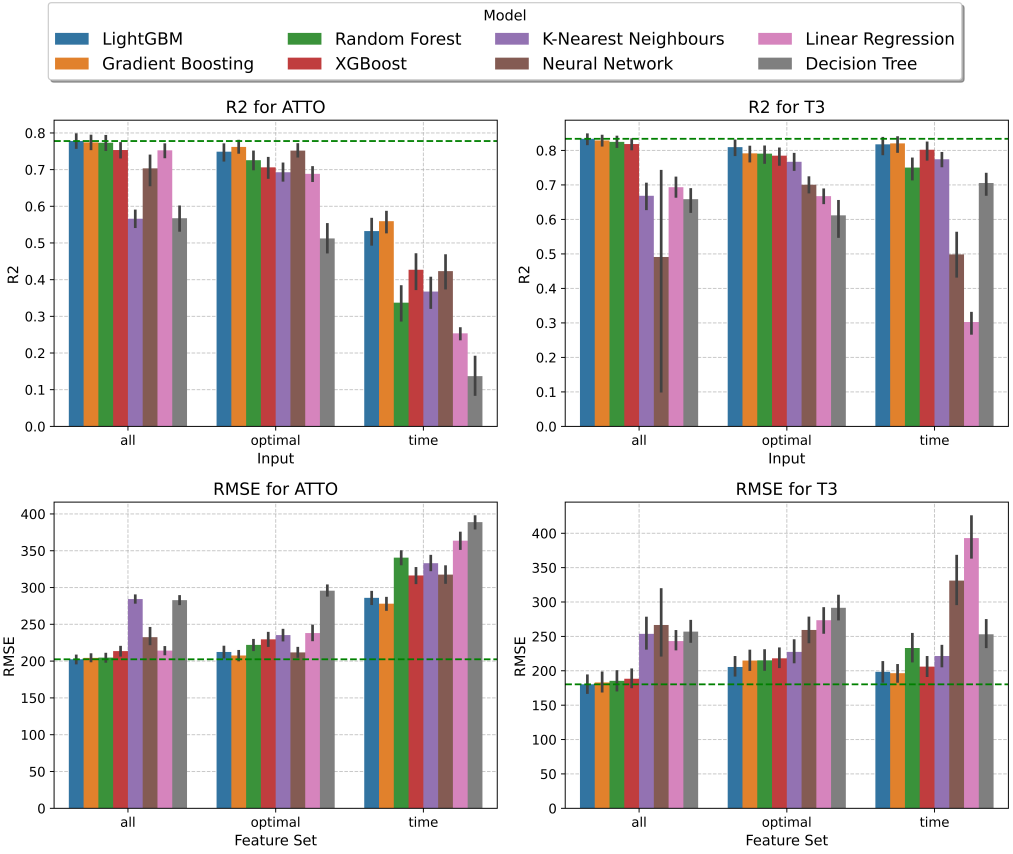


Figure 4.4: Root Mean Squared Error (RMSE, top panels) and R^2 (bottom panels) results for ATTO (left) and T3 (right) for all models and input feature sets for the hold-out test months. Models are divided into groups based on the input feature sets *Time* (*Year, Month, Day, Hour*), *Optimal* (features obtained by RFE process) and *All* (all available features), and juxtaposed with the evaluation metrics for ERA-5 predictions for the same test data. The colour of the bar indicates the machine learning algorithm being tested. Horizontal green lines indicate the best performing model.

Table 4.2: Model results for LightGBM across different input types and periods at T3 and ATTO sites.

Model	**Input**	**Period**	R^2	**RMSE**	**nRMSE**	**MAE**	**MBE**
			T3 Site				
LightGBM	All	Day	0.79	210.84	0.45	146.54	1.36
		Night	0.24	145.18	0.87	110.96	3.59
		Both	0.84	181.95	0.40	129.10	2.43
LightGBM	Optimal	Day	0.72	239.59	0.52	170.84	6.93
		Night	0.24	148.65	0.87	114.52	6.43
		Both	0.80	200.42	0.44	143.21	6.69
LightGBM	Time	Day	0.75	242.30	0.50	168.23	-14.78
		Night	0.21	144.62	0.89	110.14	1.80
		Both	0.82	200.63	0.42	139.67	-6.62
			ATTO Site				
LightGBM	All	Day	0.82	193.71	0.42	139.23	-2.72
		Night	0.62	205.60	0.61	157.76	20.57
		Both	0.79	200.25	0.46	148.63	9.04
LightGBM	Optimal	Day	0.79	202.41	0.45	147.18	-4.82
		Night	0.51	228.50	0.70	172.78	10.69
		Both	0.74	216.42	0.51	160.19	3.04
LightGBM	Time	Day	0.49	309.97	0.71	238.88	-1.04
		Night	0.31	262.75	0.83	202.73	15.05
		Both	0.52	287.31	0.69	220.65	7.05

Note: R^2 = coefficient of determination; RMSE = root mean square error (m); nRMSE = normalized root mean square error; MAE = mean absolute error (m); MBE = mean bias error (m).

consistency as there are cases, as shown in Figure 4.4, where Gradient Boosting performs slightly better but this is outweighed by LightGBM's speed in training and inference. Across both sites and all input features the performance of models in predicting night-time data is considerably lower in terms of R^2. This may be due to complications in determining ceilometer *zi* at night, as discussed in Section 4.3.1. In examining MBE results it is noted that LightGBM models tend to overestimate *zi* and this is also observed for other model types. The best performing models in terms of RMSE and R^2 are those that used all available input features for training and prediction. However it is noted that the difference in R^2 scores between the best models and the models using the optimal feature set as input is approximately 0.05 for both sites. This difference in performance may be considered negligible, particularly for applications requiring simpler models. The advantages of simpler models are that they are more likely to generalise well and are easier to understand and explain. In addition they depend on fewer surface measurements, making them more robust and cost efficient in terms of the instrumentation needed to make measurements for prediction. Overall it is observed that models performed better at T3 than at ATTO in terms of R^2 and *nRMSE* scores. This may be due to the higher continuity of the T3 data. While the ATTO data covers a longer time period, it is sparse for similar dates across multiple years, leading to under-representation of certain months in the training data.

Additional figures analysing the spread and goodness of fit of LightGBM models across all CV folds have been add to scatter plots in figures C.1 and C.2 in the appendix. In addition, seasonal analyses of the daily and weekly daytime mean and max *zi* values have been added in Figures C.3, C.5, C.4, C.6 C.7, C.7, C.8 and C.10. ML predictions maintain good agreement with ceilometer measurements across these different temporal scales, while ERA5 reanalysis tends to show systematic differences in *zi* estimation.

4.4.1 Input Feature Sets

When examining the groups of input features in Figure 4.4, we observe a discrepancy between the ATTO and T3 results. Models trained on only the constructed temporal features (*Year , Month, Hour, Day*) performs similarly well to models trained on all available features for certain ML algorithms. This may be due to the fact that there is a clearer and more consistent diurnal pattern at the T3 site compared to the ATTO site (Figures 4.2 and 4.3). It is also noteworthy that the largest difference in both RMSE and R^2 values is observed for the models trained on temporal features only. Here it is observed that Decision Tree and Random Forest models performed extremely poorly while more complex models such as LightGBM

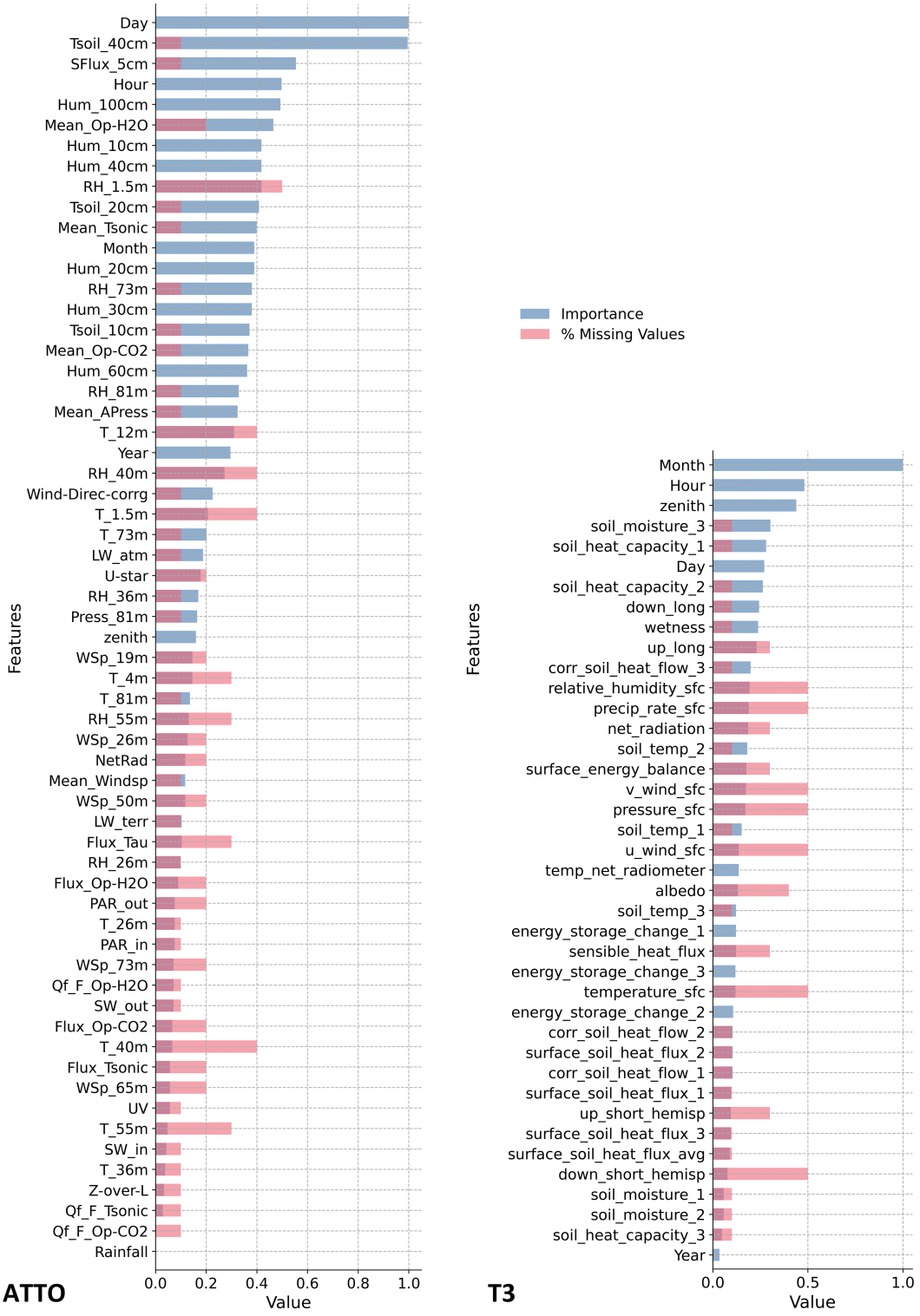


Figure 4.5: Relative feature importance (feature importance divided by the maximum feature importance score, blue) for a LightGBM model trained on all available features at the ATTO site and T3 site with % of missing values plotted alongside (red). For a description of the feature names refer to Tables C.2 and C.3.

and Gradient Boosting performed better than models trained on all available features at the T3 site. This demonstrates that there is a Pareto optimality and a trade-off between increasing the complexity of the models and keeping them as simple as possible to fit the data without over-fitting and reducing the models' ability generalise to unseen data. It also indicates that sufficiently well performing models can be trained in conditions in the absence of any measurements other than the target variable. This would imply sufficient information about the underlying processes is already contained in the time series of *zi* in this case. It also highlights the importance of ML model selection in the low data limit whilst it may be noted that the choice of model is not as critical given sufficient training data.
There are multiple elements that make it difficult to tease out the key factors that differentiate these two sites and the reasons for the discrepancies in key predictor variables. One aspect to consider are the physical differences between the sites in terms of heterogeneity and biological and physical makeup of the local environments. For example, site T3 is surrounded by large rivers as well as man-made perturbations (e.g. roads, agriculture), while ATTO is a site with almost 100% forest cover, and the topography of the two sites is quite different (Figure 4.1). There are also differences in the availability and characteristics of the data at the two sites. The variables measured do not match exactly and there are a greater variety of sensor measurements available at the ATTO site. There are also differences in the number of granular measurements for certain variables (for example there are 9 air temperature measurements available at ATTO to 1 surface air temperature measurement at T3). In addition there are differences in the temporal range of the data (where the ATTO campaign ranges from 2014 to 2023, the T3 data only covers 2014 to 2015).

4.4.2 Recursive Feature Elimination

The RFE process is highly successful in determining the optimal feature sets, reducing from 57 to 6 features at ATTO and from 40 to 5 at T3 with approximately equal R^2 and RMSE results for the best performing models using only the optimal features. The optimal features for each site found by the RFE process were the following:

1. **ATTO**: *Soil temperature (40cm), Day, Hour, Soil moisture (20cm), Soil temperature (20cm), Relative Humidity (73m).*

2. **T3**: *Zenith angle, Month, Soil Heat Capacity 1, Soil Heat Capacity 2, Soil Temperature 2, Soil Moisture 3.*

This is a notable finding as there are key variables missing from the optimal feature sets that have a known relationship with the diurnal cycle of *zi*, such as sen-

sible heat flux. However, as the transfer of heat from the surface to the atmosphere is influenced by characteristics such as soil temperature, soil moisture, vegetation cover and surface albedo, the non-linear interactions between these features might cause the elimination of variables from the optimal sets. It is noteworthy that the features obtained at the end of the RFE process do not match the most important features displayed in Figure 4.5. It may be that the RFE process removes features that are useful predictors only in combination with other features and therefore a more robust RFE procedure that looked at features in combination may identify different optimal feature sets. An example of a feature that is of high importance in the model using *All* input features is the mean water vapour concentration (*Mean_ Op-H20*) at the ATTO site (Figure 4.5) which may have ranked highly due to its known influence on atmospheric stability and thus, the rate of the boundary layer growth. However, it is removed in later stages of the RFE process, meaning it is not present in the *Optimal* feature set.

In examining the features at the ATTO site, it is notable that soil temperature at 40cm depth appears as the most important variable, ranking above soil temperature at 20cm depth which also appears in the *Optimal* feature set and in the absence of 10cm soil temperature which is closest to the surface. While intuition suggests that near-surface soil temperature would have the most direct influence on *zi*, analysis of soil temperature behavior under different rainfall conditions reveals a more nuanced relationship, as displayed in Figure 4.6. During periods without rainfall, temperature values at both 20cm and 40cm depths show minimal fluctuation. However, during rainfall events, temperatures decrease more dramatically in the near-surface layers, while the 40cm depth maintains greater thermal inertia. This deeper layer retains heat and continues to maintain positive heat fluxes, albeit at lower intensity, supporting continued PBL development even during precipitation events. The greater importance of the 40cm temperature measurements may therefore reflect the role of deeper soil layers as a more stable influence on PBL growth across varying weather conditions.

There are other observations that are more difficult to explain in terms of the biophysical significance of the results and may instead point to limitations in the RFE algorithm used. At the T3 site there are three redundant sets of soil sensors (3× each: Soil Heat Capacity, Soil Temperature, and Soil Heat Flow) installed spread over 1–2 m to account for heterogeneity in the soil properties. The actual suffix number is arbitrary beyond that each variable with the same suffix are co-located (i.e. Soil Heat Capacity 1, Soil Temperature 1, and Soil Heat Flow 1 are all near each other). While there may be some physical significance to these variables that requires further investigation, the appearance of both features could be due to the feature importance and RFE methods used. As feature importance is calculated

from the sum of the gains in performance of individual tree models that used a given feature that make up the ensemble, it may be that there were many independent trees that used one or the other of these measurements as they are interchangeable. This could be resolved by averaging across redundant sensors to see if this resulted in equally well performing models as well as investigating the significance of the differences between these measurements as features. In general it is best practice to provide the raw data to the models and discover which features are useful, rather than risk losing useful hidden information by transforming variables before testing.

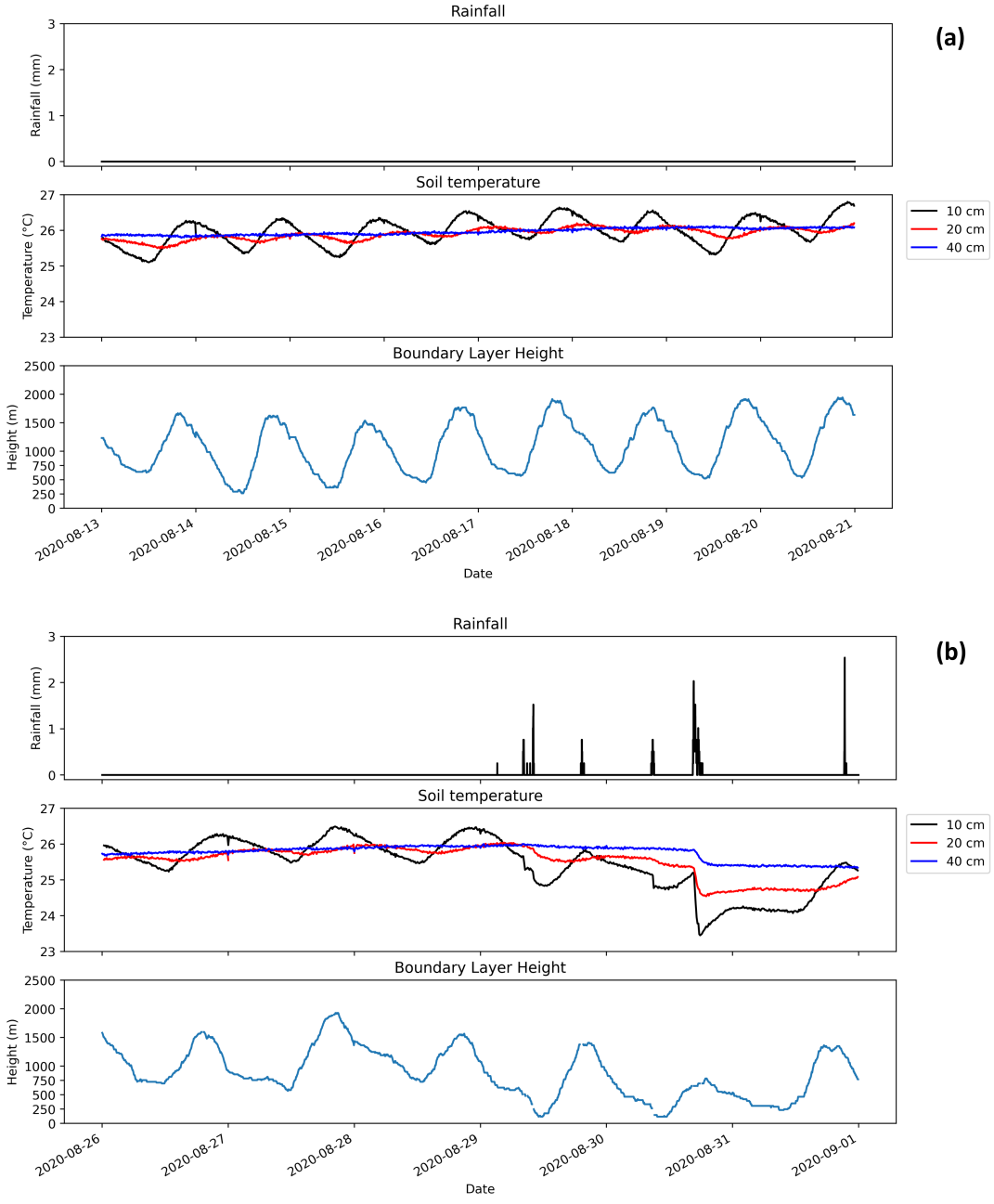


Figure 4.6: Soil temperature measurements: (a) during periods without rainfall and (b) during periods with rainfall.

4.4.3 Day vs Night Predictions

Figures 4.7 and 4.8 display the probability density estimation of predictions and measurements for day-time (Day) and night-time (Night) for the hold-out test periods at the ATTO and T3 sites respectively. This is a critical comparison when evaluating models of *zi* in the Amazon region as these periods have distinct characteristics. Negative values do not indicate physical measurements but the extension of the best fit distribution and should therefore be ignored as being non-physical. The figures shown include data only for timestamps that included radiosonde measurements (i.e. if a radiosonde measurement was missing, this timestamp and associated data were omitted from the analysis).

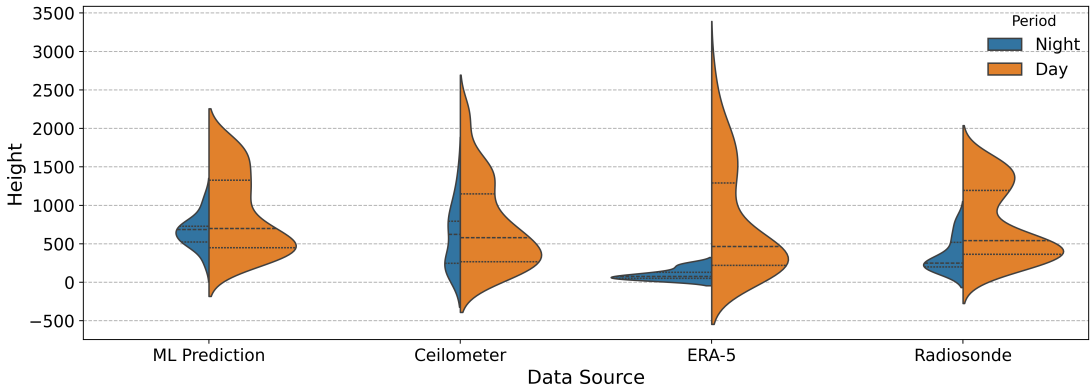


Figure 4.7: Estimated probability density distributions for predictions and measurements at the ATTO site for the out-of-sample test periods *only* where radiosonde data were available. Predictions are taken from a LightGBM model trained on all available input features. The imbalance noted between the width of the Day and Night are due to imbalanced numbers of samples.

The daytime distributions in Figures 4.7 and 4.7 exhibit bimodality in both ML predictions and ceilometer measurements, with this pattern being more pronounced at the T3 site. This bimodality likely emerges from two distinct physical regimes in the boundary layer evolution: (1) rapid growth periods during morning and evening transitions, and (2) relatively stable periods during mid-day when the boundary layer height plateaus. The sampling frequency of radiosonde measurements may also influence this distribution, as launches typically occur at fixed times that might preferentially capture certain stages of the boundary layer evolution. The absence of this bimodality in the full dataset (Figure C.11) suggests that the pattern observed in the radiosonde-matched subset may be partially attributed to sampling bias rather than purely physical phenomena. Examining the distributions of Night data at ATTO it is observed that ML models significantly overestimate *zi*. This observation is supported by positive MBE values at night in Table 4.2. The differences between Figures 4.7 and 4.8 highlight that ML models are only as good as their train-

ing data. At T3 the ML models produce a distribution that is much more similar to that of the radiosonde measurements, only because night-time ceilometer measurements are better aligned to radiosonde and do not suffer from the same difficulties in determining *zi* that is observed at ATTO. The comparison of *zi* estimates across different measurement techniques and models presents inherent challenges due to their fundamentally different detection principles. While ceilometers rely on aerosol distribution patterns to determine PBLH, radiosondes and models like ERA-5 utilize thermodynamic properties such as temperature and humidity gradients. The underlying assumption that aerosol layers correspond directly to thermodynamic boundaries isn't always valid, as aerosol distributions can be decoupled from temperature and humidity profiles due to various atmospheric processes. This methodological disparity may partially explain the significant differences observed between ceilometer, radiosonde, and ERA-5 *zi* estimates at both the ATTO and T3 sites. Previous studies have shown that such discrepancies are particularly pronounced in regions with complex atmospheric conditions ((Seidel et al., 2010; Guo et al., 2021), where elevated aerosol layers or residual layers can complicate the interpretation of ceilometer data, while thermodynamic-based methods may detect different vertical structures altogether. These fundamental differences in detection principles should be considered when interpreting cross-platform comparisons of PBLH estimates.

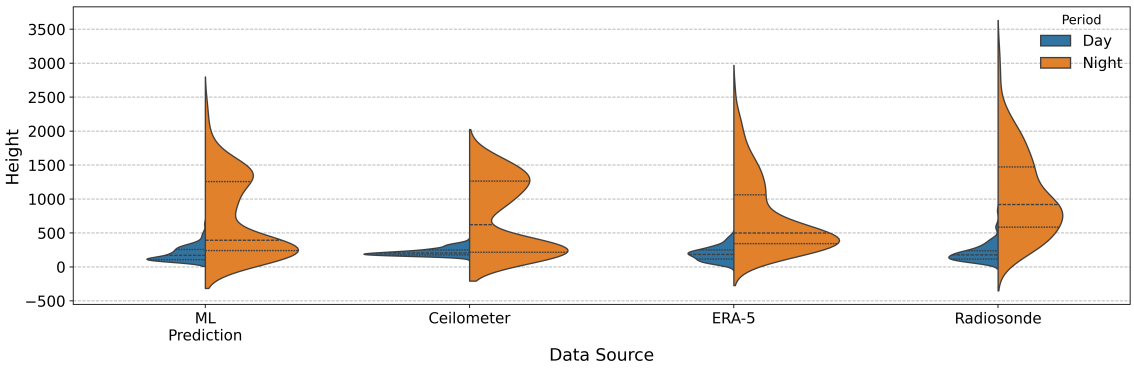


Figure 4.8: Estimated probability density distributions for predictions and measurements at the T3 site for the out-of-sample test periods *only* where radiosonde data were available. Radiosonde boundary layer heights are estimated via 4 different methods. Predictions are taken from a LightGBM model trained on all available input features. The imbalance noted between the width of the Day and Night are due to imbalanced numbers of samples.

4.4.4 Limitations and Suggestions for Future Work

While the ML predictions model *zi* well with respect to ceilometer measurements they are still limited by the underlying issues in the data upon which they are trained. Specifically, these models are trained to predict the *ceilometer* measure-

ments of *zi* which are known to have discrepancies with the true *zi* both with respect to radiosonde measurements and with respect to the nocturnal *zi* at ATTO (de Souza et al., 2023). Therefore, this paper demonstrates the applicability of machine learning models to predicting *zi* accurately in a tropical rainforest, but further methodological refinements will be needed to obtain more accurate estimates for *zi* that can be used for downstream applications such as meteorological and climate predictions. The models trained in this study are site specific and therefore can only generalise to unseen periods at that site and would not be expected to generalise to new sites. In order to obtain models that generalise well not only to unseen data but also to new sites more robust identification of optimal feature sets and understanding of the factors determining the differences in performance of models between sites will be crucial. Potential avenues for research include the application of Explainable Artificial Intelligence (XAI) and explainability techniques such as Shapely values that would aid in understanding the interdependence of the input features, as discussed in Section 4.4. Ablation studies may also aid in understanding these relationships and make the RFE process more robust. Physics Informed Machine Learning (PI-ML) techniques are also suggested for future work both for the discovery of new models of the biochemical and physical interactions that lead to the complex behaviour of the PBL as well as the incorporation of the wealth of knowledge that already exists on the PBL into data-driven models. Multi-site studies that contain the same measurements would allow for the training of a large, general model for *zi* that may allow for the identification of general factors that can predict *zi*. Due to the complexities and differences in PBL behaviour in different ecosystems it may be the case that site specific models such as those employed in this study may produce better results. In either case, a complimentary approach that utilises a combination of general models with ecosystem specific models will likely lead to the best results as an ensemble prediction. One potential application of this methodology is that of interpolation or gap-filling of *zi* measurements at these sites where there exist sufficient surface measurements but data for *zi* are missing. Another advantage of this methodology is that, due to the optimal feature sets being reduced to such a small number there are very few instruments required to take measurements needed for *zi* prediction. Further work should investigate the ability of machine learning models to extrapolate to new sites using few-shot learning (where models are trained on a limited number of *zi* samples) and transfer learning or domain adaptation (leveraging knowledge from abundant ceilometer data to improve predictions of radiosonde *zi*) in order to overcome the limitations of low data volume for radiosonde measurements, the limitations of data coverage across the Amazon region and assess the ability of ML models to assist in large scale accurate mapping of *zi*. In addition, generic models for *zi* trained on many sites such

as those introduced by Su and Zhang and Guo et al. should be trained with data from the Amazon region in order to test the performance of these large multi-site models when compared to site-specific models as demonstrated in this work. This would determine which models generalise better to unseen time periods and those models that are more robust. Should this approach deliver accurate results, it is plausible that short time duration campaigns could be executed across multiple sites in the Amazon region in order to collect sufficient training data for *zi*. From there it would be necessary to maintain a small number of inexpensive ground-based sensors, such as those measuring soil temperature and humidity, in order to accurately predict *zi* at a local level. Should further data become available and be unified into a single dataset, it may then be possible to train one large model to predict *zi* for heterogeneous sites across the Amazon region. The inclusion of satellite data for *zi*, such as those described by Teixeira et al., along with surface measurements and ground-based remote sensing techniques will be vital for developing accurate, large scale general models for *zi* across the region.

4.5 Conclusion

This study demonstrates the applicability of machine learning to model *zi* and identify optimal feature sets for model building. It is noted that the error (in terms of *RMSE*, approximately 200 metres) on predictions of *zi* is on the order of twice that of the measurement error for ceilometers in tropical regions (Carneiro and Fisch, 2020) but on the same order as the maximum observed measurement error for ceilometers as observed in arid regions (Uzan et al., 2020). Therefore, we believe that these results are positive in terms of their accuracy while room still exists for some level of improvement. It was found that gradient boosted ensemble models using all available features perform best. However, the decrease in performance when using a small subset of algorithmically identified features is on the order of 2% for the best performing models. This decrease is negligible for the advantage of having simpler, more efficient models that require fewer features to train. The RFE process demonstrated that accurate predictions of *zi* can be achieved using a minimal set of 5-7 surface measurements. This finding has significant implications for the spatial expansion of *zi* monitoring networks, suggesting that accurate estimates could be obtained through the strategic deployment of cost-effective surface sensors in conjunction with short-term measurement campaigns. Furthermore, this study identified previously unrecognized variables influential in determining *zi*, such as deep soil temperature measurements (40cm), highlighting novel mechanistic relationships in the drivers of PBL development. These findings extend current understanding of the physical drivers of planetary boundary layer height and suggest

new avenues for the investigation of land-atmosphere interactions.

Chapter 5

Gross Primary Productivity

The following chapter is in preparation for submission to the Journal of Geophysical Research: Biogeosciences
This chapter builds on the findings of the previous studies using an ensemble model for a dataset of much larger volume having validated the efficiency and accuracy of this category of model in the previous studies. This final study retains the ecosystem focus of Chapter 5 in the Amazon rainforest while expanding on the spatial extent to encompass the entire Amazon basin over a 20-year period (though at a lower temporal resolution of monthly rather than half-hourly). This chapter further develops the methodological framework by including Shapely Additive Explanations as the method of explanation. Being the most complex application this Chapter harnesses the learnings and validation carried out in the previous chapters to obtain rich scientific insight into a complex and highly important process for one of Earth's most vital ecosystems and natural carbon sinks.

5.1 Abstract

The Amazon biome represents a critical component of the global carbon cycle, contributing substantial emissions through deforestation and forest degradation while potentially serving as a significant carbon sink through old-growth forest productivity. Understanding the spatial heterogeneity of this sink function and identifying factors that determine ecosystem resilience or vulnerability to climate change is essential for predicting future carbon dynamics. We applied unsupervised k-means clustering to classify regions based on a 20-year monthly time series of Gross Primary Productivity (GPP) in primary forest obtained from satellite measurements of solar-induced chlorophyll fluorescence (SIF) observed by the Orbiting Carbon Observatory-2 (OCO-2). Subsequently, we employed machine learning models with SHapley Additive exPlanations (SHAP) analysis to identify and quantify the relative importance of atmospheric and environmental drivers influencing GPP patterns in each identified cluster. Our analysis identified 5 distinct regions varying GPP dynamics across the international Amazon. In particular mountainous regions in the South and Eastern periphery were identified with decreased mean GPP and high fluctuations between yearly maxima and minima. This region was most sensitive

to the influence of forest cover loss and water availability dynamics. Most alarmingly large areas of the Central Amazon, Eastern periphery along the Andes and in the Northeast towards the Atlantic, were identified with decreased GPP function where SHAP analysis identified the prevalence of forest cover loss and degradation. The combination of clustering analysis and explainable machine learning provides novel insights into the regional differences in photosynthesis, carbon uptake and the environmental controls on these processes. These findings have important implications for understanding large-scale responses to future climate change and informing conservation strategies for maintaining the Amazon's carbon sink function.

5.2 Introduction

The terrestrial biosphere plays a fundamental role in the global carbon cycle, storing approximately 450 GtC in living biomass and 1700 GtC in soils (Friedlingstein et al., 2023). Within this system, tropical forests, particularly the Amazon rainforest, represent critical components for carbon absorption and storage, with their capacity significantly influenced by both bioclimatic and biophysical factors (Bonan, 2015).

The Amazon rainforest exhibits complexity and heterogeneity across multiple spatial scales. At the stand level, local areas demonstrate extremely high diversity of flora, with individual hectares containing hundreds of tree species (Ter Steege et al., 2013, 2015). This diversity creates substantial variability in functional traits, canopy structure, and biogeochemical cycling even at scales of tens of meters (Asner et al., 2014). At the regional level, topology, climatic gradients, and weather systems all shape differences between regions at scales of hundreds of kilometers (Davidson et al., 2012; Marengo et al., 2018; Baker et al., 2021). The basin spans diverse precipitation regimes, from the wet western Amazon receiving over 3000 mm annually to the drier eastern regions with pronounced dry seasons (Espinoza et al., 2009). Additionally, the complex interplay between land surface processes and atmospheric dynamics, including the South American monsoon system and moisture recycling, creates emergent behaviors that are challenging to capture in traditional modeling frameworks (Marengo et al., 2010; Boulton et al., 2017; Staal et al., 2020; Argles et al., 2022; Zou et al., 2023). This multi-scale heterogeneity, combined with strong feedbacks between vegetation, hydrology, and climate, makes the Amazon an ideal testbed for ML approaches that can identify patterns and relationships across these complex spatial hierarchies.

Historically the Amazon has acted as a critical carbon sink (Brienen et al., 2015), with the net carbon emissions of nations in the basin greatly offset by the absorption of carbon from primary forest (Phillips et al., 2017). Biogenic carbon emissions in the Amazon basin have occured through two primary mechanisms. Firstly fires, drought

and their positive feedback on each other have resulted in rapid and widespread increases in tree mortality (Brando et al., 2014). Human activities have resulted in deforestation and forest degradation, with any declines in carbon emissions from decreases in deforestation in the 21st century counteracted by increases in emissions due to fires (Aragão et al., 2018). Forest degradation emissions in the Amazon have risen dramatically, sometimes exceeding deforestation emissions (Assis et al., 2020; Rosan et al., 2024; Aragão et al., 2018). Drought conditions amplify these challenges by increasing fire susceptibility (Morton et al., 2013; Andela et al., 2022) and potentially negating CO_2 fertilization benefits (Chen et al., 2024b).

Most alarmingly, top-down atmospheric measurements indicate the eastern Amazon has already transitioned from carbon sink to source due to combined effects of climate change, Land Use and Cover Change (LUCC), and fire emissions (Gatti et al., 2021). The Amazon has also been shown to be weakening in its resilience (Forzieri et al., 2022; Chen et al., 2024a) and ability to self-support via evapotranpiration driven water recycling across the basin (Salati et al., 1979; Staal et al., 2018; Baker et al., 2021), exacerbated by anthropogenic disturbances (Wang et al., 2024). Long-term projections indicate potential transitions from high-biomass canopy to low-biomass vegetation more susceptible to water deficits (Castro et al., 2022).

However, certain regions demonstrate resilient greening, such as shallow-water-table forests in Southern Amazonia (with greater availability of both water resources and access to sunlight) as well as in lower-fertility northern Amazonia, with slower-growing but hardier trees (or, alternatively, tall forests, with deep-rooted water access) (Chen et al., 2024a).

The balance between these opposing trends varies substantially across the biome, influenced by climatic gradients, soil properties, and anthropogenic pressures. Understanding this spatial heterogeneity is crucial for predicting future carbon dynamics under changing environmental conditions.

Gross Primary Production (GPP) is the rate of CO_2 uptake via photosynthesis (Chapin III et al., 2006). GPP is important for understanding carbon dynamics as it represents all carbon capture by plants and therefore the primary component of the Amazon as biological carbon sink. GPP in tropical forests responds to multiple environmental drivers whose relative importance varies across spatial and temporal scales. Previous studies have identified that, in drier regions, GPP correlates strongly with water availability, while in humid regions, energy-related variables such as photosynthetically active radiation (PAR) and temperature become primary limiting factors (Nemani et al., 2003; Wang et al., 2023). This gradient in controlling factors creates distinct functional regions within the Amazon that may respond differently to climate perturbations.

Extreme climate events reveal additional complexity in these relationships. Dur-

ing the 2015-2016 El Niño event, even humid Amazonian forests experienced GPP limitations due to soil water deficits (Longo et al., 2018; Meng et al., 2022). Conversely, La Niña events (2008-2009) reduced GPP through decreased radiation from increased cloud cover, though drought effects proved more persistent due to structural changes including increased mortality and reduced stomatal conductance (Restrepo-Coupe et al., 2024).

Knowledge Gaps and Study Objectives

Despite growing understanding of Amazon carbon dynamics, significant knowledge gaps remain regarding:

1. The spatial patterns of GPP vulnerability across the heterogeneous Amazon biome
2. The relative importance of different environmental drivers in determining regional GPP patterns
3. The identification of regions most at risk of losing carbon sink function
4. The quantitative relationships between environmental variables and GPP across different Amazon sub-regions

Traditional statistical approaches often fail to capture the complex, non-linear relationships between environmental drivers and ecosystem productivity. Machine Learning (ML) methods offer powerful alternatives for identifying patterns and relationships in high-dimensional environmental data. However, the "black box" nature of many machine learning algorithms limits their utility for scientific understanding and policy applications.

Contribution

This study addresses these challenges by combining unsupervised ML for regional classification with supervised ML with Shapely explanations to predict GPP patterns and drivers across the international Amazon rainforest. Specifically, we aim to:

1. Identify distinct regions within the Amazon based on GPP temporal dynamics using k-means clustering
2. Develop ML models to predict GPP based on atmospheric and environmental variables
3. Apply Shapely analysis to quantify the relative importance of different drivers across identified regions
4. Assess regional vulnerability to climate change based on GPP patterns and driver relationships

By integrating clustering analysis with explainable ML, this study provides novel insights into the spatial heterogeneity of Amazon forest productivity and its environmental drivers, with important implications for predicting regional responses to future climate change.

5.3 Materials and Methods

5.3.1 Data Sources

Th study encompasses the international Amazon, covering approximately 5.5 million km^2 and containing the world's largest continuous tropical forest.

GPP Data

Satellite measurements of Solar Induced Fluorscence (SIF) from the Orbiting Carbon Observatory-2 (OCO-2) have enabled new methods for the determination of global fine resolution estimates of photosynthesis (Li and Xiao, 2019). The GPP data used in this research were obtained from the OCO-2-based SIF product (GOSIF) which utilises linear mapping between the SIF and GPP observations (Li and Xiao, 2019). Monthly data were obtained for the period January 2001 through to December 2021 from the Global Ecology Data Repository (https://globalecology.unh.edu/data/GOSIF-GPP.html, accessed January 2025).

Environmental Variables

This study integrates multiple high-resolution spatiotemporal and spatial datasets to investigate which environmental and climatic variables are acting as important drivers of GPP across the region. A summary of these datasets is given in Table 5.1 and descriptions of the most commonly important variables identified by SHAP analysis are available in table 5.2. Meteorological forcing variables were obtained from the ERA5 reanalysis dataset (Hersbach et al., 2020), which provides comprehensive monthly atmospheric data at 0.1° spatial resolution covering 1950-2022, including air temperature, specific humidity, wind components, solar radiation, evapotranspiration (termed total evaporation in ERA-5), and soil moisture and temperature profiles. High-quality precipitation data were derived from the MERGE product (Rozante et al., 2010, 2020), a sophisticated dataset that combines Tropical Rainfall Measuring Mission (TRMM) and Global Precipitation Measurement (GPM) satellite estimates with in-situ surface observations to generate optimized gridded precipitation fields specifically calibrated for South American conditions at approximately 10 km resolution.

Land surface characteristics were captured through annual land use and land cover classifications from the MapBiomas initiative (Souza et al., 2020), providing detailed 30-meter resolution maps spanning 1985 to present that enable tracking of land cover transitions across the Amazon basin. Forest degradation dynamics were characterized using multiple disturbance variables including drought stress, selective logging intensity, forest edge effects, and fire occurrence patterns, derived from the comprehensive analysis by (Lapola et al., 2023) at 0.5° resolution covering the period 2001–2018. Biodiversity metrics encompassing tree density, alpha diversity indices, and species richness were obtained from the Amazon Tree Diversity Network (ter Steege et al., 2023), providing 0.1° resolution estimates based on standardized analysis of 2,046 forest inventory plots distributed across the basin.

Soil biogeochemical properties were represented through multiple phosphorus pools (plant-available, organic, and total phosphorus concentrations) derived from machine learning-based reference maps (Darela-Filho et al., 2024) at 5 arcminute resolution (approximately 0.083°), developed using Random Forest models trained on extensive soil sampling data from 108 sites within the RAINFOR network. Community-weighted wood density data were obtained from Mo et al. (2023). Additional environmental context was provided by topography Brazil's PRODES monitoring system (PRODES-INPE, 2022), and deforestation data from Hansen et al. (2013) tracking forest loss patterns from 2000-2021 at 0.008° resolution. All datasets underwent careful spatial and temporal harmonization to a common 0.1° grid to ensure compatibility for integrated machine learning analysis while preserving the original data quality and temporal resolution characteristics of each source. Spatial regridding and interpolation were performed using the xESMF (xarray Earth System Model Exchange) library and xarray's built-in interpolation functions to ensure consistent spatial resolution across all datasets (Hoyer and Hamman, 2017; Zhuang et al., 2020).

5.3.2 Data Preparation

To ensure data quality and statistical validity, correlation analysis and Variance Inflation Factor (VIF) diagnostics were employed to identify and remove redundant or multicolinear predictor variables. A mask was applied to all data in order to isolate those areas that correspond to primary forest. This mask is created by taking (ter Steege et al., 2023) only class 3 from 2001 in the MapBiomas mapping of land cover classes (Souza et al., 2020; MapBiomas Amazonía, 2024; RAISG, 2024). This first year in the dataset is used to separate out the forest as we hypothesise that the changes in GPP caused by LUCC will be detected by the clustering algorithm which can be later verified via the explicit encoding of deforestation as an input

Table 5.1: Summary of spatiotemporal datasets used for GPP prediction in the Brazilian Amazon

Dataset	**Variables**	**Resolution**	**Temporal Coverage**	**Key Reference**
ERA5	Temperature, humidity, wind, radiation, evaporation, soil moisture/temperature, pressure	0.1°	1950–2022 (monthly)	(Hersbach et al., 2020)
MERGE	Precipitation (total, max, min)	∼10 km	2001–2018 (daily aggregated)	(Rozante et al., 2010)
MapBiomas	Land use/land cover classes	∼30 m (0.008°)	1985–present (annual)	(Souza et al., 2020)
Steege et al.	Tree density, diversity, richness	0.1°	Static (2000 - ?)	(ter Steege et al., 2023)
Lapola et al.	Drought, logging, edge effects, fire	0.5°	Static (2001–2018)	(Lapola et al., 2023)
Darela-Filho et al.	Soil phosphorus forms (available, organic, total)	5 arcmin (∼0.083°)	Static	(Darela-Filho et al., 2024)
Mo et al.	Community wood density	∼0.083°	Static	(Mo et al., 2023)
Hansen Global Forest Cover	Deforestation	30m (0.008°)	2001–2023 (annual)	(Hansen et al., 2013)

Table 5.2: Description of key predictive features.

Label	**Source**	**Description**	**Units**
Lat/Lon	ERA5	Geographical coordinates	deg
Year/Month	ERA5	Temporal dimensions	-
Air Temp	ERA5	2m air temperature	°C
Dewpoint	ERA5	2m dewpoint temperature	°C
Surface Pressure	ERA5	Surface atmospheric pressure	Pa
Wind	ERA5	10m wind components (U/V)	m/s
Evaporation	ERA5	Total evaporation (soil + open water + canopy + vegetation components)	m
Pot Evaporation	ERA5	Potential evaporation	m
Precipitation	MERGE	Total, max, and min precipitation	mm
Solar Radiation	ERA5	Net and downward solar radiation	J/m^2
Thermal Radiation	ERA5	Net and downward thermal radiation	J/m^2
Heat Fluxes	ERA5	Latent and sensible heat fluxes	J/m^2
Albedo	ERA5	Forecast albedo	-
Soil Temperature	ERA5	Soil temperature layers 1 (0-7cm) and 4 (100-289cm)	°C
Soil Moisture	ERA5	Soil water layers 1 (0-7cm) and 3 (28-100cm)	m^3/m^3
Soil Type	ERA5	Dominant soil classification	-
Phosphorus	RF Model	Available, organic, and total phosphorus	mg/kg
Runoff	ERA5	Total, surface, and subsurface runoff	m
LAI	ERA5	Leaf Area Index for high and low vegetation	m^2/m^2
Veg Cover	ERA5	High and low vegetation cover fractions	-
Veg Type	ERA5	High and low vegetation type classifications	-
Wood Density	Mo et al.	Community wood density	g/cm^3
Tree Metrics	Steege	Density (stems/ha), diversity index, richness (species/ha)	various
Disturbance	Lapola	Edge effect, fire, drought, logging indices	various
Topography	PRODES	Terrain characteristics	m

feature.

5.3.3 K-means Clustering Analysis

To identify distinct regional patterns in ecosystem productivity, we applied an unsupervised clustering approach to full time series of monthly Gross Primary Productivity (GPP) data at the pixel level. K-means clustering was selected for its computational efficiency, interpretability, and proven effectiveness in identifying spatial patterns in environmental data (**?**). While k-means assumes spherical clusters and requires pre-specification of cluster number, these limitations are manageable in our application. The Euclidean distance metric is appropriate for time-series similarity in this context as methods like dynamic time warping may obfuscate factors like the inherent shift in seasonality across the Amazon basin. The combination of the elbow and silhouette method provides objective cluster number selection.

The classification procedure consisted of the following steps:

First, all GPP time series were standardized using scikit-learn's StandardScaler to prevent larger values from skewing the clustering algorithm. This transformation centered each feature around zero with unit variance.

The optimal number of clusters ($\underline{k}$) was determined through a combination of quantitative metrics. Firstly the elbow method, which uses Inertia (Eq. 5.1, within-cluster sum-of-squares) to determine how similar members of each cluster are to each other. Secondly silhouette analysis evaluates how well each data point fits within its assigned cluster over neighboring clusters (measures between-cluster goodness of fit). Notably the silhouette score will always increase with increasing $\underline{k}$ whereas the silhouette score has local and global optima.

Silhouette and Inertia scores are calculated as follows:

Inertia (Within-Cluster Sum of Squares):

$$\text{Inertia} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{5.1}$$

where k is the number of clusters, C_i represents the i-th cluster, x is a data point in cluster C_i, μ_i is the centroid of cluster C_i, and $\|\cdot\|$ denotes the Euclidean norm.

Silhouette Score:

For an individual data point x_j:

$$s(x_j) = \frac{b(x_j) - a(x_j)}{\max(a(x_j), b(x_j))} \tag{5.2}$$

where:

$$a(x_j) = \frac{1}{|C_i| - 1} \sum_{x_k \in C_i, x_k \neq x_j} d(x_j, x_k) \tag{5.3}$$

$$b(x_j) = \min_{l \neq i} \left\{ \frac{1}{|C_l|} \sum_{x_k \in C_l} d(x_j, x_k) \right\} \tag{5.4}$$

The overall silhouette score is:

$$\text{Silhouette Score} = \frac{1}{n} \sum_{j=1}^{n} s(x_j) \tag{5.5}$$

where $a(x_j)$ is the mean intra-cluster distance for point x_j, $b(x_j)$ is the mean nearest-cluster distance for point x_j, $d(x_j, x_k)$ is the distance between points x_j and x_k, $|C_i|$ is the number of points in cluster C_i, and n is the total number of data points.

The chosen $\underline{k}$ was obtained by balancing the two metrics with the objective of producing a number of clusters that makes meaningful divisions between regions of the rainforest while still remaining parsimonious and interpretable.

Finally we applied the mini-batch $\underline{k}$-means algorithm with the number of clusters $\underline{k}$ determined from the previous step. Mini-batch $\underline{k}$-means was selected for its computational efficiency with large spatial datasets, as it processes random subsets of the data iteratively rather than the entire dataset simultaneously, while maintaining clustering quality comparable to standard k-means.

5.3.4 Machine Learning Model Development

For each cluster an individual XGBoost (Chen and Guestrin, 2016) model were trained as a single-output multivariate regressor with 58 environmental variables as input feature and GPP values (per pixel, per month) as target. XGBoost was selected for its speed and notable high performance across a large range of ML prediction tasks (Nielsen, 2016). All code has been implemented in python.

Models were trained using an 80:20 train-test split where 80% of the data was used to train the model and 20% of the data was used as hold-out validation data to confirm the ability of the models to generalise to unseen data. The data are randomly sampled per pixel-month. This is the simplest sampling that could be selected, not requiring that the models extrapolate to completely unseen areas or times. The random sampling ensures that there is a high likelihood that all pixels will be sampled for at least one month out of the 240 that make up the full duration of the dataset.

In order to evaluate and compare model performance, the following metrics are used: coefficient of determination (R^2); Root Mean Squared Error ($RMSE$); Mean

Average Error (MAE) and Mean Bias Error (MBE). The notation used in the equations is as follows: $\hat{y}_i$ denotes the predicted value of the i-th sample; $\hat{Y}$ is the vector of all $\hat{y}_i$; y_t is the corresponding true value for N total samples in the test set; Y is the vector of all y_t;

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_t - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_t - \bar{y}^2)} \tag{5.6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_t - \hat{y}_i)}{N}} \tag{5.7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_t - \hat{y}_i| \tag{5.8}$$

$$MBE = \frac{1}{N}\sum_{i=1}^{N}(y_t - \hat{y}i) \tag{5.9}$$

5.3.5 Model Explanations

In order to understand the underlying function of the models and how features are contributing to GPP predictions, we apply Shapely analyis utiling the SHAP (SHapley Additive exPlanations) package (Lundberg and Lee, 2017a) . Shapely analysis leverages theories of coalitions in game theory, treating each feature as a member of a coalition or team. The analysis considers all sets of combinations of features, thus quantifying the global contribution of each feature accounting for the context of their combination with other features. The exact SHAP values are defined by the following formula

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{5.10}$$

where ϕ_i is the SHAP value for feature i, F is the set of all features, and f_S is the model trained with feature subset S. In practice the tree-based structure of the XGBoost models are leveraged to approximate the SHAP values and avoid the computational challenges of calculating the exact SHAP values, as described in (Lundberg and Lee, 2017a).

SHAP analysis was performed separately for each identified cluster to understand regional differences in GPP drivers.

SHAP values have the following desirable properties:

- SHAP values are additive and each feature has a contribution (with both magnitude and direction) for each sample in the dataset and that all contributions

sum together to give the final prediction ("local accuracy").

- SHAP values are calculated on a sample-by-sample basis, meaning that explanations can be obtained for each individual prediction, from which global measures of feature importance and distributions of influence are obtained.

5.4 Results & Discussion

5.4.1 Clustering

Utilising the silhouette and elbow methods two candidates for optimal number of clusters are apparent in Figure 5.1. The Elbow method is less clear in this analysis as there is no value of k where the drop-off in Inertia very clearly slows down as the number of clusters increases. A candidate is k=5 as the decreases in inertia thereafter are at least half of the decreases preceding. The Silhouette method however provides two clear candidates at k=3 where the Silhouette score is mazimised and at k=5 where the Silhouette score is higher than all other candidates (other than k=2 or k=3). Domain knowledge and the objective of the study (i.e. to uncover regional patterns in GPP and understand their drivers) are also taken to account when selecting the optimal number of clusters. While 3 clusters may be adequate in terms of clustering metrics, the patterns of GPP may not be sufficiently different to tease out regions under stress from large regions whose primary differences can be attributed to location, phenological timing and other well understood drivers.

Using Figures 5.1a and 5.1b, k=5 was selected as the optimal number of clusters for the task due the minimisation of Inertia with a preferentially high Silhouette score.

5.4.2 Spatial and Temporal Patterns within Clusters

Having identified 5 distinct regions from the k-means clustering, we assign descriptive names to each based on the spatial distributions visible in Figure 5.2, ordered in terms of ascending median GPP (Figure 5.3). The labels for the clusters are as follows: Peripheral (Cluster 1), Central Amazon A (Cluster 2), Central Amazon B (Cluster 3), Southern Belt (Cluster 4), Northern Belt (Cluster 5). Figure 5.3 displays the box plots of GPP for each cluster and Figure 5.4 displays the mean monthly GPP across the entire study period. Examining Figure 5.3, the Peripheral region stands out as having the lowest median, maximum, minimum and the highest spread. Notably in Figure 5.4 this can be seen to manifest in greater differences between approximately half-yearly peaks and troughs whose occurrence are delayed relative to same peaks troughs in other regions. Also notable is the lack of intra-seasonal

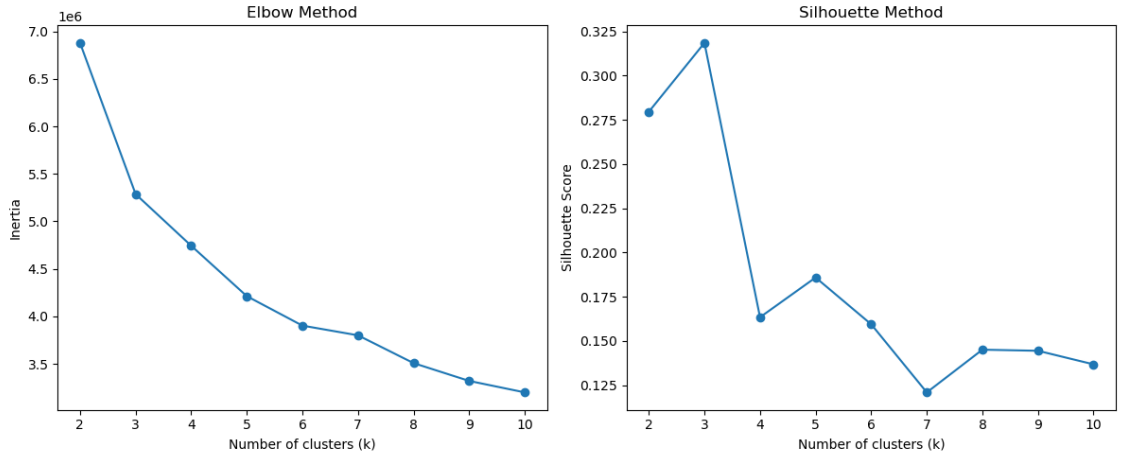


Figure 5.1: Comparison of results from (a) Elbow method (i.e. inertia analysis) and (b) Silhouette method used to determine the optimal number of clusters. The Inertia score measures distance between members of clusters and therefore is to be minimised up to the point of diminishing returns. The Silhouette score measures distance between cluster members and other cluster centres and therefore should be maximised to ensure samples are in their optimal cluster. Two clear candidates emerge at k=3 and k=5 as global and local maxima of the Silhouette score, with the lower Inertia score at k=5 indicating optimality.

Table 5.3: Performance Metrics by Cluster

Cluster	**R^2**	**RMSE**	**MSE**	**MAE**	**MAPE (%)**	**Bias**
Periphery	0.894	0.759	0.577	0.558	9.23	0.006
Central A	0.753	0.387	0.150	0.277	3.26	0.001
Central B	0.782	0.386	0.149	0.281	3.04	0.000
Northern Belt	0.837	0.528	0.279	0.388	3.86	-0.001
Southern Belt	0.792	0.437	0.191	0.331	3.29	-0.001

variation between these extremes, with GPP values oscillating nearly linearly from peak to trough. Both Central Amazonian regions as well as the Northern Belt have similar interquartile ranges with seasonal and temporal peaks and troughs aligned temporally. The Southern belt has interannual minima that co-occur with the minima of the three aforementioned regions with the notable difference being that the peaks in GPP are delayed and last longer (Figure 5.3. This also manifests in the Southern Belt having a greater interquartile range, though less than that of the Peripheral region.

5.4.3 Machine Learning Predictions & Drivers

Table 5.3 presents the test set metrics for each individual model trained for each cluster for the task of predicting GPP. The test set is comprised of randomly selected location-month pairs.

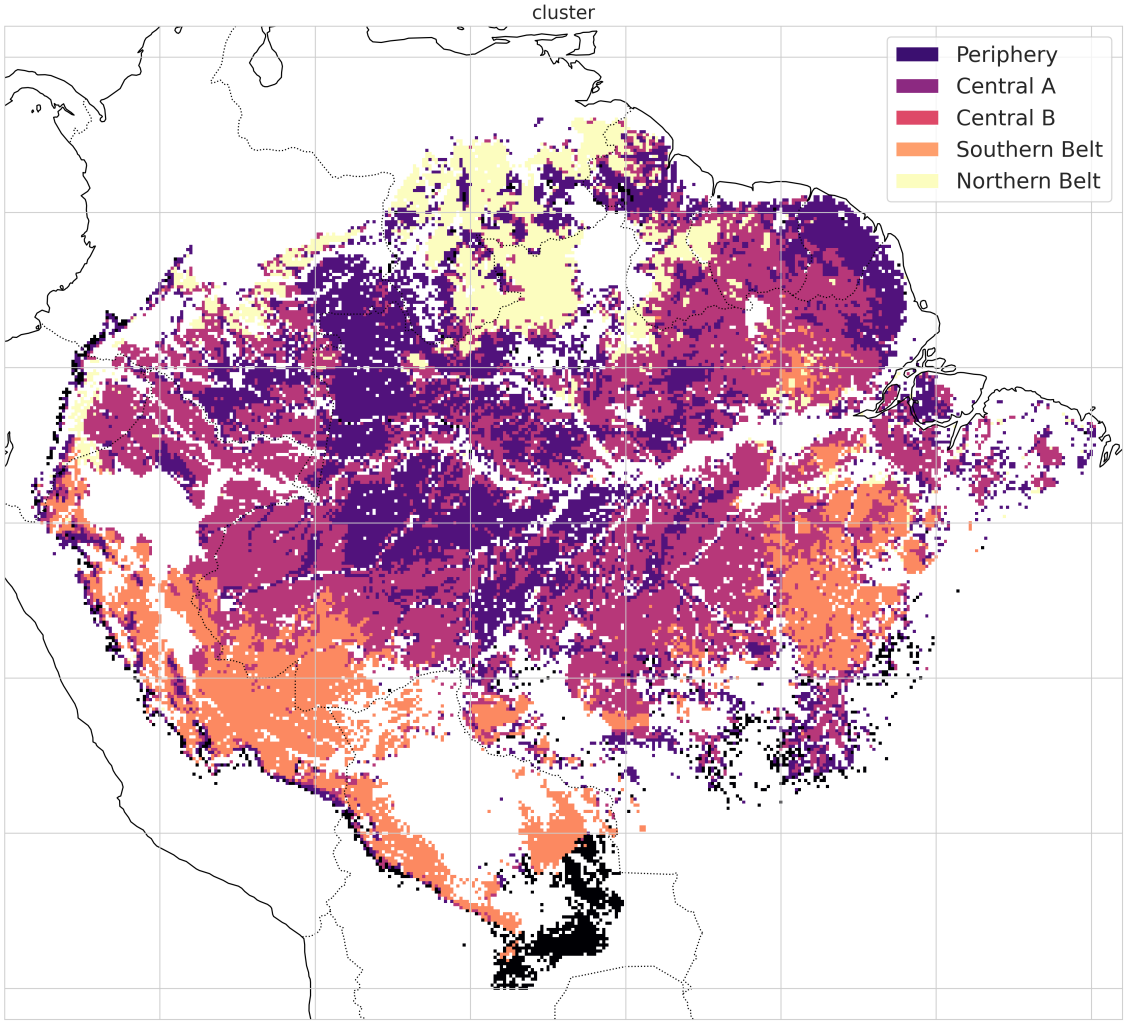


Figure 5.2: Spatial distribution of regional GPP clusters identified through k-means clustering with k = 5 clusters. Each color represents a distinct cluster characterized by similar GPP patterns and variability. Cluster boundaries reflect underlying ecological and climatic gradients that influence patterns of GPP across the study region.

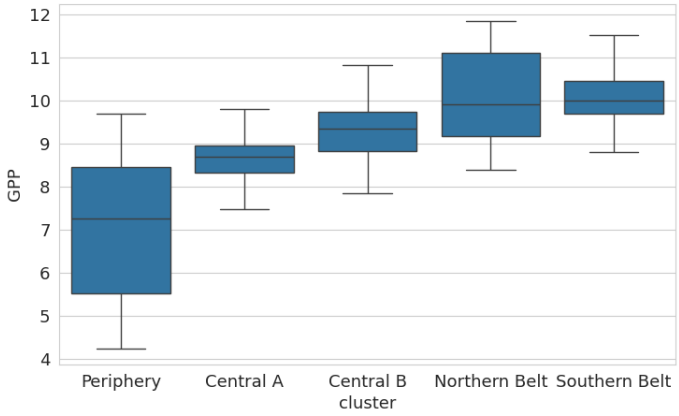


Figure 5.3: Distribution of GPP values per cluster comparing (a) k = 3 and (b) k = 5 clustering results. The box plots illustrate the variability and central tendencies within each cluster, demonstrating how increased cluster numbers capture more nuanced patterns in GPP distributions.

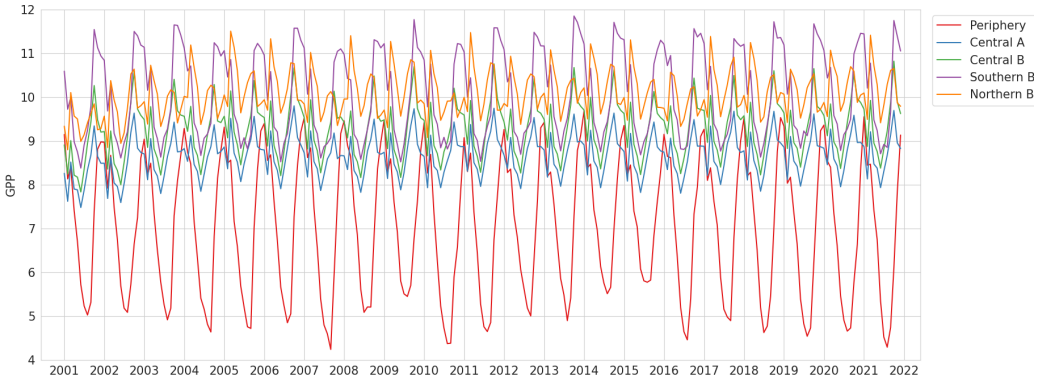


Figure 5.4: Mean monthly GPP per cluster. The temporal patterns show distinct seasonal trajectories for each cluster, with the k = 5 analysis providing finer resolution of GPP dynamics across different ecological zones.

In each cluster a R^2 score greater than 0.75 is obtained, indicating that the models are well-performing. The model with the highest R^2 score is that for the Periphery, notably also the model with the highest RMSE. This region can be seen in Figures 5.3 and 5.4 to have the highest variability and range. Models for the Periphery and Central regions have a small positive bias whereas models for the Northern and Southern Belts have a small negative bias.

5.4.4 SHAP Explanations

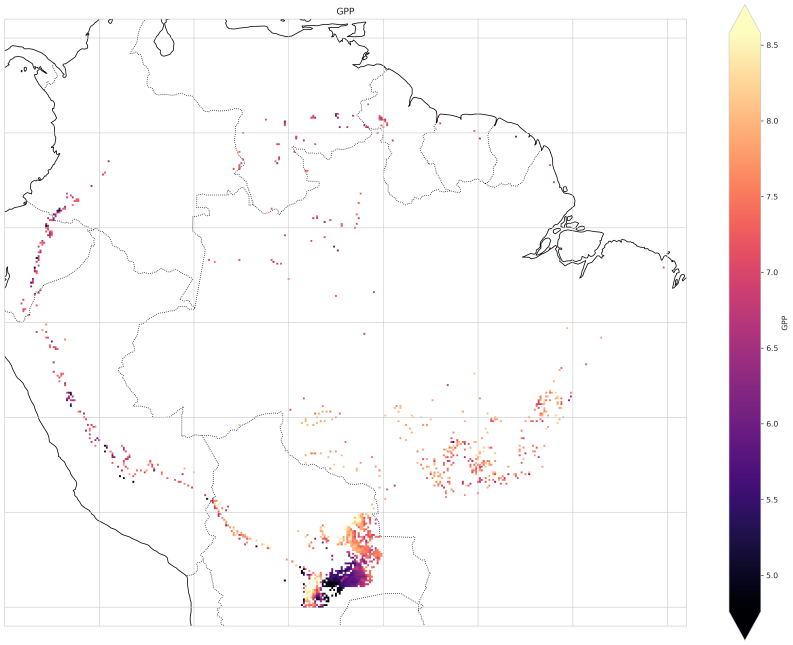
Figures 5.5b, 5.6b, 5.7b, 5.8b and 5.9b display the mean of SHAP values for each feature across all samples in the test set, termed Feature Importance as this gives a quantitative measure of the average impact of each feature on the predicted value of GPP. Figures 5.5c, 5.6c, 5.7c, 5.8c and 5.9c display the distribution of SHAP values for each sample in the test set, giving a better indication of how the influence of a feature varies across the entire region in a cluster and in conjunction with other features (remembering the SHAP values take into account all possible sets of feature combinations and thus the effect of feature interactions). In interpreting the violin plots the term positive influence will be used to describe where higher values of a feature lead to increases in the prediction of GPP and lower values lead to decreases in GPP. Vice versa, negative influence will be used to describe an inverse relationship, where lower values of a feature lead to increases in GPP predictions, and higher values lead to decreases in GPP predictions. This interpretation follows the mathematical foundation of SHAP described in Section 5.3.5 that each feature contributes a value, positive or negative, that sum together to form the final prediction for any sample of features and the target, GPP.

In order to better interpret the results of the SHAP explanations, we divide the top 20 most important features into three groups: High Importance (top 5), Medium Importance (top 6-10) and Lower Importance (top 11-20). Examining the distribution of feature importance across the five clusters reveals clear common patterns. In every cluster, Month, Albedo and either Latitude or Longitude appear in the top 5 most important features, suggesting that temporal and basic spatial patterns are the largest contributors to the underlying distribution of GPP. Forecast albedo in ERA-5 is produced by the ECMWF's land surface model (H-TESSEL), which uses a "tiling" approach where each grid cell represents different surface types. Since the data has been upscaled, this variable may be capturing phenological timings or plant coverage rather than meaningful vegetation health information, especially in conjunction with Month as an important predictor. The universally positive relationship between Albedo and GPP across all clusters indicates that phenological timing is indeed being captured, supported by the non-linear Month-GPP relation-

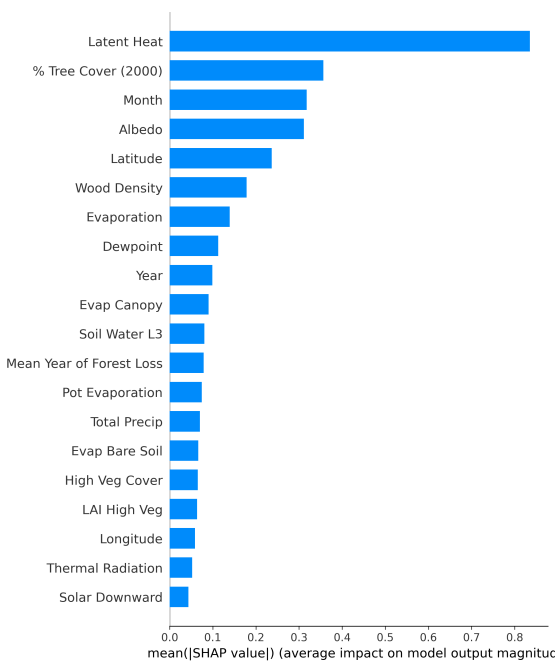
ships observed in SHAP distributions. Across all clusters common themes emerge around incoming solar radiation, water availability and forest cover characteristics. It is in the differences between the specific features that appear as important and the distribution of their influence on increasing or decreasing predictions of GPP that allows us to tease out some indicators as to the complex underlying system of causation in each region.

Cluster 1: Periphery

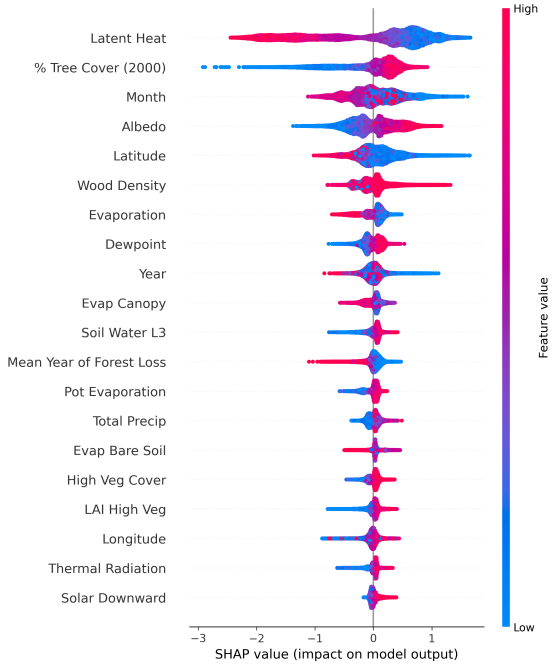
The Peripheral region (Cluster 1) appears to represent a water and heat-stressed, degraded or deforested environment, as evidenced by latent heat and percentage tree cover emerging as the two most important predictor variables in this region—features that are notably less important in other regions. The region is characterised by having higher elevation, and dwepoint than other clusters, higher instances of fire and lower precipitation, lower Forest Cover (in the year 2000), higher low vegetation cover and lower high vegetation cover. The inverse relationship between latent heat and GPP suggests that plants may be limiting photosynthesis under thermal stress conditions (Novick et al., 2016). This interpretation is further supported by the positive influence of dewpoint temperature on GPP predictions, indicating that increased atmospheric moisture availability enhances photosynthesis, thereby highlighting the atmospheric water availability sensitivity of vegetation in this region rather than soil water limitations as highlighted by recent studies (Novick et al., 2016). The significance of canopy evaporation as an important feature underscores the predominant influence of atmospheric conditions over soil water conditions on plant responses in this region. At deeper soil levels (L3), soil water content exhibits a positive influence on GPP predictions, suggesting that deeper root systems may be actively extracting water under stress conditions, with this deeper soil water limitation potentially connected to overall water availability constraints (Chitra-Tarak et al., 2021; Kühnhammer et al., 2023). The importance of total precipitation as a predictor variable, coupled with the absence of subsurface runoff and topography (which appear as important in all other regions), indicates that water shortage may be the underlying mechanism for decreased GPP in this region, as we can rule out soil water saturation or excess runoff. The emergence of potential evaporation as a significant feature further suggests heightened water loss intensity, collectively indicating stressed soil conditions and root systems. The presence of High Vegetation Cover as an important feature with a positive influence on GPP further supports this as taller stands have deeper roots that are able to access water resource (Chitra-Tarak et al., 2021; Kühnhammer et al., 2023). Features related to incoming solar energy available for photosynthesis (Thermal Radiation and Solar Downward - see table D.1 appear lower in importance.



(a) Mean GPP distribution



(b) Feature importance

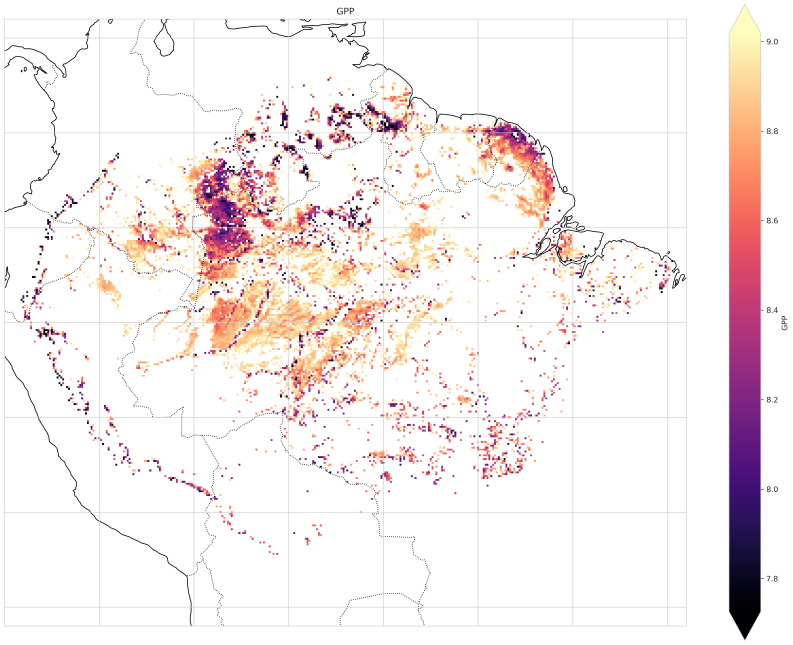


(c) SHAP value distribution

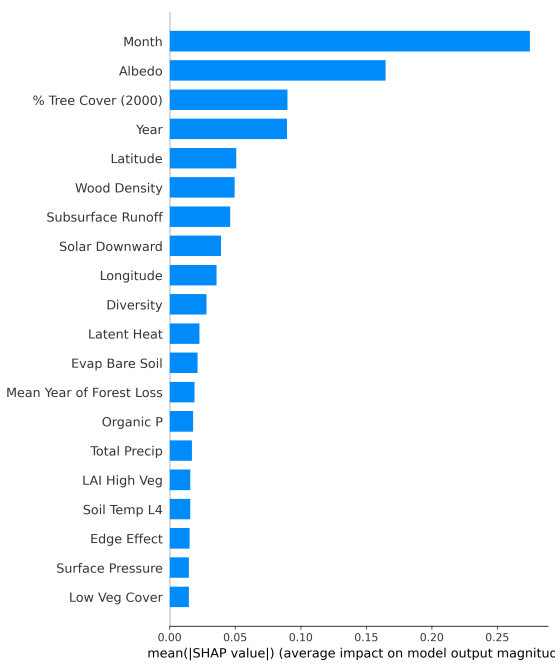
Figure 5.5: Cluster 1 analysis showing (a) the mean GPP distribution across the region, (b) feature importance measured by mean SHAP values indicating average impact on GPP prediction, and (c) distribution of positive and negative feature impacts across all test samples.

Clusters 2 & 3: Central Amazon

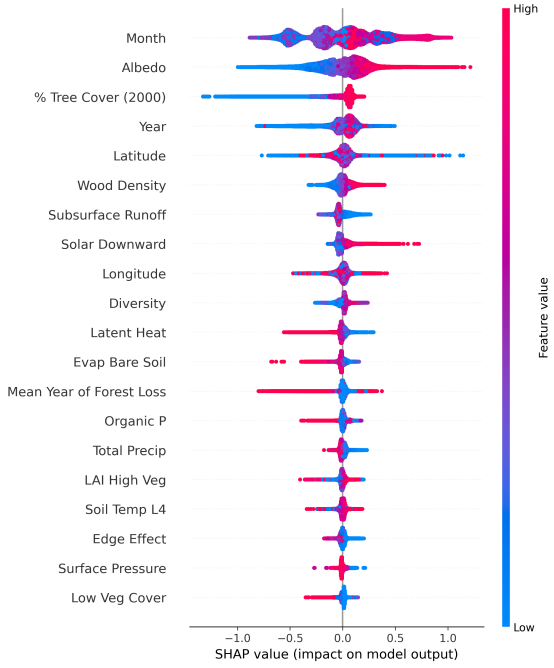
Clusters 2 and 3 warrant comparative examination as they represent contrasting forest conditions within the central Amazon basin as well as some fringe areas towards the foothills of the Andes. Cluster 2 (Central A), represents the degraded or deforested areas of the central Amazon characterized by percentage tree cover as a highly important predictor variable, while this feature appears with reduced importance in Cluster 3 (Central B), despite their spatial proximity and interwoven nature. This distinction is reinforced by wood density emerging as a medium-importance variable in Cluster 2 but being absent from the important features in Cluster 3. The positive relationship between wood density and GPP in degraded areas may indicate either that less deforested regions maintain superior carbon capture capacity, or that areas with thicker, better-adapted trees exhibit enhanced carbon uptake efficiency. This interpretation is supported by diversity appearing as a medium-importance feature, suggesting that diverse forest stands provide better resilience under external pressures such as climate change and anthropogenic degradation and deforestation. The presence of edge effects as a lower-importance feature further indicates that microclimate changes and species compositional shifts associated with forest fragmentation are present, with higher values of the distance from the forest edge being associated with positive SHAP values. The diversity component suggests that balanced species assemblages with different physiological adaptations help maintain GPP under deforestation, degradation, or environmental stress. In contrast, Cluster 3 represents healthy, intact forest, where subsurface runoff emerges as a medium-importance variable alongside factors determining nutrient availability, such as total phosphorus. The SHAP distribution analysis (Figure 5.7c) reveals a bimodal response to subsurface runoff, where lower values can produce either positive or negative effects on GPP, indicating spatial heterogeneity in soil drainage conditions. This suggests that some areas experience inadequate drainage leading to root oxygen limitation, while others suffer from excessive drainage resulting in insufficient root water availability—reflecting the critical balance between drainage and water storage necessary for optimal moisture and oxygen conditions (Silver, 2000; Aragão et al., 2007; Davidson et al., 2012). Phosphorus dynamics differ markedly between clusters, with Cluster 2 showing only organic phosphorus (out of total, organic and available phosphorous as the features given to the model) as an important feature. This suggests soil conditions that limit phosphorus cycling processes such as organic matter mineralization and microbial enzyme activity (Mabagala et al., 2022; Jindo et al., 2023). This limitation is corroborated by bare soil evaporation appearing as an important feature, where lower evaporation rates correlate with increased GPP predictions, indicating that the soil may be degraded or exposed in deforested areas. Conversely, Cluster



(a) Mean GPP distribution

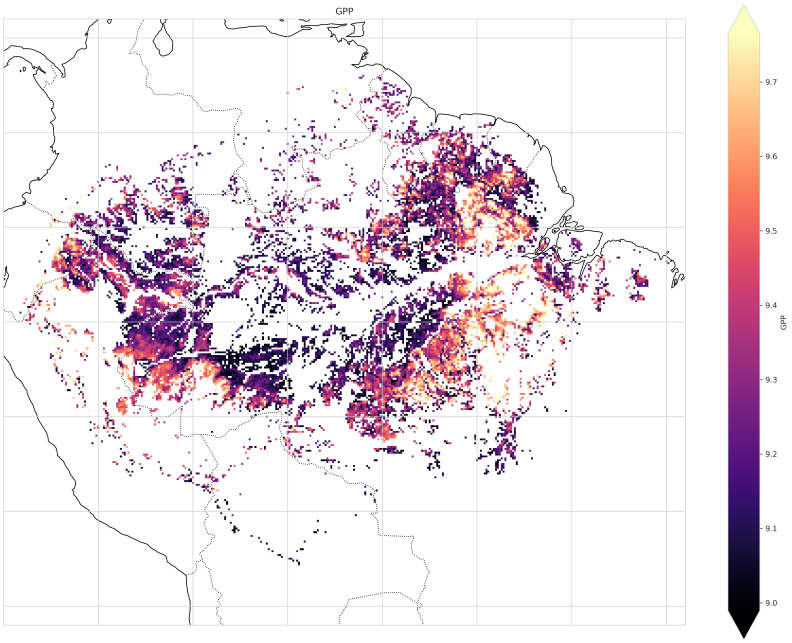


(b) Feature importance

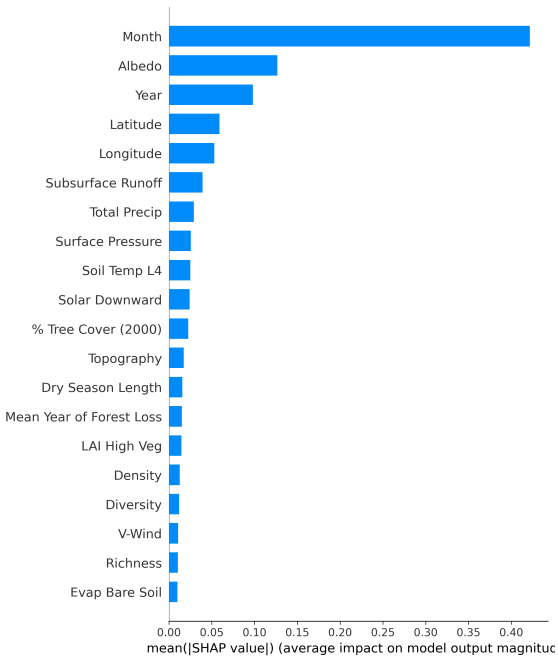


(c) SHAP value distribution

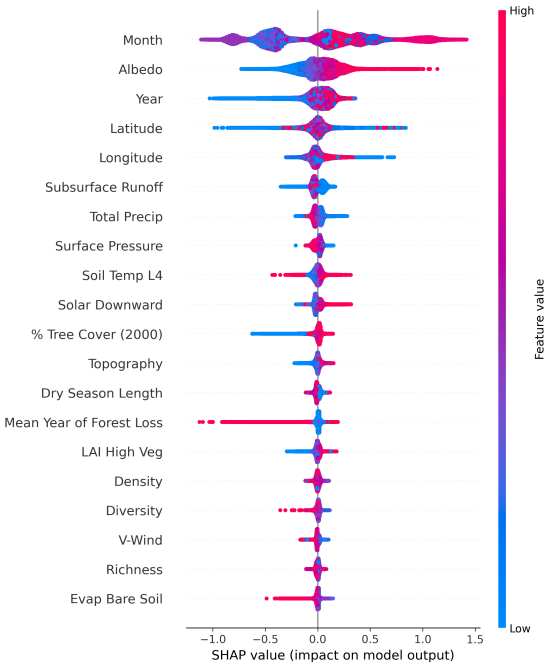
Figure 5.6: Cluster 2 analysis showing (a) the mean GPP distribution across the region, (b) feature importance measured by mean SHAP values indicating average impact on GPP prediction, and (c) distribution of positive and negative feature impacts across all test samples.



(a) Mean GPP distribution



(b) Feature importance



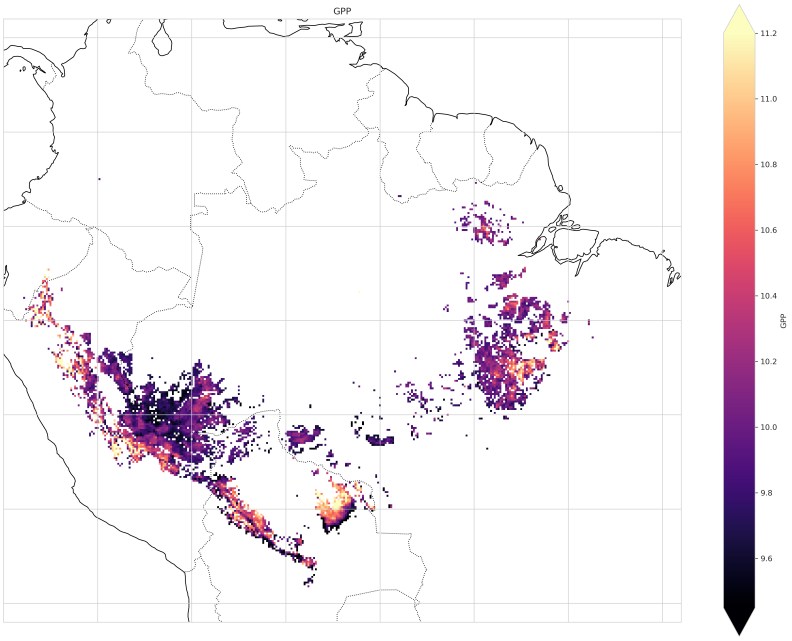
(c) SHAP value distribution

Figure 5.7: Cluster 3 analysis showing (a) the mean GPP distribution across the region, (b) feature importance measured by mean SHAP values indicating average impact on GPP prediction, and (c) distribution of positive and negative feature impacts across all test samples. Evaporation and latent heat emerge as primary drivers alongside temporal variables, indicating energy balance processes are critical for GPP prediction in this cluster.

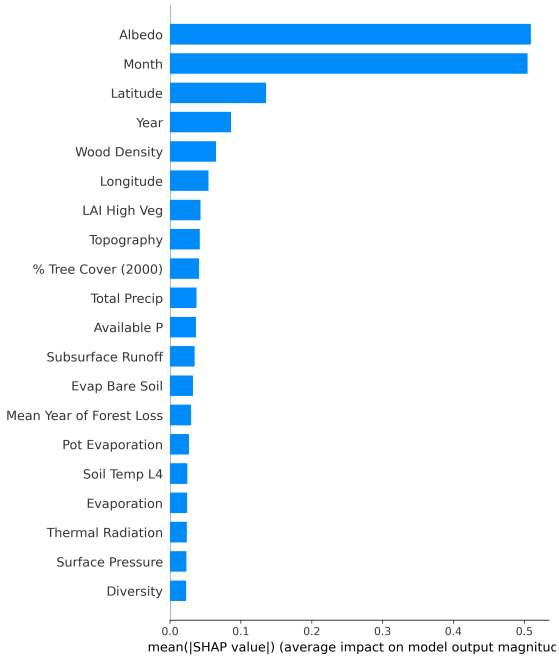
3 exhibits soil temperature at level 4 as a feature of medium importance, suggesting root zone thermal stability considerations. The bimodal SHAP response to soil temperature—with higher values producing both positive and negative effects while lower values yield neutral responses—may indicate connections to soil respiration rates and net carbon balance dynamics. Surface pressure stands out as a variable whose influence warrants further investigation, appearing as a medium-importance variable in Cluster 3 and a lower importance variable in Cluster 2. Examining the SHAP distributions in Figures 5.6c and 5.6c it is noted that surface pressure has a negative influence on GPP predictions, with lower surface pressure contributing to increased GPP. The influence of Topography in Cluster 3 may lend some clues to the underlying processes, with lower elevations producing negative SHAP values and higher elevations yielding positive values. This finding in combination with the distribution of surface pressure SHAP values suggests the presence of distinct thermal zones including cool highlands, warm lowlands, and intermediate thermal stress zones where species may be maladapted to local temperature conditions.

Cluster 4: Southern Belt

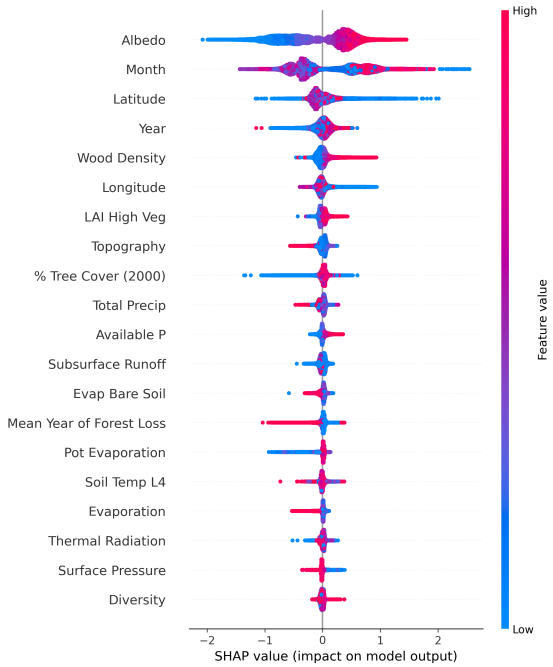
Cluster 4, representing the southern belt region, is distinguished by the appearance of wood density as a highly important predictive feature, where higher wood density values correspond to increased SHAP values. In conjunction, percentage tree cover appears only as a medium-importance feature with a complex bimodal distribution pattern: lower tree cover percentages yield both higher and lower SHAP values, while higher tree cover percentages produce median SHAP responses. These patterns suggests that there may be multiple processes influencing GPP in this region. This region represents neither degraded forest nor optimal forest conditions, but rather the transition zone between the Amazon rainforest and the Cerrado (Brazilian savanna) influenced by growth rates and mortality patterns that vary with stand density and structure and heavily influenced by dry seasons with lower precipitation (Ackerly et al., 1989). The region has the highest dry season length, though this does not appear as an important predictive feature, most likely as it is not a differentiating factor as there is a low spread of dry season length within this cluster. The region has the highest median available and organic phosphorous content and the lowest total precipitation after Cluster 1. Thermal radiation is lower in this region though other forms of radiation are comparable to those of other clusters. Longitude emerges as a medium-importance variable, with lower longitudinal values (stands closer to the Atlantic coast) associated with reduced SHAP values, with region having lower diversity than the Central Amazon regions but higher than that of the Peripheral and Northern Belt regions. Notably it is the LAI of high vegetation that appears as an important feature rather than the total cover of high vegetation as in other clusters.



(a) Mean GPP distribution



(b) Feature importance



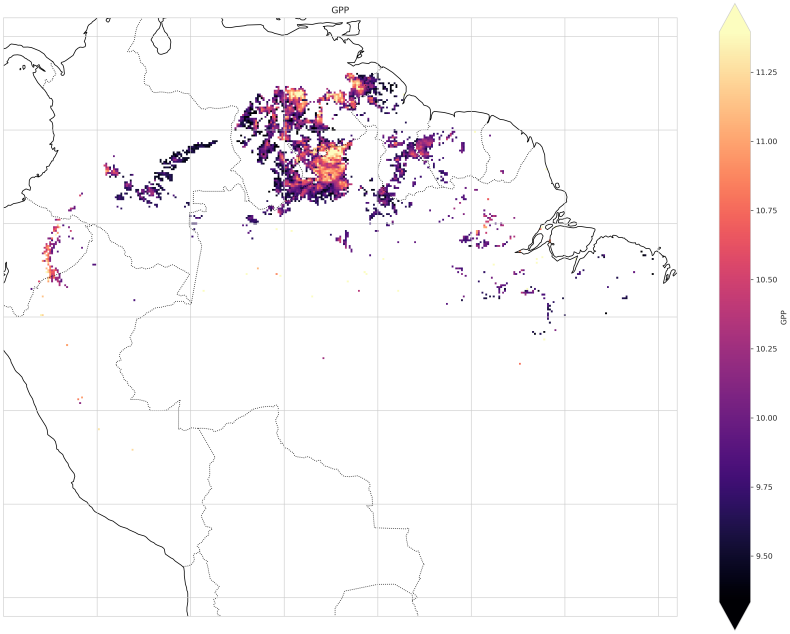
(c) SHAP value distribution

Figure 5.8: Cluster 4 analysis showing (a) the mean GPP distribution across the region, (b) feature importance measured by mean SHAP values indicating average impact on GPP prediction, and (c) distribution of positive and negative feature impacts across all test samples.

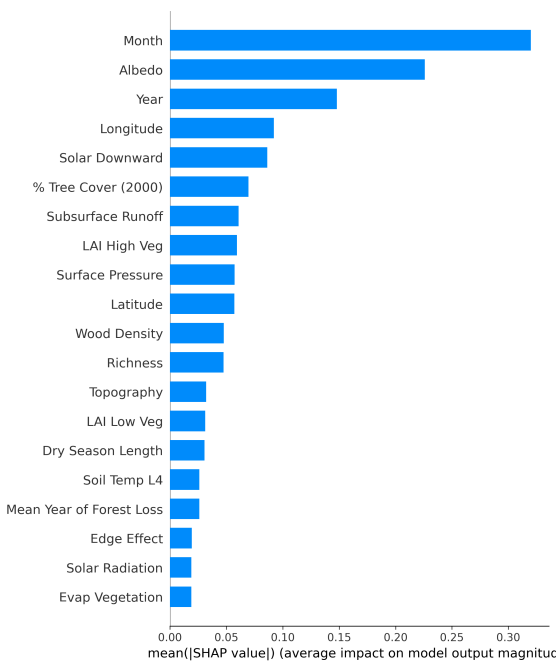
REWORD: Examining the distribution of SHAP values for some of these important features, it can be seen that elevation and bare soil evaporation have a negative influence on GPP predictions. Available phosphorus and LAI of high vegetation have a positive influence. Total precipitation, diversity and lower soil temperatures display more complex patterns of influence that require further examination. The combined importance of leaf area index of high vegetation and topography, along with diversity as a lower-importance feature, suggests that species composition and structural characteristics of trees significantly influence productivity patterns in this region. The importance of evaporation from bare soils well as available phosphorus as the dominant form may indicate either highly weathered exposed soils with lower vegetation with shallow roots or mountainous regions where environmental constraints limit forest productivity. It is indeed the case that Cluster 4 has a higher median elevation than Clusters 2 or 3. This is conflated by the observation that surface presusure has a negative influence on GPP predictions as well as the observation that the region has higher organic and available phosphorus. The phosphorus dynamics in Cluster 4 are characterized by available phosphorus as the predominant form, appearing in conjunction with subsurface runoff and bare soil evaporation as important features. This nutrient-hydrological coupling may indicate either young soils with readily available phosphorus or weathered systems where exposed soils lead to phosphorus availability becoming a critical limiting factor.

Cluster 5: Northern Belt

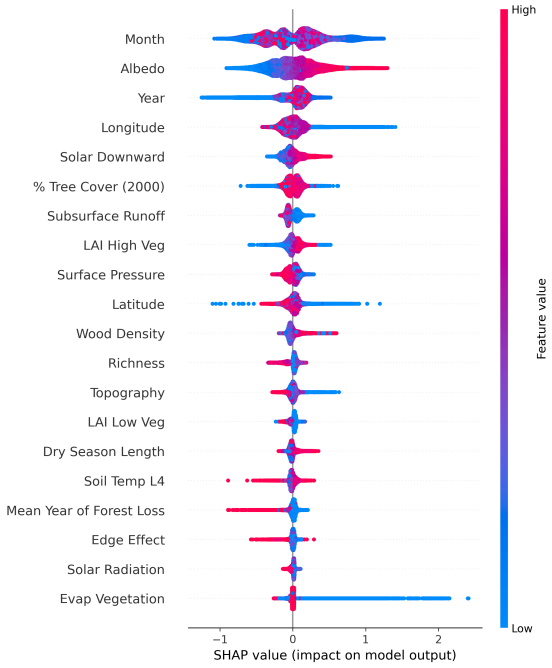
Cluster 5, representing the northern belt region, is distinctively characterized by longitude and solar downward radiation emerging among the highest importance variables, indicating that shortwave radiation availability for photosynthesis constitutes a primary limiting factor. Furthermore, evaporation from vegetation appears as an important variable for the first time, with lower values of this feature producing higher SHAP values than any other feature, suggesting water use efficiency being optimised where evaporation from the leaf is lower. The combination of the negative influence of solar radiation and elevation (topography) suggests that these may be mountainous regions where certain stands benefit from favorable solar aspect orientations that optimize light access without excess evaporation limiting photosynthesis. The complex relationship between solar radiation variables — where lower solar radiation values produce positive SHAP contributions while higher solar downward radiation values also yield positive responses — may indicate that reflection effects and atmospheric attenuation factors account for differences between incident solar radiation and net available photosynthetically active radiation. Given the northern belt's equatorial proximity, latitude becomes less influential while longitude emerges as a critical spatial predictor, with lower longitudinal values (closer to the Atlantic)



(a) Mean GPP distribution



(b) Feature importance



(c) SHAP value distribution

Figure 5.9: Cluster 5 analysis showing (a) the mean GPP distribution across the region, (b) feature importance measured by mean SHAP values indicating average impact on GPP prediction, and (c) distribution of positive and negative feature impacts across all test samples. Month and albedo emerge as the most influential predictors, with temporal patterns driving primary variation in GPP predictions.

leading to higher SHAP values and increased GPP. Among medium-importance features, low subsurface runoff has a negative influence on GPP, indicating that water availability at root zones represents a limiting factor rather than waterlogging, emphasizing water limitation rather than excess moisture as the primary hydrological constraint. Wood density and species richness appear as similarly important features and the region is characterised by lowest species richness and diversity of any cluster, sometimes 0 in some regions indicating the dominance of one species. The bimodal influence of percentage tree cover and dry season length indicate heterogeneous ecological behaviors across this cluster, while lower species richness leading to positive SHAP values suggests either competitive exclusion processes or inefficient species adaptations at various altitudes within these mountainous regions. The negative influence of topography response — where lower elevation areas yield positive GPP contributions while high elevation areas produce negative effect s—indicates distinct low-elevation and high-elevation ecological zones with contrasting productivity patterns. Vegetation structural dynamics reveal that higher values of leaf area index for high vegetation correlate with positive SHAP contributions, while lower values for low vegetation LAI also produce positive responses. This pattern indicates that taller tree stands without understory competition optimize resource utilization of scarce water and light resources (Laurans et al., 2014). Interestingly we observe the opposite influence of edge effects to the Central Amazon, where here higher distance from the forest edge indicates lower SHAP values.

5.5 Limitations

The methodology employed identifies distinct functional regions within the Amazon basin in terms of GPP patterns over a 20 year period and identifies the key drivers of GPP within each of these spatial regions. However, the analysis does not explicitly handle the time evolution of these drivers and the evolution of their influence on GPP. Future work should seek to either employ methods that account for memory effects or an explicit temporal representation. This could be achieved by adding lagged variables as input features, utilising ML architecture that encode memory through recurrent connections and/or convolutions, or physics-informed methods that either discover or incorporate physical laws in the form of systems of dynamical equations to enable forward integration and the prediction of future GPP states (Gasparrini, 2016; Brunton et al., 2016; Karniadakis et al., 2021; Yu et al., 2024). In addition, as many of the features utilised in this study come from data products that involve modelling and reanalysis, these data come with their own limitation as to how accurately they represent the *true* systems they are modelling. A direct example from this study was the replacement of ERA-5 measurements of

precipitation with the more accurate MERGE product. There are many other examples where modelling could be ameliorated by the inclusion of more accurate, fine resolution measurements. In addition, as the aim of this study was to identify large-scale patterns within the region the data were coarsened before analysis. This may miss out on some of the fine scale patterns and interactions that could be key to understanding GPP dynamics in the region.

5.6 Conclusions

This study demonstrated the utility of ML methods to identify similar regions of GPP function over a 20 year period with the use of SHAP explanations to identify suitable hypotheses as to the underlying biogeochemical and physical processes underlying these complex ecological patterns. Though SHAP explanations are powerful in highlighting the potential connections between environmental drivers and GPP, they do not constitute a causal analysis and therefore the hypotheses as to the underlying functioning of the systems in each region must be verified. However the overall trends and the identification of similar regions that are under stress and under threat of losing carbon capture function is of significant importance for the future of the Amazon as a significant carbon source over the coming decades and the potential to reverse this trend through conservation.

The Peripheral region in the mountainous Southern and Eastern Amazon emerges as particularly vulnerable, consistent with recent observations of carbon source behavior. Even more concerning is the identification of an extremely large body of the Central rainforest, along with regions along the Andes and in the North-East towards the Atlantic (Cluster 2) where the influence of degradation and deforestation indicates vulnerability to external pressures. These central regions are those that are most in need of attention as their continued degradation could lead large-scale dieback and the rapid savannisation of the Amazon (Nobre and Borma, 2009), especially considering their interwoven nature with other clusters and the increased influence of edge effects in this region. Addressing these vulnerabilities require rapid and widespread systemic solutions that occur at the level of national government conservation strategies (Brando et al., 2025). These insights can inform the targeting of specific areas for protection due to their increased vulnerability and help improve Earth system model representations of tropical forest carbon dynamics.

5.7 Data Availability Statement

The datasets supporting this study are derived from multiple sources with varying access requirements. All data sources are described below with appropriate access

information and citations.

5.7.1 Publicly Available Data

ERA5 Reanalysis. Atmospheric, land surface, and energy balance variables were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis dataset (Hersbach et al., 2020). Monthly data (1950–2022) at 0.1° spatial resolution are freely available through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/) following user registration.

MERGE Precipitation. High-resolution precipitation estimates combining gauge observations with satellite data were obtained from the MERGE dataset (Rozante et al., 2010). Data are available through the Center for Weather Prediction and Climate Studies (CPTEC/INPE) at http://ftp.cptec.inpe.br/modelos/tempo/MERGE/.

MapBiomas Land Cover. Land use and land cover classifications (Collection 6.0) were derived from the MapBiomas project (Souza et al., 2020). Annual maps (1985–2020) at 30 m resolution are freely available at https://mapbiomas.org/ for non-commercial research use.

SRTM Topography. Terrain characteristics were derived from the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global digital elevation model, distributed by the U.S. Geological Survey (**?**). Data are freely available through https://earthexplorer.usgs.gov/.

PRODES Deforestation. Annual deforestation data (2000–2017) were obtained from Brazil's National Institute for Space Research (INPE) Program for Calculating Deforestation in the Amazon (PRODES). Data are publicly available at http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/amazon/.

5.7.2 Restricted Access Data

Forest Inventory Data. Tree density, species diversity, and richness estimates were derived from the Amazon forest inventory database compiled by ter Steege et al. (2023). Access to plot-level data requires approval from the Amazon Tree Diversity Network. Data requests should be submitted through http://www.biodiversityresearch.org/research.php with a detailed research proposal.

Soil Phosphorus. Gridded estimates of available, organic, and total soil phosphorus were obtained from Random Forest model predictions developed for tropical South America (Darela-Filho et al., 2024). Access to these data requires permission from the original modeling team.

Wood Density. Community-weighted mean wood density estimates were derived from Mo et al. (2024). Processed gridded data are available upon request from

the corresponding author.

Environmental Disturbance. Edge effects, fire occurrence, drought stress, and logging intensity data were obtained from the environmental disturbance dataset of Lapola et al. (2023).

Gross Primary Productivity. GPP estimates used as the target variable were obtained from the GOSIF product (Li and Xiao, 2019) and can be accessed with permission from the authors from the Global Ecology Data Repository (https://globalecology.unh.edu/data/GOSIF-GPP.html, accessed January 2025).

Chapter 6

Discussion

The fundamental question that drives this research is how ML models can be meaningfully integrated into environmental science workflows to enhance our understanding and modelling of complex Earth system processes.

This thesis has presented four applications of Machine Learning (ML) in environmental science across a range of diverse application objectives. These works demonstrate the broad applicability of these techniques for accurate modelling and the discovery of new scientific hypotheses when combined with explanations ranging from simple gain-based feature importance to more robust SHapley Additive exPlanations (SHAP) values.

The diversity of these applications is evident in the fundamentally different research goals addressed: gap-filling of meteorological data (Ch. 2), partitioning of evapotranspiration (ET) (Ch. 3), modeling Boundary Layer Height (BLH) dynamics (Ch. 4), and understanding spatial patterns in Gross Primary Productivity (GPP) (Ch. 5). These applications span different temporal and spatial scales, ecosystem types, and modeling objectives, demonstrating the flexibility and power of ML approaches when coupled with appropriate interpretation methods.

In addition, these applications are all diverse in terms of the characteristics of the data. The key differences between the data are outlined in Table 6.1 based on data volume, coverage, and the availability of ground-truth labels.

The methodology employed across applications has been developed progressively from basic feature importance analysis in the gap-filling application, adding recursive feature elimination for the ET partitioning and BLH modeling, to cluster-based analysis with SHAP values for the GPP application. This evolution reflects both the increasing complexity of the research questions and a deepening understanding of how interpretation methods can be tailored to specific scientific objectives.

6.1 Summary of Findings

This thesis has demonstrated that systematic application of ML methods with explanations can discover previously unknown environmental drivers and generate testable scientific hypotheses. While the tools are well-established, their strategic application across diverse data scenarios — from gap-filling with complete ground truth

Table 6.1: Categorisation of thesis applications by data characteristics and ML task complexity

Application	**Volume**	**Coverage**	**Complexity**	**Ground Truth**	**Key Challenges**
Gap-filling (Ch. 2)	Moderate	Moderate	Mixed	Available	Standard supervised learning, some variables more complex than others
ET partitioning (Ch. 3)	Moderate	Low	Moderate	Unavailable	Proxy ground-truth via data augmentation, validation challenges
BLH modeling (Ch. 4)	Moderate	Low	Moderate	Available	Limited spatial representation
GPP modeling (Ch. 5)	High	High	High	Available	High heterogeneity, system complexity

to unsupervised regional classification — required careful methodological choices about algorithm selection, feature engineering, and explanation methods. The value lies not in the tools themselves, but in demonstrating when and how they reliably produce scientific insights across diverse environmental science applications.

Several concrete examples from this research illustrate the power of this approach. The identification of unexpected features as predictors for ET partitioning (methane flux at the Mayberry site and carbon dioxide concentration at the East End site) led to hypotheses that require further investigation. Similarly, the identification of deeper soil temperature relationships at the ATTO site led to a hypothesis that could be immediately tested from available data, revealing that thermal inertia may play a role in modulating BLH. This finding warrants further investigation and demonstrates how ML-driven discovery can generate testable hypotheses.

Encouragingly, for all applications, the ML models successfully identified features with known relationships to target variables without expert input, which is promising for building trust in these approaches (Reichstein et al., 2019; Ali et al., 2023). However, certain variables with established relationships to BLH were not available for model training, such as latent heat flux. This limitation suggests both that ML models may not always be the optimal choice for all applications but also that ML models may be useful where data are not available as it has been shown that sufficiently accurate models can still be produced from the available data. Care should be taken to ensure that there are sufficient data to represent known important relationships, even if this is via intermediary correlated variables.

The operational challenges inherent in setting up and maintaining remote sensing networks mean that noise and missing data are ubiquitous in environmental

science applications. For this reason, the most robust approach may be to combine ML and process-based models in multi-model ensemble frameworks where computational cost or prediction time are less critical than accuracy. Where time constraints are paramount, ML models offer distinct advantages due to their rapid inference capabilities. Where robustness is paramount process-based models may still be preferable though the promise of causal learning and physics-informed methods will be discussed in Section 6.3

In the context of filling gaps in automatic weather station data, feature importance scores enabled the discovery that data from neighboring weather stations, regardless of distance, ranked more importantly for prediction across temperature, relative humidity, and longwave radiation variables. Interestingly, the debiasing of ERA-5 reanalysis data alone performed comparably well with ML methods, indicating that proper understanding and treatment of the data remains paramount to any modeling approach.

The ET partitioning (Ch. 3) and BLH (Ch. 4) modeling applications demonstrated the applicability of feature importance scores as a heuristic for recursive feature elimination. This approach enabled the identification of optimal feature sets with minimal features without significant loss in predictive performance, providing an efficient pathway for model development and interpretation. This reduction of features in particular identified a small subset of 5-6 features that could be obtained with inexpensive equipment for the prediction of the BLH. Given the operational cost and significant challenges of taking BLH measurements in the Amazon and the lack of BLH data across this expansive and heterogeneous ecosystem there is potential for the combination of ML techniques with observational campaigns to provide more extensive coverage of the BLH. The GPP modeling work (Ch. 5) discovered regions with similar gross primary productivity patterns over a 20-year period, but it was not until explanations were applied that meaningful insights into the underlying processes driving these clusters could be ascertained. The initial intention of this research was to understand factors influencing GPP and changes in GPP over time across different Amazon regions, and this study enabled much greater specificity in understanding these processes. In this case, the primary value of modeling was understanding enabled by explanations over the ML models, somewhat independent of the model's predictive utility.

This thesis has demonstrated strong evidence for the use of accessible XAI tools that environmental science researchers could readily adopt for their datasets. The ultimate value resides in the data being collected, where the principal work of informing our understanding of these systems has been accomplished. ML systems are fundamentally built upon these datasets and possess no intrinsic knowledge about environmental systems. The value lies in the data itself, with information contained

within the data and knowledge extracted by identifying relationships hidden within these datasets.

These tools are accessible to researchers (Flora et al., 2024) and demonstrate the broad applicability of ensemble methods (such as LightGBM and XGBoost) in environmental science with minimal computational requirements, making them suitable for widespread adoption across the research community.

6.2 Methodological Considerations & Limitations

Several methodological limitations emerged during this research that warrant discussion. An overview of those identified for the individual publications will be given first before discussing general themes that are connected across the works.

6.2.1 Gap-filling

With such a large volume of experiments (1,720) conducted across different target variables (x5), gap lengths (x4) and gap-filling techniques (ML x3, statistical x2) the analysis and presentation of these results for practical recommendations becomes unwieldly. A systematic framework for benchmarking and comparing gap-filling methodologies would be of great benefit to future research on gap-filling techniques and allow for the comparison of techniques across applications. In addition, the Debiasing and ML methods utilised different artificial gap-creation methods. Future work should apply these techniques to the same gap profiles to ensure consistency and fair and accurate comparison of techniques.

6.2.2 Evapotranspiration Partitioning

The evapotranspiration partitioning technique developed in this thesis faces significant practical limitations when considering extending the methodology to other ecosystems, for example the Amazon tropical rainforest system examined in the BLH and GPP modelling work (Chapters 4 and 5). There is no clear period where transpiration is known to be negligible. Therefore the method needs to be adapted in some way so that there is a suitable target for supervised learning.

The methodology from Eichelmann et al. (2021b) and the code from Stapleton et al. (2022) have been adapted for the partitioning of ET in deciduous broadleaf forest (Foley, 2023). It was found that the method did not extend to this more complex system, identified principally by the authors to be due to lack of precise direct measurements of E and T.

It may be possible to filter other components of ET in rainforest systems using this technique, such as separating out soil, understory, intercepted and above canopy

ET. However, some of the same limitations still apply. The method requires extensive measurement within the canopy at multiple levels as well as above the canopy. Given these requirements, one must question whether the approach is worthwhile for Amazon applications without substantial infrastructure development.

Due to these measurement limitations in the Amazon region, the modelling of evapotranspiration within the rainforest using supervised ML is suggested for future work. This thesis proves the utility of the approach and creates a framework from which to build ML models for ET partitioning in other, less complex environments.

However, it must also be noted that the difficulties in extending this approach are universal for applications where the phenomenon being studied is not measured directly but only proxy measurements and data for correlated variables exist. This is a widespread issue not only in the environmental sciences but in many other fields (Zhou, 2018). The strategy of data augmentation or selective data sampling based on domain knowledge is innovative and ingenious, though requires further research to determine whether this technique is extensible to other applications. One of the major issues is validation - without a ground truth to test against, there is no means to ascertain whether the models have learned to predict the target as designed by the experiment. In the case of the ET predictions, leaf-level measurements were used as a proxy in work by Eichelmann et al. (2021b) to determine whether the predictions were following the same trends as those of local-level transpiration predictions. Another challenge is in determining suitable data augmentation or selection strategies in order to identify suitable target data for supervised learning. This requires some ingenuity on the part of the domain expert and is likely beyond the reach of most ML engineers without an intimate knowledge of the domain. In this way the method provides an opportunity that requires scientific creativity. Semi-supervised learning is also suggested for future research, wherein there exist ground-truth for only a subset of the data. It is the case for the ET research that both the leaf-level measurements and the selected periods for training are only proxy data, not the bona fide ground-truth for a subset of periods. However the methodological considerations may still be useful and techniques modified to suit this scenario, such as treating the data as having a high proportion of noise and selecting methods specifically for this.

6.2.3 Boundary Layer Height Modelling

The primary aims of the BLH research were to: (a) demonstrate that it is possible to develop accurate models for the complete diurnal course of the BLH using only a few ground sensors; (b) improve upon diurnal representation of BLH development and address the discrepancies identified in Dias-Júnior et al. (2022). Data limitations due

to operational challenges in the Amazon are one of the main barriers to widespread, accurate modelling of the BLH that takes into account the high heterogeneity of this ecosystem. Though there are many drivers with known relationships with the BLH and influence on its development, we have demonstrated that not all of them are needed for accurate BLH predictions.

The recursive feature elimination process for BLH modeling revealed duplicate soil temperature sensors as important features, suggesting that feature importance scores may not be as robust as initially assumed, as has been found in the literature (Lundberg et al., 2018). This finding could indicate that important processes are captured in the spatial variability of soil temperatures, but more likely represents a shortcoming of the methodology itself.

Feature importance scores in ensemble methods are obtained through statistics aggregated over all weak learners, specifically the sum of gains in predictive performance (reduction in RMSE). It may be that different weak learners randomly selected different soil temperature sensors, with no additional sensors needed for any particular learner. Improved feature engineering could potentially mitigate this issue. This example demonstrates the value of applying simple explanation methods first to identify flaws in the ML pipeline that can be readily addressed, showcasing the utility of explanations for model diagnosis and quality assurance (Saarela and Podgorelec, 2024; Huang et al., 2025).

Subsequent work since the publication of this research utilised random forests at the GoAmazon site using only expert selected variables as input, namely: air temperature, relative humidity, wind velocity, net radiation, and surface turbulent fluxes, such as sensible and latent heat fluxes though there were significant missing data for latent heat and net radiation (Silva et al., 2025). Their models exhibited significantly decreased performance, with an increase of RMSE of approximately 200 metres for best performing models and a decrease of 0.24 in R^2 scores. The authors also noted a statistically significant improvement in performance of approximately 1% decrease of RMSE values of random forests over LightGBM and XGBoost for their models. The argument for retaining LightGBM as the model of choice throughout the thesis is not that Random Forests cannot match or outperform LightGBM, but that LightGBM optimises both predictive accuracy and speed in training and inference. With the difference of 1% being negligible in practice this makes it an optimal choice for operational scenarios. Silva et al. (2025) used SHAP values as their method of explanation which should give more robust estimates of feature importance and other advantages in model interpretation such as distribution of feature influence and feature interactions. However with feature selection carried out apriori it is difficult to draw a meaningful comparison between the interpretation results.

6.2.4 Gross Primary Productivity

The SHAP explanations presented in this research, while highly informative as to the predictive factors associated with GPP in the regions identified by clustering, do not constitute a causal analysis. There is a high probability, based on the interpretation and analysis of these factors and their relationship with existing theory, that there is in fact a causal relationship between these variables. However, there is always the possibility that the models may be learning spurious correlations. Therefore these hypotheses must be verified by suitable methods that will be discussed in Section 6.3. One of the major drawbacks of this methodology is that it does a better job of explaining *spatial* differences across the Amazon and not necessarily the time evolution. One of the major outstanding questions in the research literature concerns the long-term evolution of the Amazon rainforest and though this research contribtutes to better understanding of this complex system, this methodology lacks an explicit focus on the time evolution of the system.

6.2.5 General Methodological Considerations & Limitations

The field of XAI has developed rapidly over the course of recent years (Ali et al., 2023). While SHAP (SHapley Additive exPlanations) has become the most popular XAI method in environmental science applications, explainability encompasses a broader suite of techniques. Feature importance metrics, permutation importance, partial dependence plots, and model-agnostic interpretation methods all contribute to understanding model behavior and extracting scientific insights from ML models. This thesis employs multiple XAI approaches appropriate to each application domain. Gain-based feature importance scores (applied in Chapters 2, 3 and 4) such as those available in random forests, XGBoost and LightGBM were previously included as explanation techniques in many works (Saarela and Jauhiainen, 2021; Ali et al., 2023; Saarela and Podgorelec, 2024). These scores have long been known to have limitations (such as bias towards higher cardinality features and inconsitency in that a features importance score decreases as its true impact increases) and in recent years these techniques have largely been replaced by more robust methods like SHAP (Strobl et al., 2007; Lundberg et al., 2018). Feature importance scores may still have a place as a diagnostic tool for troubleshooting models and as a first point of contact for environmental researchers that lack the appropriate expertise to employ more advanced methods. These scores are as easily produced as the models are trained and have been shown in the works on ET partitioning and BLH modelling to produce scientific hypotheses of merit. The learning curve for researchers to employ such techniques is a factor not often considered and simpler techniques have value in their ease-of-access as an entry point for environmental researchers look-

ing to employ these techniques. Notwithstanding this, more robust methods like SHAP should be given preference wherever possible and feature importance scores should be avoided for more in-depth analyses or debiased to account for their known limitations (Strobl et al., 2007; Lundberg et al., 2018).

We must also consider the substantial environmental impact of larger ML models. Wherever possible, researchers should utilise efficient tools such as LightGBM that reduce the energy consumption and emissions of applying ML (Strubell et al., 2019; Schwartz et al., 2020). The general approach throughout this thesis has been to see what useful scientific insight can be gained from using the simplest and most computationally efficient methods first and then expanding on this complexity as required. However the threshold of this increase in complexity and "satsificing" for a given set of research goals is somewhat arbitrary and task dependent.

Early experiments employed Bayesian hyperparameter optimisation in order to fine tune models for optimum predictive performance using the HyperOpt and Optuna packages. The difference in results was generally found to be negligible when compared to those obtained by feature set optimisation and therefore hyperparameter tuning was omitted for all papers. The reasoning is similar to the above, that the substantial increase in the number of experiments and therefore both the computational cost and energy consumption is not justified by these marginal gains in performance. In addition, hyperparameter tuning is not straightforward and is a step that could provide a barrier to entry for environmental scientists looking to applying ML models. Should these models be developed for deployment, hyperparameter tuning should be reintroduced as part of the production pipeline.

The application of recursive feature elimination allowed for the discovery of optimal feature sets with minimal reduction in prediction metrics. However, the feature sets idenified in Chapters 3 and 4 may not be truly optimal as SHAP is a more robust method of explanation over gain-based feature importance which has known limitations (bias toward certain feature types, lack of mathematical guarantees) (Strobl et al., 2007; Lundberg and Lee, 2017b; Lundberg et al., 2018). Further research should employ the techniques outlined in the GPP modeling chapter and avoid the use of gain-based feature importance.

A limitation across all applications in this thesis is the absence of formal uncertainty quantification (UQ). While the ML models provide point predictions with associated performance metrics (RMSE, R^2, MAE) averaged across predictions, they do not quantify the uncertainty or confidence associated with individual predictions. This limitation is particularly critical in environmental science applications where model outputs inform decision-making under conditions of data scarcity, high system complexity, and potential distribution shift due to climate change (Liu et al., 2023).

The ensemble methods employed throughout this thesis (Random Forests, Gradient Boosting, XGBoost, LightGBM) inherently contain information that could be leveraged for UQ, though this was not implemented in the present work. Several approaches are available which could address this limitation in future research. Random Forests naturally provide a basis for uncertainty estimation through the variance of predictions across individual trees (Coulston et al., 2016; Mentch and Hooker, 2016). More sophisticated approaches include quantile regression forests, which estimate conditional distributions rather than point predictions (Meinshausen and Ridgeway, 2006).

6.3 Future Research Directions

Multiple authors have emphasised that XAI is needed to build confidence in the underlying function of ML models and ensure they are robust before becoming operational (Reichstein et al., 2019; McGovern et al., 2019; Roscher et al., 2020). Trust and the ability to diagnose models to ensure correct function are paramount for operational deployment in environmental applications. I argue that the methodologies presented in this thesis should not only be used for the virtue of incremental advances in understanding of well-modelled systems, but could readily be applied to other complex systems that are not fully understood and whose components are manifold and whose behaviors are complex, such as the Amazon rainforest. The Amazon rainforest is but one example of an extremely complex, highly heterogeneous system in Earth System Science (ESS) and environmental science, whose behaviour may be poorly modeled by traditional process-based models (Bonan and Doney, 2018; Fisher and Koven, 2020).

Examples of these include clouds, ice sheets and vegetation dynamics (Hourdin et al., 2017; Fisher et al., 2018; Pattyn et al., 2018; Satoh et al., 2019). The stability of models under changing climate conditions represents another critical area requiring immediate attention as we face unprecedented environmental challenges (Reichstein et al., 2019). XAI approaches alongside causal discovery and physics-informed approaches stand out among the best means to address the limitations of process-based models in capturing sufficient complexity without exceeeding computational resource limits as well as well as ensuring better domain generalisation over ML models with explanations (Peters et al., 2017; Raissi et al., 2019; Runge et al., 2019; Schölkopf et al., 2021; Reichstein et al., 2019; Karniadakis et al., 2021).

Causal research in particular could identify which variables are fundamental drivers and which are intermediary, though the lack of observational data for certain variables (such as latent heat exchange in the BLH application) may still make the intermediary variables better correlative predictors. Physics-informed models can

incorporate existing knowledge on what is already understood about a system and learn "higher order" processes (residual terms in a system of dynamical equations) from the data. Schemes for the direct learning of a set of dynamical equations such as Sparse Identification of Non-Linear Dynamics (SINDy) and its variants have been successful in many complex problems as well as physics-informed NNs that incorporate physical laws as part of the loss function of the NN (Brunton et al., 2016; Raissi et al., 2019; Lawal et al., 2022; Toscano et al., 2025).

The flexibility and scalability (in terms of computational resource requirements) of ML models with explanations over causal learning and physics-informed approaches, at present, may make these models more viable for large scale climate and weather models (Rasp et al., 2018; Balaji, 2021; Harder et al., 2022). However these are areas of active research with great promise and will likely supplant XAI techniques in years to come in many domains due to their combination of direct data integration with explicit, directly interpretable representations that provide practitioners with greater trust and ensures better domain generalisation (Reichstein et al., 2019; Beucler et al., 2021; Kashinath et al., 2021). Several avenues of potential future research also present themselves that are specific to the applications studied in this thesis.

6.3.1 ET Partitioning

The ET partitioning framework should be applied to other wetland sites from the AmeriFlux network as well as other flux networks and the results compared with existing partitioning models inlcuding but not limited to those using measurements of water isotopes, underlying water use efficiency, leaf-level transpiration and sapflow (Stoy et al., 2019; Nelson et al., 2020; Rothfuss et al., 2021; AmeriFlux Management Project, 2024). Testing both local and global models (i.e. site specific models versus those trained on all available data) would improve understanding of general versus site-specific processes, with SHAP providing more robust explanations and feature selection compared to the feature importance methods used in the work presented in this thesis (Ch. 3).

6.3.2 Boundary Layer Height Modelling

Future research should identify suitable modification the methods presented in Ch. 4 to utilise existing datasets (which may not include direct measurements of the BLH) alongside collection of new data in order to extend the spatial coverage of these models and to include different rainforest sites. General, global, multi-site models coupled with site-specific models as presented in this work may be the best means to identify what is generically predictive across the Amazon and what is

specific to different regions, with feedback in learnings from each model used to improve the others.

In order to assess the fidelity of the ML models future work should compare ML predictions of the BLH with numerical and computational models examining the BLH under different regimes such as stable, neutral and convective conditions (Hanna, 1969; Zilitinkevich, 1972; Deardorff, 1974; Vilà-Guerau de Arellano et al., 2015).

For BLH modeling, learning a mathematical representation of boundary layer height dynamics using sparse identification of nonlinear dynamics (SINDy) (Brunton et al., 2016) could improve process-based models. The use of many sites, including those outside the Amazon, and combining radiosonde measurements for accuracy and ceilometer data for temporal resolution represents the most robust approach for extrapolating to other rainforest regions. While the BLH is already well studied and represented by complex theoretical and numerical models (Vilà-Guerau de Arellano et al., 2015), data-driven learning of dynamics could help identify gaps in the underlying theory, confirm or support existing theory by learning representations from the data, and identify higher-order behaviors present in real-world dynamics that may not be captured by theoretical models.

Applications with low data coverage could benefit from the addition of techniques specifically designed for domain generalisation to account for the fact that other sites may have different underlying data distributions. Domain generalisation approaches such as domain-invariant feature learning (Muandet et al., 2013; Li et al., 2018a), meta-learning frameworks (Sun et al., 2024), and domain randomisation techniques (Tobin et al., 2017) could enable the models to better generalise to unseen Amazon sites. Additionally, causal inference methods show promise for improving generalisation across heterogeneous environments (Schölkopf et al., 2021). The addition of sites *without* direct measurements of the BLH would move this problem into the realm of semi-supervised learning, where approaches such as psuedo-labelling (Lee et al., 2013) consistency regularisation (Sohn et al., 2020) could be employed.

6.3.3 GPP Modelling

Future work on GPP research could utilise Convolutional NNs to capture spatial context or architectures like Recurrent Convolutonal NNs or Convolutional LSTM to capture spatial and temporal context (**?**Liang and Hu, 2015; Shi et al., 2015). As the aim of the work presented in Chapter 5 was primarily to understand large-scale patterns and "take a step back" to see what is going on at the whole ecosystem level, the added complexity of teleconnections and memory effects could equally help or hinder understanding of the underlying processes. Temporal influences between

months and local spatial context could provide valuable improvements in model performance, though it is uncertain whether this would improve interpretability given that the models are already challenging to interpret with the current variable set. With this work complete and setting an initial frame of understanding of the large scale spatial patterns of GPP, further work could expand on this to understand the more complex interactions involving local spatial and temporal context first and then larger scale teleconnections and longer-term temporal influences. More granular approaches could also be taken to account for the fact that the current approach already misses heterogeneity spatially at the gridded resolution employed to uncover more local influences between forest stands. The current approach does not account for memory effects in the system. While time is encoded in the temporal features (*Month* and *Year*), this does not confer knowledge of previous states and their influence on the current state. An alternative approach would be to develop a predictor for the entire Amazon region and compare its performance and interpretability with the cluster-based models presented in this work. This comparison could reveal whether the addition of spatial clustering provides meaningful insights or if a unified model would be better.

Future research could expand on the prediction of GPP to include forecasting of future GPP. Adapting the ML architecture to focus on step-ahead forecasting rather than same-timestep prediction could assist in understanding the patterns in these time series better. In particular modelling of the time evolution of the Amazon rainforest under land use cover and cover change (LUCC) and anthropogenic and climate change driven deforestation will be critical to understanding and modelling this critical carbon sink as it transitions into a source. The potential for restoration and prevention of large-scale dieback is significant but the time left to act is diminshing rapidly (Bastin et al., 2019). Understanding GPP dynamics is not only vital for assessing the health of the rainforest but also for assessing vulnerability to climate change and deforestation pressures. Future research could harness probabilistic methods (such as Markov Random Fields) or generative AI techniques for video (such as diffusion models) to simulate future states of the Amazon rainforest combined with physics-informed ML models to predict the GPP of the forest under future LUCC scenarios (Fischer et al., 2020; Gao et al., 2023).

6.3.4 Integration and Operational Deployment

A critical gap exists between ML research and operational implementation in environmental sciences (Maskey et al., 2019; Irrgang et al., 2021; Jebeile et al., 2023). This issue is not unique to environmental science but also notable in healthcare for example, where models have demonstrated significant potential but practical

adoption has been slow (Chen and Asch, 2017; Seneviratne et al., 2020).

Direct implementation by or collaboration with modelers responsible for large-scale operational models represents the primary pathway for ML integration; otherwise, valuable research risks being lost to the academic literature without practical impact. For this reason the primary contribution of this thesis lies in the knowledge gained about these systems and the propensity to identify new hypotheses for study, rather than the operational integration of these models.

Greater communication, collaboration, and integration between the ML and environmental science communities is essential for advancing the field. However, the most accessible entry point lies with data owners – those who have painstakingly collected, assimilated, and processed environmental datasets. These researchers are best positioned to apply easily accessible ML models such as those presented in this thesis. The models presented require sufficiently low computational resources that they can be run locally even for large datasets, with RAM capacity being the primary limiting factor for throughput. The interpretation of explanations requires intimate knowledge of the field of study, making principal investigators and domain experts ideally positioned for understanding both the data and the underlying processes. However, approaching problems with some degree of naivety offers the advantage of starting without preconceptions and maintaining openness to discovering unexpected relationships in the data. In addition the challenges of training, validating, deploying and maintaining ML models in an operational context should not be underestimated. This is a common role for ML practitioners in industry but perhaps less common in a research context and presents an opportunity for collaboration and knowledge sharing with industry practitioners that are well versed in operationalising ML models in a variety of contexts for diverse objectives. Additionally, greater emphasis should be placed on developing frameworks for integrating ML models with existing operational systems, bridging the gap between research and practical implementation.

The ML modelling of the BLH is taken as a specific example of how the research presented in this thesis could potentially be integrated into an existing environmental science workflow. The methodology must first be refined and tested further, utilising SHAP values for more robust recursive feature elimination, better preprocessing of the data and testing of features to ensure that there are no spurious feature selections (such as may have been the case for the duplicate soil sensors) and to better understand why certain features are being selected. This should then be combined with domain generalisation and semi-supervised learning techniques to ensure that the models generalise to other sites with few or no measurements for the BLH. In addition, global models trained on BLH data from around the globe should be combined with the local Amazon models to provide ensemble predictions.

In addition care should be taken when applying global models to ensure that the predictive features make sense in terms of the underlying system and the chain of causation that are locally relevant to of each site. Short-term campaigns should be carried out to measure the BLH at a suitably selected range of sites across the Amazon both for calibration and validation of the models developed. From there the models should be packaged into a utility in a suitable open-source software environment or online tool that other researchers could access for predictions of the BLH across the Amazon. An example of a potential user would be the atmospheric chemists at the ATTO site who require accurate predictions of the BLH in order to model the development of volatile organic compounds and their dispersion in the atmosphere.

Chapter 7

Conclusion

This thesis has presented four applications of machine learning (ML) in environmental science with scientific discovery enabled by the application of explanations. The comparative analysis of multiple ML models against statistical methods for the gap-filling of automated weather station data (Ch. 2) demonstrated the power of ML approaches over statistical methods. Environmental data have many sources of gaps and adequate methods for the filling of gaps were critical for other applications presented in this thesis (Ch. 3 and 4). The results of this study also demonstrated that statistical methods such as debiasing can perform comparably well and highlighted the importance of proper data preparation and feature selection. Simple explanations via feature importance scores revealed the preferability of data from neighbouring AWS stations over ERA-5 reanalysis data for ML models. This finding has implications for many applications that utilise ERA-5 and warrants further attention in examining the drawbacks of reanalysis data for other applications where local accuracy may be more important than coverage. One of the major contributions of this work to the existing literature was the creation of a novel scheme for gap creation that ensures that the full dataset is covered while avoiding real gaps in the data and with fairly weighted sampling between the training and test sets used for ML model development and validation. The framework presented in Chapter 2 for the comparison of ML models for ET partitioning laid important groundwork for the training and evaluation of multiple ML models as well as feature selection in a systematic fashion. This methodology had important learnings for all other works presented in this thesis, including demonstrating the utility of recursive feature elimination with feature importance as a heuristic in order to obtain parsimonious models. This work demonstrated that simpler models (such as linear models) or more computationally efficient models (such as ensembles like LightGBM or XGBoost) can match or outperform neural networks which were used as the baseline for the partitioning method (Eichelmann et al., 2021b; Stapleton et al., 2022). This work also demonstrated the utility of simple explanations, again in the form of gain-based feature importance, in quantifying and identifying relationships between evaporation, transpiration and local environmental processes. These relationships included those with known relationships with evapotranspiration, such as E as the energy available for evaporation, the moisture gradient driving evapora-

tion, the turbulent processes transporting water vapor away from the surface and the temporal patterns. Two unexpected variables that were were identified by the feature importance ranking were carbon dioxide concentration and methane flux. The identification of these relationships led to the generation of new scientific hypotheses that can be tested in order to better inform understanding of the underlying process and improve models of evapotranspiration. The learnings from Chapter 3 were then used to apply an adapted version of the methodology in Chapter 4 in order to compare ML models for the prediction of the atmospheric boundary layer height over the central Amazonian rainforest. This study further validated the utility of ensembles in producing well-performing models with low computational resource requirements. Here the recursive feature algorithm identified feature sets with no more than six features needed at each of the two study sites in order to produce predictions with no more than a 1% reduction in accuracy metrics. This identification opens up the possibility for low-cost prediction of the boundary layer height across the Amazon region which notably suffers from low data coverage for ground-based boundary layer height measurements. Explanations using feature importance again identified a novel scientific hypothesis that could be verified by collaborators from the available data. This involved the identification of deeper soil temperature as an important feature which was found to relate to thermal inertia in deeper soil layers with regards to rain events. These deeper layers are therefore proposed to modulate the thermodynamic influence of the soil on air masses that produce the fluctuations in boundary layer height. The final application modelling gross primary productivity (GPP) across the entire Amazon basin was the most advanced in terms of the scale, heterogeneity and complexity of the system being studied as well as the volume and variety of the data. The primary aims of this study were to identify spatial patterns of GPP and understand what climatic factors may be influencing differences in GPP across these regions. This final work utilised findings from the previous studies implementing an ensemble model (XGBoost) to predict GPP with the key addition of SHapely Additive exPlanations (SHAP) to improve the robustness of the explanations of the model predictions. By clustering the data based on the 20 year monthly time series of GPP, regions with similar ecological function were identified that resulted of meaningful division of the basin into belts in the north and south, two regions of the central Amazon and a peripheral region. The explanations allowed for detailed understanding of the factors differing between each region and identified that the peripheral region shows the highest vulnerability to loss of carbon capture function. Disconcertingly the study also identified large regions of the central Amazon that may be at risk of loss or reduction in carbon function due to a variety of factors such as deforestation and degradation pressures and drought. The importance of this research is in its coverage and specificity with

an extremely comprehensive suite of explanatory variables considered in modelling GPP and an extensive coverage in terms of the spatial area and time period. The learnings from this research as well as the harmonisation of a range of important dataset leads into the larger goal of modelling the long-term evolution of the Amazon rainforest, including potential large-scale dieback and pathways for prevention, by understanding critical component systems and how they can be modelled using ML.

The publications presented have made several key contributions to the academic literature on applications of machine learning to environmental science problems, including the modelling of diverse phenomena and discovery of new scientific hypotheses.

The research presented in this thesis has implications extending beyond the specific applications demonstrated. The methodologies developed here can be applied to numerous complex environmental systems that remain poorly understood. All four applications, though they all fall within the domain of biosphere-atmosphere exchange are significantly diverse, with different predictive objectives and represent scenarios where traditional process-based models struggle to capture the full complexity of natural processes.

However, it is crucial to acknowledge that all modeling efforts, regardless of their sophistication or accuracy, become irrelevant without the necessary policy frameworks and political will to implement meaningful changes to the human systems causing the degradation and destruction of so many aspects of the natural environment. The scientific community's responsibility extends beyond model development to effective communication and advocacy for evidence-based environmental policies.

Appendix A

Meteorological Gap Filling Supplementary Material

Both the data used for experiments and the complete set of 1,720 results are available on the Zenodo repository. Data from the 10 AWS sites is located here (Koci, 2022) while the results are available as both a CSV file and MySQL database dump with queries to reproduce all tables (Roantree, 2024).

ID	Region	Location	Plant	Year Range	Longitude	Latitude	Elevation (m)
DAD	Kikinda	Kikinda	Plum	2013-2021	20.47	45.83	73
E9A	Novi Sad	Cenej	Apple	2013-2021	19.58	45.28	78.12
E8E	Novi Sad	Cerevic	Peach	2013-2021	19.66	45.21	151
E98	Ruma	JazakIrig-Kudos	Apple	2014-2021	19.76	45.05	140.5
E94	Sombor	Ridjica	Apple	2013-2021	19.09	45.98	83
74D	Subotica	Ljutovo	Apple	2014-2021	19.51	46.06	121.41
A1D	Subotica	Backi Vinogradi	Peach	2014-2021	19.88	46.11	90.49
EAE	Vrsac	Vrsac	Grape	2013-2021	21.32	45.11	128
EA9	Vrsac	Crvena Crkva	Plum	2013-2021	21.28	45.00	83
EA1	Zrenjanin	Sutjeska	Plum	2013-2021	20.75	45.46	84

Table A.1: Weather Station Metadata sorted by Region.

Table A.2: Model performance averaged across target variables (TA, RH, and DP). Results are ranked by normalized root mean square error (nRMSE). Model abbreviations: RF = Random Forest; LGB = Light Gradient Boosting Machine; LR = Linear Regression; Debias = ERA5 with debiasing; Spatial = simple spatial algorithm. Gap size represents the length of gaps in hours (1, 4, 36, or 288).

Model	**Gap**	**nRMSE**	**RMSE**	**MAE**	R^2
RF	1	0.2491	4.6648	2.7555	0.9259
RF	4	0.2901	5.5086	3.3324	0.8977
LGB	1	0.2971	5.7039	3.7865	0.8909
LGB	4	0.3084	5.9430	3.9185	0.8818
LGB	36	0.3294	6.3853	4.1699	0.8632
RF	36	0.3337	6.4540	3.9568	0.8600
LGB	288	0.3636	7.0229	4.5767	0.8271
RF	288	0.3757	7.2713	4.5252	0.8150
Debias	4	0.3782	7.3039	5.3959	0.8260
Debias	1	0.3806	7.3838	5.3742	0.8266
Debias	36	0.3857	7.5895	5.5522	0.8180
Debias	288	0.4002	7.9370	5.7913	0.8035
LR	1	0.4729	9.2689	6.9179	0.7146
LR	4	0.4733	9.2753	6.9238	0.7141
LR	36	0.4751	9.2989	6.9484	0.7113
LR	288	0.4841	9.3521	7.0290	0.6988
Spatial	1	0.6819	13.5154	9.8796	0.3957
Spatial	4	0.6820	13.5136	9.8795	0.3954
Spatial	36	0.6828	13.5003	9.8796	0.3926
Spatial	288	0.6865	13.3750	9.8463	0.3782

Table A.3: Model performance for air temperature (TA) only. Results are ranked by normalized root mean square error (nRMSE). Model abbreviations: RF = Random Forest; LGB = Light Gradient Boosting Machine; LR = Linear Regression; Debias = ERA5 with debiasing; Spatial = simple spatial algorithm. Gap size represents the length of gaps in hours (1, 4, 36, or 288).

Model	**Gap**	**nRMSE**	**RMSE**	**MAE**	R^2
RF	1	0.1424	1.3480	0.9501	0.9792
RF	4	0.1546	1.4629	1.0492	0.9754
LGB	4	0.1567	1.4825	1.0859	0.9747
LGB	1	0.1569	1.4842	1.0806	0.9734
LGB	36	0.1626	1.5340	1.1189	0.9726
RF	36	0.1657	1.5628	1.1248	0.9714
LGB	288	0.1741	1.6133	1.1624	0.9677
RF	288	0.1786	1.6553	1.1773	0.9657
Debias	4	0.2304	2.2260	1.7157	0.9467
LR	1	0.2319	2.1904	1.5668	0.9386
LR	4	0.2321	2.1912	1.5680	0.9385
LR	36	0.2328	2.1926	1.5715	0.9378
LR	288	0.2369	2.1925	1.5834	0.9341
Debias	36	0.2379	2.2296	1.7064	0.9432
Debias	288	0.2446	2.3242	1.7814	0.9400
Debias	1	0.2464	2.3035	1.7498	0.9384
Spatial	288	0.2856	2.6487	1.8499	0.9056
Spatial	36	0.2874	2.7073	1.8557	0.9069
Spatial	4	0.2874	2.7137	1.8556	0.9074
Spatial	1	0.2874	2.7148	1.8557	0.9075

Appendix B

Evapotranspiration Partitioning Supplementary Material

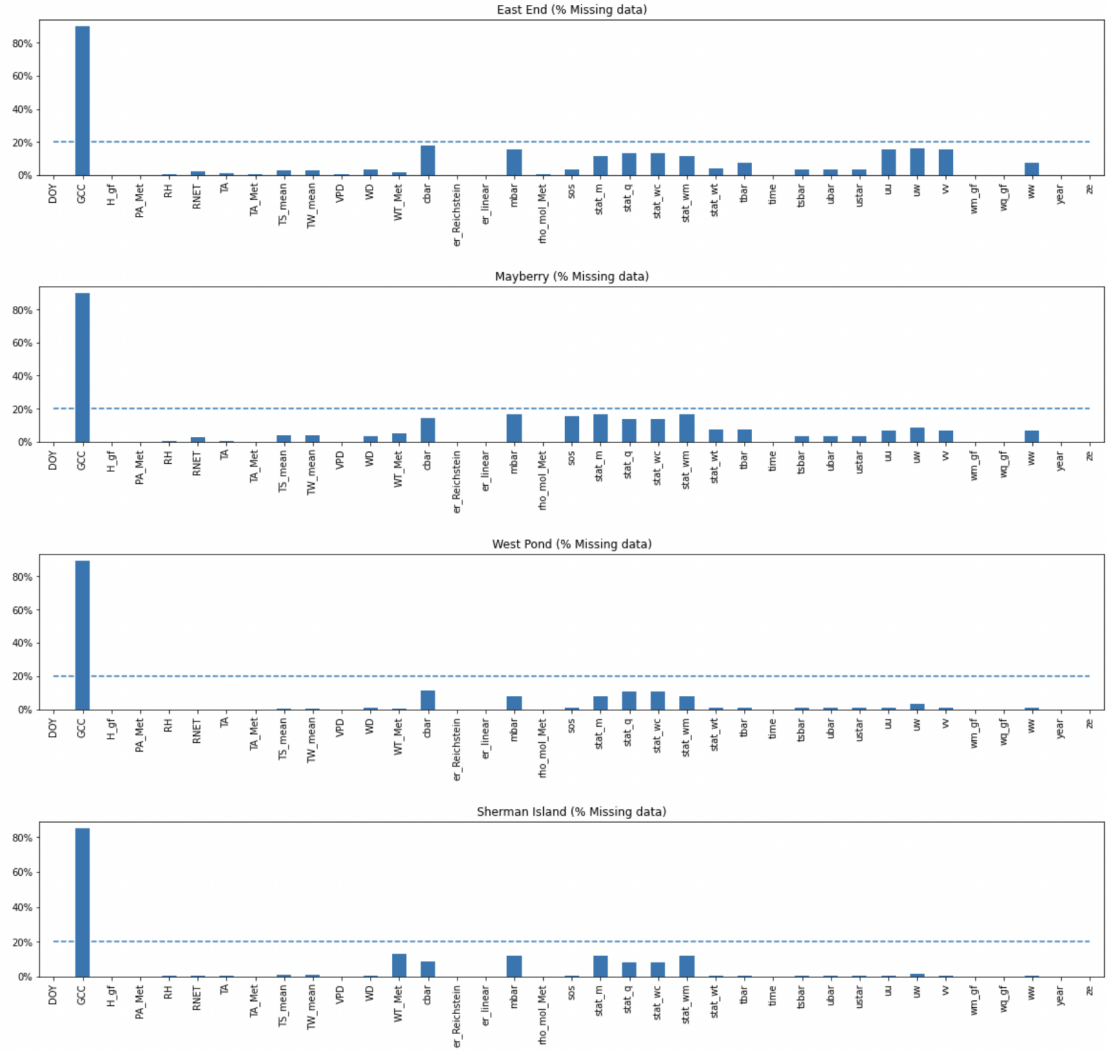


Figure B.1: Percentage of the data that were missing before linear interpolation at each site for the F_{25} feature set. The threshold at which other features were discarded (greater than 20% of the data) is illustrated by the dashed blue line. GCC was not discarded as the missing data can be explained by the fact that GCC is observed daily across a 4 hour interval and the gaps can be suitably filled with linear interpolation.

$\mathbf{F_{25}}$	**Meaning**
DOY	Day of Year
GCC	Vegetation greenness index from camera data
H	Sensible heat exchange
PA	Air Pressure
RH	Relative Humidity
RNET	Net incoming and outgoing shortwave and longwave radiation; net radiation
TA	Air temperature (from thermometer)
TS (mean)	Mean soil temperature
TW (mean)	Mean water temperature
VPD	Vapor Pressure Deficit
WD	Wind Direction
WT	Water Table Depth
c (mean)	Mean carbon dioxide concentration
$ER_{Reichstein}$	Ecosystem Respiration derived from Reichstein method
ER_{linear}	Ecosystem Respiration derived from Linear method
m (mean)	Mean methane concentration
rho_{mol}	Molar air density
sos	Speed of Sound
$stat_{m}$	Degree of instationarity for mean methane concentration
$stat_{q}$	Degree of instationarity for mean water concentration
$stat_{wc}$	Degree of instationarity for wc covariance
$stat_{wm}$	Degree of instationarity for wm covariance
$stat_{wt}$	Degree of instationarity for wt covariance
t (mean)	Average real air temperature calculated from the sonic anemometer
time	Hour of the day
ts (mean)	Average sonic temperature
u (mean)	Mean wind velocity
u*	Friction velocity
uu	Variance of the streamwise (horizontal) wind speed (u)
uw	Covariance between streamwise wind speed and vertical wind speed
vv	Variance of the cross-wind wind speed (v)
wm	Methane flux
wq	Water flux
ww	Variance of the vertical wind speed (w)
year	Calendar year
ze	Zenith angle

Table B.1: Meaning of labels of all predictive features included in the F_{25} feature set.

Feature Set	Model	Adj. R2 Night	R2 Night	Adj. R2 Winter	R2 Winter	Adj. R2 Flood	R2 Flood	Total Time (s)
F_E	DecisionTreeRegressor	0.89	0.89	-0.71	-0.71	0.37	0.37	5.39
F_E	GradientBoostingRegressor	0.91	0.91	0.06	0.06	0.50	0.50	110.40
F_E	KNNRegressor	0.85	0.85	-2.38	-2.37	-1.61	-1.61	8.43
F_E	LGBMRegressor	0.94	0.94	-0.03	-0.03	0.60	0.60	3.26
F_E	LinearRegression	0.77	0.77	-0.28	-0.28	-0.81	-0.81	0.13
F_E	Neural Network	0.89	0.89	0.45	0.45	0.51	0.51	
F_E	Ridge	0.77	0.77	-0.28	-0.28	-0.82	-0.82	0.12
F_E	XGBRegressor	0.95	0.95	-0.25	-0.25	0.53	0.53	22.76
F_{25}	DecisionTreeRegressor	0.90	0.90	-0.29	-0.28	0.52	0.52	17.45
F_{25}	GradientBoostingRegressor	0.92	0.92	0.33	0.33	0.70	0.70	374.38
F_{25}	KNNRegressor	0.88	0.88	-2.00	-1.99	-1.44	-1.42	559.75
F_{25}	LGBMRegressor	0.95	0.95	0.20	0.21	0.75	0.75	5.38
F_{25}	LinearRegression	0.82	0.82	0.33	0.34	0.63	0.64	0.25
F_{25}	Ridge	0.82	0.82	0.34	0.34	0.63	0.63	0.19
F_{25}	XGBRegressor	0.95	0.95	0.22	0.23	0.74	0.74	55.47
F_{RFE}	DecisionTreeRegressor	0.90	0.90	-0.33	-0.32	0.54	0.54	10.05
F_{RFE}	GradientBoostingRegressor	0.92	0.92	0.26	0.26	0.69	0.69	216.50
F_{RFE}	KNNRegressor	0.88	0.88	-2.01	-2.01	-1.43	-1.42	551.62
F_{RFE}	LGBMRegressor	0.95	0.95	0.13	0.14	0.74	0.74	4.08
F_{RFE}	LinearRegression	0.79	0.79	0.23	0.23	0.46	0.46	0.17
F_{RFE}	Ridge	0.79	0.79	0.23	0.23	0.46	0.46	0.13
F_{RFE}	XGBRegressor	0.95	0.95	0.06	0.07	0.73	0.73	37.72

Table B.2: East End, R2 and Adjusted R2 results

Feature Set	Model	Adj. R2 Night	R2 Night	Adj. R2 Winter	R2 Winter	Adj. R2 Flood	R2 Flood	Total Time (s)
F_E	DecisionTreeRegressor	0.86	0.86	-0.55	-0.55	0.50	0.50	2.43
F_E	GradientBoostingRegressor	0.91	0.91	0.83	0.83	0.65	0.65	50.16
F_E	KNNRegressor	0.83	0.83	0.11	0.11	-0.36	-0.36	8.59
F_E	LGBMRegressor	0.94	0.94	0.79	0.79	0.61	0.61	2.18
F_E	LinearRegression	0.83	0.83	0.71	0.71	0.82	0.82	0.09
F_E	Neural Network	0.89	0.89	0.69	0.69	0.53	0.53	
F_E	Ridge	0.83	0.83	0.71	0.71	0.82	0.82	0.09
F_E	XGBRegressor	0.94	0.94	0.53	0.53	0.55	0.55	12.81
F_{25}	DecisionTreeRegressor	0.86	0.87	0.14	0.15	0.32	0.32	7.39
F_{25}	GradientBoostingRegressor	0.92	0.92	0.84	0.85	0.42	0.42	168.42
F_{25}	KNNRegressor	0.85	0.85	-0.18	-0.18	-0.38	-0.38	161.62
F_{25}	LGBMRegressor	0.95	0.95	0.81	0.81	0.42	0.42	3.50
F_{25}	LinearRegression	0.88	0.88	0.62	0.63	0.83	0.83	0.16
F_{25}	Ridge	0.88	0.88	0.60	0.60	0.84	0.84	0.13
F_{25}	XGBRegressor	0.95	0.95	0.71	0.71	0.40	0.41	28.17
F_{RFE}	DecisionTreeRegressor	0.86	0.86	0.16	0.17	0.28	0.28	4.24
F_{RFE}	GradientBoostingRegressor	0.92	0.92	0.85	0.85	0.42	0.42	95.89
F_{RFE}	KNNRegressor	0.85	0.85	-0.23	-0.23	-0.47	-0.46	158.34
F_{RFE}	LGBMRegressor	0.95	0.95	0.81	0.81	0.42	0.42	2.87
F_{RFE}	LinearRegression	0.87	0.87	0.67	0.68	0.82	0.82	0.13
F_{RFE}	Ridge	0.87	0.87	0.68	0.68	0.82	0.82	0.10
F_{RFE}	XGBRegressor	0.95	0.95	0.66	0.66	0.30	0.30	18.42

Table B.3: East End, RMSE and Slope results

Feature Set	Model	Adj. R2 Night	R2 Night	Adj. R2 Winter	R2 Winter	Adj. R2 Flood	R2 Flood	Total Time (s)
F_E	DecisionTreeRegressor	0.87	0.87	-0.39	-0.39	0.08	0.08	8.63
F_E	GradientBoostingRegressor	0.87	0.87	0.54	0.54	0.58	0.58	167.40
F_E	KNNRegressor	0.82	0.82	-0.11	-0.11	-0.10	-0.10	18.99
F_E	LGBMRegressor	0.92	0.92	0.54	0.54	0.71	0.71	4.66
F_E	LinearRegression	0.66	0.66	0.32	0.33	0.52	0.53	0.20
F_E	Neural Network	0.89	0.89	0.43	0.43	0.56	0.56	
F_E	Ridge	0.66	0.66	0.33	0.33	0.52	0.52	0.17
F_E	XGBRegressor	0.93	0.93	-0.08	-0.08	0.58	0.58	34.80
F_{25}	DecisionTreeRegressor	0.89	0.89	-0.50	-0.50	0.27	0.28	27.50
F_{25}	GradientBoostingRegressor	0.91	0.91	0.67	0.67	0.73	0.73	570.01
F_{25}	KNNRegressor	0.87	0.87	-0.10	-0.10	-0.08	-0.07	1119.35
F_{25}	LGBMRegressor	0.94	0.94	0.58	0.58	0.70	0.71	8.38
F_{25}	LinearRegression	0.78	0.78	0.60	0.60	0.60	0.61	0.36
F_{25}	Ridge	0.78	0.78	0.60	0.60	0.60	0.61	0.25
F_{25}	XGBRegressor	0.95	0.95	0.12	0.12	0.61	0.61	83.23
F_{RFE}	DecisionTreeRegressor	0.89	0.89	-0.56	-0.56	0.24	0.24	15.58
F_{RFE}	GradientBoostingRegressor	0.91	0.91	0.64	0.64	0.72	0.73	325.46
F_{RFE}	KNNRegressor	0.86	0.86	-0.17	-0.17	-0.06	-0.06	1104.44
F_{RFE}	LGBMRegressor	0.94	0.94	0.57	0.58	0.70	0.70	5.25
F_{RFE}	LinearRegression	0.73	0.73	0.47	0.47	0.63	0.63	0.22
F_{RFE}	Ridge	0.73	0.73	0.47	0.47	0.62	0.63	0.17
F_{RFE}	XGBRegressor	0.95	0.95	-0.09	-0.09	0.55	0.55	56.48

Table B.4: Mayberry, R2 and Adjusted R2 results

Feature Set	Model	RMSE Night	RMSE Winter	RMSE Flood	Slope Night	Slope Winter	Slope Flood	Slope Day	Total Time (s)
F_E	DecisionTreeRegressor	0.37	0.91	1.09	0.93	0.59	0.59	0.29	8.63
F_E	GradientBoostingRegressor	0.36	0.53	0.74	0.84	0.63	0.77	0.29	167.40
F_E	KNNRegressor	0.42	0.82	1.19	0.83	0.22	0.19	0.07	18.99
F_E	LGBMRegressor	0.29	0.53	0.61	0.91	0.58	0.64	0.30	4.66
F_E	LinearRegression	0.59	0.64	0.78	0.66	0.83	0.46	0.34	0.20
F_E	Neural Network	0.24	0.60	0.57	0.88	0.59	0.48		
F_E	Ridge	0.59	0.64	0.78	0.66	0.83	0.45	0.33	0.17
F_E	XGBRegressor	0.27	0.80	0.74	0.92	0.48	0.72	0.28	34.80
F_{25}	DecisionTreeRegressor	0.33	0.95	0.96	0.94	0.57	0.75	0.37	27.50
F_{25}	GradientBoostingRegressor	0.30	0.45	0.59	0.88	0.72	0.74	0.34	570.01
F_{25}	KNNRegressor	0.37	0.81	1.18	0.87	0.26	0.26	0.10	1119.35
F_{25}	LGBMRegressor	0.24	0.50	0.62	0.94	0.66	0.73	0.34	8.38
F_{25}	LinearRegression	0.47	0.49	0.71	0.78	0.86	0.65	0.41	0.36
F_{25}	Ridge	0.47	0.49	0.71	0.78	0.86	0.65	0.41	0.25
F_{25}	XGBRegressor	0.23	0.72	0.71	0.95	0.57	0.75	0.35	83.23
F_{RFE}	DecisionTreeRegressor	0.33	0.97	0.99	0.94	0.55	0.77	0.36	15.58
F_{RFE}	GradientBoostingRegressor	0.30	0.47	0.60	0.88	0.71	0.76	0.33	325.46
F_{RFE}	KNNRegressor	0.38	0.84	1.17	0.87	0.27	0.27	0.09	1104.44
F_{RFE}	LGBMRegressor	0.24	0.51	0.62	0.94	0.67	0.73	0.33	5.25
F_{RFE}	LinearRegression	0.52	0.57	0.69	0.73	0.78	0.55	0.30	0.22
F_{RFE}	Ridge	0.52	0.57	0.70	0.73	0.77	0.54	0.30	0.17
F_{RFE}	XGBRegressor	0.23	0.81	0.76	0.95	0.53	0.73	0.32	56.48

Table B.5: Mayberry, RMSE and Slope results

Feature Set	Model	Adj. R2 Night	R2 Night	Adj. R2 Winter	R2 Winter	Adj. R2 Flood	R2 Flood	Total Time (s)
F_E	DecisionTreeRegressor	0.86	0.86	-0.55	-0.55	0.50	0.50	2.43
F_E	GradientBoostingRegressor	0.91	0.91	0.83	0.83	0.65	0.65	50.16
F_E	KNNRegressor	0.83	0.83	0.11	0.11	-0.36	-0.36	8.59
F_E	LGBMRegressor	0.94	0.94	0.79	0.79	0.61	0.61	2.18
F_E	LinearRegression	0.83	0.83	0.71	0.71	0.82	0.82	0.09
F_E	Neural Network	0.89	0.89	0.69	0.69	0.53	0.53	
F_E	Ridge	0.83	0.83	0.71	0.71	0.82	0.82	0.09
F_E	XGBRegressor	0.94	0.94	0.53	0.53	0.55	0.55	12.81
F_{25}	DecisionTreeRegressor	0.86	0.87	0.14	0.15	0.32	0.32	7.39
F_{25}	GradientBoostingRegressor	0.92	0.92	0.84	0.85	0.42	0.42	168.42
F_{25}	KNNRegressor	0.85	0.85	-0.18	-0.18	-0.38	-0.38	161.62
F_{25}	LGBMRegressor	0.95	0.95	0.81	0.81	0.42	0.42	3.50
F_{25}	LinearRegression	0.88	0.88	0.62	0.63	0.83	0.83	0.16
F_{25}	Ridge	0.88	0.88	0.60	0.60	0.84	0.84	0.13
F_{25}	XGBRegressor	0.95	0.95	0.71	0.71	0.40	0.41	28.17
F_{RFE}	DecisionTreeRegressor	0.86	0.86	0.16	0.17	0.28	0.28	4.24
F_{RFE}	GradientBoostingRegressor	0.92	0.92	0.85	0.85	0.42	0.42	95.89
F_{RFE}	KNNRegressor	0.85	0.85	-0.23	-0.23	-0.47	-0.46	158.34
F_{RFE}	LGBMRegressor	0.95	0.95	0.81	0.81	0.42	0.42	2.87
F_{RFE}	LinearRegression	0.87	0.87	0.67	0.68	0.82	0.82	0.13
F_{RFE}	Ridge	0.87	0.87	0.68	0.68	0.82	0.82	0.10
F_{RFE}	XGBRegressor	0.95	0.95	0.66	0.66	0.30	0.30	18.42

Table B.6: Sherman Island, R2 and Adjusted R2 results

Feature Set	Model	RMSE Night	RMSE Winter	RMSE Flood	Slope Night	Slope Winter	Slope Flood	Slope Day	Total Time (s)
F_E	DecisionTreeRegressor	0.42	1.07	1.62	0.92	1.25	0.65	0.53	2.43
F_E	GradientBoostingRegressor	0.34	0.36	1.35	0.89	0.96	0.61	0.41	50.16
F_E	KNNRegressor	0.46	0.82	2.67	0.84	0.41	0.09	0.05	8.59
F_E	LGBMRegressor	0.27	0.40	1.43	0.93	0.98	0.59	0.40	2.18
F_E	LinearRegression	0.46	0.47	0.97	0.83	1.09	0.72	0.47	0.09
F_E	Neural Network	0.24	0.54	0.56	0.88	0.87	0.43		
F_E	Ridge	0.46	0.47	0.98	0.83	1.09	0.72	0.47	0.09
F_E	XGBRegressor	0.27	0.59	1.53	0.94	0.98	0.65	0.44	12.81
F_{25}	DecisionTreeRegressor	0.41	0.79	1.89	0.93	0.99	0.43	0.52	7.39
F_{25}	GradientBoostingRegressor	0.31	0.34	1.74	0.90	0.92	0.41	0.45	168.42
F_{25}	KNNRegressor	0.43	0.94	2.69	0.86	0.38	0.16	0.25	161.62
F_{25}	LGBMRegressor	0.25	0.37	1.74	0.94	0.92	0.43	0.48	3.50
F_{25}	LinearRegression	0.38	0.53	0.94	0.88	1.15	0.71	0.61	0.16
F_{25}	Ridge	0.39	0.55	0.91	0.88	1.15	0.72	0.61	0.13
F_{25}	XGBRegressor	0.25	0.46	1.77	0.95	0.91	0.47	0.52	28.17
F_{RFE}	DecisionTreeRegressor	0.41	0.78	1.94	0.93	0.98	0.40	0.51	4.24
F_{RFE}	GradientBoostingRegressor	0.31	0.34	1.74	0.90	0.92	0.41	0.46	95.89
F_{RFE}	KNNRegressor	0.43	0.96	2.77	0.86	0.36	0.13	0.24	158.34
F_{RFE}	LGBMRegressor	0.25	0.38	1.74	0.94	0.94	0.44	0.47	2.87
F_{RFE}	LinearRegression	0.40	0.49	0.96	0.87	1.15	0.70	0.56	0.13
F_{RFE}	Ridge	0.40	0.49	0.96	0.87	1.15	0.70	0.56	0.10
F_{RFE}	XGBRegressor	0.25	0.50	1.91	0.95	0.90	0.44	0.49	18.42

Table B.7: Sherman Island, RMSE and Slope results

Feature Set	Model	Adj. R2 Night	R2 Night	Adj. R2 Winter	R2 Winter	RMSE Night	RMSE Winter	Slope Night	Slope Winter	Slope Day	Total Time (s)
F_E	DecisionTreeRegressor	0.82	0.82	-0.84	-0.84	0.14	1.04	0.91	0.23	0.02	7.71
F_E	GradientBoostingRegressor	0.86	0.86	-0.72	-0.72	0.12	1.01	0.83	0.20	0.04	145.07
F_E	KNNRegressor	0.83	0.83	-1.04	-1.04	0.14	1.10	0.84	0.16	-0.03	21.64
F_E	LGBMRegressor	0.91	0.91	-0.59	-0.59	0.10	0.97	0.89	0.23	0.04	3.54
F_E	LinearRegression	0.79	0.79	0.75	0.75	0.15	0.38	0.79	0.74	0.48	0.14
F_E	Neural Network	0.89	0.89	0.36	0.36	0.24	0.21	0.88	0.17		
F_E	Ridge	0.79	0.79	0.76	0.76	0.15	0.38	0.79	0.74	0.48	0.12
F_E	XGBRegressor	0.92	0.92	-0.66	-0.66	0.09	0.99	0.92	0.22	0.04	30.10
F_{25}	DecisionTreeRegressor	0.82	0.82	-0.89	-0.89	0.14	1.06	0.91	0.20	0.01	24.05
F_{25}	GradientBoostingRegressor	0.87	0.87	-0.74	-0.73	0.12	1.01	0.84	0.19	0.05	487.44
F_{25}	KNNRegressor	0.86	0.86	-1.32	-1.31	0.13	1.17	0.84	0.07	0.01	832.04
F_{25}	LGBMRegressor	0.92	0.92	-0.56	-0.55	0.09	0.96	0.90	0.23	0.06	6.42
F_{25}	LinearRegression	0.83	0.83	0.78	0.78	0.14	0.36	0.83	0.70	0.48	0.29
F_{25}	Ridge	0.83	0.83	0.78	0.78	0.14	0.36	0.83	0.70	0.48	0.20
F_{25}	XGBRegressor	0.93	0.93	-0.40	-0.40	0.09	0.91	0.93	0.25	0.06	74.57
F_{RFE}	DecisionTreeRegressor	0.83	0.83	-0.83	-0.82	0.14	1.04	0.92	0.23	0.03	13.62
F_{RFE}	GradientBoostingRegressor	0.87	0.87	-0.74	-0.73	0.12	1.01	0.83	0.19	0.05	279.67
F_{RFE}	KNNRegressor	0.86	0.86	-1.42	-1.42	0.13	1.20	0.84	0.08	0.02	818.09
F_{RFE}	LGBMRegressor	0.92	0.92	-0.57	-0.57	0.10	0.96	0.90	0.23	0.06	4.67
F_{RFE}	LinearRegression	0.81	0.81	0.77	0.78	0.15	0.36	0.81	0.70	0.47	0.19
F_{RFE}	Ridge	0.81	0.81	0.78	0.78	0.15	0.36	0.81	0.70	0.47	0.15
F_{RFE}	XGBRegressor	0.93	0.93	-0.52	-0.52	0.09	0.95	0.93	0.24	0.06	49.50

Table B.8: West Pond results

Appendix C

Boundary Layer Height Supplementary Material

C.0.1 Background on Machine Learning Models

Table C.1: Comparison of Regression Models

Model Name	**Type**	**Strengths**	**Weaknesses**
MLP Regressor	Neural Network	Captures complex patterns Handles large datasets	Requires large data Sensitive to feature scaling
LGBM Regressor	Ensemble	Efficient and fast Handles large datasets Supports categorical features	Can overfit Requires careful tuning
XGB Regressor	Ensemble	High performance Customizable Regularization to prevent overfitting	Computationally intensive Requires careful tuning
Gradient Boosting Regressor	Ensemble	Reduces bias and variance High accuracy	Can overfit Slow training time
Random Forest Regressor	Ensemble	Reduces overfitting Robust to outliers Provides feature importance	Can be slow Less interpretable
Decision Tree Regressor	Trees	Easy to interpret Handles both numerical and categorical data	Prone to overfitting Unstable with small changes in data
Linear Regression	Linear Model	Simple and fast Easy to interpret No hyperparameters to tune	Assumes linearity Sensitive to outliers

C.0.2 Glossary of Feature Labels

Table C.2: Description of feature labels used at the ATTO site

Label	**Explanation**
blh	Boundary layer height in meters
Day	Day of the month
Flux_Op-CO2	Flux of carbon dioxide
Flux_Op-H2O	Flux of water vapor (humidity)
Flux_Tau	Corrected momentum flux
Flux_Tsonic	Flux of temperature
Hour	Hour of the day
Hum_100cm	Soil humidity at 100 cm depth
Hum_10cm	Soil humidity at 10 cm depth
Hum_20cm	Soil humidity at 20 cm depth
Hum_30cm	Soil humidity at 30 cm depth
Hum_40cm	Soil humidity at 40 cm depth
Hum_60cm	Soil humidity at 60 cm depth
LW_atm	Incoming longwave radiation from the atmosphere
LW_terr	Outgoing longwave radiation from the surface
Mean_APress	Mean atmospheric pressure
Mean_Op-CO2	Mean carbon dioxide concentration
Mean_Op-H2O	Mean water vapor (humidity) concentration
Mean_Tsonic	Mean temperature measured by sonic anemometer
Mean_Windsp	Mean wind speed
Month	Month of the year
NetRad	Net radiation (balance of all incoming and outgoing radiation)
PAR_in	Incoming Photosynthetically Active Radiation
PAR_out	Outgoing Photosynthetically Active Radiation
Press_81m	Air pressure at 81 meters
Rainfall	Accumulated rainfall
RH_1.5m	Relative humidity at 1.5 meters
RH_26m	Relative humidity at 26 meters
RH_36m	Relative humidity at 36 meters
RH_40m	Relative humidity at 40 meters
RH_55m	Relative humidity at 55 meters
RH_73m	Relative humidity at 73 meters
RH_81m	Relative humidity at 81 meters
SFlux_5cm	Soil temperature flux at 5 cm depth

SW_in	Incoming shortwave radiation
SW_out	Outgoing shortwave radiation
T_1.5m	Temperature at 1.5 meters
T_12m	Temperature at 12 meters
T_26m	Temperature at 26 meters
T_36m	Temperature at 36 meters
T_4m	Temperature at 4 meters
T_40m	Temperature at 40 meters
T_55m	Temperature at 55 meters
T_73m	Temperature at 73 meters
T_81m	Temperature at 81 meters
Tsoil_10cm	Soil temperature at 10 cm depth
Tsoil_20cm	Soil temperature at 20 cm depth
Tsoil_40cm	Soil temperature at 40 cm depth
U-star	Friction velocity
UV	Ultraviolet radiation
Wind-Direc-corrg	Corrected wind direction
WSp_19m	Wind speed at 19 meters
WSp_26m	Wind speed at 26 meters
WSp_50m	Wind speed at 50 meters
WSp_65m	Wind speed at 65 meters
WSp_73m	Wind speed at 73 meters
Year	Year of measurement
Z-over-L	Monin-Obukhov stability parameter

Table C.3: Description of feature labels used at the T3 site

Label	**Explanation**
albedo	Surface reflectance of solar radiation
corr_soil_heat_flow_1	Corrected soil heat flow (1)
corr_soil_heat_flow_2	Corrected soil heat flow (2)
corr_soil_heat_flow_3	Corrected soil heat flow (3)
Day	Day of the month
down_long	Downwelling longwave radiation
down_short_hemisp	Downwelling shortwave radiation
energy_storage_change_1	Change in energy storage (1)
energy_storage_change_2	Change in energy storage (2)
energy_storage_change_3	Change in energy storage (3)
Hour	Hour of the day
Month	Month of the year

net_radiation	Net radiation
precip_rate_sfc	Surface precipitation rate
pressure_sfc	Surface air pressure
relative_humidity_sfc	Surface relative humidity
sensible_heat_flux	Sensible heat flux
soil_heat_capacity_1	Soil heat capacity (1)
soil_heat_capacity_2	Soil heat capacity (2)
soil_heat_capacity_3	Soil heat capacity (3)
soil_moisture_1	Soil moisture (1)
soil_moisture_2	Soil moisture (2)
soil_moisture_3	Soil moisture (3)
soil_temp_1	Soil temperature (1)
soil_temp_2	Soil temperature (2)
soil_temp_3	Soil temperature (3)
surface_energy_balance	Surface energy balance
surface_soil_heat_flux_1	Surface soil heat flux (1)
surface_soil_heat_flux_2	Surface soil heat flux (2)
surface_soil_heat_flux_3	Surface soil heat flux (3)
surface_soil_heat_flux_avg	Average surface soil heat flux
temp_net_radiometer	Temperature from net radiometer
temperature_sfc	Surface air temperature
up_long	Upwelling longwave radiation
up_short_hemisp	Upwelling shortwave radiation
u_wind_sfc	Surface u-component of wind (east-west)
v_wind_sfc	Surface v-component of wind (north-south)
wetness	Wetness of the surface
Year	Year of measurement
zenith	Zenith angle of the sun

C.0.3 Figures

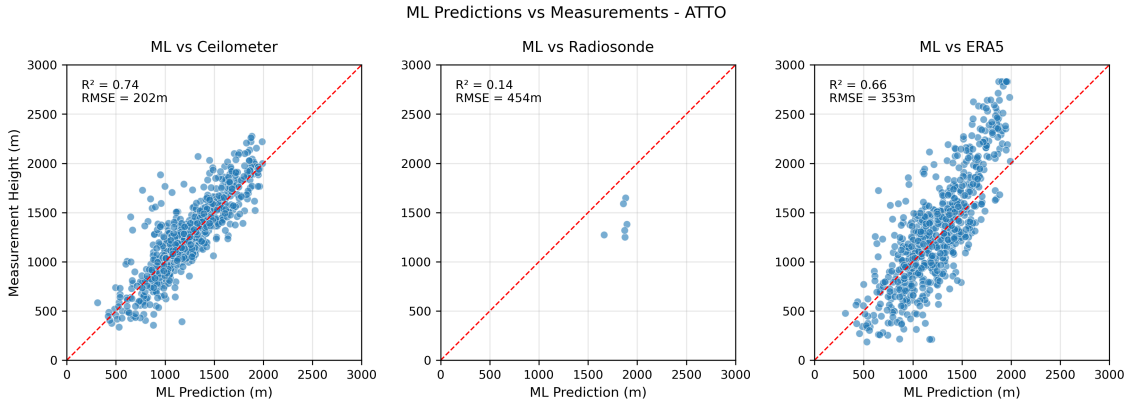


Figure C.1: Scatter plots of ML predictions against ceilometer (training data) and radiosonde measurements as well as ERA-5 predictions. R^2 and RMSE metrics are included to measure goodness of fit. Predictions are taken from all 10 CV folds used in model training to ensure a good distribution across the test set.

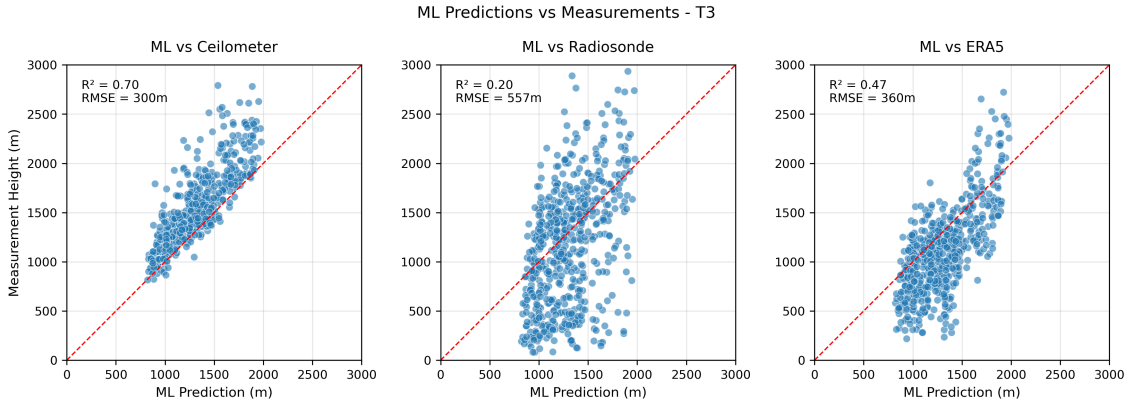


Figure C.2: Scatter plots of ML predictions against Ceilometer (training data) and radiosonde measurements as well as ERA-5 predictions. R^2 and RMSE metrics are included to measure goodness of fit. Predictions are taken from all 10 CV folds used in model training to ensure a good distribution across the test set.

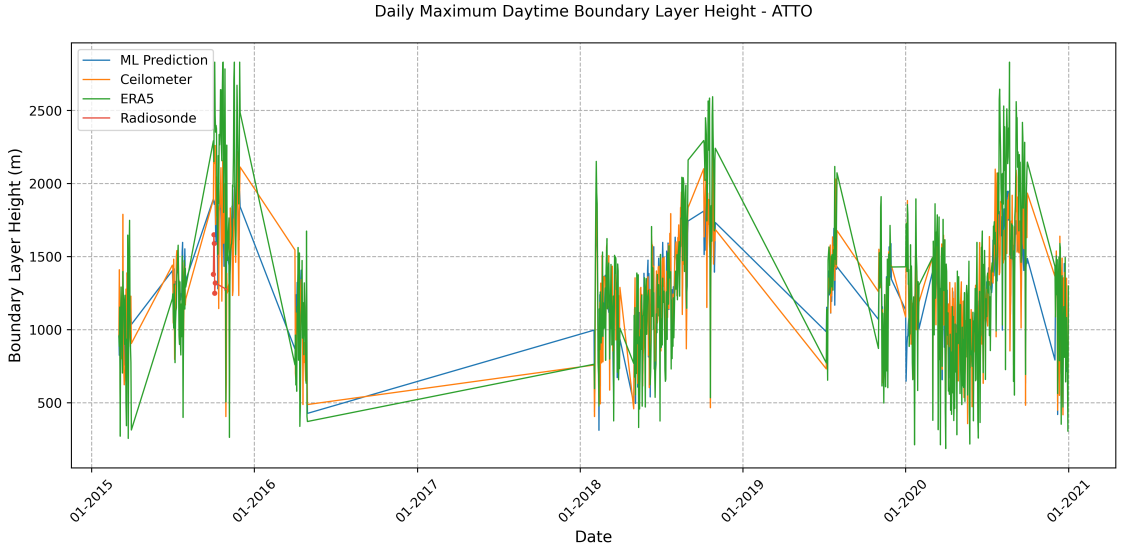


Figure C.3: Daily maximum daytime boundary layer height at the ATTO site. The plot compares measurements from ceilometer (orange), machine learning predictions (blue), ERA5 reanalysis (green), and radiosonde observations (red points). The machine learning model shows strong agreement with ceilometer measurements, while ERA5 tends to underestimate the maximum daily height.

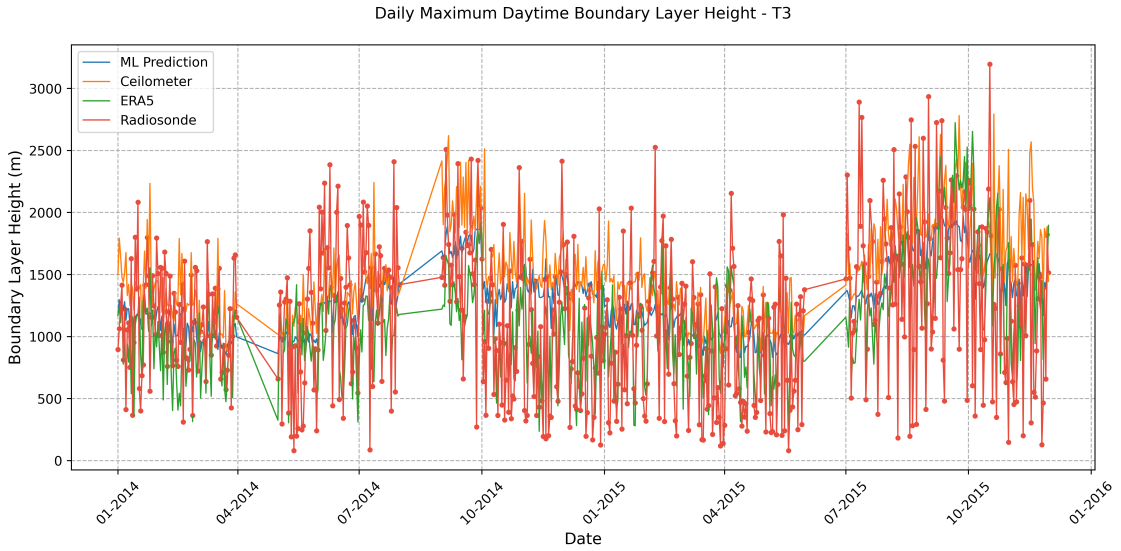


Figure C.4: Daily maximum daytime boundary layer height at the T3 site. Comparison between different measurement methods shows closer agreement between all methods compared to the ATTO site, potentially due to the site's proximity to urban areas and resulting stronger aerosol signals for the ceilometer measurements.

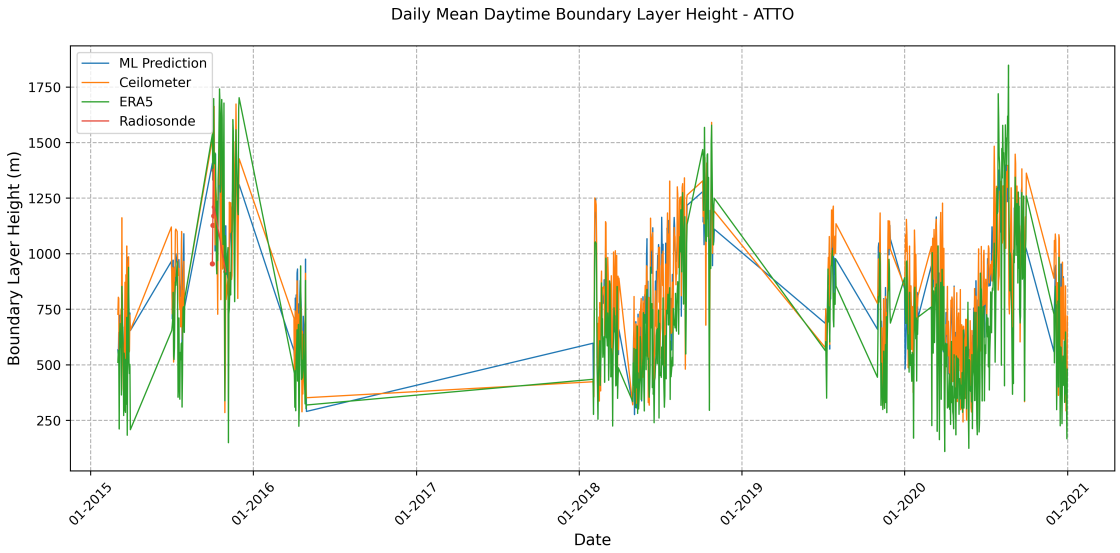


Figure C.5: Daily mean daytime boundary layer height at the ATTO site. The mean values show less variability than the maximum values, with consistent patterns in the diurnal cycle. The machine learning predictions closely track the ceilometer measurements, while ERA5 shows systematic differences in the estimation of mean boundary layer height.

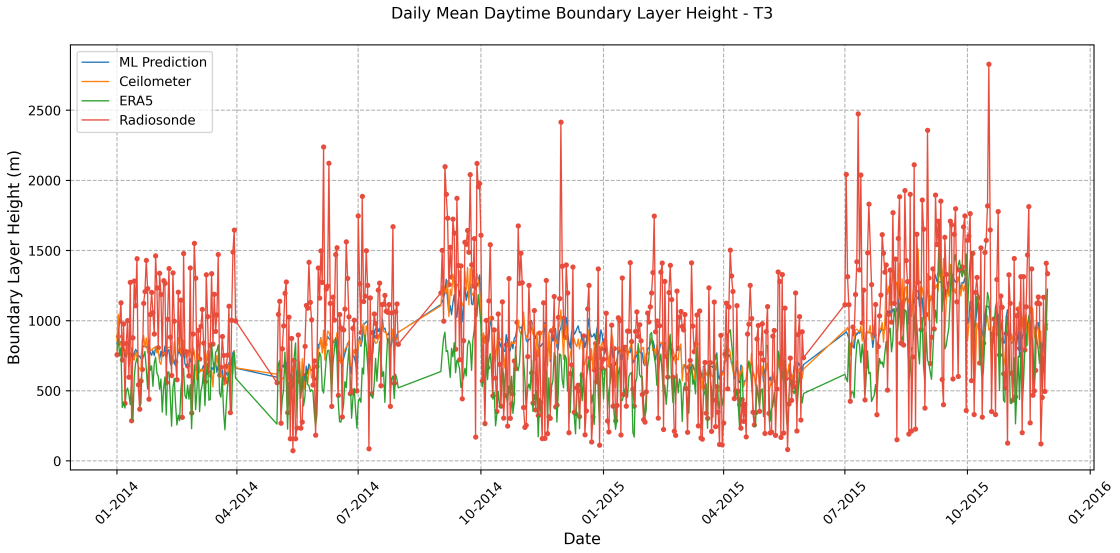


Figure C.6: Daily mean daytime boundary layer height at the T3 site. The mean values demonstrate the typical boundary layer evolution patterns in the Amazon region, with the machine learning model capturing the seasonal variations observed in the ceilometer data. Radiosonde measurements provide additional validation points showing good agreement with both ceilometer and model predictions.

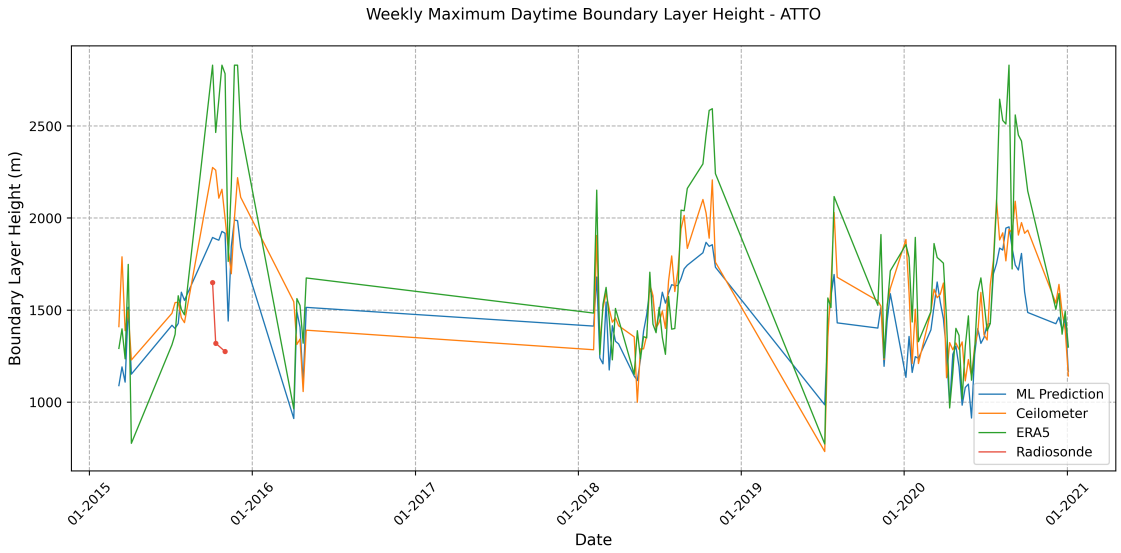


Figure C.7: Weekly maximum daytime boundary layer height at the ATTO site. The weekly aggregation smooths out daily fluctuations, revealing longer-term patterns in boundary layer development. The machine learning predictions maintain good agreement with ceilometer measurements at this temporal scale, while ERA5 consistently shows lower maximum heights.

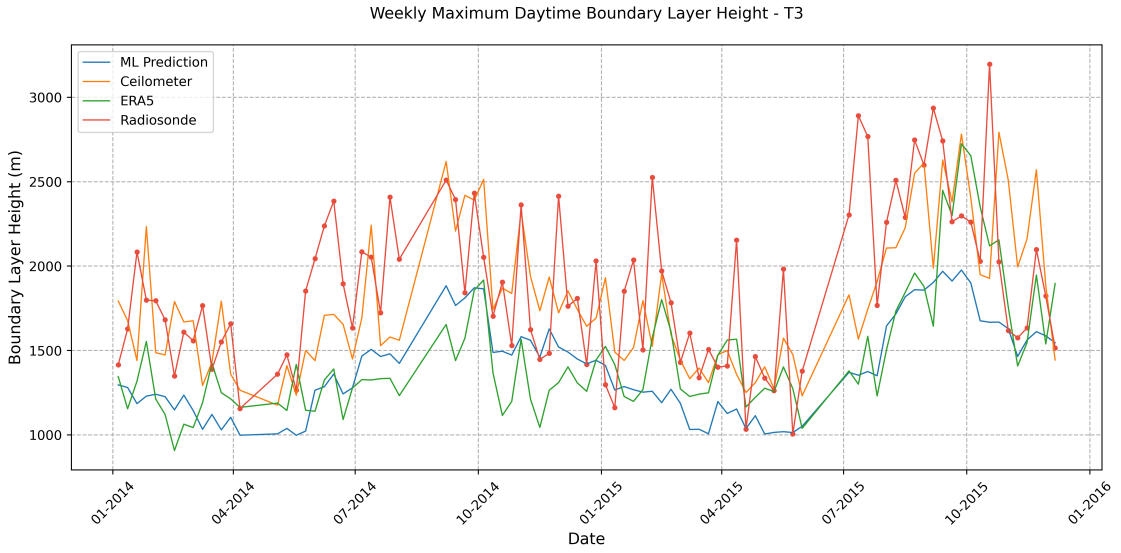


Figure C.8: Weekly maximum daytime boundary layer height at the T3 site. The weekly maximum values highlight seasonal patterns in boundary layer development, with the machine learning model successfully capturing these longer-term variations. The agreement between different measurement methods suggests robust characterization of maximum boundary layer heights at this temporal scale.

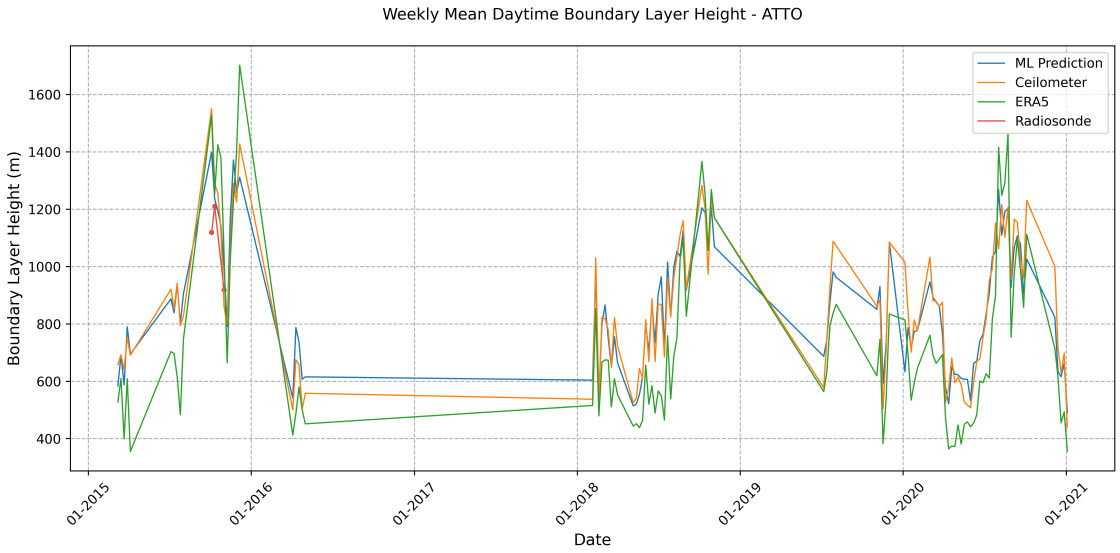


Figure C.9: Weekly mean daytime boundary layer height at the ATTO site. The weekly averaging reveals seasonal patterns in boundary layer development while maintaining the distinction between different measurement methods. The machine learning predictions demonstrate consistent tracking of ceilometer measurements across the entire observation period.

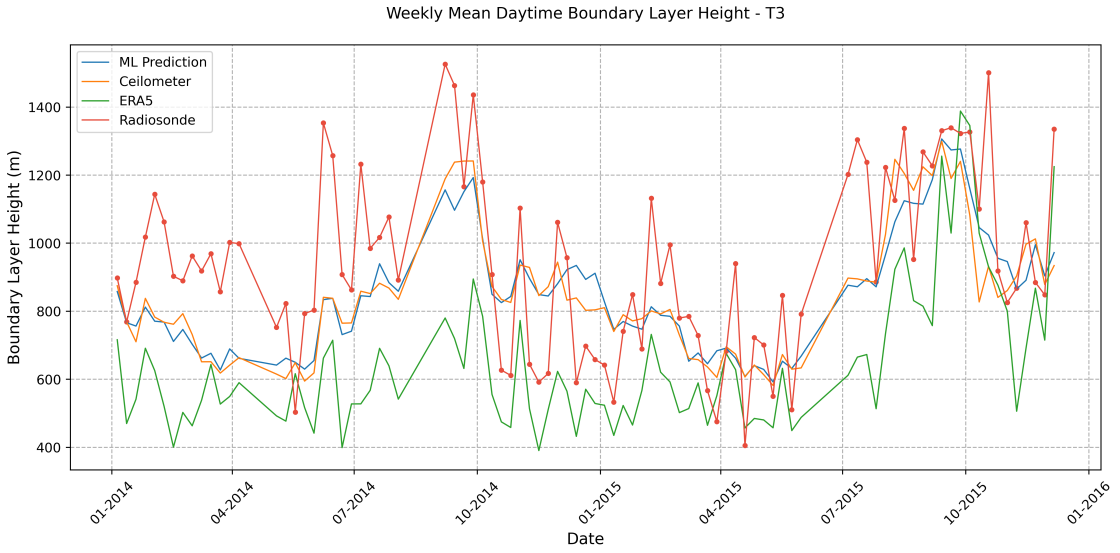


Figure C.10: Weekly mean daytime boundary layer height at the T3 site. The weekly averages show clear patterns in boundary layer evolution, with good agreement between machine learning predictions and ceilometer measurements. This temporal scale effectively captures seasonal variations while smoothing out daily fluctuations, providing insights into longer-term boundary layer dynamics at the site.

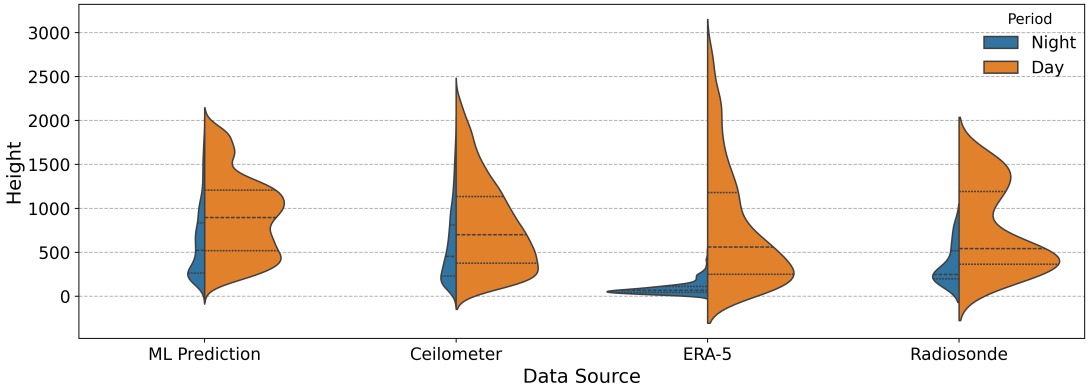


Figure C.11: Estimated probability density distributions for predictions and measurements at the ATTO site for the out-of-sample test periods, including periods where no radiosonde data were available. Predictions are taken from a LightGBM model trained on all available input features. The imbalance noted between the width of the Day and Night are due to imbalanced numbers of samples.

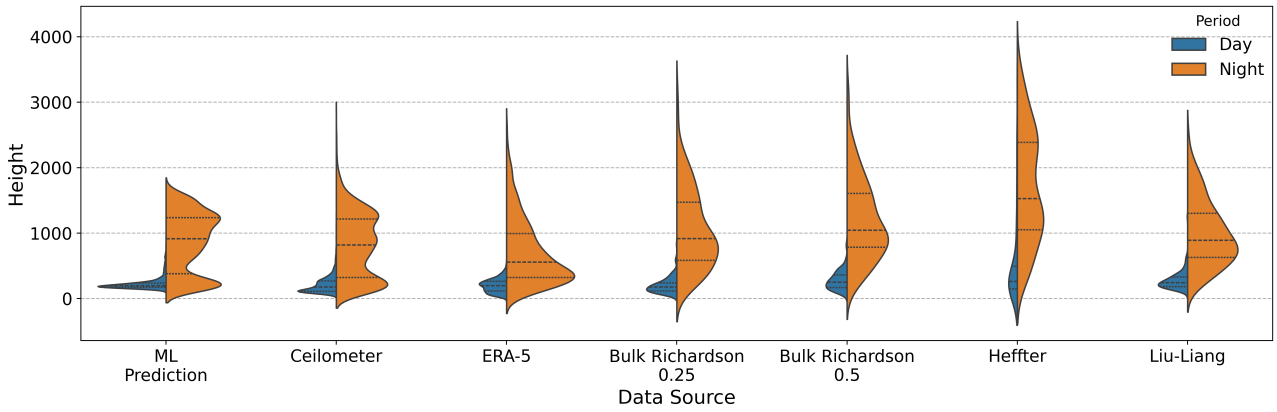


Figure C.12: Estimated probability density distributions for predictions and measurements at the ATTO site for the out-of-sample test periods, including periods where no radiosonde data were available. Predictions are taken from a LightGBM model trained on all available input features. The imbalance noted between the width of the Day and Night are due to imbalanced numbers of samples.

Appendix D

Gross Primary Productivity Supplementary Material

Table D.1: Detailed Variables for Amazon GPP Prediction Model

Label	**Source**	**Detailed Description**	**Units**	**Assigned Label**	**Original Label**
Latitude	ERA5	Geographical latitude coordinate. Critical for understanding latitudinal gradients in solar radiation, temperature, and precipitation patterns across the Amazon basin.	deg N	Latitude	lat
Longitude	ERA5	Geographical longitude coordinate. Important for capturing east-west precipitation gradients and continental effects across the Amazon.	deg E	Longitude	lon
Year	ERA5	Year of observation. Temporal dimension capturing interannual variability, climate cycles (El Niño/La Niña), and long-term trends in GPP.	-	Year	Year
Month	ERA5	Month of observation. Seasonal component critical for capturing wet/dry season dynamics that strongly control Amazon GPP patterns.	-	Month	Month

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Air Temp	ERA5	2-meter air temperature. Fundamental driver of photosynthesis rates, plant respiration, and enzyme activity. Controls temperature stress on vegetation.	°C	Air Temp	t2m
Dewpoint	ERA5	2-meter dewpoint temperature. Temperature at which air becomes saturated. Used to calculate relative humidity and atmospheric moisture content, affecting plant water stress.	°C	Dewpoint	d2m
Surface Pressure	ERA5	Surface atmospheric pressure. Affects gas exchange rates in plant stomata and atmospheric density for radiation calculations.	Pa	Surface Pressure	sp
U-Wind	ERA5	10-meter zonal wind speed. East-west wind component affecting boundary layer mixing, heat/-moisture transport, and mechanical stress on vegetation.	m/s	U-Wind	u10
V-Wind	ERA5	10-meter meridional wind speed. North-south wind component. Combined with u-wind affects turbulent transport and regional climate patterns.	m/s	V-Wind	v10
Evaporation	ERA5	Total evaporation. Cumulative water loss to atmosphere. Key component of water cycle and energy balance. Should be exactly related to Latent Heat by conversion factor.	m	Evaporation	e
Evap Bare Soil	ERA5	Evaporation from bare soil. Direct soil evaporation component. Important in areas with low vegetation cover or during dry seasons.	m	Evap Bare Soil	evabs

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Evap Water	ERA5	Evaporation from open water. Evaporation from rivers, lakes, and flooded areas. Significant in wetland regions of the Amazon.	m	Evap Water	evaow
Evap Canopy	ERA5	Evaporation from canopy interception. Water loss from precipitation intercepted by leaves and branches. Important component in tropical forests.	m	Evap Canopy	evatc
Evap Vegetation	ERA5	Transpiration from vegetation. Water loss through plant stomata during photosynthesis. Directly linked to GPP through water-use efficiency.	m	Evap Vegetation	evavt
Pot Evaporation	ERA5	Potential evaporation. Theoretical maximum evaporation under current atmospheric conditions. Indicates atmospheric demand for water.	m	Pot Evaporation	pev
Total Precip	MERGE	Total precipitation. Cumulative rainfall providing water supply for photosynthesis and plant growth. Primary limiting factor in many tropical ecosystems.	mm	Total Precip	precip_total
Max Precip	MERGE	Maximum precipitation. Peak precipitation intensity, indicating extreme weather events that can affect plant physiology and soil processes.	mm	Max Precip	precip_max
Min Precip	MERGE	Minimum precipitation. Indicates dry period intensity, important for understanding drought stress on vegetation.	mm	Min Precip	precip_min

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Solar Radiation	ERA5	Net solar radiation. Net shortwave radiation absorbed by surface. Primary energy source for photosynthesis and the key driver of GPP.	J/m²	Solar Radiation	ssr
Solar Downward	ERA5	Downward solar radiation. Incoming shortwave radiation from sun. Raw energy available for photosynthesis before surface reflection.	J/m²	Solar Downward	ssrd
Thermal Radiation	ERA5	Net thermal radiation. Net longwave radiation exchange. Important for energy balance and temperature regulation of vegetation.	J/m²	Thermal Radiation	str
Thermal Downward	ERA5	Downward thermal radiation. Incoming longwave radiation from atmosphere. Contributes to surface warming, especially at night.	J/m²	Thermal Downward	strd
Latent Heat	ERA5	Surface latent heat flux. Energy used for evapotranspiration. Should be exactly: slhf = e × 2.45×10 J/m²/m with perfect correlation expected.	J/m²	Latent Heat	slhf
Sensible Heat	ERA5	Surface sensible heat flux. Energy transfer as heat between surface and atmosphere. Affects air temperature and atmospheric boundary layer.	J/m²	Sensible Heat	sshf
Albedo	ERA5	Surface albedo. Fraction of solar radiation reflected by surface. Vegetation type and health strongly influence albedo values.	-	Albedo	fal

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Soil Temp L1	ERA5	Soil temperature layer 1 (0-7cm). Near-surface soil temperature affecting root respiration, nutrient mineralization, and microbial activity.	°C	Soil Temp L1	stl1
Soil Temp L4	ERA5	Soil temperature layer 4 (100-289cm). Deep soil temperature indicating thermal stability and long-term temperature trends affecting deep root systems.	°C	Soil Temp L4	stl4
Soil Water L1	ERA5	Soil moisture layer 1 (0-7cm). Surface soil water content directly affecting plant water uptake and surface evaporation.	m^3/m^3	Soil Water L1	swvl1
Soil Water L3	ERA5	Soil moisture layer 3 (28-100cm). Root zone soil moisture. Critical for tree water uptake and drought stress assessment.	m^3/m^3	Soil Water L3	swvl3
Soil Type	ERA5	Dominant soil type classification. Categorical soil classification affecting water retention, nutrient availability, and root development.	-	Soil Type	slt
Available P	RF Model	Plant-available phosphorus. Often the primary limiting nutrient in tropical soils, directly affecting GPP through photosynthetic capacity constraints.	mg/kg	Available P	avail_p
Organic P	RF Model	Organic phosphorus. Phosphorus bound in organic matter. Reservoir of P that becomes available through mineralization processes.	mg/kg	Organic P	org_p

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Total P	RF Model	Total phosphorus. Sum of all phosphorus forms in soil. Represents total P reservoir available to ecosystem over time.	mg/kg	Total P	total_p
Runoff	ERA5	Total runoff. Combined surface and subsurface runoff. Indicates water loss from ecosystem and flooding potential affecting root zone conditions.	m	Runoff	ro
Surface Runoff	ERA5	Surface runoff. Water flowing over surface when precipitation exceeds infiltration capacity. Fast hydrological response to precipitation events.	m	Surface Runoff	sro
Subsurface Runoff	ERA5	Subsurface runoff. Lateral water flow through soil layers indicating soil drainage and aeration conditions critical for root function.	m	Subsurface Runoff	ssro
LAI High Veg	ERA5	Leaf Area Index for high vegetation. Total leaf area per ground area for trees and tall vegetation. Directly related to photosynthetic capacity and GPP.	m^2/m^2	LAI High Veg	lai_hv
LAI Low Veg	ERA5	Leaf Area Index for low vegetation. Total leaf area per ground area for grasses and low shrubs. Important in savanna and disturbed areas.	m^2/m^2	LAI Low Veg	lai_lv
High Veg Cover	ERA5	High vegetation cover fraction. Fraction of grid cell covered by trees and tall vegetation. Key structural parameter affecting ecosystem processes.	-	High Veg Cover	cvh

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Low Veg Cover	ERA5	Low vegetation cover fraction. Fraction of grid cell covered by low vegetation. Important for understanding land use and disturbance effects.	-	Low Veg Cover	cvl
High Veg Type	ERA5	High vegetation type classification. Categorical classification of dominant tree/tall vegetation types affecting ecosystem functioning.	-	High Veg Type	tvh
Low Veg Type	ERA5	Low vegetation type classification. Classification of dominant low vegetation types (grasses, crops, shrubs) in the grid cell.	-	Low Veg Type	tvl
Wood Density	Mo et al.	Community wood density. Average wood density of tree community. Relates to carbon storage, growth rates, and mortality patterns affecting GPP.	g/cm^3	Wood Density	wood_density
Density	Steege	Tree density. Number of trees per hectare. Structural metric affecting competition, light availability, and ecosystem productivity.	stems/ha	Density	Density
Diversity	Steege	Species diversity index. Shannon or Simpson diversity index. Higher diversity may increase ecosystem stability and productivity through complementarity.	-	Diversity	Diversity
Richness	Steege	Species richness. Number of tree species per hectare. Biodiversity metric potentially affecting ecosystem functioning and GPP through niche complementarity.	species/ha	Richness	Richness-ha

Continued on next page

Table D.1 – continued from previous page

Label	Source	Detailed Description	Units	Assigned Label	Original Label
Edge Effect	Lapola	Distance to forest edge. Proximity to forest-nonforest boundaries. Edge effects alter microclimate, species composition, and carbon dynamics.	km	Edge Effect	edge
Fire	Lapola	Fire occurrence/intensity. Fire history or risk index. Fire dramatically reduces GPP and alters forest structure and composition.	-	Fire	fire
Drought	Lapola	Drought stress index. Measure of water stress conditions. Drought reduces photosynthesis and can cause tree mortality affecting long-term GPP.	-	Drought	drought
Logging	Lapola	Logging intensity/history. Selective logging pressure. Reduces forest biomass and alters canopy structure affecting light availability and GPP.	-	Logging	logging
Topography	Topo	Terrain characteristics. Elevation, slope, aspect affecting drainage, microclimate, and soil development. Influences local GPP patterns through multiple pathways.	-	Topography	TOPO
Soil Database	Steege	Soil and terrain database identifier. Red Amazónica de Información Socioambiental Georreferenciada classification for soil properties.	-	Soil Database	SoterRaisg
GPP	Target	Gross Primary Productivity. Total carbon fixation by photosynthesis in the ecosystem. The target variable representing ecosystem photosynthetic capacity.	$gC/m^2/day$	GPP	GPP

Appendix E

Declaration of Authorship

Declaration of Authorship

Candidates are required to submit a separate **Declaration of Authorship** form for each co-authored paper submitted for examination as part of a PhD by Publication thesis. Further information is available from the accompanying guideline document.[1]

Section 1: Candidate's details	
Candidate's Name	Adam Stapleton
DCU Student Number	20214892
School	Computing
Principal Supervisor	Mark Roantree
Title of PhD by Publication Thesis	Explainable Machine Learning for Knowledge Discovery in Environmental Science
Section 2: Paper details	
Title of co-authored paper included in the thesis under examination	A comparative analysis of machine learning approaches to gap filling meteorological datasets
Publication Status	Published in the journal *Environmental Earth Sciences*, Springer-Verlag GmbH Germany.
ISSN and link to URL (where available)	1866-6280 (print) 1866-6299 (online) https://doi.org/10.1007/s12665-024-11982-8
This paper is one of 4 co-authored papers to be submitted as part of the PhD by publication thesis submitted for examination	
Section 3: Candidate's contribution to the paper	
Provide details below of the **nature** and **extent of your contribution** to the paper (include both your intellectual and practical contributions) and your overall contribution in **percentage terms**:	
Developed the novel gap creation methodology and experimental design (including feature set comparisons), implemented all machine learning models (Random Forest, LightGBM, Linear Regression) and reference spatial interpolation, conducted data pre-processing and primary data analysis including feature importance assessment and drafted the manuscript including introduction, related research, methodology, results and initial discussion sections. Contribution: 60%	

[1] 'Guidelines for candidates, supervisors and examiners on the 'PhD by Publication' format': https://www.dcu.ie/graduatestudies/A_Z-of-GSO-Policies.shtml

Where a paper has joint or multiple authors, list the names of all other authors who contributed to the work (this can be appended in a separate document, where necessary)**:**

Branislava Lalic, Thomas Vergauwe, Steven Caluwaerts, Elke Eichelmann, Mark Roantree

Section 4: Signature and Validation

I confirm that the following statements are true:

(a) the information I have provided in this form is correct

(b) this paper is based on research undertaken during my candidature at DCU

Signature of PhD Candidate: Adam Stapleton Date: 10/10/25

I confirm that the information provided by the candidate is correct:

Signature of Principal Supervisor: Mark Roantree Date: 10th October 2025

In some cases, it may be appropriate for verification to be given by both the principal supervisor **and** the lead/corresponding author of the work (where the lead/corresponding author of the work is not the candidate or the principal supervisor):

Signature of Lead/Corresponding Author: Branislava Lalic **Date** 14th October 2025

Nature of Current Post/Responsibilities Professor of Meteorology and Biophysics

Home institution Faculty of Agriculture, University of Novi Sad, Novi Sad, Serbia

Declaration of Authorship

Candidates are required to submit a separate **Declaration of Authorship** form for each co-authored paper submitted for examination as part of a PhD by Publication thesis. Further information is available from the accompanying guideline document.[1]

Section 1: Candidate's details	
Candidate's Name	Adam Stapleton
DCU Student Number	20214892
School	Computing
Principal Supervisor	Mark Roantree
Title of PhD by Publication Thesis	Explainable Machine Learning for Scientific Discovery in Environmental Science
Section 2: Paper details	
Title of co-authored paper included in the thesis under examination	A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning.
Publication Status	Published in Applied Computing and Geosciences, Elsevier.
ISSN and link to URL (where available)	ISSN: 2590-1974 DOI: 10.1016/j.acags.2022.100094
This paper is one of 4 co-authored papers to be submitted as part of the PhD by publication thesis submitted for examination	
Section 3: Candidate's contribution to the paper	
Provide details below of the **nature** and **extent of your contribution** to the paper (include both your intellectual and practical contributions) and your overall contribution in **percentage terms**:	
Designed the novel machine learning framework for ET partitioning and feature selection methodology, collected and preprocessed all data from AmeriFlux sites, implemented the recursive feature elimination algorithm with feature importance heuristics, developed and compared all ML models (linear, ridge, KNN, decision tree, gradient boosting, LightGBM, XGBoost), conducted all experimental analysis including discovery of methane flux relationships, and drafted the complete manuscript. Contribution: 80%	
Where a paper has joint or multiple authors, list the names of all other authors who contributed to the work (this can be appended in a separate document, where necessary)**:** Eichelmann, E., & Roantree, M.	

[1] 'Guidelines for candidates, supervisors and examiners on the 'PhD by Publication' format': https://www.dcu.ie/graduatestudies/A_Z-of-GSO-Policies.shtml

Section 4: Signature and Validation

I confirm that the following statements are true:

(a) the information I have provided in this form is correct

(b) this paper is based on research undertaken during my candidature at DCU

Signature of PhD Candidate: Adam Stapleton **Date:** 10/10/25

I confirm that the information provided by the candidate is correct:

Signature of Principal Supervisor: Mark Roantree **Date:** 10th October 2025

In some cases, it may be appropriate for verification to be given by both the principal supervisor **and** the lead/corresponding author of the work (where the lead/corresponding author of the work is not the candidate or the principal supervisor):

Signature of Lead/Corresponding Author: Adam Stapleton **Date**: 10/10/25

Nature of Current Post/Responsibilities: Ph.D. Researcher

Home institution: Dublin City University

Declaration of Authorship

Candidates are required to submit a separate **Declaration of Authorship** form for each co-authored paper submitted for examination as part of a PhD by Publication thesis. Further information is available from the accompanying guideline document.[1]

Section 1: Candidate's details	
Candidate's Name	Adam Stapleton
DCU Student Number	20214892
School	Computing
Principal Supervisor	Mark Roantree
Title of PhD by Publication Thesis	Explainable Machine Learning for Scientific Discovery in Environmental Science
Section 2: Paper details	
Title of co-authored paper included in the thesis under examination	Intercomparison of machine learning models to determine the planetary boundary layer height over Central Amazonia.
Publication Status	Published in Journal of Geophysical Research: Atmospheres, American Geophysical Union.
ISSN and link to URL (where available)	ISSN: 2169-8996 URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JD042488 DOI: 10.1029/2024JD042488
This paper is one of 4 co-authored papers to be submitted as part of the PhD by publication thesis submitted for examination	
Section 3: Candidate's contribution to the paper	
Provide details below of the **nature** and **extent of your contribution** to the paper (include both your intellectual and practical contributions) and your overall contribution in **percentage terms**:	
Developed the machine learning framework and experimental design for comparing multiple ML algorithms, implemented all data preprocessing and harmonization procedures, trained and evaluated all models (linear regression, ridge regression, KNN, decision tree, gradient boosting, LightGBM, XGBoost, neural networks), developed and executed the recursive feature elimination methodology, conducted all statistical analysis and model validation, created all figures and tables, and drafted the manuscript including all sections. Contribution: 75%	
Where a paper has joint or multiple authors, list the names of all other authors who contributed to the work (this can be appended in a separate document, where necessary)**:**	

[1] 'Guidelines for candidates, supervisors and examiners on the 'PhD by Publication' format': https://www.dcu.ie/graduatestudies/A_Z-of-GSO-Policies.shtml

Dias-Junior, C. Q., Von Randow, C., D'Oliveira, F. A. F., Pöhlker, C., de Araújo, A. C., Roantree, M., & Eichelmann, E. (2025)

Section 4: Signature and Validation

I confirm that the following statements are true:

(a) the information I have provided in this form is correct

(b) this paper is based on research undertaken during my candidature at DCU

Signature of PhD Candidate: Adam Stapleton **Date:** 10/10/25

I confirm that the information provided by the candidate is correct:

Signature of Principal Supervisor: Mark Roantree Date: 10th October 2025

In some cases, it may be appropriate for verification to be given by both the principal supervisor **and** the lead/corresponding author of the work (where the lead/corresponding author of the work is not the candidate or the principal supervisor):

Signature of Lead/Corresponding Author: Adam Stapleton **Date**: 10/10/25

Nature of Current Post/Responsibilities: Ph.D. Researcher

Home institution: Dublin City University

Declaration of Authorship

Candidates are required to submit a separate **Declaration of Authorship** form for each co-authored paper submitted for examination as part of a PhD by Publication thesis. Further information is available from the accompanying guideline document.[1]

Section 1: Candidate's details	
Candidate's Name	Adam Stapleton
DCU Student Number	20214892
School	Computing
Principal Supervisor	Mark Roantree
Title of PhD by Publication Thesis	Explainable Machine Learning for Scientific Discovery in Environmental Science
Section 2: Paper details	
Title of co-authored paper included in the thesis under examination	Discovering Regional Vulnerability Patterns of Gross Primary Productivity in Amazon Rainforests with XAI
Publication Status	In preparation for submission to Journal of Geophysical Research: Biogeosciences.
ISSN and link to URL (where available)	
This paper is one of 4 co-authored papers to be submitted as part of the PhD by publication thesis submitted for examination	
Section 3: Candidate's contribution to the paper	
Provide details below of the **nature** and **extent of your contribution** to the paper (include both your intellectual and practical contributions) and your overall contribution in **percentage terms**:	
Conceptualized the combined k-means clustering and SHAP explanation approach, harmonized all multi-source datasets to common resolution, implemented the clustering analysis and optimal cluster selection, developed and trained XGBoost models for each region, conducted all SHAP analysis for identifying regional drivers, created all visualizations and figures, and drafted the complete manuscript including interpretation of results. Contribution: 70%	
Where a paper has joint or multiple authors, list the names of all other authors who contributed to the work (this can be appended in a separate document, where necessary)**:** Aline Anderson de Castro, Celso Von Randow, Mark Roantree, and Elke Eichelmann (2026)	

[1] 'Guidelines for candidates, supervisors and examiners on the 'PhD by Publication' format': https://www.dcu.ie/graduatestudies/A_Z-of-GSO-Policies.shtml

Section 4: Signature and Validation

I confirm that the following statements are true:

(a) the information I have provided in this form is correct

(b) this paper is based on research undertaken during my candidature at DCU

Signature of PhD Candidate: Adam Stapleton **Date:** 10/10/25

I confirm that the information provided by the candidate is correct:

Signature of Principal Supervisor: Mark Roantree **Date:** 10th October 2025

In some cases, it may be appropriate for verification to be given by both the principal supervisor **and** the lead/corresponding author of the work (where the lead/corresponding author of the work is not the candidate or the principal supervisor):

Signature of Lead/Corresponding Author: Adam Stapleton **Date**: 10/10/25

Nature of Current Post/Responsibilities: Ph.D. Researcher

Home institution: Dublin City University

Bibliography

Ackerly, D. D., Thomas, W. W., Ferreira, C. C., and Pirani, J. R. (1989). The forest-cerrado transition zone in southern amazonia: results of the 1985 projeto flora amazônica expedition to mato grosso. Brittonia, 41(2):113–128.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access, 6:52138–52160.

Alduchov, O. A. and Eskridge, R. E. (1996). Improved magnus form approximation of saturation vapor pressure. Journal of Applied Meteorology (1988-2005), pages 601–609.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. Information fusion, 99:101805.

AmeriFlux Management Project (2024). Ameriflux data sharing platform. Funding for the AmeriFlux data portal was provided by the U.S. Department of Energy Office of Science.

Andela, N., Morton, D. C., Schroeder, W., Chen, Y., Brando, P. M., and Randerson, J. T. (2022). Tracking and classifying amazon fire events in near real time. Science Advances, 8(30):eabd2713.

Andreae, M. O., Acevedo, O. C., Araújo, A., Artaxo, P., Barbosa, H. M. J., et al. (2015). The amazon tall tower observatory (atto): Overview of pilot measurements on ecosystem ecology, meteorology, trace gases, and aerosols. Atmospheric Chemistry and Physics, 15:10723–10776.

Andreae, M. O., Rosenfeld, D., Artaxo, P., Costa, A. A., Frank, G. P., Longo, K. M., and Silva-Dias, M. A. F. (2004). Smoking rain clouds over the amazon. Science, 303(5662):1337–1342.

Aragão, L. E., Malhi, Y., Metcalfe, D. B., Silva-Espejo, J. E., Jiménez, E., Navarrete, D., Almeida, S., Costa, A. C., Salinas, N., Phillips, O. L., et al. (2007). Above-and below-ground net primary productivity across ten amazonian forests on contrasting soils. Biogeosciences, 4(5):1209–1220.

Aragão, L. E. O. C., Anderson, L. O., Fonseca, M. G., Rosan, T. M., Vedovato, L. B., Wagner, F. H., Silva, C. V. J., et al. (2018). 21st century drought-related fires counteract the decline of amazon deforestation carbon emissions. Nature Communications, 9(1):536.

Argles, A. P., Moore, J. R., and Cox, P. M. (2022). Dynamic global vegetation models: Searching for the balance between demographic process representation and computational tractability. PLOS Climate, 1(9):e0000068.

ARM Research Facility (2014a). Armbeatm data from the goamazon campaign. https://www.arm.gov/research/campaigns/amf2014goamazon. [Dataset].

ARM Research Facility (2014b). Armbecldrad data from the goamazon campaign. https://www.arm.gov/research/campaigns/amf2014goamazon. [Dataset].

Aslan, S. (2010). Comparison of missing value imputation methods for meteorological time series data. Master's thesis, Middle East Technical University.

Asner, G. P., Martin, R. E., Tupayachi, R., Anderson, C. B., Sinca, F., Carranza-Jiménez, L., and Martinez, P. (2014). Amazonian functional diversity from forest canopy chemical assembly. Proceedings of the National Academy of Sciences, 111(15):5604–5609.

Assis, T. O., Aguiar, A. P. D. d., Randow, C. v., Gomes, D. M. d. P., Kury, J. N., Ometto, J. P. H. B., and Nobre, C. A. (2020). Co2 emissions from forest degradation in brazilian amazon. Environmental Research Letters, 15(10):104035.

ATTO Data Repository (2024). 5-minute averaged and upsampled zi data for the atto site. https://www.attodata.org/. [Dataset].

Aubinet, M., Vesala, T., and Papale, D. (2012). Eddy covariance: a practical guide to measurement and data analysis. Springer Science & Business Media.

Avissar, R. and Pielke Sr, R. A. (2006). Representation of heterogeneity effects in earth system modeling: Experience from land surface modeling. Reviews of Geophysics, 44(2).

Baker, J. C., Garcia-Carreras, L., Gloor, M., Marsham, J. H., Buermann, W., da Rocha, H. R., Nobre, A. D., de Araujo, A. C., and Spracklen, D. V. (2021). Evapotranspiration in the amazon: spatial patterns, seasonality, and recent trends in observations, reanalysis, and climate models. Hydrology and Earth System Sciences, 25(4):2279–2300.

Balaji, V. (2021). Climbing down charney's ladder: machine learning and the post-dennard era of computational climate science. Philosophical Transactions of the Royal Society A, 379(2194):20200085.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., et al. (2001). Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society, 82(11):2415–2434.

Baldocchi, D., Sturtevant, C., and Contributors, F. (2015). Does day and night sampling reduce spurious correlation between canopy photosynthesis and ecosystem respiration? Agricultural and Forest Meteorology, 207:117–126.

Bamber, J. L., Westaway, R. M., Marzeion, B., and Wouters, B. (2018). The land ice contribution to sea level during the satellite era. Environmental Research Letters, 13(6):063008.

Barlow, J. F., Dunbar, T., Nemitz, E., Wood, C. R., Gallagher, M., Davies, F., O'Connor, E., and Harrison, R. (2011). Boundary layer dynamics over london, uk, as observed using doppler lidar during repartee-ii. Atmospheric Chemistry and Physics, 11(5):2111–2125.

Bastin, J.-F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., Zohner, C. M., and Crowther, T. W. (2019). The global tree restoration potential. Science, 365(6448):76–79.

Beamesderfer, E. R., Buechner, C., Faiola, C., Helbig, M., Sanchez-Mejia, Z. M., Yáñez-Serrano, A. M., Zhang, Y., and Richardson, A. D. (2022). Advancing cross-disciplinary understanding of land-atmosphere interactions. Journal of Geophysical Research: Biogeosciences, 127(2):e2021JG006707.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., et al. (2010). Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science, 329(5993):834–838.

Benediktsson, J. A., Swain, P. H., and Ersoy, O. K. (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. Technical report.

Bessenbacher, V., Seneviratne, S. I., and Gudmundsson, L. (2022). Climfill v0. 9: a framework for intelligently gap filling earth observations. Geoscientific Model Development, 15(11):4569–4596.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G., Essery, R., Ménard, C., Edwards, J., Hendry, M., Porson, A., Gedney, N., et al. (2011). The joint uk land environment simulator (jules), model description–part 1: energy and water fluxes. Geoscientific Model Development, 4(3):677–699.

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., et al. (2024). Climate-invariant machine learning. Science Advances, 10(6):eadj7250.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., and Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. Physical review letters, 126(9):098302.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., and Reichstein, M. (2018). Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. Earth System Science Data, 10(3):1327–1365.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bonan, G., Williams, M., Fisher, R., and Oleson, K. (2014). Modeling stomatal conductance in the earth system: linking leaf water-use efficiency and water transport along the soil–plant–atmosphere continuum. Geoscientific Model Development, 7(5):2193–2222.

Bonan, G. B. (2008). Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science, 320(5882):1444–1449.

Bonan, G. B. (2015). Ecological Climatology: Concepts and Applications. Cambridge University Press, Cambridge, England, 3rd edition.

Bonan, G. B. and Doney, S. C. (2018). Climate, ecosystems, and planetary futures: The challenge to predict life in earth system models. Science, 359(6375):eaam8328.

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., et al. (2015). Clouds, circulation and climate sensitivity. Nature geoscience, 8(4):261–268.

Boomgard-Zagrodnik, J. P. and Brown, D. J. (2022). Machine learning imputation of missing mesonet temperature observations. Computers and Electronics in Agriculture, 192:106580.

Boulton, C. A., Booth, B. B., and Good, P. (2017). Exploring uncertainty of amazon dieback in a perturbed parameter earth system ensemble. Global Change Biology, 23(12):5032–5044.

Brando, P. M., Balch, J. K., Nepstad, D. C., Morton, D. C., Putz, F. E., Coe, M. T., Silvério, D., Macedo, M. N., Davidson, E. A., Nóbrega, C. C., et al. (2014). Abrupt increases in amazonian tree mortality due to drought–fire interactions. Proceedings of the National Academy of Sciences, 111(17):6347–6352.

Brando, P. M., Barlow, J., Macedo, M. N., Silvério, D. V., Ferreira, J. N., Maracahipes, L., Anderson, L., Morton, D. C., Alencar, A., Paolucci, L. N., et al. (2025). Tipping points of amazonian forests: Beyond myths and toward solutions. Annual Review of Environment and Resources, 50.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Brienen, R. J., Phillips, O. L., Feldpausch, T. R., Gloor, E., Baker, T. R., Lloyd, J., Lopez-Gonzalez, G., Monteagudo-Mendoza, A., Malhi, Y., Lewis, S. L., et al. (2015). Long-term decline of the amazon carbon sink. Nature, 519(7543):344–348.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences, 113(15):3932–3937.

Brutsaert, W. and Parlange, M. (1998). Hydrologic cycle explains the evaporation paradox. Nature, 396(6706):30–30.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V. (2018). Observed fingerprint of a weakening atlantic ocean overturning circulation. Nature, 556(7700):191–196.

Carneiro, R., Fisch, G., Kaufmann, T., et al. (2016). Determination of planetary boundary layer heights in the forest amazon using a ceilometer. Ciência e Natura, 38(Special Edition):460–466.

Carneiro, R., Fisch, G., Neves, T., Santos, R., Santos, C., and Borges, C. (2021). Nocturnal boundary layer erosion analysis in the amazon using large-eddy simulation during goamazon project 2014/5. Atmosphere, 12(2):240.

Carneiro, R. G. and Fisch, G. (2020). Observational analysis of the daily cycle of the planetary boundary layer in the central amazon during a non-el niño year and el niño year (goamazon project 2014/5). Atmospheric Chemistry and Physics, 20(9):5547–5558.

Castro, A. A. d., Randow, C. v., Randow, R. d. C. S. v., and Bezerra, F. G. S. (2022). Evaluating carbon and water fluxes and stocks in brazil under changing climate and refined regional scenarios for changes in land use. Frontiers in Climate, 4.

Castro, J. L., Mantas, C. J., and Benítez, J. M. (2002). Interpretation of artificial neural networks by means of fuzzy rules. IEEE Transactions on Neural Networks, 13(1):101–116.

Caughey, S. J. (1984). Observed characteristics of the atmospheric boundary layer. In Atmospheric Turbulence and Air Pollution Modelling: A Course held in The Hague, 21–25 September, 1981, pages 107–158. Springer.

Cerlini, P. B., Silvestri, L., and Saraceni, M. (2020). Quality control and gap-filling methods applied to hourly temperature observations over central italy. Meteorological Applications, 27(3):e1913.

Chapin III, F. S., Woodwell, G. M., Randerson, J. T., Rastetter, E. B., Lovett, G. M., Baldocchi, D. D., Clark, D. A., Harmon, M. E., Schimel, D. S., Valentini, R., Wirth, C., Aber, J. D., Cole, J. J., Goulden, M. L., Harden, J. W., Heimann, M., Howarth, R. W., Matson, P. A., McGuire, A. D., Melillo, J. M., Mooney, H. A., Neff, J. C., Houghton, R. A., Pace, M. L., Ryan, M. G., Running, S. W., Sala, O. E., Schlesinger, W. H., and Schulze, E.-D. (2006). Reconciling carbon-cycle concepts, terminology, and methods. Ecosystems, 9(7):1041–1050.

Chapin III, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., et al. (2000). Consequences of changing biodiversity. Nature, 405(6783):234–242.

Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B., and Thomas, C. D. (2011). Rapid range shifts of species associated with high levels of climate warming. Science, 333(6045):1024–1026.

Chen, J. H. and Asch, S. M. (2017). Machine learning and prediction in medicine—beyond the peak of inflated expectations. The New England journal of medicine, 376(26):2507.

Chen, S., Stark, S. C., Nobre, A. D., Cuartas, L. A., de Jesus Amore, D., Restrepo-Coupe, N., Smith, M. N., Chitra-Tarak, R., Ko, H., Nelson, B. W., et al. (2024a). Amazon forest biogeography predicts resilience and vulnerability to drought. Nature, 631(8019):111–117.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794.

Chen, Z., Wang, W., Forzieri, G., and Cescatti, A. (2024b). Transition from positive to negative indirect co2 effects on the vegetation carbon uptake. Nature Communications, 15(1):1500.

Chitra-Tarak, R., Xu, C., Aguilar, S., Anderson-Teixeira, K. J., Chambers, J., Detto, M., Faybishenko, B., Fisher, R. A., Knox, R. G., Koven, C. D., et al. (2021). Hydraulically-vulnerable trees survive on deep-water access during droughts in a tropical forest. New Phytologist, 231(5):1798–1813.

CloudRoots Campaign (2024). Cloudroots campaign data. https://cloudroots.eu/data. [Dataset].

Cohn, S. A. and Angevine, W. M. (2000). Boundary layer height and entrainment zone thickness measured by lidars and wind-profiling radars. Journal of Applied Meteorology, 39(8):1233–1247.

Coulston, J. W., Blinn, C. E., Thomas, V. A., and Wynne, R. H. (2016). Approximating prediction uncertainty for random forest regression models. Photogrammetric Engineering & Remote Sensing, 82(3):189–197.

Crutzen, P. J. (2002). Geology of mankind. Nature, 415(6867):23–23.

Curtius, J., Heinritzi, M., Beck, L. J., Pöhlker, M. L., Tripathi, N., Krumm, B. E., Holzbeck, P., Nussbaumer, C. M., Hernández Pardo, L., Klimach, T., et al. (2024). Isoprene nitrates drive new particle formation in amazon's upper troposphere. Nature, 636(8041):124–130.

Dai, C., Wang, Q., Kalogiros, J., Lenschow, D., Gao, Z., and Zhou, M. (2014). Determining boundary-layer height from aircraft measurements. Boundary-layer meteorology, 152:277–302.

Daly, S., Davis, R., Ochs, E., and Pangburn, T. (2000). An approach to spatially distributed snow modelling of the sacramento and san joaquin basins, california. Hydrological Processes, 14(18):3257–3271.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., et al. (2020). The community earth system model version 2 (cesm2). Journal of Advances in Modeling Earth Systems, 12(2):e2019MS001916.

Darela-Filho, J. P., Rammig, A., Fleischer, K., et al. (2024). Reference maps of soil phosphorus for the pan-amazon region. Earth System Science Data, 16(2):715–729.

Davidson, E. A., de Araújo, A. C., Artaxo, P., Balch, J. K., Brown, I. F., Bustamante, M. M., Coe, M. T., DeFries, R. S., Keller, M., Longo, M., et al. (2012). The amazon basin in transition. Nature, 481(7381):321–328.

de Arellano, J. V.-G., Hartogensis, O., de Boer, H., Moonen, R., González-Armas, R., Janssens, M., Adnew, G., Bonell-Fontás, D., Botía, S., Jones, S., et al. (2024). Cloudroots-amazon22: Integrating clouds with photosynthesis by crossing scales. Bulletin of the American Meteorological Society.

de Arruda Moreira, G., Sánchez-Hernández, G., Guerrero-Rascado, J. L., Cazorla, A., and Alados-Arboledas, L. (2022). Estimating the urban atmospheric boundary layer height from remote sensing applying machine learning techniques. Atmospheric Research, 266:105962.

de Souza, C. M. A., Júnior, C. Q. D., Martins, H. d. S., D'Oliveira, F. A. F., Machado, L. A. T., Carneiro, R. G., and Fisch, G. F. (2023). Climatology of the height of the atmospheric boundary layer in the central amazon. In XII Workshop Brasileiro de Micrometeorologia. UFSM Revista Ciência e Natura.

Deardorff, J. W. (1974). Three-dimensional numerical study of the height and mean structure of a heated planetary boundary layer. Boundary-Layer Meteorology, 7(1):81–106.

Dengel, S., Zona, D., Sachs, T., Aurela, M., Jammet, M., Parmentier, F., Oechel, W., and Vesala, T. (2013). Testing the applicability of neural networks as a gap-filling method using ch 4 flux data from high latitude wetlands. Biogeosciences, 10(12):8185–8200.

Deser, C., Knutti, R., Solomon, S., and Phillips, A. S. (2012). Uncertainty in climate change projections: the role of internal variability. Climate dynamics, 38(3-4):527–546.

Detto, M., Baldocchi, D., and Katul, G. G. (2010). Scaling properties of biologically active scalar concentration fluctuations in the atmospheric surface layer over a managed peatland. Boundary-layer meteorology, 136(3):407–430.

Dias-Júnior, C. Q., Carneiro, R. G., Fisch, G., D'Oliveira, F. A. F., Sörgel, M., Botía, S., Machado, L. A. T., Wolff, S., Santos, R. M. N. d., and Pöhlker, C. (2022). Intercomparison of planetary boundary layer heights using remote sensing retrievals and era5 reanalysis over central amazonia. Remote Sensing, 14(18):4561.

Dias-Júnior, C. Q., Dias, N. L., dos Santos, R. M. N., Sörgel, M., Araújo, A., Tsokankunku, A., et al. (2019). Is there a classical inertial sublayer over the amazon forest? Geophysical Research Letters, 46(10):5614–5622.

Dumitrescu, A., Brabec, M., and Cheval, S. (2020). Statistical gap-filling of seviri land surface temperature. Remote Sensing, 12(9):1423.

Eichelmann, E., Hemes, K. S., Knox, S. H., Oikawa, P. Y., Chamberlain, S. D., Sturtevant, C., Verfaillie, J., and Baldocchi, D. D. (2018). The effect of land cover type and structure on evapotranspiration from agricultural and wetland sites in the sacramento–san joaquin river delta, california. Agricultural and Forest Meteorology, 256:179–195.

Eichelmann, E., Knox, S., Rey Sanchez, C., Valach, A., Sturtevant, C., Szutu, D., Verfaillie, J., and Baldocchi, D. (2021a). AmeriFlux US-Tw4 Twitchell East End Wetland, Ver. 11-5, AmeriFlux AMP, (Dataset). "https://doi.org/10.17190/AMF/1246151".

Eichelmann, E., Mantoani, M. C., Chamberlain, S. D., Hemes, K. S., Oikawa, P. Y., Szutu, D., Valach, A., Verfaillie, J., and Baldocchi, D. D. (2021b). A novel approach to partitioning evapotranspiration into evaporation and transpiration in flooded ecosystems. bioRxiv.

Eresmaa, N., Karppinen, A., Joffre, S., Räsänen, J., and Talvitie, H. (2006). Mixing height determination by ceilometer. Atmospheric Chemistry and Physics, 6(6):1485–1493.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. University of Montreal, 1341(3):1.

Espinoza, J. C., Ronchail, J., Guyot, J. L., Cochonneau, G., Filizola, N., Fraizy, P., Labat, D., de Oliveira, E., Ordoñez, J. J., and Vauchel, P. (2009). Spatio-temporal rainfall variability in the amazon basin countries (brazil, peru, bolivia, colombia, and ecuador). International Journal of Climatology, 29(11):1574–1594.

et. al., J. I. (2021). Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at fluxnet-ch4 wetlands. Agricultural and Forest Meteorology, 308:108528.

EU Cost Action (2021). Fair network of micrometeorological measurements. https://www.fairness-ca20108.eu/, (CA20108).

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. Geoscientific Model Development, 9(5):1937–1958.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., et al. (2001). Gap filling strategies for defensible annual sums of net ecosystem exchange. Agricultural and forest meteorology, 107(1):43–69.

Fan, Z., Yan, Z., and Wen, S. (2023). Deep learning and artificial intelligence in sustainability: a review of sdgs, renewable energy, and environmental health. Sustainability, 15(18):13493.

Fischer, R., Piatkowski, N., Pelletier, C., Webb, G. I., Petitjean, F., and Morik, K. (2020). No cloud on the horizon: Probabilistic gap filling in satellite image series. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pages 546–555. IEEE.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. The Journal of Machine Learning Research, 20(1):177–206.

Fisher, J. B., Lee, B., Purdy, A. J., Halverson, G. H., Dohlen, M. B., Cawse-Nicholson, K., Wang, A., Anderson, R. G., Aragon, B., Arain, M. A., et al. (2020). Ecostress: Nasa's next generation mission to measure evapotranspiration from the international space station. Water Resources Research, 56(4):e2019WR026058.

Fisher, R. A. and Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. Journal of Advances in Modeling Earth Systems, 12(4):e2018MS001453.

Fisher, R. A., Koven, C. D., Anderegg, W. R., Christoffersen, B. O., Dietze, M. C., Farrior, C. E., Holm, J. A., Hurtt, G. C., Knox, R. G., Lawrence, P. J., et al. (2018). Vegetation demographics in earth system models: A review of progress and priorities. Global change biology, 24(1):35–54.

Flora, M. L., Potvin, C. K., McGovern, A., and Handler, S. (2024). A machine learning explainability tutorial for atmospheric sciences. Artificial Intelligence for the Earth Systems, 3(1):e230018.

Foley, P. (2023). Evaluating a novel approach to partitioning evapotranspiration into evaporation and transpiration in deciduous broadleaf forest ecosystems. MSc thesis, University College Dublin, Dublin, Ireland. MSc Environmental Sustainability, ENVB40520-Practicum (Research).

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE international conference on computer vision, pages 3429–3437.

Forzieri, G., Dakos, V., McDowell, N. G., Ramdane, A., and Cescatti, A. (2022). Emerging signals of declining forest resilience under climate change. Nature, 608(7923):534–539.

French, M. N., Krajewski, W. F., and Cuykendall, R. R. (1992). Rainfall forecasting in space and time using a neural network. Journal of hydrology, 137(1-4):1–31.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., et al. (2023). Global carbon budget 2023. Earth System Science Data, 15(12):5301–5369.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232.

Gad, I., Manjunatha, B., et al. (2017). Performance evaluation of predictive models for missing data imputation in weather data. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 1327–1334. IEEE.

Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., Zhu, Y., Li, M., and Wang, Y. (2023). Prediff: Precipitation nowcasting with latent diffusion models. In Advances in Neural Information Processing Systems, volume 36.

Garen, D. C., Johnson, G. L., and Hanson, C. L. (1994). Mean areal precipitation for daily hydrologic modeling in mountainous regions 1. JAWRA Journal of the American Water Resources Association, 30(3):481–491.

Garratt, J. R. (1994). The atmospheric boundary layer. Earth-Science Reviews, 37(1-2):89–134.

Gasparrini, A. (2016). Modelling lagged associations in environmental time series data: a simulation study. Epidemiology, 27(6):835–842.

Gatti, L. V., Basso, L. S., Miller, J. B., Gloor, M., Domingues, L. G., Cassol, H. L. G., Tejada, G., et al. (2021). Amazonia as a carbon source linked to deforestation and climate change. Nature, 595(7867):388–393.

Gauss, C. F. (1877). Theoria motus corporum coelestium in sectionibus conicis solem ambientium, volume 7. FA Perthes.

Geiß, A., Wiegner, M., Bonn, B., Sch"afer, K., Forkel, R., Schneidemesser, E., M"unkel, C., Chan, K., and Nothard, R. (2017). Mixing layer height as an indicator for urban air quality? Atmospheric Measurement Techniques, 10:2969–2988.

Gerken, T., Bromley, G. T., and Stoy, P. C. (2018). Surface moistening trends in the northern north american great plains increase the likelihood of convective initiation. Journal of Hydrometeorology, 19(1):227–244.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). Random forests for land cover classification. Pattern recognition letters, 27(4):294–300.

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., Pulignano, V., et al. (2023). A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. Artificial intelligence review, 56(4):3473–3504.

Green, J. K., Konings, A. G., Alemohammad, S. H., Berry, J., Entekhabi, D., Kolassa, J., Lee, J.-E., and Gentine, P. (2017). Regionally strong feedbacks between the atmosphere and terrestrial biosphere. Nature Geoscience, 10(6):410–414.

Guo, J., Zhang, J., Shao, J., Chen, T., Bai, K., Sun, Y., Li, N., Wu, J., Li, R., Li, J., et al. (2024). A merged continental planetary boundary layer height dataset based on high-resolution radiosonde measurements, era5 reanalysis, and gldas. Earth System Science Data, 16(1):1–14.

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Lv, Y., Shao, J., Yu, T., Tong, B., et al. (2021). Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first results and comparison with era5, merra-2, jra-55, and ncep-2 reanalyses. Atmospheric Chemistry and Physics, 21(22):17079–17097.

Guo, M.-H., Xu, J., Zhang, Y., Song, J., Peng, H., Deng, Y.-X., Dong, X., Nakayama, K., Geng, Z., Wang, C., et al. (2025). R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. arXiv preprint arXiv:2505.02018.

Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. IEEE intelligent systems, 24(2):8–12.

Han, J., Pei, J., and Kamber, M. (2011). Data mining: concepts and techniques, volume 3. Elsevier.

Hanna, S. R. (1969). The thickness of the planetary boundary layer. Atmospheric Environment (1967), 3(5):519–536.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., et al. (2013). High-resolution global maps of 21st-century forest cover change. science, 342(6160):850–853.

Hao, Z., Liu, S., Zhang, Y., Ying, C., Feng, Y., Su, H., and Zhu, J. (2023). Physics-informed machine learning: A survey on problems, methods and applications.

Harder, P., Watson-Parris, D., Stier, P., Strassel, D., Gauger, N. R., and Keuper, J. (2022). Physics-informed learning of aerosol microphysics. Environmental Data Science, 1:e20.

Hartkamp, A., de Beurs, K., Stein, A., and White, J. (1999). Interpolation techniques for climate variables. Geographic Information Systems Series 99-01. International Maize and Wheat Improvement Center (CIMMYT), Mexico 1999. ISSN: 1405-7484.

Hatala, J. A., Detto, M., Sonnentag, O., Deverel, S. J., Verfaillie, J., and Baldocchi, D. D. (2012). Greenhouse gas (co2, ch4, h2o) fluxes from drained and flooded agricultural peatlands in the sacramento-san joaquin delta. Agriculture, ecosystems & environment, 150:1–18.

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. Bulletin of the American Meteorological Society, 90(8):1095–1108.

Heffter, J. L. (1980). Transport layer depth calculations. In Second joint conference on applications of air pollution meteorology, pages 24–27. Louisiana New Orleans.

Heffter, J. L. (1983). Branching atmospheric trajectory (bat) model. NOAA Technical Memo ERL ARL, 121:1–16.

Helbig, M., Gerken, T., Beamesderfer, E. R., Baldocchi, D. D., Banerjee, T., Biraud, S. C., Brown, W. O., Brunsell, N. A., Burakowski, E. A., Burns, S. P., et al. (2021). Integrating continuous atmospheric boundary layer and tower-based flux measurements to advance understanding of land-atmosphere interactions. Agricultural and Forest Meteorology, 307:108509.

Hennemuth, B. and Lammert, A. (2006). Determination of the atmospheric boundary layer height from radiosonde and lidar backscatter. Boundary-Layer Meteorology, 120:181–200.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730):1999–2049.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359–366.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., et al. (2017). The art and science of climate model tuning. Bulletin of the American Meteorological Society, 98(3):589–602.

Hoyer, S. and Hamman, J. (2017). xarray: N-D labeled arrays and datasets in Python. Journal of Open Research Software, 5(1):10.

https://pypi.org/project/haversine/ (Released: Jan 16, 2024). Haversine formula. Technical report, Python Package Index.

Huang, F., Jiang, S., Li, L., Zhang, Y., Zhang, Y., Zhang, R., Li, Q., Li, D., Shangguan, W., and Dai, Y. (2025). Applications of explainable artificial intelligence in earth system science. Under Review.

Huntington, T. G. (2006). Evidence for intensification of the global water cycle: review and synthesis. Journal of hydrology, 319(1-4):83–95.

IPCC (2023). Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E. A., Kadow, C., Staneva, J., and Saynisch-Wagner, J. (2021). Towards neural earth system modelling by integrating artificial intelligence in earth system science. Nature Machine Intelligence, 3(8):667–674.

Iwata, H., Hirata, R., Takahashi, Y., Miyabara, Y., Itoh, M., and Iizuka, K. (2018). Partitioning eddy-covariance methane fluxes from a shallow lake into diffusive and ebullitive fluxes. Boundary-Layer Meteorology, 169(3):413–428.

Jebeile, J., Lam, V., Majszak, M., and Räz, T. (2023). Machine learning and the quest for objectivity in climate model parameterization. Climatic Change, 176(8):101.

Jenkins, C. N., Pimm, S. L., and Joppa, L. N. (2013). Global patterns of terrestrial vertebrate diversity and conservation. Proceedings of the National Academy of Sciences, 110(28):E2602–E2610.

Jiang, Y., Tang, R., and Li, Z.-L. (2022). A physical full-factorial scheme for gap-filling of eddy covariance measurements of daytime evapotranspiration. Agricultural and Forest Meteorology, 323:109087.

Jindo, K., Audette, Y., Olivares, F. L., Canellas, L. P., Smith, D. S., and Paul Voroney, R. (2023). Biotic and abiotic effects of soil organic matter on the phytoavailable phosphorus in soils: A review. Chemical and Biological Technologies in Agriculture, 10(1):29.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. Nature, 596(7873):583–589.

Jung, M., Reichstein, M., and Bondeau, A. (2009). Towards global empirical upscaling of fluxnet eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. Biogeosciences, 6(10):2001–2013.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., De Jeu, R., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. Nature, 467(7318):951–954.

Kaimal, J. C. and Finnigan, J. J. (1994). Atmospheric boundary layer flows: their structure and measurement. Oxford University Press.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6):422–440.

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al. (2021). Physics-informed machine learning: case studies for weather and climate modelling. Philosophical Transactions of the Royal Society A, 379(2194):20200093.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kim, Y., Johnson, M. S., Knox, S. H., Black, T. A., Dalmagro, H. J., Kang, M., Kim, J., and Baldocchi, D. (2020). Gap-filling approaches for eddy covariance methane fluxes: A comparison of three machine learning algorithms and a traditional method with principal component analysis. Global Change Biology, 26(3):1499–1518.

Klosterhalfen, A., Graf, A., Brüggemann, N., Drüe, C., Esser, O., González-Dugo, M. P., Heinemann, G., Jacobs, C. M., Mauder, M., Moene, A. F., et al. (2019). Source partitioning of h 2 o and co 2 fluxes based on high-frequency eddy covariance data: a comparison between study sites. Biogeosciences, 16(6):1111–1132.

Knox, S. H., Sturtevant, C., Matthes, J. H., Koteen, L., Verfaillie, J., and Baldocchi, D. (2015). Agricultural peatland restoration: effects of land-use change on greenhouse gas (co2 and ch4) fluxes in the sacramento-san joaquin delta. Global change biology, 21(2):750–765.

Koci, I. (2022). Micromet data from aws 00000e88 from 21.4.2010 to 31.12.2022. (https://zenodo.org/records/7944501).

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. International conference on machine learning, pages 1885–1894.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild

distribution shifts. In International conference on machine learning, pages 5637–5664. PMLR.

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., and Dean, R. (2025). Ai 2027. webpage, April, 3.

Kørner, P., Kronenberg, R., Genzel, S., and Bernhofer, C. (2018). Introducing gradient boosting as a universal gap filling tool for meteorological time series. Meteorol. Z, 27(5):369.

Kotthaus, S., O'Connor, E., Münkel, C., Charlton-Perez, C., Haeffelin, M., Gabey, A., and Grimmond, C. (2016). Recommendations for processing atmospheric attenuated backscatter profiles from Vaisala CL31 ceilometers. Atmospheric Measurement Techniques, 9:3769–3791.

Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D. (2021). On the estimation of boundary layer heights: a machine learning approach. Atmospheric Measurement Techniques, 14(6):4403–4424.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., et al. (2018). The low-resolution version of hadgem3 gc3. 1: Development and evaluation for global climate. Journal of Advances in Modeling Earth Systems, 10(11):2865–2888.

Kühnhammer, K., van Haren, J., Kübert, A., Bailey, K., Dubbert, M., Hu, J., Ladd, S. N., Meredith, L. K., Werner, C., and Beyer, M. (2023). Deep roots mitigate drought impacts on tropical trees despite limited quantitative contribution to transpiration. Science of the Total Environment, 893:164763.

Lahsen, M. and Nobre, C. A. (2007). Challenges of connecting international science and local level sustainability efforts: the case of the large-scale biosphere–atmosphere experiment in Amazonia. Environmental Science & Policy, 10(1):62–74.

Lalic, B., Marcic, M., Sremac, A. F., Eitzinger, J., Koci, I., Petric, T., Ljubojevic, M., and Jezerkic, B. (2020). Landscape phenology modelling and decision support in serbia. Landscape Modelling and Decision Support, pages 567–593.

Lalic, B., Stapleton, A., Vergauwen, T., Caluwaerts, S., Eichelmann, E., and Roantree, M. (2024). A comparative analysis of machine learning approaches to gap filling meteorological datasets. Environmental Earth Sciences, 83(24):679.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning skillful medium-range global weather forecasting. Science, 382(6677):1416–1421.

Lapola, D. M., Pinho, P., Barlow, J., et al. (2023). The drivers and impacts of amazon forest degradation. Science, 379(6630):eabp8622.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1):1096.

Laurans, M., Hérault, B., Vieilledent, G., and Vincent, G. (2014). Vertical stratification reduces competition for light in dense tropical forests. Forest Ecology and Management, 329:79–88.

Lawal, Z. K., Yassin, H., Lai, D. T. C., and Che Idris, A. (2022). Physics-informed neural network (pinn) evolution and beyond: A systematic literature review and bibliometric analysis. Big Data and Cognitive Computing, 6(4):140.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., Kennedy, D., et al. (2019). The community land model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. Journal of Advances in Modeling Earth Systems, 11(12):4245–4287.

Lawrence, P. J. and Chase, T. N. (2007). Representing a new modis consistent land surface in the community land model (clm 3.0). Journal of Geophysical Research: Biogeosciences, 112(G1).

Lawrence, P. J., Feddema, J. J., Bonan, G. B., Meehl, G. A., O'Neill, B. C., Oleson, K. W., Levis, S., Lawrence, D. M., Kluzek, E., Lindsay, K., et al. (2012). The climate impacts of land surface change and carbon management, and the implications for climate-change mitigation policy. Climate dynamics, 39(6):1331–1348.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521(7553):436–444.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, volume 3, page 896. Atlanta.

Lee, J., Weger, R. C., Sengupta, S. K., and Welch, R. M. (1990). A neural network approach to cloud classification. IEEE Transactions on Geoscience and Remote Sensing, 28(5):846–855.

Legendre, A. M. (1806). Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805. Courcier.

LeMone, M. A., Angevine, W. M., Bretherton, C. S., Chen, F., Dudhia, J., Fedorovich, E., Katsaros, K. B., Lenschow, D. H., Mahrt, L., Patton, E. G., et al. (2019). 100 years of progress in boundary layer meteorology. Meteorological Monographs, 59:9–1.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. (2018a). Domain generalization with adversarial feature learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5400–5409.

Li, O., Liu, H., Chen, C., and Rudin, C. (2018b). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In Proceedings of the AAAI conference on artificial intelligence, volume 32.

Li, X. and Xiao, J. (2019). Mapping photosynthesis solely from solar-induced chlorophyll fluorescence: A global, fine-resolution dataset of gross primary production derived from oco-2. Remote Sensing, 11(21):2563.

Liang, M. and Hu, X. (2015). Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3367–3375.

Lin, Z., Mo, X., Li, H., and Li, H. (2002). Comparison of three spatial interpolation methods for climate variables in china. Acta Geographica Sinica, 57(1):47–56.

Lipson, M., Grimmond, S., Best, M., Chow, W. T., Christen, A., Chrysoulakis, N., Coutts, A., Crawford, B., Earl, S., Evans, J., et al. (2022). Harmonized gap-filled datasets from 20 urban flux tower sites. Earth system science data, 14(11):5157–5178.

Liu, S. and Liang, X.-Z. (2010). Observed diurnal cycle climatology of planetary boundary layer height. Journal of Climate, 23(21):5790–5809.

Liu, S., Lu, D., Painter, S. L., Griffiths, N. A., and Pierce, E. M. (2023). Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. Frontiers in Water, 5:1150126.

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., Wehner, M., Collins, W., et al. (2016). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv preprint arXiv:1605.01156.

Liu, Z., Chang, J., Li, H., Chen, S., and Dai, T. (2022). Estimating boundary layer height from lidar data under complex atmospheric conditions using machine learning. Remote Sensing, 14(2):418.

Lompar, M., Lalić, B., Dekić, L., and Petrić, M. (2019). Filling gaps in hourly air temperature data using debiased era5 data. Atmosphere, 10(1):13.

Longo, M., Knox, R. G., Levine, N. M., Alves, L. F., Bonal, D., Camargo, P. B., Fitzjarrald, D. R., et al. (2018). Ecosystem heterogeneity and diversity mitigate amazon forest resilience to frequent extreme droughts. The New Phytologist, 219(3):914–931.

Lucas-Moffat, A. M., Schrader, F., Herbst, M., and Brummer, C. (2022). Multiple gap-filling for eddy covariance datasets. Agricultural and Forest Meteorology, 325:109114.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lundberg, S. M. and Lee, S.-I. (2017a). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. Curran Associates, Inc.

Mabagala, F. S. et al. (2022). On the tropical soils; the influence of organic matter (om) on phosphate bioavailability. Saudi Journal of Biological Sciences, 29(5):3635–3641.

Mahabbati, A., Beringer, J., Leopold, M., McHugh, I., Cleverly, J., Isaac, P., and Izady, A. (2021). A comparison of gap-filling algorithms for eddy covariance fluxes and their drivers. Geoscientific Instrumentation, Methods and Data Systems, 10(1):123–140.

Malone, T. F. (1955). Application of statistical methods in weather prediction. Proceedings of the National Academy of Sciences, 41(11):806–815.

MapBiomas Amazonía (2024). Mapbiomas amazonía project – collection [version] of annual land cover and land use maps. http://amazonia.mapbiomas.org. Accessed on [date]. The MapBiomas Amazonía project is a multi-institutional initiative coordinated by the Amazonian Network of Geo-referenced Socio-environmental Information (RAISG) to generate annual land cover and land use maps from automatic classification processes applied to satellite imagery.

MapBiomas Project (2022). Collection 8 of the annual land use land cover maps of brazil. https://storage.googleapis.com/mapbiomas-public/brasil/sentinel/lclu/coverage/brasil_sentinel_coverage_2022.tif. [Dataset].

Marengo, J. A., Liebmann, B., Grimm, A., Misra, V., Silva Dias, P. d., Cavalcanti, I., Carvalho, L. M. V. d., Berbery, E., Ambrizzi, T., Vera, C. S., et al. (2010). Recent developments on the south american monsoon system.

Marengo, J. A., Souza Jr, C. M., Thonicke, K., Burton, C., Halladay, K., Betts, R. A., Alves, L. M., and Soares, W. R. (2018). Changes in climate and land use over the amazon region: current and future variability and trends. Frontiers in Earth Science, 6:228.

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Andreae, M. O., Barbosa, H., Fan, J., et al. (2016). Introduction: observations and modeling of the green ocean amazon (goamazon2014/5). Atmospheric Chemistry and Physics, 16(8):4785–4797.

Maskey, M., Ramachandran, R., Gurung, I., Freitag, B., Miller, J., Ramasubramanian, M., Bollinger, D., Mestre, R., Cecil, D., Molthan, A., et al. (2019). Machine learning lifecycle for earth science application: a practical insight into production deployment. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, pages 10043–10046. IEEE.

Materia, S., García, L. P., van Straaten, C., O, S., Mamalakis, A., Cavicchia, L., Coumou, D., de Luca, P., Kretschmer, M., and Donat, M. (2024). Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. Wiley Interdisciplinary Reviews: Climate Change, 15(6):e914.

Matthes, J. H., Sturtevant, C., Oikawa, P., Chamberlain, S. D., Szutu, D., Ortiz, A. A., Verfaillie, J., and Baldocchi, D. (2021). AmeriFlux US-Myb Mayberry Wetland, Ver. 11-5, AmeriFlux AMP, (Dataset). "https://doi.org/10.17190/AMF/1246139".

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. Bulletin of the American Meteorological Society, 100(11):2175–2199.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an ai system for breast cancer screening. Nature, 577(7788):89–94.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. Journal of machine learning research, 7(6).

Melfi, S., Spinhirne, J., Chou, S.-H., and Palm, S. (1985). Lidar observations of vertically organized convection in the planetary boundary layer over the ocean. Journal of Applied Meteorology and Climatology, 24(8):806–821.

Meng, L., Chambers, J., Koven, C., Pastorello, G., Gimenez, B., Jardine, K., Tang, Y., et al. (2022). Soil moisture thresholds explain a shift from light-limited to water-limited sap velocity in the central amazon during the 2015–16 el niño drought. Environmental Research Letters, 17(6):064023.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. Journal of Machine Learning Research, 17(26):1–41.

Menut, L., Flamant, C., Pelon, J., and Flamant, P. H. (1999). Urban boundary-layer height determination from lidar measurements over the paris area. Applied Optics, 38(6):945–954.

Miller, E. J. (1997). Towards a 4d gis: Four-dimensional interpolation utilizing kriging. Innovations in GIS, page 181.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, New York.

Mitra, S. and Hayashi, Y. (2000). Neuro-fuzzy rule generation: survey in soft computing framework. IEEE Transactions on neural networks, 11(3):748–768.

Mo, L., Crowther, T. W., Maynard, D. S., Van den Hoogen, J., Ma, H., Bialic-Murphy, L., Liang, J., De-Miguel, S., Nabuurs, G.-J., Reich, P. B., et al. (2024). The global distribution and drivers of wood density and their impact on forest carbon stocks. Nature Ecology & Evolution, 8(12):2195–2212.

Mo, L., Zohner, C. M., Reich, P. B., Liang, J., De Miguel, S., Nabuurs, G.-J., Renner, S. S., Van Den Hoogen, J., Araza, A., Herold, M., et al. (2023). Integrated global assessment of the natural forest carbon potential. Nature, 624(7990):92–101.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., et al. (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and Forest Meteorology, 147(3-4):209–232.

Molero, F., Barragán, R., and Artíñano, B. (2022). Estimation of the atmospheric boundary layer height by means of machine learning techniques using ground-level meteorological data. Atmospheric Research, 279:106401.

Morales-Moraga, D., Meza, F. J., Miranda, M., and Gironás, J. (2019). Spatio-temporal estimation of climatic variables for gap filling and record extension using reanalysis data. Theoretical and Applied Climatology, 137:1089–1104.

Morris, V. (2016). Ceilometer instrument handbook. Technical report, DOE Office of Science Atmospheric Radiation Measurement (ARM) User Facility.

Morton, D. C., Le Page, Y., DeFries, R., Collatz, G. J., and Hurtt, G. C. (2013). Understorey fire frequency and the fate of burned forests in southern amazonia. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368(1619):20120163.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In International conference on machine learning, pages 10–18. PMLR.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., et al. (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9):4349–4383.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., et al. (2013). Anthropogenic and natural radiative forcing. Climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change, pages 659–740.

NASA (2013). Shuttle radar topography mission (srtm) global. https://doi.org/10.5069/G9445JDF. [Dataset].

NCAR Earth Observing Laboratory (2024). Cafe brazil campaign data. https://data.eol.ucar.edu/project/CAFE-Brazil. [Dataset].

Nelson, J. A., Pérez-Priego, O., Zhou, S., Poyatos, R., Zhang, Y., Blanken, P. D., Gimeno, T. E., Wohlfahrt, G., Desai, A. R., Gioli, B., et al. (2020). Ecosystem transpiration and evaporation: Insights from three water flux partitioning methods across fluxnet sites. Global change biology, 26(12):6916–6930.

Nemani, R. R., Keeling, C. D., Hashimoto, H., Jolly, W. M., Piper, S. C., Tucker, C. J., Myneni, R. B., and Running, S. W. (2003). Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science, 300(5625):1560–1563.

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU.

Nobre, C. A. and Borma, L. D. S. (2009). 'tipping points' for the amazon forest. Current Opinion in Environmental Sustainability, 1(1):28–36.

Novick, K. A., Biederman, J., Desai, A., Litvak, M., Moore, D. J., Scott, R., and Torn, M. (2018). The ameriflux network: A coalition of the willing. Agricultural and Forest Meteorology, 249:444–456.

Novick, K. A., Ficklin, D. L., Stoy, P. C., Williams, C. A., Bohrer, G., Oishi, A. C., Papuga, S. A., Blanken, P. D., Noormets, A., Sulman, B. N., et al. (2016). The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. Nature climate change, 6(11):1023–1027.

Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) (2024). Lba data archive. https://daac.ornl.gov/LBA/. [Dataset].

Oki, T. and Kanae, S. (2006). Global hydrological cycles and world water resources. Science, 313(5790):1068–1072.

OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI.

O'Connell, J. L., Byrd, K. B., and Kelly, M. (2015). A hybrid model for mapping relative differences in belowground biomass and root: shoot ratios using spectral reflectance, foliar n and plant biophysical data within coastal marsh. Remote Sensing, 7(12):16480–16503.

Padial-Iglesias, M., Pons, X., Serra, P., and Ninyerola, M. (2022). Does the gap-filling method influence long-term (1950–2019) temperature and precipitation trend analyses? GeoFocus. International Review of Geographical Information Science and Technology, (29):5–33.

Palomas, S., Acosta, M. C., Utrera, G., and Tourigny, E. (2024). Reducing time and computing costs in ec-earth: An automatic load-balancing approach for coupled esms. Geoscientific Model Development Discussions, 2024:1–25.

Pape, R., Wundram, D., and Löffler, J. (2009). Modelling near-surface temperature conditions in high mountain environments: an appraisal. Climate Research, 39(2):99–109.

Pascolini-Campbell, M., Reager, J. T., Chandanpurkar, H. A., and Rodell, M. (2021). A 10 per cent increase in global land evapotranspiration from 2003 to 2019. Nature, 593(7860):543–547.

Pattyn, F., Ritz, C., Hanna, E., Asay-Davis, X., DeConto, R., Durand, G., Favier, L., Fettweis, X., Goelzer, H., Golledge, N. R., et al. (2018). The greenland and antarctic ice sheets under 1.5 c global warming. Nature climate change, 8(12):1053–1061.

Peatier, S., Sanderson, B. M., and Terray, L. (2024). Exploration of diverse solutions for the calibration of imperfect climate models. Earth System Dynamics, 15(4):987–1014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011a). Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Peters, J., Janzing, D., and Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms. The MIT Press.

Peterson, T. C. and Vose, R. S. (1997). An overview of the global historical climatology network temperature database. Bulletin of the American Meteorological Society, 78(12):2837–2850.

Phillips, O. L., Brienen, R. J., and Collaboration, R. (2017). Carbon uptake by mature amazon forests has mitigated amazon nations' carbon emissions. Carbon Balance and Management, 12(1):1.

Pielke, R. A., Avissar, R., Raupach, M., Dolman, A. J., Zeng, X., and Denning, A. S. (1998). Interactions between the atmosphere and terrestrial ecosystems: influence on weather and climate. Global change biology, 4(5):461–475.

Pitman, A. (2003). The evolution of, and revolution in, land surface schemes designed for climate models. International Journal of Climatology, 23(5):479–510.

Poulter, B., MacBean, N., Hartley, A., Khlystova, I., Arino, O., Betts, R., Bontemps, S., Boettcher, M., Brockmann, C., Defourny, P., et al. (2015). Plant functional type classification for earth system models: results from the european space agency's land cover climate change initiative. Geoscientific Model Development, 8(7):2315–2328.

PRODES-INPE (2022). Satellite monitoring of the amazon forest. http://www.obt.inpe.br/prodes.

Python Software Foundation (2024). Python Language Reference, version 3.

Qin, Y., Wang, D., Cao, Y., Cai, X., Liang, S., Beck, H. E., and Zeng, Z. (2023). Sub-grid representation of vegetation cover in land surface schemes improves the modeling of how climate responds to deforestation. Geophysical Research Letters, 50(15):e2023GL104164.

RAISG (2024). Red amazónica de información socioambiental georreferenciada - mapbiomas amazonía. https://www.raisg.org/en/project/mapbiomas-amazonia/. Amazonian Network of Geo-referenced Socio-environmental Information.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics, 378:686–707.

Rasp, S., Pritchard, M. S., and Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. Proceedings of the national academy of sciences, 115(39):9684–9689.

Razavi, A. R., Nassiri Mahallati, M., Koocheki, A., and Beheshti, A. (2018). Applicability of agmerra for gap-filling of afghanistan in-situ temperature and precipitation data. Water and Soil, 32(3):601–616.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., et al. (2013). Climate extremes and the carbon cycle. Nature, 500(7462):287–295.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. Nature, 566(7743):195–204.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., et al. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global change biology, 11(9):1424–1439.

Restrepo-Coupe, N., Campos, K. S., Alves, L. F., Longo, M., Wiedemann, K. T., de Oliveira Jr, R. C., Aragao, L. E. O. C., et al. (2024). Contrasting carbon cycle responses to dry (2015 el niño) and wet (2008 la niña) extreme events at an amazon tropical forest. Agricultural and Forest Meteorology, 353:110037.

Reynolds, S. A., Beery, S., Burgess, N., Burgman, M., Butchart, S. H., Cooke, S. J., Coomes, D., Danielsen, F., Di Minin, E., Durán, A. P., et al. (2025). The potential for ai to revolutionize conservation: a horizon scan. Trends in ecology & evolution, 40(2):191–207.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you? explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144.

Richardson, L. F. (1920). The supply of energy from and to atmospheric eddies. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 97(686):354–373.

Rieutord, T., Aubert, S., and Machado, T. (2021). Deriving boundary layer height from aerosol lidar using machine learning: Kabl and adabl algorithms. Atmospheric Measurement Techniques, 14(6):4335–4353.

Roantree, M. (2024). A comparative analysis of machine learning approaches to gap filling meteorological datasets (results only). (10.5281/zenodo.12818855).

Roantree, M., Lalic, B., Savic, S., Milosevic, D., and Scriney, M. (2023). Constructing a searchable knowledge repository for fair climate data. https://arxiv.org/pdf/2304.05944, pages 1–5.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., et al. (2022). Tackling climate change with machine learning. ACM Computing Surveys (CSUR), 55(2):1–96.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695.

Rosan, T. M., Sitch, S., O'Sullivan, M., Basso, L. S., Wilson, C., Silva, C., Gloor, E., et al. (2024). Synthesis of the land carbon fluxes of the amazon region between 2010 and 2020. Communications Earth & Environment, 5(1).

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. Ieee Access, 8:42200–42216.

Rosenfeld, D., Zheng, Y., Hashimshoni, E., Pöhlker, M. L., Jefferson, A., Pöhlker, C., Yu, X., Zhu, Y., Liu, G., Yue, Z., Fischman, B., Li, Z., Giguzin, D., Goren, T., Artaxo, P., Barbosa, H. M. J., Pöschl, U., and Andreae, M. O. (2016). Satellite retrieval of cloud condensation nuclei concentrations by using clouds as ccn chambers. Proceedings of the National Academy of Sciences, 113(21):5828–5834.

Rothfuss, Y., Quade, M., Brüggemann, N., Graf, A., Vereecken, H., and Dubbert, M. (2021). Reviews and syntheses: Gaining insights into evapotranspiration partitioning with novel isotopic monitoring methods. Biogeosciences, 18(12):3701–3732.

Rozante, J. R., Gutierrez, E. R., Fernandes, A. d. A., and Vila, D. A. (2020). Performance of precipitation products obtained from combinations of satellite and surface observations. International Journal of Remote Sensing, 41(19):7585–7604.

Rozante, J. R., Moreira, D. S., de Gonçalves, L. G. G., and Vila, D. A. (2010). Combining trmm and surface observations of precipitation: Technique and validation over south america. Weather and Forecasting, 25(3):885–894.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019). Inferring causation from time series in earth system sciences. Nature communications, 10(1):1–13.

Russell, S. J. and Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Pearson, Boston, 4 edition.

Saarela, M. and Geogieva, L. (2022). Robustness, stability, and fidelity of explanations for a deep skin cancer classification model. Applied Sciences, 12(19):9545.

Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2):1–12.

Saarela, M. and Podgorelec, V. (2024). Recent applications of explainable ai (xai): A systematic literature review. Applied Sciences, 14(19):8884.

Saeed, W. and Omlin, C. (2023). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. Knowledge-based systems, 263:110273.

Salati, E., Dall'Olio, A., Matsui, E., and Gat, J. R. (1979). Recycling of water in the amazon basin: an isotopic study. Water resources research, 15(5):1250–1258.

Saleem, M. U. and Ahmed, S. R. (2016). Missing data imputations for upper air temperature at 24 standard pressure levels over pakistan collected from aqua satellite. Journal of Data Analysis and Information Processing, 4(3):132–146.

Saleem, U., Akram, M. S., Ullah, M. F., and Rehman, F. (2018). Accurate imputation for relative humidity over pakistan gathered from aqua satellite. Open Journal of Geology, 8(10):987–1001.

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., and Düben, P. (2019). Global cloud-resolving models. Current Climate Change Reports, 5(3):172–184.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., et al. (2020). The global methane budget 2000–2017. Earth System Science Data, 12(3):1561–1623.

Sawyer, V. and Li, Z. (2013). Detection, variations and intercomparison of the planetary boundary layer depth from radiosonde, lidar and infrared spectrometer. Atmospheric environment, 79:518–528.

Scanlon, T. M. and Kustas, W. P. (2010). Partitioning carbon dioxide and water vapor fluxes using correlation analysis. Agricultural and Forest Meteorology, 150(1):89–99.

Scanlon, T. M. and Sahu, P. (2008). On the correlation structure of water vapor and carbon dioxide in the atmospheric surface layer: A basis for flux partitioning. Water Resources Research, 44(10).

Scanlon, T. M., Schmidt, D. F., and Skaggs, T. H. (2019). Correlation-based flux partitioning of water vapor and carbon dioxide fluxes: Method simplification and estimation of canopy water use efficiency. Agricultural and Forest Meteorology, 279:107732.

Schiller, J., Stiller, S., and Ryo, M. (2025). Artificial intelligence in environmental and earth system sciences: explainability and trustworthiness. Artificial Intelligence Review, 58(10):1–23.

Schlesinger, W. H. and Jasechko, S. (2014). Transpiration in the global water cycle. Agricultural and Forest Meteorology, 189:115–117.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61:85–117.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634.

Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction?, philos. In Roy. Soc. A, volume 379, pages 10–1098.

Schwalbe, G. and Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38(5):3043–3101.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. Communications of the ACM, 63(12):54–63.

See, S. and Adie, J. (2021). Challenges and opportunities for a hybrid modelling approach to earth system science. CCF Transactions on High Performance Computing, 3(3):320–329.

Seidel, D. J., Ao, C. O., and Li, K. (2010). Estimating climatological planetary boundary layer heights from radiosonde observations: Comparison of methods and uncertainty analysis. Journal of Geophysical Research: Atmospheres, 115(D16).

Seneviratne, M. G., Shah, N. H., and Chu, L. (2020). Bridging the implementation gap of machine learning in healthcare. Bmj Innovations, 6(2).

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28.

Shortt, R., Hemes, K., Szutu, D., Verfaillie, J., and Baldocchi, D. (2021). AmeriFlux US-Sne Sherman Island Restored Wetland, Ver. 7-5, AmeriFlux AMP, (Dataset). "https://doi.org/10.17190/AMF/1418684".

Silva, P. R. P., Carneiro, R. G., Moraes, A. O., Dias-Junior, C. Q., and Fisch, G. (2025). Estimating planetary boundary layer height over central amazonia using random forest. Atmosphere, 16(8):941.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489.

Silver, W. L. (2000). Global patterns in tropical forest soil phosphorus availability. Ecology, 81:95–103.

Silvestro, D., Goria, S., Sterner, T., and Antonelli, A. (2022). Improving biodiversity protection through artificial intelligence. Nature sustainability, 5(5):415–424.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Skaggs, T. H., Anderson, R. G., Alfieri, J., Scanlon, T., and Kustas, W. (2018). Fluxpart: Open source software for partitioning carbon dioxide and water vapor fluxes. Agricultural and Forest Meteorology, 253:218–224.

Sleeman, J., Halem, M., Yang, Z., Caicedo, V., Demoz, B., and Delgado, R. (2020). A deep machine learning approach for lidar based boundary layer height detection. In IGARSS 2020-2020 IEEE international geoscience and remote sensing symposium, pages 3676–3679. IEEE.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33:596–608.

Souza, C. M., Shimbo, J. Z., Rosa, M. R., et al. (2020). Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. Remote Sensing, 12(17):2735.

Souza, C. M. A., Dias-Júnior, C. Q., D'Oliveira, F. A. F., Martins, H. S., Carneiro, R. G., Portela, B. T. T., and Fisch, G. (2023). Long-term measurements of the atmospheric boundary layer height in central amazonia using remote sensing instruments. Remote Sensing, 15(13):3261.

Staal, A., Fetzer, I., Wang-Erlandsson, L., Bosmans, J. H., Dekker, S. C., van Nes, E. H., Rockström, J., and Tuinenburg, O. A. (2020). Hysteresis of tropical forests in the 21st century. Nature communications, 11(1):4978.

Staal, A., Tuinenburg, O. A., Bosmans, J. H., Holmgren, M., van Nes, E. H., Scheffer, M., Zemp, D. C., and Dekker, S. C. (2018). Forest-rainfall cascades buffer against drought across the amazon. Nature Climate Change, 8(6):539–543.

Stahl, K., Moore, R., Floyer, J., Asplin, M., and McKendry, I. (2006). Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. Agricultural and forest meteorology, 139(3-4):224–236.

Stapleton, A., Dias-Junior, C. Q., Von Randow, C., Farias D'Oliveira, F. A., Pöhlker, C., de Araújo, A. C., Roantree, M., and Eichelmann, E. (2025). Intercomparison of machine learning models to determine the planetary boundary layer height over central amazonia. Journal of Geophysical Research: Atmospheres, 130(6):e2024JD042488.

Stapleton, A., Eichelmann, E., and Roantree, M. (2022). A framework for constructing machine learning models with feature set optimisation for evapotranspiration partitioning. Applied Computing and Geosciences, 16:100105.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., De Vries, W., De Wit, C. A., et al. (2015). Planetary boundaries: Guiding human development on a changing planet. Science, 347(6223):1259855.

Stoy, P., El-Madany, T., Fisher, J., Gentine, P., Gerken, T., Good, S., Liu, S., Miralles, D., Perez-Priego, O., Skaggs, T., et al. (2019). Reviews and syntheses: Turning the challenges of partitioning ecosystem evaporation and transpiration into opportunities. biogeosciences discuss.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, 8(1):25.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3645–3650.

Stull, R. B. (2012). An introduction to boundary layer meteorology, volume 13. Springer Science & Business Media.

Su, T. and Zhang, Y. (2024). Deep-learning-derived planetary boundary layer height from conventional meteorological measurements. Atmospheric Chemistry and Physics, 24(11):6477–6493.

Summa, D., Di Girolamo, P., Stelitano, D., and Cacciani, M. (2013). Characterization of the planetary boundary layer height and structure by raman lidar: comparison of different approaches. Atmospheric Measurement Techniques, 6(12):3515–3525.

Sun, M., Li, Y., Wang, Y., and Wang, X. (2024). Towards domain-aware stable meta learning for out-of-distribution generalization. ACM Transactions on Knowledge Discovery from Data, 18(8):1–24.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S. M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W. H., Ma, X., et al. (2022). A review of earth artificial intelligence. Computers & Geosciences, 159:105034.

Taoka, T., Iwata, H., Hirata, R., Takahashi, Y., Miyabara, Y., and Itoh, M. (2020). Environmental controls of diffusive and ebullitive methane emissions at a subdaily time scale in the littoral zone of a midlatitude shallow lake. Journal of Geophysical Research: Biogeosciences, 125(9):e2020JG005753.

Teixeira, J., Piepmeier, J. R., Nehrir, A. R., Ao, C. O., Chen, S. S., Clayson, C. A., Fridlind, A. M., Lebsock, M., McCarty, W., Salmun, H., et al. (2021). Toward a global planetary boundary layer observing system: The nasa pbl incubation study team report. Toward a Global Planetary Boundary Layer Observing System: The NASA PBL Incubation Study Team Report.

ter Steege, H., Pitman, N. C., do Amaral, I. L., et al. (2023). Mapping density, diversity and species-richness of the amazon tree flora. Communications Biology, 6(1):1130.

Ter Steege, H., Pitman, N. C., Killeen, T. J., Laurance, W. F., Peres, C. A., Guevara, J. E., Salomão, R. P., Castilho, C. V., Amaral, I. L., de Almeida Matos, F. D., et al. (2015). Estimating the global conservation status of more than 15,000 amazonian tree species. Science advances, 1(10):e1500936.

Ter Steege, H., Pitman, N. C., Sabatier, D., Baraloto, C., Salomão, R. P., Guevara, J. E., Phillips, O. L., Castilho, C. V., Magnusson, W. E., Molino, J.-F., et al. (2013). Hyperdominance in the amazonian tree flora. Science, 342(6156):1243092.

Tobin, C., Nicotina, L., Parlange, M. B., Berne, A., and Rinaldo, A. (2011). Improved interpolation of meteorological forcings for hydrologic applications in a swiss alpine region. Journal of Hydrology, 401(1-2):77–89.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pages 23–30. IEEE.

Toscano, J. D., Oommen, V., Varghese, A. J., Zou, Z., Ahmadi Daryakenari, N., Wu, C., and Karniadakis, G. E. (2025). From pinns to pikans: Recent advances in physics-informed machine learning. Machine Learning for Computational Science and Engineering, 1(1):1–43.

Towell, G. G. and Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. Machine learning, 13(1):71–101.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., et al. (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. Biogeosciences, 13(14):4291–4313.

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T. F., Camps-Valls, G., Ogee, J., Verrelst, J., and Papale, D. (2020). Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. Global change biology, 26(9):5235–5253.

Trenberth, K. E., Fasullo, J. T., and Kiehl, J. (2009). Earth's global energy budget. Bulletin of the American Meteorological Society, 90(3):311–324.

Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., Parker, D., Rahimzadeh, F., Renwick, J. A., Rusticucci, M., et al. (2007). Observations: surface and atmospheric climate change. Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, pages 235–336.

Tucker, S. C., Senff, C. J., Weickmann, A. M., Brewer, W. A., Banta, R. M., Sandberg, S. P., Law, D. C., and Hardesty, R. M. (2009). Doppler lidar estimation of mixing height using turbulence, shear, and aerosol profiles. Journal of atmospheric and oceanic technology, 26(4):673–688.

Uzan, L., Egert, S., Khain, P., Levi, Y., Vadislavsky, E., and Alpert, P. (2020). Ceilometers as planetary boundary layer height detectors and a corrective tool for cosmo and ifs models. Atmospheric Chemistry and Physics, 20(20):12177–12192.

Valach, A., Szutu, D., Eichelmann, E., Knox, S., Verfaillie, J., and Baldocchi, D. (2021a). AmeriFlux US-Tw1 Twitchell Wetland West Pond, Ver. 9-5, AmeriFlux AMP, (Dataset). "https://doi.org/10.17190/AMF/1246147".

Valach, A. C., Kasak, K., Hemes, K. S., Anthony, T. L., Dronova, I., Taddeo, S., Silver, W. L., Szutu, D., Verfaillie, J., and Baldocchi, D. D. (2021b). Productive wetlands restored for carbon sequestration quickly become net co2 sinks with site-level factors driving uptake variability. PloS one, 16(3):e0248398.

van der Kamp, D. and McKendry, I. (2010). Diurnal and seasonal trends in convective mixed-layer heights estimated from two years of continuous ceilometer observations in vancouver, bc. Boundary-layer meteorology, 137:459–475.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Vilà-Guerau de Arellano, J., Ney, P., Hartogensis, O., De Boer, H., Van Diepen, K., Emin, D., De Groot, G., Klosterhalfen, A., Langensiepen, M., Matveeva, M., et al. (2020). Cloudroots: integration of advanced instrumental techniques and process modelling of sub-hourly and sub-kilometre land–atmosphere interactions. Biogeosciences, 17(17):4375–4404.

Vilà-Guerau de Arellano, J., van Heerwaarden, C. C., van Stratum, B. J., and van den Dries, K. (2015). Atmospheric Boundary Layer: Integrating Air Chemistry and Land Interactions. Cambridge University Press, Cambridge, UK.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. nature, 575(7782):350–354.

Vuichard, N. and Papale, D. (2015). Filling the gaps in meteorological continuous data measured at fluxnet sites with era-interim reanalysis. Earth System Science Data, 7(2):157–171.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology, 31:841.

Wang, H., Ciais, P., Sitch, S., Green, J. K., Tao, S., Fu, Z., Albergel, C., Bastos, A., Wang, M., Fawcett, D., et al. (2024). Anthropogenic disturbance exacerbates resilience loss in the amazon rainforests. Global Change Biology, 30(1):e17006.

Wang, S., Hu, M., Li, Q., Safari, M., and Yang, X. (2025). Capabilities of gpt-5 on multimodal medical reasoning. arXiv preprint arXiv:2508.08224.

Wang, Y., Liu, J., Wennberg, P. O., He, L., Bonal, D., Köhler, P., Frankenberg, C., Sitch, S., and Friedlingstein, P. (2023). Elucidating climatic drivers of photosynthesis by tropical forests. Global Change Biology, 29(17):4811–4825.

Williams, K., Copsey, D., Blockley, E., Bodas-Salcedo, A., Calvert, D., Comer, R., Davis, P., Graham, T., Hewitt, H., Hill, R., et al. (2018). The met office global coupled model 3.0 and 3.1 (gc3. 0 and gc3. 1) configurations. Journal of Advances in Modeling Earth Systems, 10(2):357–380.

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., and Jiang, Y.-G. (2024). A survey on video diffusion models. ACM Computing Surveys, 57(2):1–42.

Ye, J., Liu, L., Wang, Q., Hu, S., and Li, S. (2021). A novel machine learning algorithm for planetary boundary layer height estimation using aeri measurement data. IEEE Geoscience and Remote Sensing Letters, 19:1–5.

Yeh, C.-K., Kim, J., Yen, I. E.-H., and Ravikumar, P. K. (2018). Representer point selection for explaining deep neural networks. Advances in neural information processing systems, 31.

Yu, M., Huang, Q., and Li, Z. (2024). Deep learning for spatiotemporal forecasting in earth system science: a review. International Journal of Digital Earth, 17(1):2391952.

Zahn, E., Bou-Zeid, E., Good, S. P., Katul, G. G., Thomas, C. K., Ghannam, K., Smith, J. A., Chamecki, M., Dias, N. L., Fuentes, J. D., et al. (2022). Direct partitioning of eddy-covariance water and carbon dioxide fluxes into ground and plant components. Agricultural and Forest Meteorology, 315:108790.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer.

Zeng, Z., Piao, S., Li, L. Z., Wang, T., Ciais, P., Lian, X., Yang, Y., Mao, J., Shi, X., and Myneni, R. B. (2018). Impact of earth greening on the terrestrial water cycle. Journal of Climate, 31(7):2633–2650.

Zhang, S. and Chen, J. (2021). Uncertainty in projection of climate extremes: A comparison of cmip5 and cmip6. Journal of Meteorological Research, 35(4):646–662.

Zhang, Y., Sun, K., Gao, Z., Pan, Z., Shook, M. A., and Li, D. (2020). Diurnal climatology of planetary boundary layer height over the contiguous united states derived from amdar and reanalysis data. Journal of Geophysical Research: Atmospheres, 125(20):e2020JD032803.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2).

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. National science review, 5(1):44–53.

Zhuang, J., Dussin, R., Huard, D., Bourgault, P., Banihirwe, A., Raynaud, S., Malevich, B., Schupfner, M., Filipe, Jüling, A., Almansi, M., Scott, R., Bosboom, J., Frölicher, T. L., Rasp, S., Lierhammer, L., Caneill, R., Vo, H., and Bourgault, P. (2020). xesmf: Universal regridder for geospatial data. https://xesmf.readthedocs.io/. Version 0.3.0.

Zilitinkevich, S. (1972). On the determination of the height of the ekman boundary layer. Boundary-Layer Meteorology, 3(2):141–145.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. In International Conference on Learning Representations.

Zou, L., Stan, K., Cao, S., and Zhu, Z. (2023). Dynamic global vegetation models may not capture the dynamics of the leaf area index in the tropical rainforests: A data-model intercomparison. Agricultural and Forest Meteorology, 339:109562.