Faculty of Science

Novel Applications of Machine Learning in Biological Sciences

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Science (PhD): Computer Science
at the University of Antwerp

Steven Mortier

Antwerpen, 2024

Supervisors
prof. dr. ir. Steven Latré
prof. dr. Tim Verdonck

Jury

Chairman
prof. dr. ir. José Antonio Oramas Mogrovejo, University of Antwerp, Belgium

Supervisors
prof. dr. ir. Steven Latré, University of Antwerp, Belgium
prof. dr. Tim Verdonck, University of Antwerp, Belgium

Members
prof. dr. Bjarni D. Sigurdsson, Agricultural University of Iceland - AUI, Iceland
prof. dr. ir. Erik Mannens, University of Antwerp, Belgium
dr. Véronique Van Vlasselaer, SAS, Belgium

Contact
Steven Mortier
University of Antwerp
Faculty of Science
Department of Computer Science
IDLab Research Group
Middelheimlaan 1, 2020 Antwerp, Belgium
M: steven.mortier@uantwerpen.be

Dutch title:

Vernieuwende Toepassingen van Machinaal Leren in Biologische Wetenschappen

Acknowledgements

A fella like me can't make it alone, you know?
- Ranwid

Completing this PhD thesis has been a tough yet rewarding journey, and I would like to take this opportunity to acknowledge the support and contributions of several individuals that have played a crucial role in making it possible.

First, I would like to express gratitude to Steven for giving me the opportunity to pursue a PhD, and his support throughout the years. My heartfelt thanks also go to Jannes for his excellent guidance and insightful feedback during my first year. When you left for EPFL, I briefly found myself without a daily advisor, and that's when Tim stepped in. Tim, despite your already busy schedule, you graciously "adopted" me as your PhD student. Your detailed —sometimes even handwritten— feedback, unwavering guidance, and continuous support have ultimately led to the successful completion of this PhD journey. You put me on a clear track with detailed objectives, and for that I will be forever grateful.

I would also like to thank the members of my PhD jury, for taking the time to read my thesis and attend the defense. Thank you for your stimulating questions, thoughtful comments and constructive feedback. Their input has significantly enhanced the quality of my research.

During my PhD, I had the pleasure of being part of the IDLab research group. It would be impossible to thank each and every one of you here for your feedback, inspiring conversations, and all-round good atmosphere at The Beacon. However, I would like to especially thank a few colleagues that were fundamental to the completion of my PhD track. Thank you, Bart, Benjamin, Fer, Iris, Laurens, Leonid, Lucas, Lynn, Renata, Sarah, Thomas D., Thomas *Tommy* S., Tom D.S., Tom V.D., and Vaidehi.

On a personal note, I would like to thank a group of friends that are very special to me, some of whom I've known since I was merely six years old. Thank you, Christiaan, Jelle, *Pally*, Quinten, Reinier, Simon, and Thomas, for being an amazing group of friends and for your never-ending support.

I would also like to thank Johan and Ine, for their support during the past years. Thank you also to Fiebe (and Fons), with whom I have lived together for the majority of my PhD. Thank you for all the fun dinners and board-game escape rooms.

I am deeply grateful to my parents, brother, and sister. Without them, none of this would have ever been possible. Thank you for enabling me to do what I like to do and supporting me no matter what.

As the saying goes, save the best for last. I would like to express my deepest gratitude to my girlfriend, Luna, for her unwavering love, patience, and support throughout this journey. Thank you for providing a listening ear when times were difficult and motivating me to keep going when I was struggling. We have been on our journey for 10 years now, and I would like to think that it is still only the beginning.

Steven Mortier
Antwerp, September 2024

Samenvatting

In de afgelopen jaren heeft de snelle toename van beschikbare data, samen met significante algoritmische vooruitgangen, ons vermogen om deze data te analyseren en er waardevolle inzichten uit te halen aanzienlijk vergroot. In dit proefschrift presenteren we nieuwe toepassingen van machinaal leren (ML) in de biologische wetenschappen, met een specifieke focus op neurowetenschappen, ecologie, en landbouwkundig onderzoek. Verder leggen we in dit proefschrift nadruk op de interpreteerbaarheid en verklaarbaarheid van onze modellen, zodat we de redenering achter hun voorspellingen begrijpen.

Het proefschrift is onderverdeeld in vier delen, elk uitgewerkt in een afzonderlijk hoofdstuk. In Hoofdstuk 2 behandelen we een binaire classificatietaak gericht op het detecteren van aandacht in elektro-encefalografie (EEG) data. Door gebruik te maken van geavanceerde ML-modellen, slagen we erin om onderscheid te maken tussen doel- en afleidingsstimuli, gebaseerd op de EEG-data die zijn verzameld tijdens een audiovisuele aandachtstaak. Daarnaast onderzoeken we in hoeverre EEG-data afhankelijk is van de proefpersoon bij dezelfde stimulus. Dit doen we door de prestaties van ML-modellen getraind op individuele proefpersonen te vergelijken met ML-modellen getraind op meerdere proefpersonen. Ten slotte passen we verklaarbare kunstmatige intelligentietechnieken toe om de kenmerken te identificeren die door de modellen worden gebruikt voor hun voorspellingen, en vinden we dat deze kenmerken overeenkomen met de verwachtingen van domeinexperts.

Hoofdstuk 3 richt zich op de relatie tussen bodemtemperatuur, verschillende meteorologische variabelen en fenologische kenmerken van de vegetatie. Door gebruik te maken van ML en verklaarbare kunstmatige intelligentie hebben we vastgesteld dat hogere bodemtemperaturen leiden tot een vroeger begin van het groeiseizoen voor planten. Onze analyse onthult bovendien dat meteorologische variabelen de meest significante invloed hebben op de fenologische kenmerken van de vegetatie, terwijl jaarlijkse variaties voornamelijk bepaald worden door veranderingen in de bodemtemperatuur.

In Hoofdstuk 4 bespreken we het probleem van ontbrekende waarden in sensorgegevens, waarbij we nadruk leggen op grootschalige draadloze sensornetwerken. We evalueren twaalf methoden voor het invullen van ontbrekende gegevens, die elk verschillende strategieën gebruiken en afkomstig zijn van verschillende achtergronden. Om het evaluatieproces te verbeteren, introduceren we een scenario dat gebruik maakt van "gemaskeerde ontbrekende gegevens". Dit laat een realistischere evaluatie toe dan de gebruikelijke methode, waarbij willekeurige ontbrekende gegevens worden gebruikt. Onze bevindingen wijzen erop dat methoden voor het invullen van ontbrekende gegevens over het algemeen het beste resultaat opleveren wanneer ze expliciet rekening houden met ruimtelijke correlaties. We gebruiken deze inzichten om richtlijnen voor toekomstig onderzoek voor te stellen.

Tot slot ontwikkelen we in Hoofdstuk 5 een nieuwe, wereldwijde dataset over het historisch gebruik van meststoffen. Hier gebruiken we ML om historische bemesting te voorspellen op zowel gewas- als landniveau. We baseren deze voorspellingen op een reeks variabelen die gerelateerd zijn aan verschillende klassen van gewassen, evenals sociaaleconomische, agrologische, en omgevingsvariabelen. Verder gebruiken we verklaarbare kunstmatige intelligentie om de meest invloedrijke factoren die het gebruik van meststoffen bepalen te identificeren.

Summary

In recent years, the rapid increase in available data, accompanied by significant algorithmic advancements, have enhanced our ability to analyze and extract valuable insights from this data. In this thesis, we present novel applications of machine learning (ML) in biological sciences, more particular in neuroscience, ecology, and agricultural research. Throughout the thesis, we pay special attention to the interpretability and explainability of our models, ensuring that we understand the reasoning behind their predictions.

The thesis is divided into four parts, each presented in a separate chapter. In Chapter 2, we address a binary classification task aimed at detecting attention in electroencephalography (EEG) data. By employing state-of-the-art ML models, we successfully differentiate between target and distractor stimuli using EEG data collected during an audiovisual attention task. Additionally, we examine subject dependence in EEG data by comparing the performance of ML models trained on individual subjects versus multiple subjects. Finally, we apply explainable AI (xAI) techniques to identify the features utilized by the models for their predictions, and find that these features align with the expectations of domain experts.

Chapter 3 focuses on the relationship between soil temperature, various meteorological variables, and vegetation phenology characteristics. Using ML and xAI, we find that rising soil temperatures result in an earlier onset of the growing season for plants. Additionally, our analysis reveals that the meteorological variables have the most significant impact on the vegetation phenology characteristics, while annual variations are primarily driven by changes in the soil temperature.

In Chapter 4, we address the issue of missing values in sensor data, with a focus on large-scale wireless sensor networks (WSNs). We evaluate twelve missing value imputations methods, each using different imputation strategies and originating from diverse backgrounds. To enhance the evaluation process, we define a "masked missings" scenario, offering a more realistic assessment compared to the standard practice of using random missings. Our findings indicate that imputation methods explicitly accounting for spatial correlations between sensors generally perform best, and we use these insights to suggest directions for future research.

Finally, in Chapter 5, we develop a new global dataset considering the historical application of fertilizers. Specifically, we use ML to predict historical fertilizer application at both the crop and country levels, based on a set of features related to crop classes, along with socioeconomic, agrological, and environmental variables. Additionally, we use xAI to identify the most relevant drivers influencing fertilizer application.

List of Publications

The following publications are part of this thesis:

- **Steven Mortier**, Renata Turkeš, Jorg De Winne, Wannes Vanransbeeck, Dick Botteldooren, Paul Devos, Steven Latré, Marc Leman, Tim Verdonck. *Classification of Targets and Distractors in an Audiovisual Attention Task based on Electroencephalography*. Sensors. 2023; 23(23):9588. https://doi.org/10.3390/s23239588
- **Steven Mortier**, Amir Hamedpour, Bart Bussmann, Ruth Phoebe Tchana Wandji, Steven Latré, Bjarni D. Sigurdsson, Tom De Schepper, Tim Verdonck. *Inferring the relationship between soil temperature and normalized difference vegetation index with machine learning*. Ecological Informatics. 2024. 82:102730. https://doi.org/10.1016/j.ecoinf.2024.102730
- **Steven Mortier**, Thomas Decorte, Jonas J. Lembrechts, Filip J. R. Meysman, Steven Latré, Erik Mannens, and Tim Verdonck. *Missing Value Imputation of Wireless Sensor Data for Environmental Monitoring*. Sensors. 2024; 24(8):2416. https://doi.org/10.3390/s24082416
- Fernando Coello, Thomas Decorte, Iris Janssens, **Steven Mortier**, Jordi Sardans, Josep Peñuelas, and Tim Verdonck. *Global Crop-Specific Fertilization Dataset from 1961-2019*. 2024. Under review.

The following publications were part of my research, but are not explicitly covered in this thesis:

- Tullia Calogiuri, Mathilde Hagens, Jan Willem Van Groenigen, Thomas Corbett, Jens Hartmann, Rick Hendriksen, Iris Janssens, Ivan A Janssens, Guillermo Ledesma Dominguez, Grant Loescher, **Steven Mortier**, Anna Neubeck, Harun Niron, Reinaldy Pratama Poetra, Lukas Rieder, Eric Struyf, Michiel Van Tendeloo, Tom De Schepper, Tim Verdonck, Siegfried E Vlaeminck, Sara Vicca, Alix Vidal. *Design and Construction of an Experimental Setup to Enhance Mineral Weathering through the Activity of Soil Organisms*. JoVe. 2023. https://doi.org/10.3791/65563
- Renata Turkeš, **Steven Mortier**, Jorg De Winne, Dick Bottelooren, Paul Devos, Steven Latré and Tim Verdonck, *Who is WithMe? EEG features for attention in a visual task, with auditory and rhythmic support*. 2024. Under review.
- Iris Janssens, Thomas Servotte, **Steven Mortier**, et al., Steven Latré and Tim Verdonck. *Machine learning as a tool to identify dominant drivers of carbon dioxide removal in enhanced weathering experiments*. 2024. Close to submission.

Furthermore, research conducted throughout this doctoral program was presented at international conferences:

- **Steven Mortier**, Bart Bussmann, Tim Verdonck. 2022. *PRINCE Out-of-Distribution Challenge Winning Solution: a Robust Neural Network Ensemble.* Oral invited presentation: ECML-PKDD 2022, Grenoble, France.
- **Steven Mortier**, Renata Turkeš, Jorg De Winne, Wannes Vanransbeeck, Dick Botteldooren, Paul Devos, Steven Latré, Marc Leman, Tim Verdonck. 2023. *Classification of Targets and Distractors in an Audiovisual Attention Task based on Electroencephalography.* Oral contributed presentation: DSSV-ECDA 2023, Antwerp, Belgium.
- **Steven Mortier**, Amir Hamedpour, Bart Bussmann, Ruth Phoebe Tchana Wandji, Steven Latré, Bjarni D. Sigurdsson, Tom De Schepper, Tim Verdonck. 2023. *Inferring the relationship between soil temperature and normalized difference vegetation index with machine learning*. Oral invited presentation: CMStatistics 2023, Berlin, Germany.
- Thomas Decorte, **Steven Mortier**, Christian Suys, Tim Verdonck. 2023. *Missing value imputation of sensor data for environmental monitoring*. Oral invited presentation: CMStatistics 2023, Berlin, Germany.

Contents

List of Figures

List of Tables

List of Acronyms

AI artificial intelligence

AKE applying k-nearest neighbor estimation

ALS amyotrophic lateral sclerosis

ANN artificial neural network

AUC area under the curve

BAM! Bio-Accelerated Mineral weathering

BCI brain computer interface

BERT Bidirectional Encoder Representations from Transformers

BO bayesian optimization

BRITS bidirectional recurrent imputation for time series

CEC cation exchange capacity

CNidT CurieuzeNeuzen in de Tuin; Nosy Parkers in the Garden

CNN convolutional neural network

CPU central processing unit

CS cross-subject

CV cross validation

DAE denoising autoencoder

DEFRA Department for Environment, Food & Rural Affairs

DESM data estimation using statistical model

DL deep learning

DNN deep neural network

DS data science

EEG electroencephalography

ERP event-related potential

EU European Union

FAMD Factor Analysis of Mixed Data

FAO Food and Agriculture Organization

FE Fertilizer Europe

FN false negatives

FP false positives

FUBC fertilizer use by crop

GDP gross domestic product

GPT generative pre-trained transformer

GPU graphics processing unit

HGB HistGradientBoosting

ICA independent component analysis

ICC Indicative Crop Classification for agricultural census

IFA International Fertilizer Association

IFDC International fertilizer Development Center

IoT Internet of Things

IS individual subject

KNN k Nearest Neighbours

LCU local currency units

LIME Local Interpretable Model-Agnostic Explanations

LLM large language model

LOO leave-one-out

M-RNN multi-directional recurrent neural network

MAE mean absolute error

MAP mean annual precipitation

MAPE mean absolute percentage error

MAR missing at random

MAT mean annual temperature

MC matrix completion

MCAR missing completely at random

MCMC Markov chain Monte Carlo

MI multiple imputation

MICE multiple imputation using chained equations

MIDA multiple imputation using denoising autoencoders

ML machine learning

MLP multilayer perceptron

MMREx minimax risk extrapolation

MNAR missing not at random

MSE mean squared error

NB-IoT Narrowband Internet of Things

NDVI Normalized Difference Vegetation Index

NETs negative emission technologies

NLP natural language processing

NN neural network

OCS organic carbon stock

OHE one-hot encoded

OOD out-of-distribution

PC Principal Components

PCE prediction coverage error

PEAK maximum annual NDVI value

PET potential evapotranspiration

POS peak of the season

PRINCE PRomoting INvariance for Classification of browsing journeys across Environments

REx risk extrapolation

RF random forest

RL reinforcement learning

RMSE root-mean-square error

RNN recurrent neural network

ROC receiver operating characteristic curve

ROC AUC area under the receiver operating characteristic curve

Rocket RandOm Convolutional KErnel Transform

SFR Social Federal Republic

SHAP SHapley Additive exPlanations

SOC soil organic carbon

SOS start of the season

SVM support vector machine

TN true negatives

TP true positives

UK United Kingdom

USA United States of America

USD United States dollars

USDA United States Department of Agriculture

USSR Union of Soviet Socialist Republics

VREx variance risk extrapolation

WSN wireless sensor network

xAI explainable AI

XGB eXtreme Gradient Boosting or XGBoost

Chapter 1

Introduction

This chapter situates the conducted research within its broader context. We begin by providing clarity regarding important terminology, followed by a brief history of machine learning (ML). Next, we give an overview of the research advances and applications of ML in biological sciences, after which we provide an important note on interpretable and explainable ML. We then clarify the structure of the remainder of this thesis, and briefly introduce each research chapter by highlighting its main contributions. Finally, we give an overview of other conducted research that is not explicitly covered within this thesis.

1.1 Machine learning

1.1.1 Terminology

In the midst of the current artificial intelligence (AI) boom, it can be difficult to distinguish between terms like AI, ML, deep learning (DL), and data science (DS). AI refers to the development of machines, usually computer systems, capable of intelligent behavior [1]. Whereas AI can include pre-defined rule-based systems, ML is a subset of AI that focuses on algorithms and (statistical) models that allow computer systems to perform specific tasks without explicit instructions. Instead, they rely on patterns learned from data to make predictions or decisions [2]. DL is the subset of ML that focuses on a specific type of models, namely, neural networks (NNs) with many layers, otherwise called deep neural networks (DNNs) [2]. DL is particularly useful for tasks that involve large-scale data, and require the model to uncover complex patterns, such as image recognition, natural language processing (NLP), and audio recognition. Finally, DS is an interdisciplinary field that uses scientific methods and algorithms to extract knowledge and insights from structured and unstructured data. Contrary to the three previous fields, DS is not confined to constructing models, as it includes the entire data processing pipeline, including data gathering, cleaning, exploration, modeling, and inference [3]. A summary of the intersections between these various fields is included in Figure 1.1.

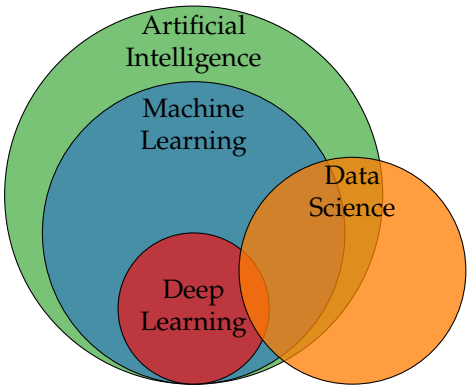


Figure 1.1: Relationships and intersections between artificial intelligence, machine learning, deep learning, and data science. Figure based on [4].

1.1.2 A brief history of machine learning

In this thesis, we will focus more on ML, and specifically its applications in biological sciences. During the last decade, there have been massive advancements in the field of ML. However, Alan Turing already set the stage for thinking about intelligent machines in 1950, as he proposed the now-called Turing Test as a criterion of machine intelligence [5]. The first instance of ML was created in 1952 by Arthur Samuel, as he developed a program that could learn to play checkers. However, it was not until 1959 that he formally introduced the term "machine learning" to describe this learning process [6]. In 1958, psychologist Frank Rosenblatt developed the perceptron, an early NN that could learn patterns from data, and eventually became the foundations for modern NNs [7]. Initially, research into perceptrons experienced a significant setback, largely due to the influential book *Perceptrons* [8]. This book emphasized the limitations of perceptrons, particularly their inability to learn nonlinear relationships. Although it was recognized by researchers that multilayer perceptrons (MLPs) did not suffer this issue, the methods for effectively training such MLPs were unknown at the time. For this, backpropagation was needed, which implements a methodology for the backwards propagation of errors from the output layer (the layer where the final prediction is made) back to the input layer (the layer to which the input data is provided). The modern version of backpropagation was first proposed in a master thesis written by Seppo Linnainmaa in 1970 [9]. However, during the 70s, the field experienced an *AI winter*, during which both funding and interest for AI research was reduced. It was not until 1986 that a paper by Rumelhart, Hinton, and Williams popularized the use of the backpropagation algorithm for training MLPs. From this point on, ML research experienced significant growth and expansion, leading to other learning algorithms such as decision trees [10] and support vector machines (SVMs) [11], but also more complex NN architectures such as convolutional neural networks (CNNs) [12].

In the early 2000s, DL was popularized, emphasizing deeper NNs and eventually leading to the AI boom. Major milestones include AlexNet in 2012 [13], as the winning solution of the ImageNet competition, and AlphaGo [14], which defeated human champions in the game of Go using reinforcement learning (RL), thereby inspiring a new era of AI systems.

In 2017, the disruptive paper *Attention is all you need* was published, which introduces the modern transformer architecture [15]. This transformer architecture led to major breakthroughs in NLP, including large language models (LLMs) such as Bidirectional Encoder Representations from Transformers (BERT) [16] and generative pre-trained transformers (GPTs) [17]. More recently, OpenAI publicly launched ChatGPT, a chatbot based on LLMs [18]. ChatGPT is credited as one of the catalysts for the most recent AI boom, which has resulted in sustained investment in the field of (generative) AI. Additionally, public attention and interest for AI techniques has also increased massively, partly due to the enhanced accessibility of AI technologies for the general public. As AI research continues at a rapid pace, new applications in industry and research are discovered and implemented every day, further driving innovation and creating exciting possibilities across various sectors.

1.2 Machine learning for biological sciences

In this thesis, we focus on novel applications of ML in biological sciences. The term "biological sciences" refers to any scientific discipline that considers the study of life and living organisms, their life cycles, adaptations, and the environment. The exponential growth in peer-reviewed publications exploring the use of ML methods in biological sciences, as illustrated in Figure 1.2, underscores the potential of ML in this broad field. Indeed, ML enables researchers to analyze vast datasets, predict behaviors and interactions, and ultimately gain deeper insights into the mechanisms that govern life at all scales. Below, we describe various applications of ML in the broad field of biological sciences, with a focus on neuroscience, ecology and agricultural research.

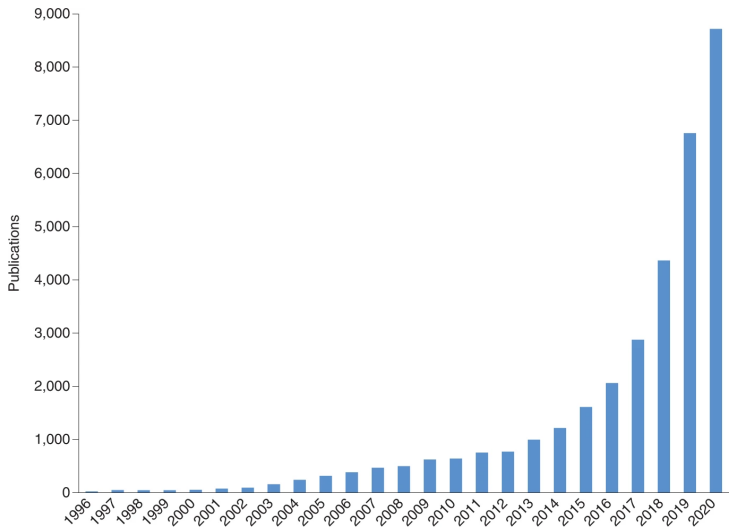


Figure 1.2: Exponential increase of machine learning (ML) publications in biology. The number of ML publications per year is based on Web of Science from 1996 onwards using the topic category for "machine learning" in combination with each of the following terms: "biolog*", "medicine", "genom*", "prote*", "cell*", "post translational", "metabolic" and "clinical". Source: [19].

1.2.1 Health sciences

In the rapidly evolving field of health sciences, ML plays a pivotal role by opening up new avenues for advancements in diagnostics, drug discovery and personalized medicine, among others. For example, ML can be used to automatically detect various types of cancer, based on different medical imaging modalities [20]. In drug discovery, ML can be of great value by predicting the 3D structures of proteins from their amino acid sequence (also known as protein folding) [21, 22], or by generating new molecules with desirable properties using graph generative models [23]. Furthermore, advances in causal ML can enable a personalized medicine approach by estimating individualized treatment effects [24].

ML also has a wide range of applications in neuroscience, which considers the study of the nervous system, from structure to function [25]. For example, ML can be used to detect and predict seizures in patients with epilepsy, based on electroencephalography (EEG) data [26], allowing for timely intervention and management. Additionally, ML can aid in the diagnosis of various neurological disorders such as Alzheimer's disease [27, 28], ADHD [29], and other neurological disorders [30]. Automatically detecting specific responses in the brain, so-called event-related potentials (ERPs), using ML, also has a wide range of applications. Primary examples of such applications are P300-based spellers [31–33] and intelligent home control systems [34, 35]. These applications can be of great help for patients suffering from amyotrophic lateral sclerosis (ALS) or spinocerebellar ataxia, as it can enable them to communicate in a daily environment [31, 33, 36, 37].

1.2.2 Biology

The ability of ML algorithms to analyze and retrieve patterns from data also facilitates biological research by providing nuanced insights into biodiversity, ecosystem dynamics, and the impact of climate change, to name a few. Specific applications include the automated detection and classification of (plant) species, thereby supporting phenology research [38–40]. Additionally, ML is helping in the fight against climate change [41], for example by improving energy efficiency using smart electricity grids [42], accelerating carbon capture, utilization, and storage research [43], enhancing the forecasting of extreme events [44], modeling climate (change) using physics-informed ML [45], and reducing agriculture-related emissions by enabling precision agriculture [46]. Within agriculture, ML is also used to perform fertilizer research [47] and optimize fertilization strategies [48]. In ecology, which studies the relationships among living organisms, and their environments [49], ML has proven to be of great value in areas such as the ecological modeling of species distribution [50, 51], ecosystem monitoring and management using interpretable ML models [52], and (benthic) habitat mapping [53]. In summary, the effective application of ML algorithms enables biologists to not only conduct comprehensive ecological assessments, but also predict future environmental and climatic conditions, thereby serving a crucial role in the study and preservation of the environment.

1.3 Interpretable and explainable machine learning

Whereas ML offers a wide range of applications across various domains, a significant challenge that is often encountered is the lack of interpretability of many ML models [54]. This is especially important when working together with domain experts (e.g., biologists) or individuals with no ML experience, as a lack of interpretability can make it difficult to build trust in the model's predictions and results. Without a clear understanding of how an ML model arrives at its conclusions, domain experts may find it challenging to validate the model's outputs, integrate them into their existing workflows, or provide valuable feedback for further model refinement [55].

Unfortunately, some models, especially complex models like deep NNs, operate as so-called "black boxes", where the decision-making processes are not readily transparent or understandable [56]. In response to this, the concepts of interpretable and explainable ML (more often referred to using the broader term explainable AI (xAI)) have become increasingly important [57]. While interpretable ML and xAI are often used interchangeably, it is important to distinguish between these two concepts. Interpretable ML refers to models that are inherently interpretable, that is, the mechanisms that occur in the input-to-output transformation are clear and understandable, for example, linear regression models or decision trees [57]. A significant disadvantage of interpretable models is that they are often less powerful compared to more complex models, which can limit their effectiveness in handling large or highly complex datasets [58]. On the other hand, xAI techniques provide insights into the workings of complex and less transparent models, enabling end-users to understand and trust the decisions made by these systems [57].

Among the various xAI techniques, SHapley Additive exPlanations (SHAP) values [59] have gained significant popularity for their ability to provide consistent and interpretable insights into model predictions [60, 61]. SHAP values are grounded in cooperative game theory and offer a unified approach to attributing feature importance. It uses an approach that measures each player's contribution to the final outcome. In ML, this means that each feature is assigned an importance value, quantifying its impact on the model's output. The contributions are additive, with the sum of all contributions of all features equaling the difference between the model's prediction for a specific instance and the average prediction across the training set [59]. This additivity property is crucial for generating intuitive explanations, and ultimately facilitates a better understanding and trust in the model's predictions. Moreover, SHAP values offer both local and global explainability [59]. Locally, they explain individual predictions, making it clear how each separate feature influences a specific model prediction. Globally, SHAP provides insights into overall feature importance and model behavior by aggregating SHAP values across many predictions. Additionally, SHAP values are model-agnostic, meaning that they can be applied to any ML model [59]. This allows for consistent interpretability across different types of models, providing an easy way to compare feature importance and model behavior. Finally, SHAP adheres to the missingness property, ensuring that features which have no impact on the model's predictions (or that are missing in the dataset) are assigned a SHAP value of zero [59].

Stimulated by the critical importance of transparency in ML models, as well as our close collaboration with domain experts, the following chapters of this work are committed to employing interpretable ML models or, when this is not feasible, to applying xAI techniques to our models. This approach ensures that our methodologies remain clear

and understandable. More specifically, given the properties and advantages of SHAP values discussed above, we will frequently use SHAP values throughout this thesis.

However, we should exercise caution when interpreting results from xAI techniques. While these methods offer valuable insights into a model's decision-making process, they can sometimes oversimplify complex behaviors [62]. For instance, xAI techniques may highlight certain features or patterns as highly influential to a model's predictions, but this does not necessarily mean they capture the full scope of the underlying relationships [63]. Moreover, xAI methods typically do not identify causal relationships [63]. Furthermore, different xAI techniques can yield varying explanations for the same model, which can lead to misinterpretations if not critically evaluated. To avoid misleading interpretations, it is essential to combine xAI insights with domain expertise and critical thinking.

1.4 Overview of research and contributions

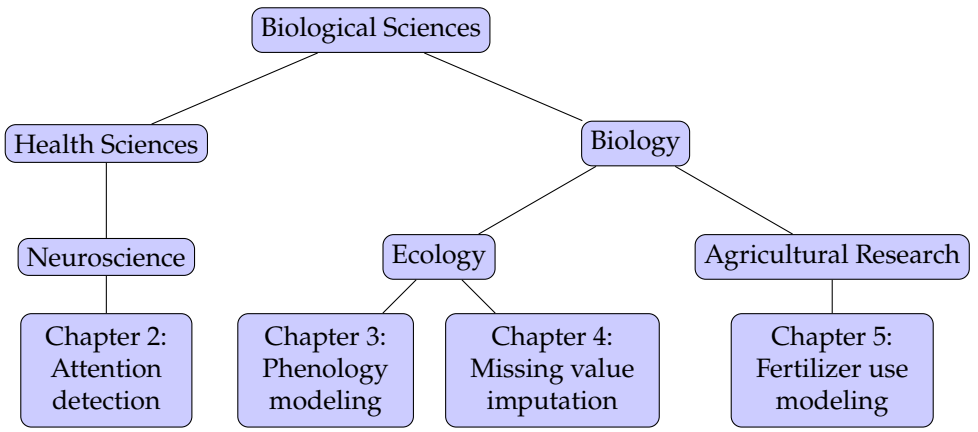


Figure 1.3: Schematic representation of the structure of this thesis, illustrating the breakdown of biological sciences into specialized research domains.

Figure 1.3 provides a schematic representation of the structure of this thesis, and details how all research contributions are integrated under the theme "ML applications for biological sciences". Below, we provide a brief overview of every chapter, highlighting its goals and key findings.

In Chapter 2, we build a model to accurately differentiate between target and distractor stimuli in an audiovisual attention task. This classification is entirely based on the subject's EEG data that was collected during the task. The main contributions of this chapter are the following:

- We use state-of-the-art ML models to classify EEG signals associated with attention-related brain activity.
- We study the difference in performance between models trained on individual

subjects and models trained on multiple subjects, emphasizing the challenges associated with generalizing models across different subjects.

- Using xAI, we investigate which EEG channels and time points the model relies on for its predictions, and verify whether these findings correspond with the expectations of domain experts.

In the end, this study demonstrated the effectiveness of ML to detect attention by successfully differentiating between targets and distractors. Furthermore, we found that the models used patterns in the EEG signals that aligned with the expectations of the domain experts to make their predictions.

This paper was published in Sensors:

Mortier, Steven, Renata Turkeš, Jorg De Winne, Wannes Van Ransbeeck, Dick Botteldooren, Paul Devos, Steven Latré, Marc Leman, and Tim Verdonck. Classification of Targets and Distractors in an Audiovisual Attention Task Based on Electroencephalography. *Sensors* 23, no. 23: 9588. 2023 [64]

In Chapter 3, we use ML techniques and SHAP values to investigate the relationship between soil temperature, a selection of meteorological variables, and vegetation phenology in subarctic grasslands. The findings from this research enhance our understanding of the mechanisms driving ecosystem dynamics in these regions and have implications for predicting and managing subarctic grasslands in the face of environmental change. More specifically, the following contributions were made:

- We derive three key parameters considering vegetation phenology based on Normalized Difference Vegetation Index (NDVI) data: the start of season, peak of season, and maximum annual NDVI value.
- We improve upon existing research methodologies by also taking nonlinear effects into account. Specifically, instead of linear models, we use ML models capable of modeling nonlinear relationships to predict the aforementioned key parameters.
- We extend previous research by taking a set of three meteorological variables into account: the air temperature, solar irradiance, and precipitation.

In summary, this study confirmed that previous research findings, indicating that the start of the season occurs earlier as soil temperature increases, also hold over longer time scales. In addition, considering meteorological variables such as precipitation, we discovered their significant influence on vegetation dynamics. As a result, we recommend that future research on vegetation dynamics incorporates these meteorological variables.

This paper was published in Ecological Informatics:

Mortier, Steven, Amir Hamedpour, Bart Bussmann, Ruth Phoebe Tchana Wandji, Steven Latré, Bjarni D. Sigurdsson, Tom De Schepper, and Tim Verdonck. Inferring the Relationship between Soil Temperature and the Normalized Difference Vegetation Index with Machine Learning. *Ecological Informatics*, 82:102730. 2024 [65]

In Chapter 4, we analyze methods for missing value imputation in environmental monitoring datasets, collected using wireless sensor networks (WSNs). Our analysis focuses on a comprehensive dataset from a large-scale citizen science project, using low-cost sensors to record microclimate data. The key contributions of our research include:

- We assess twelve different imputation methods, ranging from simple statistical methods to complex DL methods, and focusing on different types of features.
- We introduce a "masked missing" scenario to get an evaluation of the methods that is more representative of a real-world scenario.

In summary, this study effectively addressed the challenge of imputing missing values in large-scale sensor data. In general, methods focusing on spatial correlations outperformed methods that focused solely on temporal correlations.

This paper was published in Sensors:

Mortier, Steven, Thomas Decorte, Jonas J. Lembrechts, Filip J. R. Meysman, Steven Latré, Erik Mannens, and Tim Verdonck. Missing Value Imputation of Wireless Sensor Data for Environmental Monitoring. *Sensors* 24, no. 8: 2416. 2024 [66]

In Chapter 5, we use ML to develop a comprehensive global dataset detailing crop-specific fertilization rates, on country level. The main contributions are the following:

- We extend upon the previously most comprehensive database [67] by adding new global datasets, country-specific datasets, and extensive crop class related data. Additionally, we add socioeconomic, environmental, and agricultural variables that are identified as possible drivers of cropland fertilization.
- Using ML, we predict the N, P_2O_5 and K_2O fertilizer application rate for different crops over 60 years, thereby filling gaps in the dataset.
- We apply xAI to identify global socioeconomic, agricultural, and environmental drivers related to fertilizer intensity for different crops.
- After taking the fraction of fertilizer use allocated to grasslands and fodder crops into account, we adjust the ML predictions to align with the total annual country-level N, P_2O_5 and K_2O use in agricultural land.

This study resulted in a global dataset of fertilizer application rates, including a tabular dataset and high-resolution maps. Thereby, we provided a valuable open-source resource for analyzing historical fertilizer consumption across different crops, countries, and fertilizer types.

This paper is currently under review:

Coello, Fernando Coello, Thomas Decorte, Iris Janssens, Steven Mortier, Jordi Sardans, Josep Peñuelas, and Tim Verdonck. Global Crop-Specific Fertilization Dataset from 1961-2019. *Under Review*. 2024 [68]

Finally, in Chapter 6, we conclude the thesis by summarizing our key findings, describing the main limitations, and suggesting interesting directions for possible future research.

1.5 Contributions not included in this thesis

Over the course of my PhD, I have made other significant contributions that are not explicitly covered within the following chapters. These contributions are briefly summarized below.

1.5.1 EEG features for attention

In Chapter 2, we successfully built a model to classify target and distractor stimuli in an audiovisual attention task. Following this, we conducted further research into which EEG data representations or features are most closely linked to attention, and assess their robustness with respect to cross-subject variability. We explored various EEG features, including time-domain and recurrence plots obtained from univariate time series, as well as representations obtained directly from multivariate time series, such as global field power or functional brain networks. Additionally, to address the cross-subject variability in EEG data, we investigated persistent homology features that are robust to different types of noise. These contributions are formalized in the following publication:

Renata Turkeš, Steven Mortier, Jorg De Winne, Dick Bottelooren, Paul Devos, Steven Latré and Tim Verdonck. Who is WithMe? EEG features for attention in a visual task, with auditory and rhythmic support. *Under review.* 2024.

1.5.2 Bio-accelerated mineral weathering

During my PhD, I was part of the Bio-Accelerated Mineral weathering (BAM!) project [69], a multidisciplinary Horizon2020 project that aims to develop new negative emission technologies (NETs). NETs allow to actively sequester CO_2 from the atmosphere, in an effort to counterbalance climate change. More specifically, the BAM!-project brings the most important natural CO_2 sequestering process on geological timescales, silicate weathering, to a controlled reactor environment [69]. This reactor combines the weathering power of prokaryotes, plants, fungi, and invertebrates in order to maximize CO_2 removal. Using ML, we identified the dominant drivers of CO_2 removal, providing insights that can be used to further enhance the process. Additionally, we designed an optimal experiment design methodology, using bayesian optimization (BO) to intelligently select combinations for subsequent batches, based on prior insights. The BAM!-project is still in progress but has already produced the following publications:

- Calogiuri, Tullia, Mathilde Hagens, Jan Willem Van Groenigen, et al., Steven Mortier, et al., Alix Vidal. Design and Construction of an Experimental Setup to Enhance Mineral Weathering through the Activity of Soil Organisms. *JoVE (Journal of Visualized Experiments)* 2023 (11 2023): e65563. [70]
- Iris Janssens, Thomas Servotte, Steven Mortier, et al., Steven Latré, and Tim Verdonck. Machine learning as a tool to identify dominant drivers of carbon dioxide removal in enhanced weathering experiments. *Close to submission*. 2024.

1.5.3 PRINCE Out-of-Distribution Challenge

I participated in the PRomoting INvariance for Classification of browsing journeys across Environments (PRINCE) challenge, organized as part of ECML-PKDD 2022 [71]. The purpose of the challenge was to use training data to learn a robust prediction function in a binary classification task, with a special focus on out-of-distribution (OOD) generalization. The data consisted of categorical features aggregated from traces in computational advertising. Participants had access to data from three training environments, while solutions were evaluated on nine hidden testing environments. Here, an environment is defined as a distinct subset of the dataset that may feature unique underlying patterns that are not necessarily present in other environments. This task is challenging for existing methods that usually focus on correlations to make a prediction, as some interesting shifts (covariate/diversity shift and concept/correlation shift) existed between these environments. The competition had two prizes: a generalization prize, which was awarded to the solution that achieved the best performance on the test set, and a robustness prize, which was awarded to the solution that achieved the "best worst performance across the test environments". Using MLPs with a modified loss function focused on OOD generalization, we won both prizes. Specifically, we used risk extrapolation (REx) as introduced by Krueger et al. [72], which has two main goals:

1. Reducing training risks
2. Increasing similarity of training risks

As shown by [72], sacrificing performance in *individual* environments in order to perform more similarly *across* environments can help with OOD generalization. We use two implementations of REx: minimax risk extrapolation (MMREx) and variance risk extrapolation (VREx). The loss function corresponding to MMREx is as follows:

$$R_{\text{MMREx}} = (1 + m\beta) \max_e R_e - \beta \sum_{e=1}^{m} R_e \tag{1.1}$$

with m the number of environments, R_e the risk of environment e and β a hyperparameter that controls how much we seek to extrapolate in the space of risk functions. VREx on the other hand, minimizes the variance of the risks of the environments together with the total risk across all environments. Its loss function is the following:

$$R_{\text{VREx}} = \beta \text{Var}(\{R_1, \ldots, R_m\}) + \sum_{e=1}^{m} R_e \tag{1.2}$$

where β controls the balance between reducing the average risk and enforcing the equality of risks. For a more in-depth discussion of these algorithms, we refer to [72].

This winning solution was presented at ECML-PKDD 2022:

Steven Mortier, Bart Bussmann, Tim Verdonck. 2022. *PRINCE Out-of-Distribution Challenge Winning Solution: a Robust Neural Network Ensemble.* Oral invited presentation: ECML-PKDD 2022, Grenoble, France.

Chapter 2

Classification of targets and distractors in an audiovisual attention task

This chapter is based on:

Mortier, Steven, Renata Turkeš, Jorg De Winne, Wannes Van Ransbeeck, Dick Botteldooren, Paul Devos, Steven Latré, Marc Leman, and Tim Verdonck. Classification of Targets and Distractors in an Audiovisual Attention Task Based on Electroencephalography. *Sensors* 23, no. 23: 9588. 2023

2.1 Introduction

The interaction between humans and artificial intelligence (AI) still lacks the level of engagement and synchronization that symbolizes the interaction between humans. The primary goal of the WithMe project[1] is to thoroughly study the processes that occur in the human brain during joint activities with another individual, such as working towards shared objectives [73]. The brain signals collected in this study are primarily indicative of attention, but also of emotion and reward. The purpose of this research is to determine relevant electroencephalography (EEG) features indicative of attention using machine learning (ML).

To this end, a specific experiment was designed. Temporal audiovisual integration and support of visual attention by sound is well demonstrated in the pip-and-pop experiment [74]. The pip-and-pop experiment is based on a visual search, which does not lead to a strong visually evoked potential. Moreover, as we expected that rhythmic presentation of target stimuli also affects working memory, the task was replaced by a modified digit-span task where five target digits had to be remembered and reported in our experiment [73]. This task involves visual attention, working memory, and sequence recall. To investigate the role of attention, we measured the brain activation directly by means of EEG. Specifically, event-related potentials (ERPs) have been shown to be excellent tools

[1]The WithMe project is a research project funded by the Research Foundation Flanders (FWO). More information can be found at `https://researchportal.be/en/project/withme-making-human-artificial-intelligence-interactions-more-entraining-and-engaging`.

for studying attention [75, 76]. Risto Näätänen was a pioneer in this domain, as he studied the connection between ERPs and attention, which led to his discovery of the (auditory) mismatch negativity ERP [77–80]. Additionally, research has shown that the amplitude of the P300 is directly related to the amount of attentional resources available for stimulus processing [80–83]. The P300 ERP is observed to be elicited for deviant stimuli in a sequence of standard stimuli, where the deviant stimuli are in some way more relevant to the presented task [84–86]. In our experiment, we thus expect that the targets elicit a P300 ERP. Research has shown that the P300 actually consists of two subcomponents: the P3a and P3b [87]. The P3a generally reaches its peak around 250ms to 280ms post-stimulus and is associated with attention-related brain activity [88]. On the other hand, the P3b peak can vary in latency, lying between 300ms and 500ms post-stimulus [87]. The P3b is elicited by improbable events, provided that the improbable event is somehow relevant to the task at hand [89]. In our experimental setting, we expect to elicit a P3a, as the target stimuli are not scarce (there are approximately 50% targets and 50% distractors) and our experiment is designed to evoke attention. We do not expect to elicit a P3a for distractors, as subjects should not pay attention to them.

The goal of this work is to accurately classify whether a target or distractor stimulus was presented to the subject, based on the subject's EEG data. For this purpose, we applied different existing ML methods to classify EEG data and investigate which method performs best on our specific use case. As we expect to elicit attention when a target is shown (and not when a distractor is shown), the trained ML will effectively be an attention detector. We expect the attention to manifest itself in the form of a P3a ERP, and therefore expect that the model will base its predictions on the presence of a P3a peak. Detecting P3a signals, and, more broadly, P300 signals, has a wide range of applications [90, 91], particularly in P300-based brain computer interfaces (BCIs) [92], for example in spellers [31–33] and intelligent home control systems [34, 35]. These applications can be of great help for patients suffering from amyotrophic lateral sclerosis (ALS) or spinocerebellar ataxia, as it can enable them to communicate in a daily environment [31, 33, 36, 37]. In the literature, a wide array of techniques are used to classify and detect P300 [93]. Some techniques rely on a data transformation and subsequently use logistic regression to classify the transformed data, for example xDAWN+RG [94–97]. Recently, deep learning approaches, primarily based on convolutional neural networks (CNNs), for example EEGNet [98–100], have also gained in popularity [101–103]. Finally, as EEG data are essentially heavily correlated multivariate time series, it is possible to apply standard time series classification techniques as well [104–106].

Building BCIs that are trained on multiple subjects and generalize well to previously unseen subjects holds significant value [107]. Indeed, BCIs often need to be re-trained or at least calibrated for the end-user [108], which is a costly and user unfriendly process [109, 110]. However, due to inter-subject variability of EEG data, training models that generalize to multiple subjects (cross-subject (CS) models) is a harder task than training models for one subject (individual subject (IS) models) [109–111]. For this reason, we also investigate the hypothesized drop in performance when transitioning from IS to CS models. Additionally, the ML models should be able to make predictions in real-time, as this is essential in real-world BCI applications.

Finally, we analyze which EEG channels and time points are used by our models to make its predictions, and check whether these align with the expected P3a attention signature. However, ML models such as CNNs are considered “black boxes", as no clear

explanation for the decisions made by these models exists [112]. The rapidly emerging and improving field of explainable AI (xAI) aims to tackle these issues by providing insights into ML models' decision-making processes. Some xAI techniques that are often used to gain insights into EEG classification models are Local Interpretable Model-Agnostic Explanations (LIME) [113, 114], DeepLIFT [98, 115, 116] and saliency maps [117–119], among others.

In summary, we aim to enhance the interaction between humans and AI and have designed a novel experiment for this purpose. Specifically, this work considers building a ML model to recognize targets shown to a subject, which equates to creating an attention detector. These models should ideally generalize well to previously unseen subjects. The primary contributions of this work are:

- Training of state-of-the-art classification methods to accurately predict target and distractor stimuli based on EEG data.
- Analysis of the performance difference between IS and CS models.
- Investigation into which EEG channels and time points were important for the model predictions, using xAI.

Ultimately, the contributions of this research collectively advance our understanding of human-AI interaction and will aid in the development of more effective BCIs and their associated applications.

The remainder of this paper is structured as follows: Section 2.2.1 introduces the WithMe experiment and dataset, while Section 2.2.2 explains the data preprocessing routine. Section 2.2.3 illustrates the classification problems and provides a description of the classification methods used in this work. Section 2.3 presents the results and provides an in-depth analysis of the best performing model. This section also contains an extensive discussion of the achieved results. Finally, in Section 2.4, we draw conclusions and provide possible directions for future research.

2.2 Materials and Methods

In this section, we describe the WithMe dataset that is analyzed using ML in this study. We then define the preprocessing steps that are applied to the EEG data. Finally, we present the classification problem and the classification methods and metrics that we will use to tackle this problem.

2.2.1 The WithMe experiment

Forty-two young adults participated in the experiment (21 females, 21 males; mean age 23.64 ± 2.69 years). They were recruited through the university network and through the social and professional network of the authors. All subjects declared to have normal or corrected to normal vision and showed normal hearing (< 25 dB hearing loss) for

the frequencies present in the experiment, based on a standard pure tone audiometry hearing test. To mitigate a potential influence of age and/or intelligence, the subject's age range was limited to young adults under 30, and they were only accepted if they were enrolled in or finished some form of higher education.

Before starting the experiment, subjects had to fill in a questionnaire that asked for general background information to identify some personal characteristics. For example, subjects were asked if they ever enrolled in some form of musical education and/or were an active musician. More details about the questionnaires and their extensive analyses are described in [73].

The experiment consisted of a modified digit-span task. A target digit was presented, followed by either no, one or two distractor digits, another target, etc. One sequence of digits always consisted of five targets and five distractors, although the subjects did not know this a priori. After one sequence of targets and distractors was presented, the subject had to report all targets in the correct order that they were presented to them. The targets and distractors were presented as an encircled number $x \in \{0, 1, 2, \ldots, 9\}$. Additionally, a distractor could also show up as an empty circle. Target digits were colored black ($rgb(0, 0, 0)$), while distractor digits were displayed in dark gray ($rgb(x, x, x), x \in [50, 75]$), with the exact value of x determined individually to make the difference between targets and distractors just noticeable. An example sequence is shown in Table 2.1. In total, 30 different sequences of targets and distractors were created, which were shown to the subject under four different conditions in a pseudo-randomized order [73]. Depending on the condition, the subject received either no support (Con1), visual rhythmic support (Con2), auditory non-rhythmic support (Con3) or visual rhythmic and auditory support (Con4), as shown in Figure 2.1b. This adds up to a total of 120 sequences shown to the subject. In conditions with auditory support (Con3 and Con4), targets were accompanied by a 500 Hz tone burst, which lasted for 50ms. In conditions with rhythmic support, targets were presented with a fixed time interval of exactly 1.25s between them. In these rhythmic conditions, the sequence of digits was preceded by five rhythm inducing stimuli to induce the subject with the rhythm. In Con2, this was done using empty black circles, while in Con4 auditory tone bursts were used to induce the rhythm. For more detailed information about the experiment, we refer to the original paper that describes the experiment and performs a behavioral analysis [73].

Table 2.1: An example of a sequence of stimuli shown to the subject, with the targets in black and distractors in gray. In conditions with rhythm (Con2 and Con4), this sequence is preceded with five empty circles to induce the rhythm. The subject is expected to report the target digits and ignore the distractors.

Sequence	Targets	Distractors
③ ⑨ ② ④ ③ ③ ◯ ① ② ⑧	39432	23 18

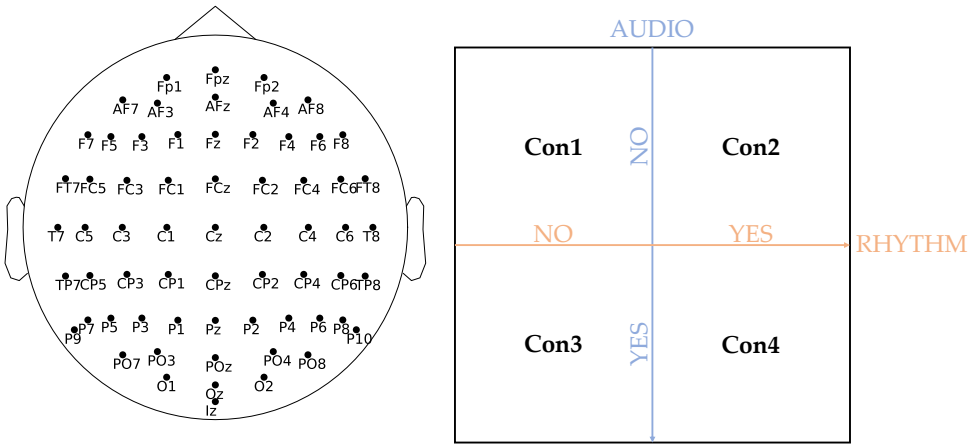


Figure 2.1: (**a**) The WithMe experiment uses 64 electrodes, organized according to the 10-10 system. Figure made using [120]. (**b**) Depending on the condition, the subject received no support (Con1), visual rhythmic support (Con2), auditory non-rhythmic support (Con3) or visual rhythmic and auditory support (Con4).

2.2.2 Dataset and preprocessing

During this experiment, EEG data was sampled at 2048 Hz using the standard 64-electrode EEG 10-10 system as shown in Figure 2.1a. Thereafter, standard EEG preprocessing techniques were applied. The data was re-referenced to the average of both earlobes, just one earlobe if the other one was too noisy, or another pair of channels, if both earlobes were badly recorded or too noisy. In the case of bad channels, these were identified and removed. EEG data was notch-filtered at the line frequency (50 Hz) and its multiples, after which a band-pass filter from 0.2 Hz to 100 Hz was applied. The data was split into epochs, ranging from 0.2 seconds pre-stimulus to 1.0 seconds post-stimulus, resulting in 1.2 second epochs. Independent component analysis (ICA) was applied to the epoched data. Any components that represent artifacts were removed by visual inspection of the ICA components.

During the previous steps, some channels were marked as bad channels. Instead of dropping these channels, we chose to interpolate them using their neighboring channels, as the former would result in an inconsistent number of channels across sequences and subjects. The interpolation was performed using the MNE-Python package [120] (All Python packages that were used in this work can be found in Table 2.7, together with their version number and citation). Finally, the data was downsampled to 50 Hz, as this reduces computation time, decreases file read/write time and saves memory, while generally leading to little or no loss of information [121]. We should however note that, based on the Nyquist theorem, this limits the highest frequency that can be accurately represented to half of the sampling frequency, i.e., 25 Hz. This preprocessing routine ideally resulted in 600 target epochs, 600 distractor epochs, and 300 induction epochs for

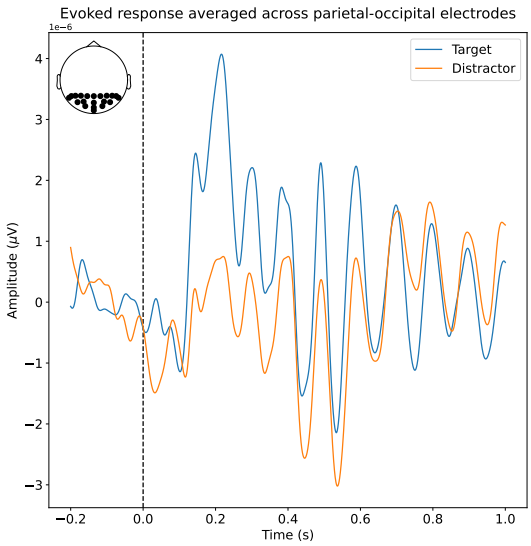


Figure 2.2: The evoked response for targets and distractors for one subject. The data is averaged over all electrodes of the parietal-occipital region in the brain, as indicated in the figure inset.

each of the 42 subjects. However, during preprocessing, some epochs were rejected for various reasons, for example, an excessive number of bad electrodes or too much noise. On average, less than 0.6% of the epochs were rejected per subject.

As mentioned in Section 2.1, we expect to observe a P3a ERP when subjects see a target stimulus. The P3a ERP is characterized by a positive voltage deflection between 250ms and 280ms post-stimulus, although the exact timing can vary [88, 122, 123]. As our experiment uses visual stimuli, we expect the P3a ERP to be the most pronounced in the parietal-occipital region of the brain [121]. Figure 2.2 shows the evoked response for one subject, averaged over all parietal-occipital electrodes, as indicated in the figure inset. We observe a clear positive deflection between 200ms and 300ms post-stimulus, which is in line with our expectations.

2.2.3 Classification problem

The models trained in this study consider a two-class classification problem (target versus distractor) and take single-trial EEG epochs as input to predict a binary label. As the data is downsampled to 50 Hz, one epoch contains 60 time steps, for 64 electrodes. This means that the input is of shape $(N, 64, 60)$ with N the number of epochs. It is important to note that it is impossible to obtain 100% accuracy for this model. Indeed, the model makes a prediction based on the subject's assessment of a stimulus, and it is possible that a subject did not correctly recognize all targets and distractors. As the ground truth labels are based on the predefined labels of the experiment, it is possible that there is

a slight mismatch between the labels and the subject's perceived class. Nevertheless, we assume that this problem is rare, meaning that commonly used metrics, for example accuracy, have a valid interpretation.

Ideally, the models should be able to generalize to previously unseen subjects. To investigate this, we train the models in two ways: models trained on IS and models trained on (nearly) all subjects, also called CS models. The former will be evaluated using a randomly sampled test set with a standard 80% train and 20% test set split, while the latter are evaluated using a leave-one-out (LOO) methodology. In general, models perform better when trained and tested on individual subjects [124]. This can be attributed to the variability in subject's EEG data elicited by the same stimuli. However, in practice, EEG classification models should ideally extrapolate to previously unseen subjects. For example, BCIs often need to be calibrated for new end-users, which usually takes 20 to 30 minutes [125–127]. Therefore, it is interesting to investigate which model architectures are best suited to build subject-independent classifiers.

2.2.3.1 Classifiers

To solve this classification problem, we train and evaluate different existing ML models. Different methodologies for classifying EEG data exist. For example, we can extract features from the EEG data and use these extracted features as input to the classifier. These features can, among others, be extracted from the time domain, frequency domain or the time-frequency domain, or through methods such as principal component analysis [128, 129]. Such methods are referred to as feature-based methods. Another common approach uses raw or preprocessed EEG data as input to the classifier. In this approach, commonly referred to as an end-to-end method, the classifier itself will extract relevant features from the data during training, and use these features to classify a sample. As both methodologies are interesting approaches, we will use methods belonging to both approaches. In this study, we apply four distinct classifiers and compare the results on a novel data set. An overview of the classifiers and their methodologies is presented in Table 2.2. First, we will apply the xDAWN pipeline, which has demonstrated significant success in several EEG classification tasks and is often used as a baseline technique [94, 95]. For example, the BCI challenge organized as part of the IEEE Neural Engineering Conference 2015 was won by an xDAWN-based approach [95]. In this study, we employ a similar approach, consisting of first estimating two sets of xDAWN spatial filters, one for each class (target and distractor) [94]. Subsequently, the grand average evoked potential of each class is filtered using the corresponding filters, after which they are concatenated to each of the trials. Then, the covariance matrix of each resulting trial is used as a feature for the next steps in the pipeline [130, 131]. The next step is to project the covariance matrices on the tangent space using a Riemannian metric, as described in [96, 97]. After these feature extraction steps, a classifier is used to make the final predictions. Based on [95] and [132], we used logistic regression [133]. For the remainder of this study, we will refer to this method as xDAWN+RG (xDAWN+Riemannian Geometry). Calculating the xDAWN covariance matrices and projection to the tangent space were done using the PyRiemann package [132].

Table 2.2: Overview of the methods that were used in this study, together with their original target domain and methodology.

Model	Target domain	Methodology
xDAWN+RG [94]	EEG	feature-based
EEGNet [98]	EEG	end-to-end
MiniRocket [105]	time series	feature-based
Rocket [104]	time series	feature-based

The second method we consider is EEGNet, as it is often considered to be the state-of-the-art solution for various EEG-based classification tasks [98]. Indeed, EEGNet exhibits strong performance on a variety of EEG-based classification tasks, such as P300 ERP classification [98, 118] and motor imagery classification [134]. Whereas the previous method used extracted features as input to the classifier, EEGNet performs both the feature extraction and classification. EEGNet is a deep learning model, more specifically a CNN. As its name suggests, EEGNet is optimized for classifying EEG data by employing a set of specific design choices. Firstly, it uses temporal convolutions to learn frequency filters [98]. As suggested by the authors, the length of the temporal kernel used in these convolutions is set to half the sampling rate, which allows the model to capture frequency information at frequencies of 2 Hz and higher [98]. Second, depthwise convolutions are used to learn frequency-specific spatial filters. In this context, depthwise convolutions have two main advantages. First, they noticeably reduce the number of trainable parameters, since these convolutions are not fully connected to the previous layer; instead, they are connected to each feature map individually. This induces the second, EEG-specific advantage: the model learns spatial filters for each temporal filter, which enables the efficient extraction of frequency-specific spatial filters [98]. The last convolutional part consists of a separable convolution, which is a combination of a depthwise convolution and pointwise convolution. The former learns how to summarize individual feature maps in time, while the latter learns how to optimally combine the feature maps [98]. Finally, all features are passed to a dense layer for classification. More details on the EEGNet architecture can be found in [98]. We use the standard EEGNet-8,2 layout, which means that the model learns 8 temporal filters and 2 spatial filters per temporal filter.

The first two methods are designed for EEG specifically. However, since EEG data is essentially a heavily correlated multivariate time series, it is interesting to study the results of a more general method designed to classify such time series. This opens the door to a much wider range of potential methods, which may be more optimized or easier to implement. To this end, we applied RandOm Convolutional KErnel Transform (Rocket) [104]. Based on the success of CNNs for time series classification, Rocket uses random convolutional kernels combined with simple linear classifiers. This novel combination achieves state-of-the-art performance on the UCR time series archive using only a fraction of the computational cost of existing methods [104, 135]. As a follow-up to Rocket, the authors also designed MiniRocket [105]. They claim MiniRocket can be trained up to 75 times faster than Rocket, while achieving nearly the same performance. MiniRocket distinguishes itself from Rocket primarily by reducing the degree of randomness that Rocket generates, resulting in MiniRocket being almost deterministic [105]. Since methods to classify EEG data, such as EEGNet, can be very computationally expensive, it is worth exploring the effectiveness of less computationally expensive methods. We

used the Rocket and MiniRocket implementations in the sktime package and combined them with the ridge regression classifier implemented in scikit-learn, as suggested by the authors [104, 105, 133, 136].

2.2.3.2 Metrics

To allow the comparison of various approaches, it is essential to have predetermined performance metrics. We will focus on three metrics that are widely used in EEG classification literature: accuracy, F1-score, and area under the receiver operating characteristic curve (ROC AUC) [137]. First, the accuracy states the number of correctly classified samples across both classes. Second, the F1-score assesses the predictive performance of a model by calculating the harmonic mean of the precision and recall metrics. The equations used to calculate the accuracy and F1-score are given in Equation (2.1) and Equation (2.4) respectively, where we use the following abbreviations: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Third, by plotting the true positive rate against the false positive rate for different classification thresholds, we obtain the ROC curve. The ROC AUC is defined as the area under this curve and provides a measure for how well a classifier can distinguish between true and false samples, or in our case, targets and distractors respectively. Finally, we also assess the required training time and model complexity of all models.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{2.3}$$

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN} \tag{2.4}$$

2.3 Results and discussion

2.3.1 Individual subject models

The performance of the models, assessed using the metrics introduced in Section 2.2.3.2, can be seen in Table 2.3 and Figure 2.3. Evidently, using an EEG-specific model architecture benefits the performance of IS models. While xDAWN+RG and EEGNet perform equally well, they demonstrate superior accuracy, F1-score and area under the curve (AUC) in comparison to MiniRocket and Rocket. As expected, MiniRocket achieves slightly inferior performance compared to Rocket. However, MiniRocket's training time was 15 times faster on our dataset. Notably, while xDAWN+RG and EEGNet exhibit

equal performance, xDAWN+RG is significantly less computationally expensive than EEGNet. On central processing units (CPUs) alone, EEGNet's training time is 9 times longer. Although training times can be accelerated for EEGNet using (expensive) graphics processing units (GPUs), even when using an NVIDIA GTX 1080 GPU, EEGNet still requires 2.5 times as long to train as xDAWN+RG.

Table 2.3: Classifier test performance for individual subject models, averaged across the 42 subjects. The best performances are indicated in boldface.

Model	Accuracy	F1-score	AUC
xDAWN+RG	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.84 ± 0.05**
MiniRocket	0.69 ± 0.05	0.70 ± 0.05	0.76 ± 0.06
Rocket	0.72 ± 0.05	0.72 ± 0.05	0.79 ± 0.06
EEGNet	**0.76 ± 0.04**	**0.76 ± 0.04**	**0.83 ± 0.04**

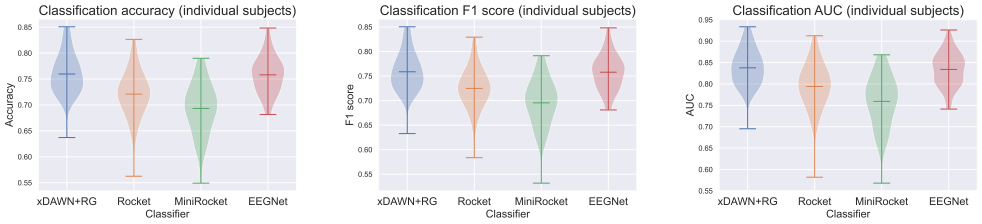


Figure 2.3: Violin plots of the test accuracy, F1-score and AUC for models trained on individual subjects.

2.3.2 Cross-subject models

Similar results are obtained for CS models, where EEG-specific approaches perform better than Rocket and MiniRocket, as shown in Table 2.4 and Figure 2.4. However, in this scenario, EEGNet outperforms xDAWN+RG. We hypothesize that this can be attributed to EEGNet's added complexity and a greater number of parameters compared to xDAWN+RG. This additional capacity is more likely to be able to learn features that extrapolate well to previously unseen data points.

Table 2.4: Classifier performance for cross-subject models. Every subject was used as a test subject once, we report the average across all test sets. The best performances are indicated in boldface.

Model	Accuracy	F1-score	AUC
xDAWN+RG	0.73 ± 0.04	0.72 ± 0.06	0.81 ± 0.05
MiniRocket	0.69 ± 0.04	0.69 ± 0.05	0.76 ± 0.05
Rocket	0.72 ± 0.05	0.71 ± 0.06	0.79 ± 0.05
EEGNet	**0.76 ± 0.04**	**0.76 ± 0.05**	**0.84 ± 0.05**

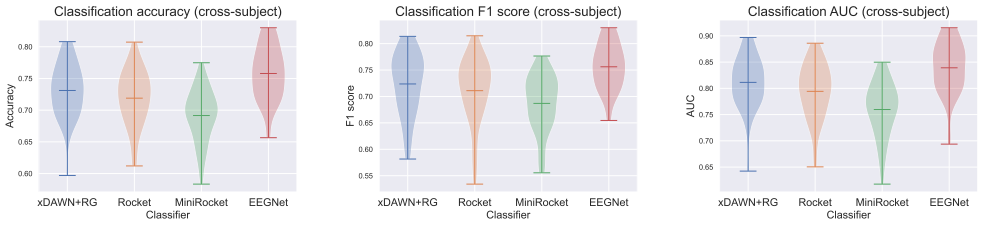


Figure 2.4: Violin plots of the test accuracy, F1-score and AUC for cross-subject models.

2.3.3 Individual subject models vs. cross-subject models

As we discussed in Section 2.2.3, we expect that the performance of the IS models is better than that of the CS models. Despite having access to a significantly larger amount of data, constructing a CS model is a considerably more challenging task. To illustrate the performance disparity between the two, we refer to Table 2.5 and Figure 2.5, which showcase the performance difference by subtracting the CS model's performance from that of the IS model. EEGNet, MiniRocket and Rocket exhibit similar performance for both IS and CS models. However, the xDAWN+RG model demonstrates a noticeable decrease in performance. Given the lower absolute performance of the (Mini)Rocket models compared to EEGNet and xDAWN+RG, we will focus on the latter for the remainder of this discussion. We hypothesize that the inferior performance on CS models when using xDAWN+RG can be attributed to its simpler and lightweight nature. Furthermore, xDAWN+RG works by first calculating the evoked responses for all classes. These can differ significantly from subject to subject, both in P3a peak height and in time [37, 138, 139]. The convolutional nature of EEGNet likely enabled it to capture the temporal dynamics of the elicited responses more effectively across different subjects. It is important to note that the CS models had access to a significantly larger corpus of training data than the IS models, which is part of the reason that they kept up reasonably well with the IS models.

Table 2.5: Drop in performance, calculated by subtracting the test performance of cross-subject models from that of individual subject models. The best performances are indicated in boldface.

Model	Accuracy	F1-score	AUC
xDAWN+RG	0.03 ± 0.02	0.04 ± 0.04	0.03 ± 0.02
MiniRocket	**0.00 ± 0.03**	0.01 ± 0.03	-0.00 ± 0.03
Rocket	**0.00 ± 0.02**	0.01 ± 0.04	-0.00 ± 0.03
EEGNet	**0.00 ± 0.02**	**0.00 ± 0.02**	**-0.01 ± 0.03**

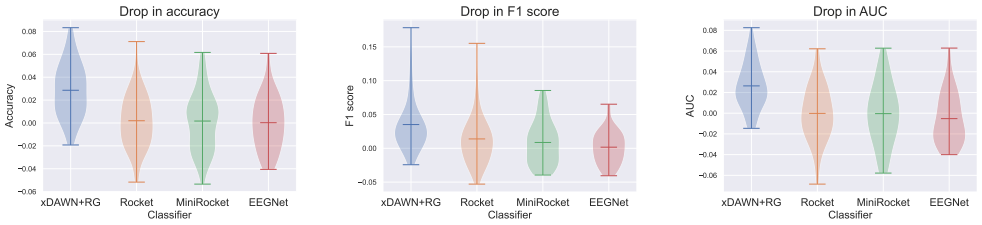


Figure 2.5: Violin plots of the drop in performance, calculated by subtracting the test performance of cross-subject models from that of individual subject models.

2.3.4 Analysis of the EEGNet cross-subject model

This section aims to conduct a further investigation into the CS EEGNet model. We conduct this analysis for the EEGNet model, as it performed the best in both the IS and CS scenario. Furthermore, we include this analysis only for the CS models, as they are the most useful in practice, due to their generalization capabilities. However, the conclusions are similar for the IS models.

2.3.4.1 Confusion matrices

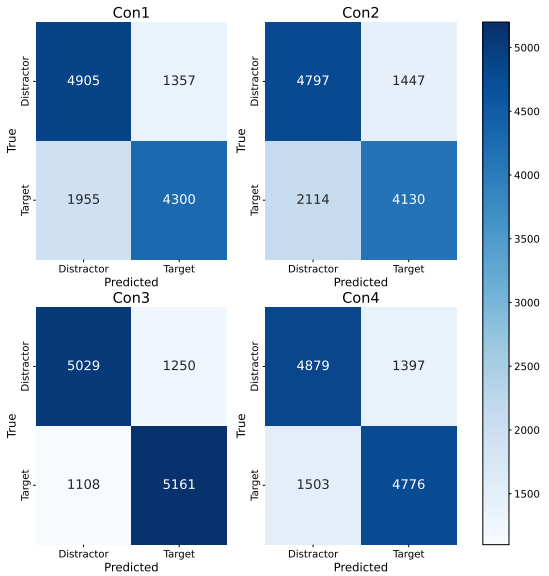


Figure 2.6: Confusion matrices for the cross-subject EEGNet model, split across the four conditions defined in Figure 2.1b. The confusion matrices are obtained by aggregating all the test predictions of the CS models.

First, we investigate whether the model focuses on the correct features to make a prediction. For example, it is possible that we trained a sound detector instead of a target/distractor model. Indeed, conditions Con3 and Con4 contain auditory clues for the target. Theoretically, the model could rely solely on the activation in the auditory stimuli processing region of the brain and achieve acceptable performance. For example, if the model performs perfectly on Con3 and Con4, while predicting all trials belonging to Con1 and Con2 to be distractors (due to the absence of auditory stimuli), it would achieve an accuracy of approximately 75%. The confusion matrices in Figure 2.6 negate this assumption. The model performs comparably in detecting distractors under all conditions. However, the model performs slightly better at identifying targets correctly for Con3 and Con4. The accuracies for specific conditions, shown in Table 2.6, also reflect this. Indeed, the accuracies for conditions Con3 and Con4 are higher than the accuracies for Con1 and Con2. We hypothesize that the inclusion of auditory support causes an additional signature in the EEG data, making it easier for the model to recognize targets. Additionally, it was already confirmed by a previous analysis that the subjects are able to recall the targets better in conditions with auditory support [73].

Table 2.6: The test accuracies of the CS EEGNet model for the different conditions.

Condition	**Accuracy**
Con1	0.74
Con2	0.71
Con3	0.81
Con4	0.77

2.3.4.2 Saliency maps

Next, we explore the electrodes and timings that are predominantly used by our models for making predictions. Trivially, we expect that the model does not use the pre-stimulus ($t < 0$) EEG data. As deep learning methods such as EEGNet are inherently black box models, we resort to xAI methods to obtain (interpretable) insights into the model. A possible technique is a saliency map, which is a visual representation that highlights the degree of importance of regions or features in an input sample on the model prediction [117]. To generate a saliency map, the gradient of the model output with respect to the input sample is computed using backpropagation [118]. More specifically, this process involves fixing the weights of the trained model and propagating the gradient with respect to the layer's inputs back to the first layer that receives the input data. Figure 2.7 shows such a saliency map. This saliency map illustrates the electrodes and timings that had the greatest average impact on the model prediction when identifying a sample as a target. It is computed by first calculating the average saliency map for each test subject individually, then normalizing these saliency maps, and ultimately taking the average across all 42 subjects. In Figure 2.8, the same information is repeated, displayed as a topographic map at five time points. From Figures 2.7 and 2.8, we can see that our model predominantly used the parietal-occipital electrodes and time points between 200ms and 300ms post-stimulus to make its prediction, which is what we expected. We also investigated the saliency maps under different conditions, but noticed no significant difference between the conditions.

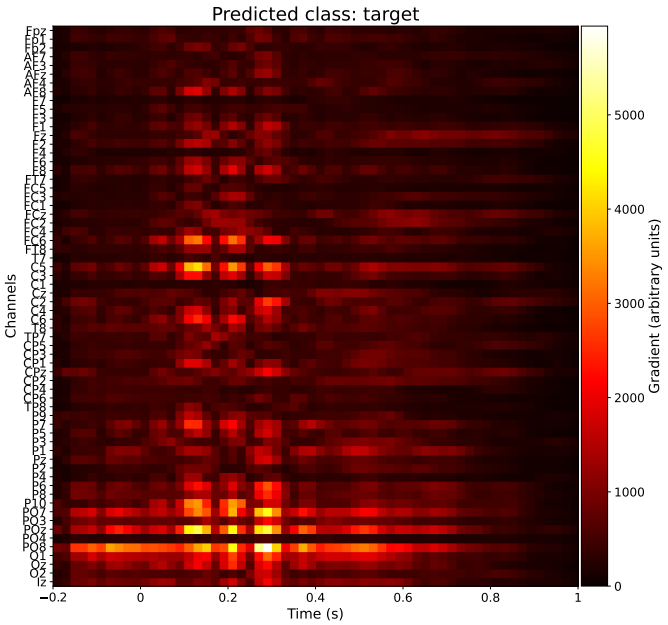
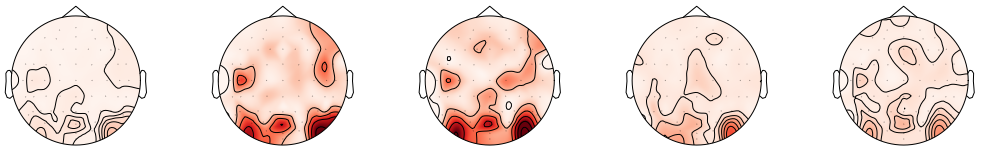


Figure 2.7: Saliency map for epochs labeled as targets by the cross-subject EEGNet model. We averaged normalized saliency maps over all 42 test subjects for the CS model.



(a) -0.200ms (b) 0.200ms (c) 0.300ms (d) 0.400ms (e) 0.700ms

Figure 2.8: The saliency map from Figure 2.7, shown as a topographic map at five different timings. A deeper shade of red indicates a larger gradient. At $t = -0.200$ms and $t = 0.700$ms, the gradients are near zero, indicating that the model does not use these timings. Contrary, for $t \in \{0.200, 0.300, 0.400\}$ms, there is a large gradient in the parietal-occipital region of the brain.

2.3.5 Real-world applicability

As discussed earlier in Section 2.3, the EEGNet model requires significantly more training time compared to the xDAWN+RG model. Despite this, all models demonstrate the ability to perform inference in near real-time. This quick inferencing capability is crucial for applications that require immediate feedback, such as ERP spellers, where any delay could negatively impact user experience and overall system responsiveness.

Fast inference ensures that users receive the feedback they need without perceptible lag, enhancing both usability and effectiveness in real-world scenarios.

2.4 Conclusions and Future work

The WithMe project has led to the collection of a large, novel EEG dataset that can be used to create ML methods to automatically detect attention using P3a ERPs in single trial data. This is of great importance to BCIs, as they often rely on the P3a, or, more broadly, the P300 ERP and have a wide range of applications.

We successfully achieved the goal of this work, which was to classify target and distractor stimuli based on the subject's EEG data. To achieve this goal, we studied four classification methods that differed significantly in origin and complexity. We investigated the performance of these methods both as IS and CS models, with the latter the most practically relevant due to its generalization capabilities. For the IS models, xDAWN+RG and EEGNet obtained an accuracy of 76%, outperforming MiniRocket and Rocket. While EEGNet was able to obtain the same accuracy of 76% in the CS case, the accuracy of xDAWN+RG dropped to 0.73%. We attribute this difference to the larger complexity of EEGNet, which likely enables it to generalize better to previously unseen subjects. The drop in performance between IS and CS models is not as pronounced as we expected it to be, and even nonexistent for EEGNet. We attribute this to the fact that the CS models had approximately 42 times more training data available. The EEGNet CS model performed slightly better on samples recorded under conditions Con3 and Con4, which are the conditions that include auditory support. While EEGNet achieved the best performance overall, it also has the highest model complexity (highest number of trainable parameters) and takes the longest time and most compute to train. However, all four models are able to make predictions in real time. This property is essential for real-world human-AI interaction experiments and applications.

Finally, the application of xAI enabled us to investigate which EEG channels and time points were used by the otherwise black-box EEGNet CS model to make its predictions. Indeed, using saliency maps, we conclude that the model primarily based its prediction on the values of the electrodes in the parietal-occipital region between 200ms and 300ms post-stimulus. This is in line with our hypotheses, as we expected to elicit an attention-related P3a ERP in the parietal-occipital region of the brain when the subject saw a target digit.

In conclusion, we achieved the goal of accurately classifying targets and distractors based on a subject's EEG data. At the same time, our work contributes to the development of more effective BCIs and their applications. Finally, we validated the EEG data collected in the WithMe experiment.

While this study provides valuable insights into attention detection using EEG data, it is important to acknowledge some limitations. For example, as mentioned in Section 2.2.3, part of the data used to train the model was labeled incorrectly, as the ground truth labels are based on the predefined labels of the experiment rather than the subject's perceived class. A possible solution is to limit the data to samples where the entire sequence was reported correctly. However, this would mean that we lose a lot of data, which would in

turn decrease the performance of the models. Alternatively, we could remove all "bad sequences", where a bad sequence would be defined as a sequence in which none of the targets were remembered correctly. This could be caused by either incorrectly identifying the stimuli, or by bad memory management, despite correctly identifying the targets and distractors. However, the number of answers that did not include at least one of the target digits (regardless of its place in the sequence) is negligible.

In future work, an experiment dedicated to attention should be used to circumvent the limitations regarding bad labels, as described earlier in this section. This would allow for labels that exactly correspond to the subject's perception of a stimulus, which would in turn lead to more accurate attention detectors. The ultimate goal could then be to use this attention detector in a BCI, to detect whether a subject paid attention. In case they did not, the BCI could repeat the sequence or stimulus, to make sure that the subject can act accordingly. This could also improve learning systems, that is, systems that know whether a student actually paid attention to the provided information [140, 141]. Regarding the training and optimization of ML models, it would be interesting to include an exhaustive feature selection procedure to allow the ML model to focus on the (most) relevant features. Additionally, we want to explore other ways to enable CS generalization, for example using transfer learning [142, 143]. This could further increase the generalization performance of all methods. In particular, this has the potential to elevate the performance of lightweight models such as xDAWN+RG to that of the computationally expensive EEGNet. While this work focuses on the detection of attention using epoched EEG data, the experiment can also be used to study working memory [73]. Indeed, the complete sequence EEG data should permit an investigation regarding working memory and whether it is influenced by auditory and/or rhythmic support.

2.5 Data availability statement

Publicly available datasets were analyzed in this study. The behavioral data can be found at: `https://osf.io/ntmy8/?view_only=88d951c394c7481dba00a1497d64797f`. The preprocessed EEG data are openly available from Figshare at `https://doi.org/10.6084/m9.figshare.24278887` [144].

2.6 Appendix

Table 2.7: The versions of the Python packages used in the project.

Package	Version	Reference
python	3.9.13	[145]
MNE	1.2.1	[120]
pyriemann	0.3	[132]
torch	1.13.1	[146]
sktime	0.17.1	[136]
scikit-learn	1.1.3	[133]

Chapter 3

Inferring the relationship between soil temperature and the normalized difference vegetation index

This chapter is based on:

Mortier, Steven, Amir Hamedpour, Bart Bussmann, Ruth Phoebe Tchana Wandji, Steven Latré, Bjarni D. Sigurdsson, Tom De Schepper, and Tim Verdonck. Inferring the Relationship between Soil Temperature and the Normalized Difference Vegetation Index with Machine Learning. *Ecological Informatics* 82:102730. 2024

3.1 Introduction

In-situ monitoring of changes in vegetation in inaccessible Arctic regions is challenging, prompting many such studies to rely on remote sensing techniques [147]. In the field of remote sensing, vegetation indices such as the Normalized Difference Vegetation Index (NDVI) are used to quantify and qualify vegetation cover [148]. This is achieved through airborne or satellite spectral methods [149, 150] or ground-level measurements, using handheld instruments [151, 152]. Vegetation activity monitoring using NDVI has shown both intra-annual and inter-annual variations that can give valuable insights into ecosystem changes [153, 154]. Some parameters that can be derived from such intra-annual seasonal NDVI curves are the start of the season (SOS), peak of the season (POS), and maximum annual NDVI value (PEAK) [155, 156].

In high latitudes, the intra-annual temperature and irradiance variation are important factors that control the cycles in the growth and reproduction of the flora [157, 158]. Over the last decades, different life-cycle events of vegetation (phenology) have been observed to change in this region [159]. This has been related to ongoing climate change [160], which has started to affect vegetation phenological cycles, productivity, and community structure [161]. Inter-annual analyses found relationships between climate change and these changes in vegetation dynamics, particularly with regard to the increase in surface temperature, resulting in an increased PEAK NDVI and with a notable impact on the

length of the growing seasons [162, 163]. Starting from the year 2000, scientists started to name this phenomenon of an increasing PEAK "Arctic greening" [164]. This phenomenon was hypothesized to persist with continued climate warming, based on the compelling evidence of increased PEAK NDVI [165], plant productivity [166], phenology [161], and vegetation composition [167] between 1980s and early 2000s [159, 168].

Interestingly, the "Arctic greening" effect has not occurred everywhere at high latitudes and since the early 2000s, the relationship between PEAK NDVI with an increase in surface temperature has weakened in many places [169, 170]. In fact, in some regions, this relationship has even become negative, introducing the term "Arctic browning" [165]. It is generally believed that the shift towards browning must indicate that other meteorological drivers (e.g., temperature, precipitation, wind, photoperiod) or biological drivers (e.g., insect grazing, drought, etc.) are in play. However, the issue still requires further study.

In Iceland, the same strong "Arctic greening" trend was shown to occur during the 1980s-2000s as in many other high-latitude regions, but with a notable stagnation of the national PEAK NDVI during 2000-2010, even if the surface temperatures continued to increase in Iceland during that period [171, 172]. What happened in Iceland after 2010 is unclear, but a recent study showed that the inter-annual variation in the national average PEAK NDVI has been large during 2001-2019 period [173]. Therefore, it is of interest to further study how the NDVI of Icelandic ecosystems responds to further warming.

The impact of climate change on high latitude vegetation is not only limited to the air temperature increases. Soil warming studies have revealed significant insights into how soil warming affects soil processes and, consequently, vegetation. Soil warming experiments in high latitudes have demonstrated that increased soil temperatures can lead to changes in nutrient availability, microbial activity, plant composition and biomass, all of which influence plant growth and ecosystem dynamics [174–176]. For example, [177] found that soil warming resulted in changes to below-ground plant biomass and fine root biomass, under different warming conditions. These changes were associated with shifts in plant community composition and soil chemistry, highlighting the complex adaptation mechanisms of subarctic grasslands to prolonged soil warming.

Continued climate change is expected to cause relatively higher increases in surface temperatures at higher latitudes in the coming decades [160], which will likely lead to relatively more ecosystem changes in plant productivity than at lower latitudes [178]. Potential changes include further temporal shifts in parameters that characterize growing seasons [161] and increases in plant productivity [179, 180]. However, it is important to further investigate the warming impacts on NDVI to better underpin such predictions for future changes. Combining data from manipulation (warming) experiments offer possibilities to study future high-latitude ecosystem NDVI responses [181, 182].

To relate changes in vegetation composition, biomass or NDVI to environmental parameters, traditional statistical methods like (non-)linear regression or linear mixed models have been most commonly used [182–187]. Additionally, multivariate methods have also been used, for example multivariate analysis of variance tests [188].

Despite massive advancements in the field of machine learning (ML) during the last decade, ML is not yet often used for vegetation studies. ML models can be used for various tasks, among which are classification, regression, and image segmentation. In ML,

models extract knowledge from data and use this knowledge to produce an output relevant to the task at hand. These models use three main learning paradigms: supervised learning, unsupervised learning or reinforcement learning. This study only considers the first paradigm, as we build a regression model. Within supervised learning, there are a multitude of model types, for example, support vector machines [189], boosted tree ensembles (e.g., XGBoost [190] or LightGBM [191]) and artificial neural networks (ANNs) [192]. This analysis will use ANNs, particularly multilayer perceptrons (MLPs), which are fully connected feedforward neural networks that consist of multiple layers of nodes that are connected with each other by weighted edges.

Recently, ML has also shown promising results in the field of ecology [51, 193], for use cases such as species identification [39, 194–196], behavioral studies [197, 198], ecological modelling and forecasting [199–201], remote sensing [202, 203] and climate change studies [204–206], among others. The utilization of ML techniques has opened new avenues for understanding complex ecological phenomena and predicting ecological responses. Considering the proven potential of ML in addressing research questions in the broad field of ecology [207, 208], we propose to apply ML methods to investigate the relationship between vegetation phenology and environmental drivers in subarctic grasslands.

Unfortunately, MLPs are black-box models. This means that, while they can approximate any function, it is nearly impossible to determine the structure of the approximated function. This led to a whole new field within ML, explainable AI (xAI), which tries to create methods that allow human users to understand the predictions made by an ML model [209]. Some popular examples include sensitivity analysis [210], Local Interpretable Model-Agnostic Explanations (LIME) [113], and SHapley Additive exPlanations (SHAP) values [59]. This study uses the last method, as it is gaining in popularity and is now often used in ecology. For example, [211] use SHAP values to investigate how inter-annual variation in the daily average temperature affected the first flowering date or the full blossom date of the Yoshino cherry trees in Japan. [212] construct a seagrass distribution model and explain the importance of environmental variables in the model and subsequent predictions. In [213], an XGBoost model is trained to predict chlorophyll concentration, and they use SHAP values to perform feature selection, as well as investigate feature importance. SHAP values have a number of advantages over other methods for understanding the output of a model. First, SHAP values are model-agnostic, which means that they can be used with any ML model [59]. Second, SHAP values are able to account for interactions between features, which is something other methods are not able to do. Third, SHAP values have an intuitive interpretation, which means that they are easy to understand and explain to others. Finally, SHAP values have some desirable mathematical properties, such as local accuracy, missingness, and consistency [214].

An earlier study was conducted by [182] at the same research sites in Iceland [215], focusing on the phenology of subarctic grasslands. They used a short-term temporal dataset from 2013 to 2015 with curve function fitting analyses based on the methodology proposed by [216] to determine seasonal (intra-annual) parameters (e.g. SOS). They found that the response towards earlier SOS in the warmed subarctic grasslands did not saturate at higher soil warming levels (*i.e.*, +10°C). Therefore they concluded that growing seasons at high-latitudes grasslands are likely to continue lengthening with future warming. However, there was still quite a large unexplained inter-annual variability in their 3-year dataset, that warranted a further study [182]. In this study, we extended the analysis period to six years, compared to the three years used by [182]. This enabled

a more comprehensive examination of the inter-annual variability in NDVI phenology and annual maximum values. Specifically, the variables used to study NDVI phenology were the annual day numbers of the SOS and POS, as well as the PEAK, in each research plot. Our primary objective was to reanalyze the soil warming effects using conventional linear statistics, as performed by [182], and to assess the robustness of these relationships over a longer timeframe. Additionally, our study extends previous research by employing ML algorithms to identify further drivers of the unexplained inter-annual variation in the studied variables. Specifically, we added a set of three meteorological variables, namely air temperature, precipitation, and irradiance. However, as predictions made by ML are often not intuitive, we used xAI methods, providing deeper insights into the model outputs.

Our objective was to study the relationship between soil temperature and vegetation phenology. More specifically, we studied this relationship using three vegetation phenology characteristics: SOS, POS and PEAK. Additionally, we investigated the effect of other meteorological variables on these characteristics. To this end, we postulated following hypotheses:

A **Soil warming**

i. A higher soil temperature will introduce significantly earlier SOS, as was found by [182] for individual years.

ii. The POS will take place at a similar time each year, regardless of the soil temperature. Plants must use some external trigger to "know" when to start to slow down growth and prepare for autumn. The prevailing theory suggests that for most plants, this is triggered by the length of the day [217, 218], which remains consistent across different years, and is mediated through the phytochrome system [219].

iii. The PEAK value will not be significantly related to soil temperature, as [220] showed that there was no difference in above-ground biomass between the warming treatments.

B **Other meteorological variables**
We expect that ML can identify other important controls for the previously observed inter-annual variability of NDVI phenology and PEAK values. Additionally, we expect that ML can identify the importance of meteorological variables compared to the soil temperature. Out of the three additional meteorological variables, we hypothesized for both phenology and PEAK values:

i. Larger impact of meteorological variables compared to the soil temperature, as they can also impact the soil temperature [221–223].

ii. Within the meteorological variables, air temperature's influence is expected to be the smallest due to its regulation of soil temperature , while precipitation may have an intermediate effect given consistently high soil water content in these areas [215]. Additionally, a substantial impact of irradiance is hypothesized, particularly in consistently cloudy sub-Arctic climates [224].

Ultimately, the contributions of this research advance our understanding of the relationships between soil temperature, other meteorological variables, and vegetation phenology. We achieve this goal by employing a methodology that exceeds standard practice, using ML and SHAP values.

3.2 Materials and Methods

3.2.1 Data

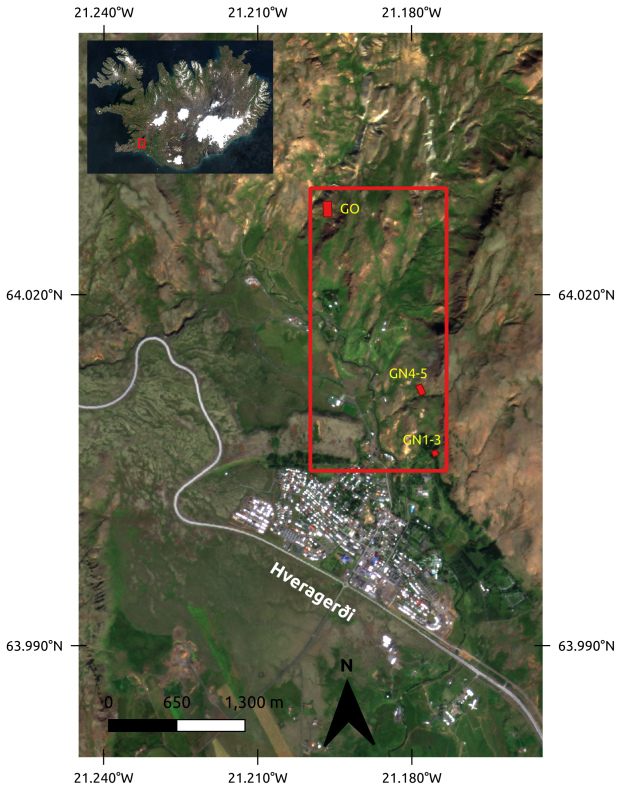


Figure 3.1: Map depicting the research site locations near the village of Hveragerdi, Iceland. "GO" (grassland old) marks the sites where the soil has been warming for over six decades, and "GN" (grassland new) denotes the sites where soil warming began following the May 2008 earthquake.

The study was carried out in the south of Iceland near the village of Hveragerdi on the ForHot site [215], as shown in Figure 3.1. Following an earthquake in May 2008, the bedrock of one unmanaged (cold) grassland field site underwent a disruption, resulting in the creation of areas with differently warmed soils. Another nearby grassland field site had had such warmed soil gradients for at least six decades, and those were not disturbed by the earthquake in 2008 [215]. In spring 2013, five transects were selected in each field site, each with five permanent plots across the natural soil temperature gradients, resulting in a total of 50 studied plots. We categorized the plots according to their annual soil temperature range, as indicated in Table 3.1.

Table 3.1: Category of the temperature range of the plots.

Category	Temperature Range
A	Ambient
B	+0.5 to 1°C
C	+2 to 3°C
D	+3 to 5°C
E	+5 to 10°C

3.2.1.1 NDVI data

To be able to estimate vegetation phenology characteristics, we measured the NDVI of all studies plots using a handheld instrument from SKYE Instruments (SpectraoSense2). From 2014 to 2019, NDVI measurements were done approximately bi-weekly from April to November, except during periods with continuous snow cover in early spring, late autumn, or winter. The measurements were always conducted on a clear day. We refer to [182] for further information about the NDVI measurements. As can be seen in Figure 3.2, the NDVI data clearly showed a seasonal pattern, with a higher NDVI in the summer months.

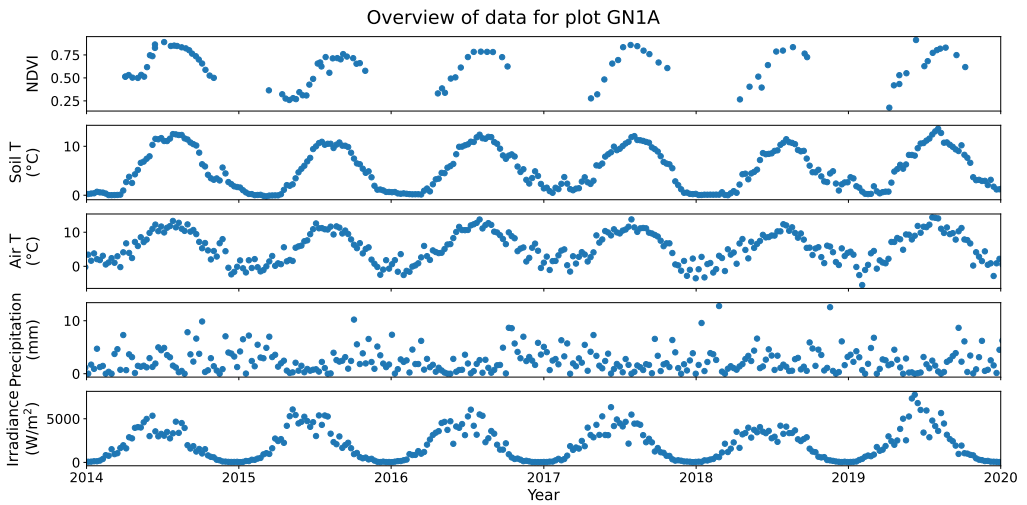


Figure 3.2: Overview of all available variables for plot GN1A (unwarmed control plot). Whereas the NDVI and soil temperature (upper two figures) are unique for all 50 plots, the meteorological variables (bottom three figures) are the same for every plot.

3.2.1.2 Soil Temperature data

The soil temperature at a depth of 10 cm was monitored in all the permanent plots using HOBO TidbiT v2 Water Temperature Data Loggers (Onset Computer Corporation, USA) since the spring of 2013 [215]. In Table 3.1, the different soil warming categories with their accompanying temperature range are given, while Figure 3.2 shows the data for one of the 50 plots used in this study. The main soil warming effect was an approximately constant shift in temperature across the seasons, as shown by [215].

3.2.1.3 Meteorological data

In addition to NDVI and soil temperature data, we also used meteorological data. As the measurement of meteorological variables such as irradiance (global radiation), precipitation, and air temperature at the Forhot site only began in 2019, we relied on data from another source. Specifically, we obtained the aforementioned meteorological variables from a weather station in Reykjavík[1],located approximately 40 km from the research site, as this is the closest station where irradiance is measured. We aggregated the data by taking the average on a weekly resolution scale, and assumed that the weather conditions are the same for all plots during each year. Given the distance between the weather station and the research plots, the data serve as a proxy for the actual weather conditions at the ForHot site. In Figure 3.2, the three bottom panes show all meteorological variables measured in the relevant period.

3.2.2 Data analysis

3.2.2.1 Estimating the NDVI seasonal characteristics

To extract the intra-annual vegetation phenology characteristics (SOS, POS and PEAK) in each plot during each growing each growing season, we first fitted a curve to the measured NDVI data. Based on the approach of [216], we used a double logistic curve. We require that the two logistic curves transition into each other continuously, such that the resulting function is differentiable at every point. These requirements result in the following formula for the estimated NDVI:

$$\widehat{NDVI}(x) = \begin{cases} \dfrac{c}{1 + e^{b_1 \cdot (x - a_1)}} + d & x \leq p \\ -\dfrac{c}{1 + e^{b_2 \cdot (x - a_2)}} + d + c & x > p \end{cases} \tag{3.1}$$

where the parameters a_1, a_2, b_1, b_2, c, d and p are fitted to a season's NDVI data and x represents the week number ($x \in 0, 1, \ldots, 52$) of the year. The parameter p has an important interpretation, as it is defined as the date of the POS, i.e., where the maximal NDVI value is reached.

[1]Data courtesy of the Icelandic Meteorological Institute.

The best fit for the curve parameters is found using the Trust Region Reflective algorithm [225]. This generally robust optimization method finds the optimal set of parameters by minimizing the mean squared error (MSE) between the predicted NDVI curve and the NDVI data points. After the curve parameters have been fitted, we extracted the start SOS, POS and PEAK for each plot in each year.

The SOS is considered to be the time of year when the NDVI increases the fastest, i.e., the curvature of the NDVI curve increases the most. This can be calculated using the second derivative of the fitted curves. As shown in Figure 3.3 , the estimated start of season is the moment in time when the second derivative of the first logistic function is maximal. Combined with the aforementioned definition of the POS, we establish the following equation for calculating the relevant vegetation phenology characteristics:

$$\widehat{SOS} = \underset{x}{\operatorname{argmax}} - \frac{cb_1^2 e^{b_1(x-a_1)}\left(-e^{b_1(x-a_1)}+1\right)}{\left(1+e^{b_1(x-a_1)}\right)^3} \tag{3.2}$$

$$\widehat{POS} = p \tag{3.3}$$

$$\widehat{PEAK} = \widehat{NDVI(p)} \tag{3.4}$$

where $\widehat{SOS}$ indicates the estimated start of the season, $\widehat{POS}$ the date of the peak of the season, and $\widehat{PEAK}$ the maximum value of the NDVI.

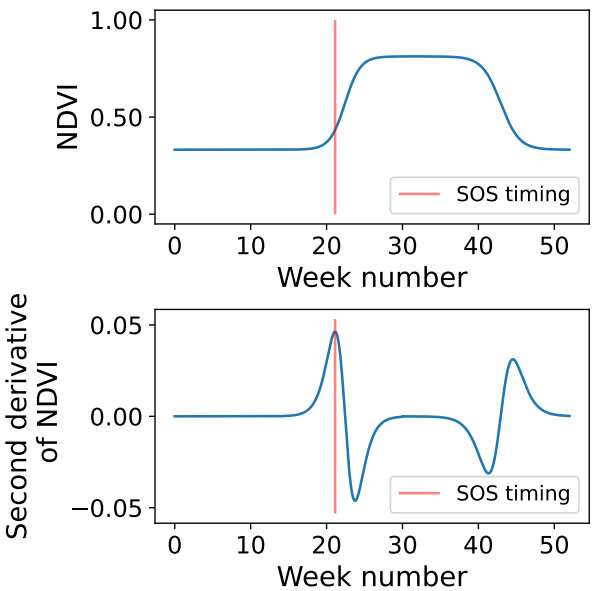


Figure 3.3: The SOS is estimated based on the second derivative of the fitted NDVI curve. The SOS is defined as the week when the NDVI curvature increases the most, and is indicated with a red line.

3.2.3 Statistical modeling and machine learning

3.2.3.1 Linear regression

After identifying the start and peak of the season for each plot and year, we performed a linear regression analysis. In this analysis, we used the SOS, POS, and PEAK as depenent variables, with the average soil temperature in each plot an year as the independent variable. We conducted this analysis using the ordinary least squares method available in the statsmodels library (version 0.13.2) for Python 3.9.13 [226]. This approach also enabled us to compute the p-values for the slope and intercept of the linear model through a t-test. These p-values help use determine the statistical significance of the relationship between soil temperature and vegetation phenology characteristics by indicating whether the observed relationships are likely due to chance or reflect a genuine underlying pattern.

3.2.3.2 Machine learning

To better understand the inter-annual variability in our results, we used ML techniques to create models predicting different vegetation phenology characteristics. Specifically, we trained three separate MLPs: one to predict the start of the season, another to predict the peak of the season, and a third to predict the height of the peak season. An MLP is a type of ANN designed to mimic the way the human brain processes information. It consists of multiple layers of nodes (neurons): an input layer, one or more hidden layers, and an output layer which is used to provide the final predictions. Each node in a layer connects to every node in the next layer, with each connection having a specific weight. During training, the MLP adjusts these weights to minimize the difference between its predictions and the actual outcomes using an algorithm called backpropagation, allowing it to learn complex patterns in the data [227].

Contrary to the linear models introduced in Section 3.2.3.1, the MLPs also take meteorological variables into account. This meant that in total, each MLP used 79 input variables, which included the average weekly air temperature, precipitation and solar irradiance for the first 26 weeks of the year, as well as the average soil temperature over the entire year. We implemented the MLPs using the MLPRegressor class from the scikit-learn package (version 1.1.3) [133]. To ensure the models were as accurate as possible, we optimized their hyperparamaters – the parameters that control the learning process – through a process called grid search, which we performed using Optuna (version 3.1.0) [228]. This process involved testing different combinations of hyperparameters to find the best settings for each of the three target variables. A description of these hyperparameters, the ranges we explored, and the optimal values we found are provided in Table 3.2.

To evaluate how well the models performed, we used three standard metrics: MSE, mean absolute error (MAE), and the coefficient of determination (r^2). For the grid search, we focused on minimizing the MSE to identify the optimal set of hyperparameters. Prior to conducting the grid search, we divided our data into a training set (80% of the data) and a test set (20% of the data). This split ensures that the models are trained on one portion of the data and tested on a separate, previously unseen portion, allowing us to assess their ability to generalize to new, unseen data accurately.

Table 3.2: Overview of the explored ranges of hyperparameters used in the Optuna grid search. The optimal values for the three different regression tasks are displayed in the right-most three columns. If the number of neurons in a layer is zero, the layer was not used in the network.

Description	Range	SOS	POS	PEAK
Number of neurons in first layer	int: 10, 20, . . . , 100	100	70	30
Number of neurons in second layer	int: 0, 10, . . . , 100	0	0	100
Strength of the L2 regularization term	float: 1e-4 — 1e-1 logscale	0.0290	0.0010	0.0606
the solver for weight optimization	adam, lbfgs	adam	adam	adam
initial learning rate	float: 1e-4 — 1e-1 logscale	0.0031	0.0003	0.0028
learning rate schedule for weight updates	constant, adaptive	constant	adaptive	adaptive
maximum number of iterations	int: 1000, 2000, . . . , 10000	8000	8000	8000
maximum number of iterations with no improvement	int, 10, 20, . . . , 100	20	50	100

3.2.3.3 SHAP values

The 79 input features we used are not equally important, and each one influences the model's predictions differently. To understand which features are most significant, and what the direction of their impact is, we use SHAP values. They are calculated by by examining how the model's predictions change when a specific feature is included or excluded, considering all possible combinations of features [59]. By averaging these effects, SHAP values provide a clear and fair measure of each feature's contribution to the final prediction. This method ensures that the importance of each feature is assessed in the context of all other features in the model. In the end, SHAP values can break down each prediction made by the model, showing the contribution of each feature. The sum of the SHAP values for all features then equals the model's output.

After training the MLP models, we computed SHAP values using the model-agnostic Kernel SHAP method to understand which features are most important in predicting the start and (height of the) peak of the greening season. We used the implementation in the Python SHAP package for this analysis [59].

3.3 Results

3.3.1 The logistic fitting

For most plots and years, good fits were found for the double logistic curves that were fitted to the intra-annual individual plot NDVI data, with an average r^2 of 0.942 (± 0.095). However, for 5.8% of all plots and years, the data did not follow a double sigmoid curve, and the r^2 value was lower than 0.80. These curves were not included in the analysis. The mean estimated SOS was week 20.41 (± 2.40), the mean estimated POS was week 29.97 (± 3.27), and the mean estimated PEAK was 0.842 (± 0.071) across all the soil warming treatments.

3.3.2 The average response to soil temperature

Figure 3.4 shows the linear relationship found between the average annual soil temperature and the three NDVI characteristics found by the double-logistic curves. The parameters of the linear model are given in Table 3.3. A significant linear relationship was found between average soil temperature and SOS ($p < 0.001$), POS ($p = 0.001$) and PEAK NDVI ($p < 0.001$) (Figure 3.4 and Table 3.3). The relationship between soil temperature and SOS was negative, with an estimated coefficient of -0.2160 (± 0.053). This means that for every 4.63 degrees of soil warming, the greening season starts a week earlier. Otherwise stated, the SOS happens 1.52 days earlier per degree of soil warming when derived across multiple years. Similarly, we see that the date of the NDVI peak shifted forward. The estimated coefficient of -0.2353 (± 0.07) indicates that for every 4.25 degrees of soil warming, the NDVI peaks a week earlier, or the POS occurs 1.65 days earlier per degree of soil warming. Finally, the PEAK value of the NDVI curve increased slightly with increasing soil temperature.

Table 3.3: The parameters describing the results of the linear models, where different variables are fitted against the average soil temperature over a whole year. The SOS and POS are measured in weeks, while the intercept is measured in degrees Celsius.

Target variable	Slope	Intercept	r^2	p-value
SOS	-0.216 ± 0.052	22.011 ± 0.454	0.06	<0.001
POS	-0.235 ± 0.070	31.755 ± 0.607	0.04	0.001
PEAK	0.005 ± 0.001	0.801 ± 0.013	0.05	<0.001

Although the linear relationships that were observed between average soil temperature and SOS, POS, and PEAK were significant (Figure 3.4), we also observed a lot of unexplained variance, which is indicated by the relatively low r^2 values in Table 3.3.

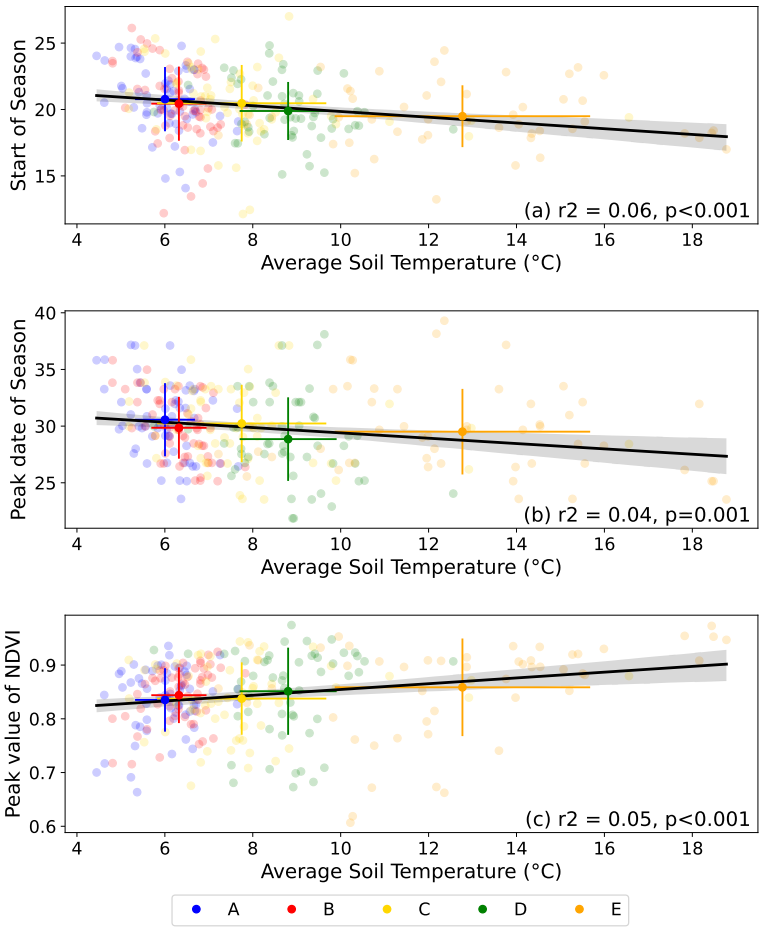


Figure 3.4: Linear model that predicts the start of the season (a), the peak date of the season (b) and the peak value of NDVI (c), based on the average annual soil temperature. The filled circles represent the mean values for each category (A to E) of average soil temperature, with error bars indicating the standard deviation. The semi-transparent circles represent individual observations. The color indicates the soil warming category where the blue points are A plots, the red points are B plots, the yellow points are C plots, the green points are D plots, and the orange points are E plots. All models had a significant relationship between the average soil temperature and the studied NDVI curve parameter (See Table 3.3).

3.3.3 The machine learning approach

To explain a larger part of the variance, the possibility of predicting characteristics of the NDVI curve using MLPs, based on both the soil temperature and meteorological variables, was investigated. The performance of the MLPs can be found in Table 3.4. From Tables 3.3 and 3.4, it becomes evident that the inclusion of the meteorological variables and the utilization of MLPs enabled us to explain a significantly larger part of

the variance compared to the linear models.

Table 3.4: Model performance of MLP after a 5-fold cross validation (CV) grid search. The test set consists of 20% of the total data, and is split evenly across the years of data taking. The naive MSE (MAE) is the MSE (MAE) when the mean of all training samples is used as the prediction.

Target	5-fold CV MSE	Test MSE (naive)	Test MAE (naive)	Test r^2
SOS	3.408	4.760 (7.102)	1.521 (2.095)	0.322
POS	7.933	8.943 (11.103)	2.473 (2.696)	0.192
PEAK	0.004	0.004 (0.006)	0.053 (0.063)	0.248

To investigate the impact of a given feature on the predictions made by the model, we calculated SHAP values for all three MLPs. These can be found in Figure 3.5, Figure 3.6 and Figure 3.7 for the SOS, POS and PEAK, respectively. In these figures, we separate the six years to investigate the annual variation in the SHAP values. To obtain the SHAP value for one meteorological variable, we summed the SHAP values of the 26 weekly averages, as shown in Equation (3.5). Next, we calculated the sum of absolute values of the SHAP values, A_SHAP, for the four remaining features for all n samples, as shown in Equation (3.6). By taking the absolute value and adding it over all years, we can investigate the total impact of a feature on the prediction, regardless of the direction of the impact. The results for the A_SHAP values are shown in Figure 3.8.

$$\text{SHAP}_{feature} = \sum_{week=1}^{26} \text{SHAP}_{feature,week} \tag{3.5}$$

$$\text{A\_SHAP}_{feature} = \sum_{i}^{n} |\text{SHAP}_{feature,i}| \tag{3.6}$$

When interpreting Figures 3.5 and 3.8a, we see that the meteorological variables had the largest impact on the prediction of the SOS. However, within each year, this impact was approximately constant. The intra-annual variation in the SOS was clearly the result of soil warming. In fact, the Pearson correlation between soil temperature and its accompanying SHAP values was -0.93, meaning that the higher the soil warming, the earlier the season started each year. All Pearson correlation values can be found in Table 3.5.

From Figures 3.8b and 3.8c, we can also conclude that the three meteorological variables also had the largest impact on the predictions of the POS and PEAK. From Table 3.5, we can see that the POS was earlier and the PEAK value of the NDVI was higher with increasing soil temperature, as they had a Pearson correlation coefficient of -0.85 and 0.91, respectively. For the POS, Figure 3.6 indicates that the size and direction of the SHAP effect for the three meteorological variables shifts significantly over the years, while the smaller effect of the soil temperature is relatively stable across the six years and drives the intra-annual variation within the dataset.

Table 3.5: Pearson correlation coefficient between the average soil temperature and its corresponding SHAP values.

Target variable	Pearson correlation
SOS	-0.93
POS	-0.85
PEAK	0.91

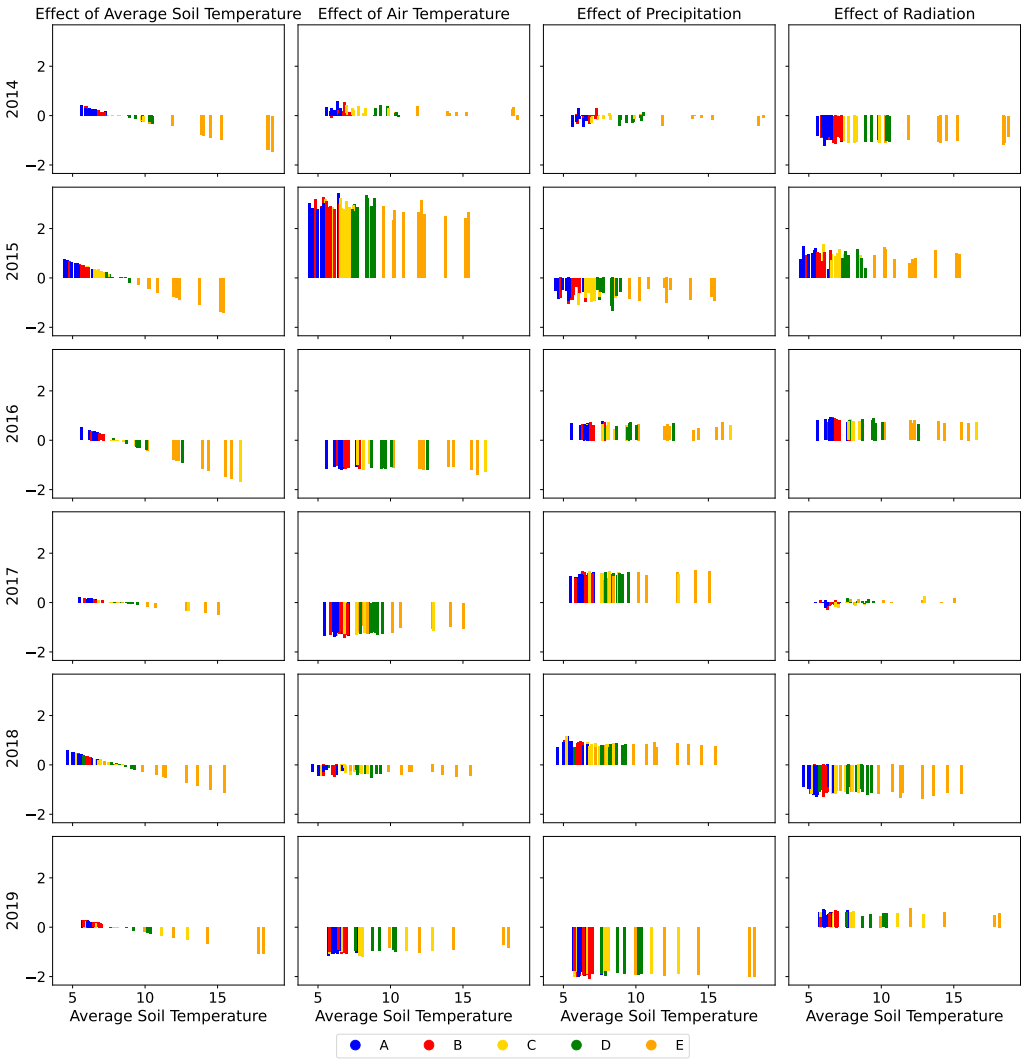


Figure 3.5: SHAP values of multi-layer perceptron that predicts the start of the greening season based on the average soil temperature, air temperature, precipitation, and radiation. The color indicates the soil warming category where the blue bars are A plots, the red bars are B plots, the yellow bars are C plots, the green bars are D plots, and the orange bars are E plots.

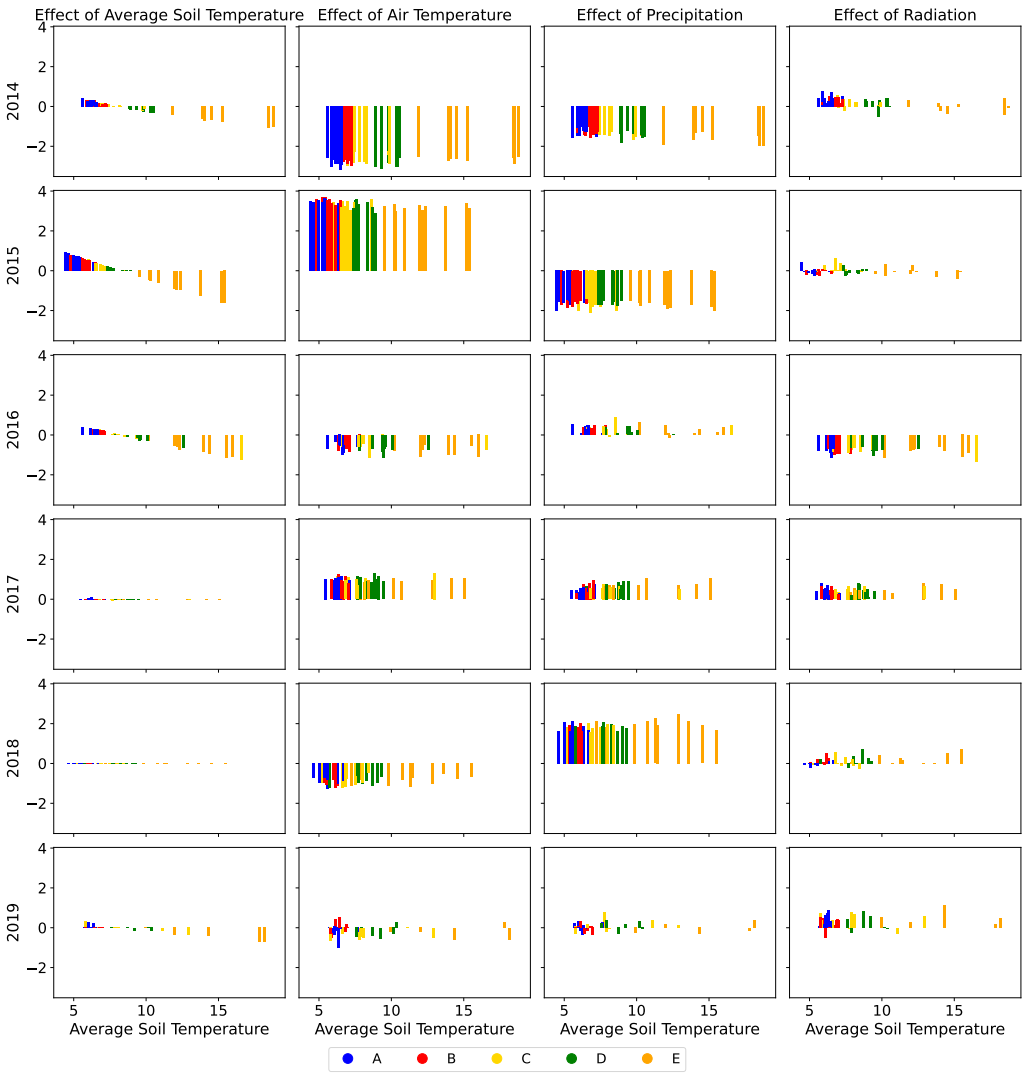


Figure 3.6: SHAP values of multi-layer perceptron that predicts the peak of the greening season (POS) based on the average soil temperature, air temperature, precipitation, and radiation. The color indicates the soil warming category where the blue bars are A plots, the red bars are B plots, the yellow bars are C plots, the green bars are D plots, and the orange bars are E plots.

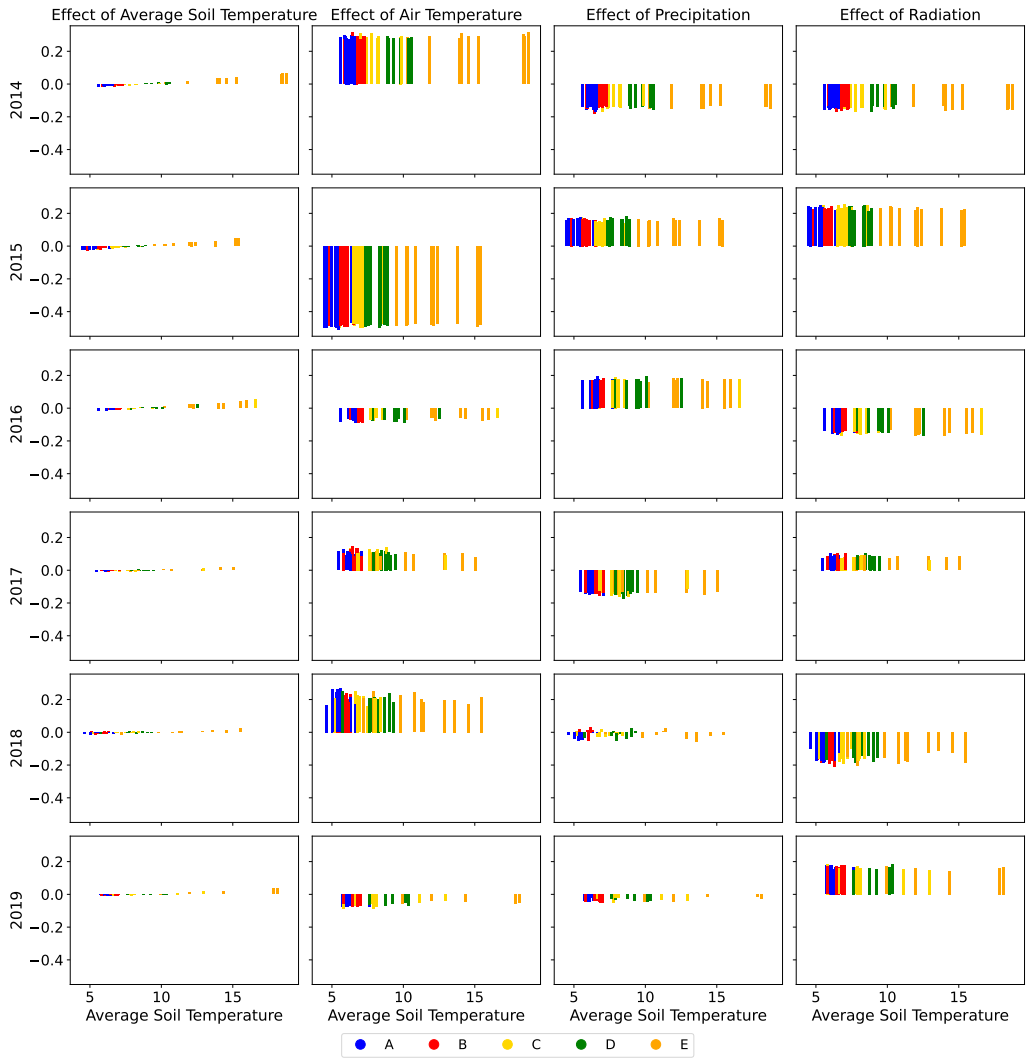
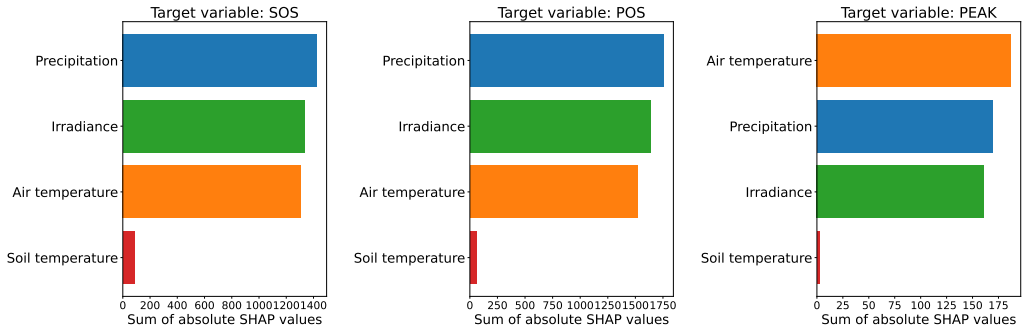


Figure 3.7: SHAP values of multi-layer perceptron that predicts the peak NDVI based on the average soil temperature, air temperature, precipitation, and radiation. The color indicates the soil warming category where the blue bars are A plots, the red bars are B plots, the yellow bars are C plots, the green bars are D plots, and the orange bars are E plots.



(a) Start of season (b) Peak of season (c) Peak NDVI value

Figure 3.8: Sum of the absolute SHAP values as defined in Equation (3.6).

3.4 Discussion

The purpose of this study was to explore the relationship between soil temperature and NDVI, along with the impact of meteorological variables, utilizing ML techniques. The discussion will focus on emphasizing the novelties of this work, addressing the hypotheses presented in the paper, discussing the findings in relation to previous research, and highlighting the implications of the results.

3.4.1 Using machine learning to study vegetation phenology

Currently, the standard practice in vegetation phenology studies using NDVI consists of using traditional statistical methods such as (non-)linear regression or linear mixed models [182–187]. However, our results clearly indicate that, after applying linear regression, a large amount of unexplained variance remains. Our study advances the traditional approach by using ML models, specifically MLPs, which integrate meteorological variables to capture nonlinear relationships. This method allowed us to explain a larger portion of inter-annual variance compared to traditional methods. The use of SHAP values further provided insights into a deeper understanding of the complex interactions between soil temperature, meteorological variables, and NDVI dynamics.

3.4.2 Effect of the soil temperature on SOS, POS, and PEAK in subarctic grasslands

The first hypothesis stated that a higher soil temperature would lead to an earlier SOS based on previous research by [182]. Such responses have also been found when past changes in NDVI have been related to changes in annual, seasonal or monthly temperatures [162, 163, 229].

The findings of this study supported this hypothesis, as a significant relationship was observed between average soil temperature and the start of the greening season. The

negative coefficient (-0.2160) indicates that SOS occurs 1.5 days earlier per degree of soil warming across the six years. This finding was consistent with a recent analysis from the International Tundra Experiment covering up to 20 years of data from 18 sites and 46 open-top chamber warming experiments across the Arctic, sub-Arctic, and alpine ecosystems [230]. They observed a 0.73-day earlier start of the greening season, in an environment where the average air warming was 1.4 °C and the soil warming approximately half of that [230]. Our finding was also consistent with previous research at the same ForHot site, as [182] found that on average, the SOS occurred 1.6 days earlier for every degree of soil warming.

Day length has traditionally been considered a dominant factor in regulating the phenology of many plant species [217, 218], particularly in high-latitude ecosystems where day length changes significantly throughout the growing season. Therefore in this study, the second hypothesis stated that the date of the POS would occur at a similar time each year, regardless of soil temperature, as the day length remains consistent across all years. However, our results indicate that temperature conditions in the soil can have a considerable influence on the timing of POS. The hypothesis was therefore rejected. This finding suggests that, in our sub-Arctic grasslands, day length might not be the primary factor influencing the timing of the POS. While previous studies have highlighted the interplay between day length and air temperature in determining phenological events [231], our study is unique in demonstrating the notable impact of soil temperature. This underscores the need to consider soil temperature as an influential factor in phenological studies, particularly in the context of climate change where both soil and air temperatures are rising.

The third hypothesis proposed that the PEAK NDVI would not be significantly related to soil temperature, based on previous research by [220], who had not found significant differences in vegetation biomass across the warming gradients. However, the findings of this study indicate a slight increase in the PEAK value with increasing soil temperature. Although the relationship was not as strong as for the SOS and POS, it suggests that higher soil temperatures may contribute to higher NDVI peak values. It is worth noting that while NDVI is often used to estimate vegetation biomass [232–235], it is not measuring it directly, but rather the amount of chlorophyll per surface area [148]. Therefore, "Arctic greening" measured using the NDVI, could occur without any changes in vegetation biomass, if the plants are getting "greener" due to a higher nutrient content in warmer soils. Further research is needed to better understand this relationship and its underlying mechanisms.

3.4.3 Effect of the other meteorological variables

Hypothesis B focused on the impact of other meteorological variables (air temperature, precipitation, and irradiance) on the inter-annual variability of the NDVI phenology and PEAK values, and the potential of ML to identify their importance. The results of the ML analysis using MLPs showed that these variables have a strong impact on the predictions of the SOS, POS, and PEAK, and the r^2 values of the MLPs were much higher than those obtained by the linear regression.

The SHAP values also provided information on the relative importance of these variables. It was noteworthy that the three meteorological variables had a much larger impact on

the predictions than the soil warming data. These findings align with other studies that emphasize the significance of climatic variables over soil conditions because of their influence on soil temperature in predicting vegetation responses [221, 222]. However, the intra-annual variation in the SOS, POS, and PEAK was found to be influenced by the soil temperature. This influence of soil temperature highlights the significant role of below-ground processes in driving vegetation phenology and productivity [175]. Studies have shown that soil temperature can affect root growth, nutrient availability, and microbial activity, all of which are crucial for plant development [174, 177, 220]. Understanding these interactions is essential for accurately predicting how the ecosystems will respond to ongoing climate change.

The SHAP values did not indicate significant differences among the meteorological parameters, making it challenging to prioritize their impact as hypothesized. This contradicts our hypothesis that air temperature's influence would be minimal due to its regulation of soil temperature, precipitation would have an intermediate effect, and irradiance would have a substantial impact, especially in cloudy sub-Arctic climates [215, 224]. However, collectively, these meteorological factors exhibited a considerably higher influence on the predictions compared to the soil warming data. Therefore, our findings not only contribute to understanding the dominant impact of meteorological parameters on vegetation dynamics, but also emphasize the need for continued research to explain the interdependencies and potential interactions between these factors.

3.4.4 Methodological considerations

It is important to note some limitations of the study. The analysis focused on a specific location in Iceland, and the results may not be directly applicable to other regions. The study period also covered a limited period of time (2014-2019), and longer-term data would provide a more comprehensive understanding of the inter-annual variation in NDVI. Furthermore, the meteorological data does not have the same spatial resolution as the NDVI or soil temperature data. Indeed, as we relied on the measurements of the nearest weather station, we had to assume that the weather conditions were the same across all plots.

The SHAP values should also be interpreted with caution. Although they are model-agnostic, we can only draw valid conclusions if the model generalizes well. That is, if it has an acceptable test set performance [236]. Furthermore, the SHAP values do not have a causal interpretation [237]. We cannot assume that if the variable X has a large impact on the prediction of Y, then X causes Y. On the contrary, Y might cause X, X and Y could both be caused by a confounding variable, or they could have no causal relationship at all.

Nevertheless, this study produces valuable insights and provides clear directions for future research. Our promising results, achieved by applying ML in a vegetation phenology study, emphasize the potential of this approach in advancing our understanding of seasonal plant characteristics based on NDVI data. They can also be viewed as a starting point for other analyses in a broader ecological context.

In the future, it would be interesting to consider other model architectures or methodologies, as this could further validate our results, or maybe even improve them. Additionally,

other xAI approaches like LIME [113] could be considered, allowing comparison between different xAI approaches.

3.5 Conclusions

Our results only partly supported our hypotheses regarding the effect of soil temperature on the timing of the SOS, timing of the POS, and peak NDVI values. We observed a significant relationship between soil warming and the timing of SOS and POS, indicating that higher soil temperatures advance the onset of the growing season. Unexpectedly, this also led to a corresponding shift in the timing of POS. Moreover, the peak NDVI values showed a slight increase with higher soil temperatures. Furthermore, we explored the impact of meteorological variables, more specifically air temperature, precipitation, and irradiance, on vegetation phenology and its inter-annual variation. The use of SHAP values allowed us to gain insight into the relative importance and contribution of each meteorological variable to the predictions. It became evident that the three meteorological variables had the largest impact on the prediction of SOS, POS, and PEAK NDVI values across the six years. However, within a given year, the impact of the three meteorological variables remained approximately equal, while the variations in phenological characteristics were primarily driven by soil temperature.

For future work, we suggest further exploration of the underlying mechanisms driving the observed relationships between soil temperature and phenology. Investigating the physiological responses of plant species to soil temperature variations and exploring the interactions between soil temperature and other environmental factors at finer temporal and spatial scales would provide a more comprehensive understanding. Additionally, collecting data considering the soil characteristics, e.g., soil chemistry or nutrient availability, could improve the performance of the ML models, and further increase the explained variance.

In addition, incorporating advanced remote sensing techniques, such as satellite imagery, in conjunction with ground-based measurements can improve the accuracy and comprehensiveness of phenological studies in subarctic grassland ecosystems. Long-term monitoring at multiple sites and the incorporation of various geographical locations would provide valuable information on the generalizability of our findings and the response of subarctic grasslands to ongoing climate change.

This study contributes to our knowledge of the relationships between soil temperature, other meteorological variables, and vegetation phenology in subarctic grassland ecosystems. The findings enhance our understanding of the mechanisms driving ecosystem dynamics in these regions and have implications for predicting and managing subarctic grasslands in the face of environmental change. Finally, this work also functions as a proof-of-concept for ML-based vegetation phenology studies, and thereby provides a solid foundation for future research in this domain.

3.6 Data availability statement

The data used in this study is made available as open data and can be found here: `http://dx.doi.org/10.17632/C9T7FX9N4H.1` [238].

Chapter 4

Missing value imputation of wireless sensor data for environmental monitoring

This chapter is based on:

Mortier, Steven, Thomas Decorte, Jonas J. Lembrechts, Filip J. R. Meysman, Steven Latré, Erik Mannens, and Tim Verdonck. Missing Value Imputation of Wireless Sensor Data for Environmental Monitoring. *Sensors* 24, no. 8: 2416. 2024

4.1 Introduction

Over the last decade, major advances in wireless communication technology, microelectronics, and (big) data analytics have caused a significant increase in the application of wireless sensor networks (WSNs) [239, 240]. A WSN comprises a network of many spatially distributed sensors that monitor certain parameters of a physical system and engage in wireless data communication. The WSN is made up of sensor nodes, sometimes also called sensor motes, which are essentially microcomputers with the ability to collect data, process these data internally, and finally transmit these data to a centralized location. WSNs have numerous applications in different fields, including environmental monitoring, health monitoring, logistics, and smart cities [241, 242]. With the increasing use of WSNs, there is a growing demand for performant data analysis techniques capable of handling the vast volumes of collected data.

An important challenge within WSN research concerns missing value imputation for the extensive spatiotemporal datasets that are generated. Unavoidably, networks tend to lose readings from sensors for reasons that are difficult or impossible to anticipate, such as sensor failure due to power depletion, network outages, and communication errors, but also destruction due to storms or vandalism [243].

These missing readings can have important consequences for real-time monitoring, for example, in an emergency setting. Likewise, environmental monitoring applications

relying on WSN data can suffer from missing data, which might lead to delayed or incorrect responses to environmental changes. Additionally, missing values can weaken the reliability of sensor data and increase the difficulty of sensor calibration. Finally, incomplete data can also compromise the performance of subsequent modeling and statistical analysis, which may result in biased conclusions or inaccurate predictions. A concrete example can be found in environmental research, where a WSN is commonly leveraged to measure variables such as temperature, humidity, atmosphere pressure, and sunlight, among others. Despite the wealth of data collected by sensor nodes, they often exist in raw form. Analytical tools commonly employed in such fields, such as support vector machines, principal component analysis, and singular value decomposition, face limitations when confronted with datasets containing missing data. Consequently, addressing the issue of missing data in these datasets presents a significant hurdle, impacting the efficacy of analyses and hindering the ability to draw meaningful conclusions [244].

The objective of this study was to evaluate the performance of missing value imputation techniques on a dataset generated by a WSN for environmental monitoring. To this end, we employ a unique dataset that originated from one of the largest citizen science projects to date involving Internet of Things (IoT) monitoring. Throughout the summers of 2021 and 2022, 4400 citizens within the region of Flanders (Belgium) installed IoT sensors in their gardens to measure the temperature and soil moisture at a high temporal frequency (every 15 min). The goal of this citizen science project, called CurieuzeNeuzen in de Tuin; Nosy Parkers in the Garden (CNidT), was to gain insight into how garden ecosystems can provide cooling for climate adaptation and mitigate the impacts of extreme weather events like heat waves. In projects like CNidT, missing values in the sensor time series are undesirable, both from a scientific and from a citizen perspective. From a scientific perspective, the data incompleteness reduces the power of the ensuing statistical analysis, which here aimed to uncover the factors that drive local garden cooling during extreme weather events. Likewise, data incompleteness was also highly unwanted from the citizen perspective: participating citizens were updated daily through personal dashboards, while society at large was informed through real-time maps on the website of a national newspaper. However, missing values were common in the recorded time series due to a combination of random sensor failure (e.g., battery problems), failed data transfers (e.g., due to network outages), and errors made by the citizens (e.g., destruction or damaging of the sensor). For these reasons, the dataset from the CNidT project was especially suitable as a case study for missing value imputation in WSN data. The CNidT dataset is an integral component of the SoilTemp project, which is a publicly available database outlined in Lembrechts et al. (2020) [245]. This extensive database comprises data from 7538 temperature sensors spanning 51 countries and encompassing diverse biomes. The primary objective of the SoilTemp project is to enhance the global comprehension of microclimates and to address discrepancies between existing climate data and the finer spatiotemporal resolutions pertinent to organisms and ecosystem dynamics [245].

Given that missing data within WSNs pose a fundamental challenge, the development of methods capable of imputing these missing values represents an active area of research. Within our study, several imputation approaches were evaluated to analyze their performance. An overview of all considered approaches and their imputation strategies is given in Table 4.1. A first approach involves techniques that take advantage of the temporal correlation between data, thus imputing missing values for a given sensor using the available data of that same sensor at different time steps. Evaluated methods for this approach include mean and linear spline imputation [246]. A second class of techniques

utilizes spatial correlation to impute values, focusing on data from other sensors in the network at the same time step to impute the missing values of one sensor. Evaluated methods for this approach include k Nearest Neighbours (KNN) imputation [247], multiple imputation (MI) techniques such as Multiple imputation using chained equations (MICE) and Markov chain Monte Carlo (MCMC) [248–250], and random forests (RFs) to replace missing data (MissForest) [251]. The last strategy combines both the spatial and temporal aspects, taking full advantage of the patterns and intricacies present within the data. For this, specific methods for WSNs have been developed, such as data estimation using statistical model (DESM) and applying k-nearest neighbor estimation (AKE). Matrix completion (MC) methods can also be exploited here as they use correlations within one sensor and across multiple sensors but assume that the data is static, i.e., they ignore the temporal component of the data [252, 253]. Other methods in this class tend to leverage deep learning to impute missing values, for example multiple imputation using denoising autoencoders (MIDA) [254] or recurrent neural network (RNN)-based approaches such as bidirectional recurrent imputation for time series (BRITS) and multi-directional recurrent neural network (M-RNN) [253, 255]. For a detailed explanation of all imputation methods evaluated in this study, we refer to Section 4.2.3.

Table 4.1: The imputation techniques that were considered in this study, together with their respective imputation strategy.

Method	**Imputation Strategy**
AKE [244]	WSN-specific
BRITS [255]	Deep learning
DESM [256]	WSN-specific
KNN [247]	Spatial correlations
MC [252]	Temporal and spatial correlations (static)
MCMC [250]	Spatial correlations
MICE [249]	Spatial correlations
MIDA [254]	Deep learning
MRNN [253]	Deep learning
Mean imputation	Temporal correlations
MissForest [251]	Spatial correlations
Spline [246]	Temporal correlations

Previous studies have conducted various comparative analyses, assessing different datasets, classes of algorithms, setups, and types or scenarios of missingness. However, in most studies, the focus is more on multivariate time series rather than on specific WSN data. [257] compared seven imputation methods across five publicly available datasets, concluding that KNN imputation exhibited the highest performance. Similarly, [258] evaluated six imputation techniques on 69 datasets, noting that random forest-based solutions generally outperformed others. Notably, their study also evaluated performance in downstream machine learning (ML) tasks, finding that the imputation rendered a 10 to 20% performance increase. [259] focused on sensor time series imputation, comparing 16 recovery algorithms on six public and two synthetic datasets, including block missings, which are more reflective of WSN data characteristics. Their findings suggested that the optimal recovery method often depends on dataset-specific characteristics. [260] assessed six imputation techniques using Turkish State Meteorological Service data, introducing

the correlation dimension technique to account for spatiotemporal dependencies in imputation evaluation. Their study indicated that the MCMC approach yielded the most favorable results.

In this study, we evaluated 12 imputation techniques for different artificial missing scenarios (by inducing 10%, 20%, 30%, 40%, and 50% data removal), as well as a more realistic scenario defined as "masked" missings. In this scenario, we replicated the missing patterns observed in sensors with incomplete data onto sensors with complete information, simulating a real-world missings scenario. In this way, we created a standardized scenario through which we can evaluate how effective every method is in compensating for the missing patterns. Comparisons are made based on the root-mean-square error (RMSE) and mean absolute error (MAE) to assess the accuracy of the imputed values. Our study advances the existing literature by conducting a comprehensive comparison of various missing value imputation methods, employing different strategies and model types. Moreover, we analyze a genuine WSN dataset featuring a substantial sensor count (1500) and expand the assessment of these techniques from random missing values to masked missing values, offering a more realistic evaluation scenario for practical deployment.

The remainder of this study is structured as follows: Section 4.2.1 introduces the CNidT project, while Section 4.2.2 describes the dataset collected in the project, as well as the preprocessing steps that were used. In Section 4.2.3, the imputation methods evaluated in this study are described, and the evaluation criteria are detailed in Section 4.2.4. Section 4.3 presents the results and discusses the implications of these results. In Section 4.4, we summarize our findings, list the most important insights and conclusions, and provide possible directions for future research.

4.2 Materials and Methods

4.2.1 The Curieuze-Neuzen Citizen Science Project

The dataset analyzed in this study originated from the citizen science project "Curieuze-Neuzen in de Tuin" (CNidT), which translates as "Nosy Parkers in the Garden" [261, 262]. The project engaged 4400 citizen participants across the strongly urbanized and densely populated region of Flanders (Belgium, Northwestern Europe) to monitor the microclimate in their garden. The scientific objective was to quantitatively assess the impact of gardens on the local microclimate and their potential role in the mitigation of extreme weather events [263]. Initially, 50,578 citizens registered as candidates to participate in the project. From this pool of registrations, 4400 sampling locations were selected using an environmental sampling algorithm, to obtain a representative subsample that covered the range of gardens in terms of size and composition but also to ensure a suitable geographical distribution across the measurement domain [264]. To this end, metadata were collected for each garden, including variables related to urbanity, garden characteristics, garden management, topography, and geography. These metadata were obtained from participants through questionnaires combined with available remote sensing data. Factor Analysis of Mixed Data (FAMD) was implemented to reduce the dimensionality of available metadata, and the selection algorithm used the first three Principal Components

(PC) from FAMD coordinates to hierarchically select candidate locations, maximizing variation between gardens in the available environmental space.

The project included two six-month measurement campaigns, spanning the growing season (spring and summer) of 2021 and 2022. Citizens received a microclimate sensor device (a "lawn dagger") that was inserted in the soil in the middle of a lawn patch in the garden. The microclimate sensor was a custom-tailored modified version of the well-established TMS logger, which provides a robust and cost-effective device to monitor temperature and soil moisture near the soil surface [265]. The TMS sensors recorded data every 15 min with three temperature sensors (DS7505 digital thermometer), one positioned at 10 cm below the soil surface, one at the soil surface, and one 12 cm above it. In addition, the TMS measures soil moisture using the time domain transmission principle in the top 15 cm of the soil [265]. The device has a large internal storage, allowing it to internally store the data collected over a period up to 10 years.

While regular TMS sensors only allow off-line data collection, a novel TMS-NB version of the instrument was specifically developed for the project, which was equipped with wireless transmission ability (collaboration between University of Antwerp, sensor development company TOMST, and telecom operator Orange Belgium). To this end, the TMS-NB was equipped with a data transfer module (BG77 Quectel with Qualcomm chipset) to send small data packages via the Narrowband Internet of Things (NB-IoT) network hosted by Orange Belgium. Measurements of temperature and soil moisture were recorded every 15 min, and the recorded data were stored in the device's internal memory. Data collected over one day were sent as one data package via NB-IoT each day at midnight. This data package included additional metrics (e.g., battery status, signal quality) and was transmitted via the LiveObjects platform of telecom operator Orange to a relational database (MS SQL) at the University of Antwerp.

The CNidT project thus gave rise to a large WSN (>4000 nodes) that performed NB-IoT-based environmental monitoring for a period of two summers (April–September). The use of low-cost sensors and reliance on citizen input occasionally led to erroneous values or missing data points due to various factors, including random sensor malfunctioning (e.g., occasional missing data), connectivity issues (i.e., data package not sent over the NB-IoT network), as well errors and accidents by the participants (e.g., sensors damaged by kid's play or robot lawn mowers). As sensor malfunctioning was virtually absent (<0.01% of data points) and connectivity issues could be solved by reading out the data manually after the end of the project, overall data availability ended up around 90% (see Section 4.2.2). Although such a percentage might be sufficient for most scientific questions, the project's goal of reporting back to individual citizens in real time about conditions in their own garden, as well as a subsequent analysis using ML methods, makes a gap-filling exercise especially appealing.

4.2.2 Dataset and Preprocessing

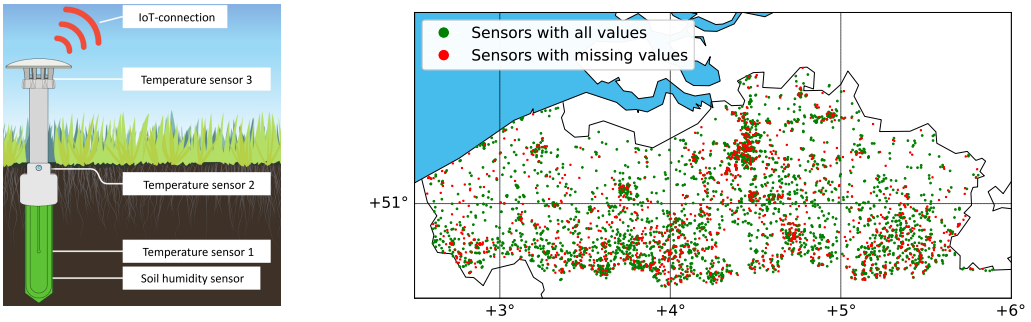
A subset of the available data was used to evaluate the missing value imputation techniques. Sensor readings were retained for one six-month measurement campaign (starting on 12 April 2021, at 00:00:00 until 30 September 2021, at 23:45:00), measured every 15 min, thus providing a maximum of 16,512 records in each sensor time series (172 days of data collection times 96 readings per day). Before the final construction of the

dataset, extensive manual data recovery measurements were done from the TMS-NB sensor to have as complete a dataset as possible. Furthermore, the additional metrics (e.g., battery status, signal quality) from the processed signal were also analyzed to generate as complete a sample as possible. Each sensor reading contained four data records (three temperatures at different heights and soil moisture). Data from the temperature sensor at 12 cm above the soil surface (expressed in degrees Celsius) were selected for the evaluation (and thus the remainder of our analysis), as these readings expressed the highest variability. This finding made sense, as it is generally known that the soil temperature becomes less variable over time as you go deeper into the soil profile [266]. Data series were available for 4163 sensor locations. Figure 4.1a illustrates the location of the sensors and whether the recorded time series for the sensor was complete or not. In total, 2978 sensors (or 71.5% of the WSN) had no missing values. Across all 4163 sensors comprising the WSN, 7.8% of records were missing. Although the missing percentage is not extremely high, about one-third of the sensors showed at least a few missing values, with some sensors missing nearly all values.

The geographical coordinates of the sensor location (uncertainty 10 m) are part of the metadata. As some methods utilize geometric distances between sensors, we calculated the haversine distance (as described by [267]) for every sensor combination as follows:

$$d_{hav}(x, y) := 2r \arcsin\left(\sqrt{\sin^2\left(\frac{y_{lat} - x_{lat}}{2}\right) + \cos(x_{lat})\cos(y_{lat})sin^2\left(\frac{y_{lon} - x_{lon}}{2}\right)}\right) \quad (4.1)$$

where x and y are the coordinates of two different sensors, and r is the radius of the Earth (6371 km). The haversine (or great circle) distance is the angular distance between two points on the surface of a sphere. We used the haversine distance rather than the Euclidean distance to account for the Earth's curvature given the scale of the measurement domain (Flanders region; ~300 km).



(a) (b)

Figure 4.1: (**a**) The TMS-NB microclimate sensor was used in a large-scale citizen science project on microclimate monitoring. The sensor measures temperature at three heights, as well as soil moisture. Data transmission occurred via NB-IoT. (**b**) The WSN covered 4400 gardens across Flanders. Sensor locations are colored based on whether time series were complete (green) or had missing records (red).

4.2.3 Missing Value Imputation

In the literature, various methodologies for imputing missing (sensor) data exist. Below, we provide a concise overview of the methods that were considered in this study. For a more detailed explanation, we refer to their respective papers and code implementations. First, we introduce different types of missing values, after which we detail how to create suitable test datasets for imputation.

4.2.3.1 Different Types of Missing Values

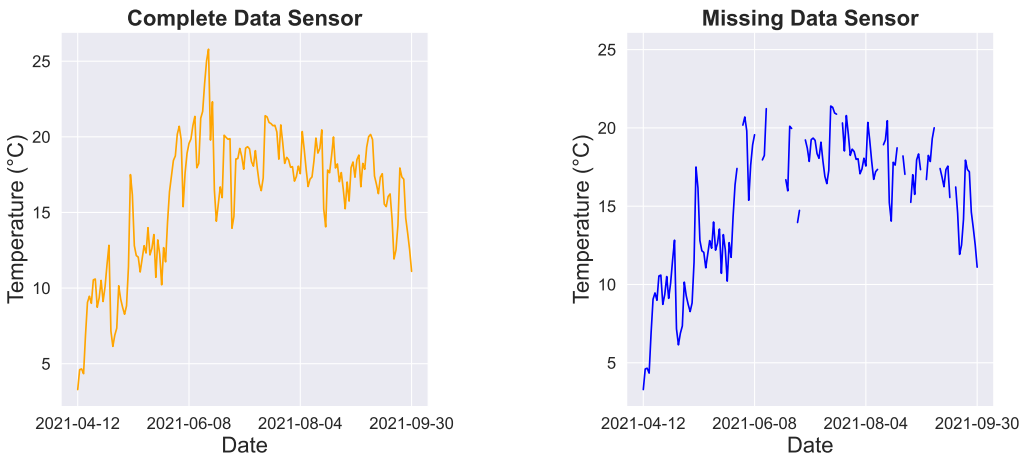
There are three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [268–270]. MCAR implies that missingness is independent of observed and unobserved data, making observed data still representative. This assumption is often strong and unrealistic. MAR means missingness is linked to observed but not unobserved data [268]. MNAR occurs when missingness is related to unobserved variables, making it the most challenging scenario to handle and non-ignorable [271].

4.2.3.2 General Approach

To test the performance of imputation methods, we restricted ourselves to data series that had complete information, as done in other comparison studies [253, 272]. From the 2978 sensors available with complete records, we selected a subset of 1500 as our basic dataset. In these complete data series, we artificially introduced data gaps that had to be resolved by imputation. Two separate approaches for missing value creation were applied.

In a first approach, we applied patterns of randomly missing values with increasing fractions of data missing (10%, 20%, 30%, 40%, and 50%) using the `numpy.random.choice` function [273]. This approach provides missing values that are MCAR, which is the missingness pattern that is most often used in the literature on missing value imputation [258]. Although it would also be possible to introduce MAR missing values, e.g., based on sensor location, we did not consider this option, as it is included in the masked missing scenario introduced below. Finally, MNAR missing values would manifest themselves by removing temperature values based on the actual temperature values themselves (e.g., low temperature values are removed). As our sensor operates within a range of $-55\,^{\circ}\mathrm{C}$ to $125\,^{\circ}\mathrm{C}$ [265], this scenario was not relevant for our dataset.

In reality, missing data patterns are not necessarily random. In order to mimic a more realistic case, we took advantage of the missing data patterns from the sensors that actually had missing data. To this end, missing patterns were imposed from incomplete sensors onto complete sensors, thus imposing a mask with missing data. In this approach, time points for which an incomplete sensor was missing data were imposed as a mask onto a sensor with a complete data series (as illustrated in Figure 4.2). This approach is further referred to as the masked missing. This approach allows for a more realistic evaluation, as potential issues such as spatial or temporal block missing, for example, due to network failure or sensor failure, will be present in the data.



(a) (b)

Figure 4.2: Example of missing data pattern in a representative sensor time series, which includes 15-min temperature recordings over a six-month period. (**a**) Time series for a sensor with complete data. (**b**) The same time series but with missing data artificially imposed. The missing time points are based on a mask derived from a different sensor with actual missing data.

For all six scenarios (10%, 20%, 30%, 40%, and 50% random missing as well as the masked missing), a 2×5 nested cross validation (CV) was run to obtain robust results. Nested CV involves two levels of CV loops: an outer loop and an inner loop. In the outer loop, the dataset is divided into training and testing sets using k-fold CV. Each fold of the outer loop trains the model on the training set and evaluates it on the testing set. Within each fold of the outer loop, an inner CV loop is employed where the training data are split into training and validation sets, also using k-fold CV. The inner loop is responsible for selecting the set of hyperparameters that performs best on the validation set. In our study, we used a 2×5 nested CV, i.e., we had two outer loops and five inner loops. For hyperparameter tuning, where applicable, we utilized a randomized grid search strategy. This involved exploring a predefined range of hyperparameters, as listed in Table 4.2, and selecting the combination that minimized the average Root-Mean-Square Error (RMSE) of the validation set of all folds. To ensure comprehensive exploration of the hyperparameter space, we conducted tuning across 50 different hyperparameter combinations. Notably, this process was carried out each time on a randomly selected subset of 500 sensors from our dataset to ensure computational efficiency while maintaining representativeness. To facilitate transparency and reproducibility, we have included the best-performing hyperparameters used in our study, which are presented in Table 4.2 in the appendix.

In this section, we use the following notation, where a given test dataset X is defined as follows:

$$X := \{X_{nt}\}, n \in \{1, \ldots, N\}; t \in \{1, \ldots, T\} \tag{4.2}$$

with N being the number of sensors and T the number of time points. The imputed value for sensor n at time point t is denoted by $\hat{X}_{nt}$. A test dataset is schematically depicted in Figure 4.3. The implementation of all evaluated methods was done in Python, with the list of packages included in Table 4.7.

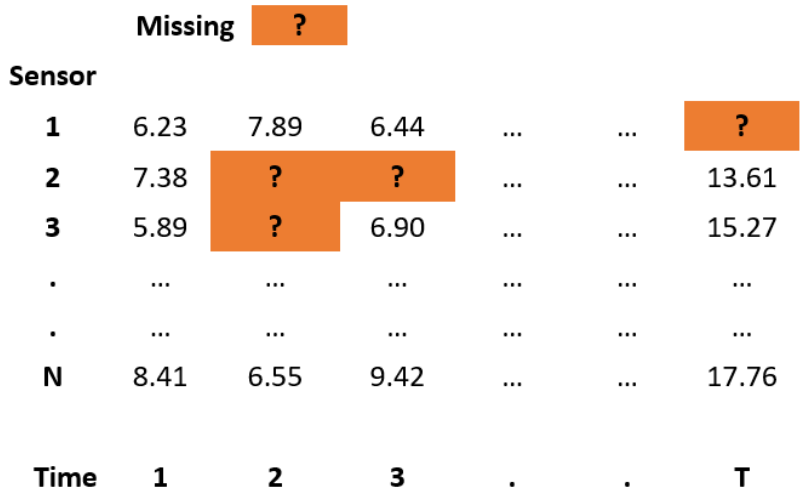


Figure 4.3: Schematic representation of a test dataset used in imputation analysis. A network of N sensors is providing data readings over T time points. Artificially induced missings are indicated by the orange fields.

4.2.3.3 Mean Imputation

In this approach, missing values are filled using the arithmetic mean. Specifically, within our application, we replace missing data points from a particular sensor with the mean value of that sensor across all available time steps. Mathematically, the imputed value is expressed as shown in Equation (4.3), where V represents the set of time points with available observations for the given sensor as follows:

$$\hat{X}_{nt} = \frac{1}{|V|} \sum_{v \in V} X_{nv}. \tag{4.3}$$

Mean imputation is effective when there is limited temporal and spatial variability and when the number of missing observations for a sensor is relatively low. Because of its straightforward approach, it serves as a baseline method for comparison within our study.

4.2.3.4 Spline Imputation

The linear spline imputation method uses temporal correlation within one sensor to impute missing values [246]. An imputed value $\hat{X}_{nt}$ for sensor n is estimated at time t by applying a linear interpolation based on the closest available time point in both directions, t_- and t_+:

$$\hat{X}_{nt} = X_{nt_-} + \frac{X_{nt_+} - X_{nt_-}}{t_+ - t_-}(t - t_-). \tag{4.4}$$

4.2.3.5 K Nearest Neighbor (KNN) Imputation

The KNN method was originally developed to estimate missing values in gene expression microarray experiments, but it can be easily applied to other use cases [247]. During imputation, data points with similar features as the data point with missing values are selected. In our case, the data points are the different sensors, while the features are the values at different time points. Thus, the KNN imputation technique leverages the spatial correlation of the dataset. This method would find k sensors that have a value present for the missing time point, where the values of the other time points are most similar to those of the sensor with the missing value. Afterward, a weighted average of the k "closest" sensors is calculated to estimate the missing value. The contribution of each of the k sensors is weighted by its similarity to the features of the sensor with missing values, where the similarity is quantified using the Euclidean distance.

4.2.3.6 Multivariate Imputation by Chained Equations (MICE)

The previously introduced methods all involve replacing missing values with a single estimation, disregarding the uncertainty and variability of the missingness. MI is a statistical technique used to handle missing data that generates multiple plausible imputations based on the distribution of the observed data [248, 249]. Estimating multiple imputations, as opposed to just one imputation, accounts for (part of) the statistical uncertainty in the imputations [274]. MICE is an example of an MI technique and generally operates under the assumption that the missing data are MAR or MCAR [249, 274–276]. When the data are not MAR, the application of MICE could result in biased or inaccurate estimates. The chained equations process used in MICE consists of the following steps [274, 276]:

1. Make an initial guess about the missing values using a simple imputation method, such as mean imputation.
2. Set the missing values for one feature f back to missing. The observed values for f are then regressed using (all) other features in the dataset.
3. Make a prediction for the missing values of f using the regression model from the previous step.
4. Repeat steps 2 and 3 for all features that contain missing values. At the end of this step, all features with missing values have been imputed.
5. Repeat steps 2, 3, and 4 for a number of cycles and update the imputations in each cycle.

The number of cycles can be chosen by the user and is task-dependent. The final imputation is the imputation found in the final cycle [274].

4.2.3.7 Markov Chain Monte Carlo (MCMC) Imputation

Another MI technique is MCMC [250], based on the Bayesian framework. In essence, MCMC leverages the principles of a Markov Chain and Monte Carlo simulation to ap-

proximate missing values by iteratively sampling from a probability distribution. The main focus is finding the desired posterior distribution defined by a set of parameters θ, from which the unobserved values X_u can be predicted using the conditional density of the observed observations X_o [250, 277]. The method starts from an initial, plausible approximation of the missing readings X_u. In the next step, the MI technique starts. Given certain parametric assumptions, the θ^i can be estimated from the posterior distribution $f(\theta^i|X_o, X_u)$, with θ^i being the estimated parameter values in step i. In a second step, the predictive distribution can be used to obtain the improved predicted values X_u^i at iteration i.

$$X_u^{i+1} \sim f(X_u|\theta^i, X_o) \tag{4.5}$$

In the next step, the θ parameter values can again be estimated from the complete data posterior distribution using the newly acquired values.

$$\theta^{i+1} \sim f(\theta|X_o, X_u^{i+1}) \tag{4.6}$$

These last two steps are iteratively executed until gradually converging to the true distribution. Due to the sequential sampling from two distributions, a Markov Chain is made, and the use of simulations renders the MCMC name [250, 277, 278]. MCMC imputation offers several advantages, including the ability to handle complex data structures and missingness patterns, as well as the flexibility of incorporating prior knowledge or constraints. The technique tends to be computationally intensive for large datasets, and it often requires careful tuning.

4.2.3.8 Matrix Completion (MC) Imputation

Another imputation technique based on iterative MI is matrix completion, based on [252]. The method uses the spatial and temporal correlations of the data to impute missing values. The main idea of MC is to handle missing values in a data matrix by imputing them with estimates based on the observed values and the low-rank structure of the data matrix. More precisely, missing readings get replaced iteratively with those obtained from a soft-thresholded singular value decomposition [252]. First, a singular value decomposition is applied to the incomplete matrix with soft-thresholding, where the nuclear norm of the matrix is used as a regularizer. In the next step, the modified singular value matrices are used to reconstruct the data matrix. Then, these two steps are iterated until convergence of the imputed values is reached. The matrix completion technique as discussed by [252] is well-suited for situations where the data matrix can easily be approximated by a lower-rank matrix, rendering an effective solution to the missing value problem for large and sparse matrices.

4.2.3.9 Data Estimation Using Statistical Model (DESM) Imputation

Similar to previously discussed methods, DESM uses temporal and spatial correlations between sensors to impute missing values [256]. The method is specifically developed for WSNs, with the sensor data specific characteristics in mind. More specifically, DESM relies on historical values of the sensor for which a value is missing (sensor n), as well

as the values of the sensor spatially located the closest (sensor m), with the requirement that the latter sensor does not have missing values around the time point that needs to be imputed. Missing values are then estimated according to

$$\hat{X}_{nt} = (1 - \alpha)\hat{Y} + (\alpha)\hat{Z}. \tag{4.7}$$

In Equation (4.7), $\hat{Y}$ is the imputed value at the previous time point $\hat{X}_{n(t-1)}$, and $\hat{Z}$ is defined as follows:

$$\hat{Z} = X_{n(t-1)} \left(1 + \frac{X_{mt} - X_{m(t-1)}}{X_{m(t-1)}}\right). \tag{4.8}$$

DESM leverages both the temporal influence of one sensor on itself, included in $\hat{Y}$, as well as the spatial attributes of the other sensors (in this case m), represented by $\hat{Z}$, to impute the missing readings. The α in Equation (4.7) is the Pearson correlation coefficient between two sensors X_n and X_m, which serves as a weight parameter that evaluates the effects of $\hat{Y}$ and $\hat{Z}$ on the estimated value. Equation (4.8) is based on the assumption that the data collected by sensors X_n and X_m are approximately similar, as they are spatially close to each other.

4.2.3.10 Applying k Nearest Neighbor Estimation (AKE) Imputation

Sensors that are located in close spatial proximity to other sensors will yield very similar measurements, which means that it is possible to impute missing values based on the neighboring sensor values for the same time point. As the exact functional relationship between two nearby sensors is unknown, the AKE method assumes that this relationship can be approximated linearly in a short time period [244]. Under this assumption, we can estimate missing values X_{nt}^m from a neighbor sensor X_{mt} for any time t using linear regression

$$\hat{X}_{nt}^m = \alpha + \beta X_{mt}, \tag{4.9}$$

where α and β are estimated using all non-missing (X_{nt}, X_{mt}) pairs. In total, k linear regression models will be fitted for every sensor, where k is a tunable hyperparameter. To obtain an imputed value, we have to combine the estimations from all k neighboring sensors. While using the arithmetic mean of all imputations is a valid option, this would disregard the strength of the linear correlation between two sensors. For this reason, AKE uses a weighted average of all k estimated values

$$\hat{X}_{nt} = \sum_{m=1}^{k} w_{nm} \cdot \hat{X}_{nt}^m, \tag{4.10}$$

where w_{nm} is the weight, for which $0 \leq w_{nm} \leq 1$ and $\sum_{m=1}^{k} w_{nm} = 1$. As we can assess the performance of a linear regression by using the determination coefficient r^2, we define the weight w_{nm} as the normalized determination coefficient as follows:

$$w_{nm} = \frac{r_{nm}^2}{\sum_{j=1}^{k} r_{nj}^2}. \tag{4.11}$$

4.2.3.11 MissForest Imputation

The MissForest method is a non-parametric imputation method that uses an RF to impute missing values [251]. To start, an initial guess is made for the missing values, using mean imputation or another imputation method. Afterward, the RF is trained on all data, including the initial guess for the missing values. The trained RF is then used to impute the missing values again. This procedure is repeated for a fixed number of iterations or until a stopping criterion is reached, whichever comes first. The stopping criterion is met as soon as the difference between the previously imputed values ($\hat{X}^{old}$) and the newly imputed values ($\hat{X}^{new}$) increases for the first time, i.e., the imputation has converged. The difference Δ is defined as follows, with p being the total number of missing values:

$$\Delta = \frac{\sum_{i=1}^{p}(\hat{X}_i^{new} - \hat{X}_i^{old})^2}{\sum_{i=1}^{p}(\hat{X}_i^{new})^2}. \tag{4.12}$$

Due to the use of random forests, the method is relatively robust against outliers; however, it can become computationally expensive on large datasets.

4.2.3.12 Multiple Imputation Using Denoising Autoencoders (MIDA)

MIDA is another MI technique that uses overcomplete denoising autoencoders (DAEs) to impute missing values [254]. An overcomplete DAE is a DAE where the input data are projected to a higher-dimensional subspace, from which the missing values are recovered. The input layer has T nodes, assuming the data have T features (time points in our use case). Then, each successive hidden layer adds Θ nodes, where Θ is a tunable hyperparameter. This is done for j encoding layers, after which $j-1$ decoding layers are added, which decrease the number of nodes from $T + j\Theta$ to T for the output layer. Empirically, $\Theta = 7$ and $j = 3$ have been found to be a good choice, but both of these parameters can be seen as tunable hyperparameters [254]. The MI part of MIDA is established by initializing the model with a different set of random weights in multiple runs, thereby providing multiple predictions. By leveraging the representational learning capabilities of denoising autoencoders, MIDA can capture the underlying patterns in the data, thus potentially generating more realistic results. However, these results heavily depend on the quality and similarity between the unknown and known observations in the training data. Furthermore, the use of autoencoders and MI also make this method computationally expensive.

4.2.3.13 Bidirectional Recurrent Imputation for Time Series (BRITS)

There are several methods based on an RNN for missing value imputation, such as BRITS [255]. The imputation method tries to learn the missing values in a bidirectional recurrent dynamical system, without any specific assumptions [255]. The method was originally developed for missing value imputation in multiple correlated time series, which we extend to the WSN framework. In BRITS, an RNN is used directly for predicting missing values, meaning that missing values are regarded as variables of the bidirectional RNN graph, leveraged in the back-propagation of the neural network. This approach ensures

that missing values receive delayed gradient updates in both the forward and backward directions, with consistency constraints. The model architecture can also be leveraged for simultaneous regression or classification jointly in one graph, rather than pure missing imputation. This can mitigate a part of the error propagation in subsequent modeling tasks. The main advantages of the BRITS imputation method are the application to general missing settings, as well as the ability to handle correlated time series and nonlinear dynamics within the data.

4.2.3.14 Multi-Directional Recurrent Neural Network (M-RNN) Imputation

Closely related to BRITS is M-RNN, as introduced by [253]. M-RNN imputes values both within and across data-streams, thus both in a temporal and spatial fashion. The original method was developed for clinical applications, yet it can be easily applied to other scenarios. The imputation technique contains both an interpolation block (temporal) and an imputation block (spatial), which are trained simultaneously. The interpolation block uses an adjusted bi-directional RNN with a lagged timing for the inputs into the hidden layers in the forward direction and advanced in the backward direction [253]. The imputation block is then a fully connected neural network with dropout. Similarly to BRITS, the method can also be used for a subsequent modeling task directly. Notably, the M-RNN tends to be less affected by both the quantity and specific nature of missing data.

4.2.4 Empirical Evaluation

To compare the performance of the different methods, it is important to have a predetermined set of performance metrics. In this study, we use the RMSE and MAE to assess the accuracy of the imputed values. Based on previous studies, we also evaluate the percentage of cases in which a missing value can be estimated, i.e., the prediction coverage error (PCE), as defined in Equation (4.13) [279, 280].

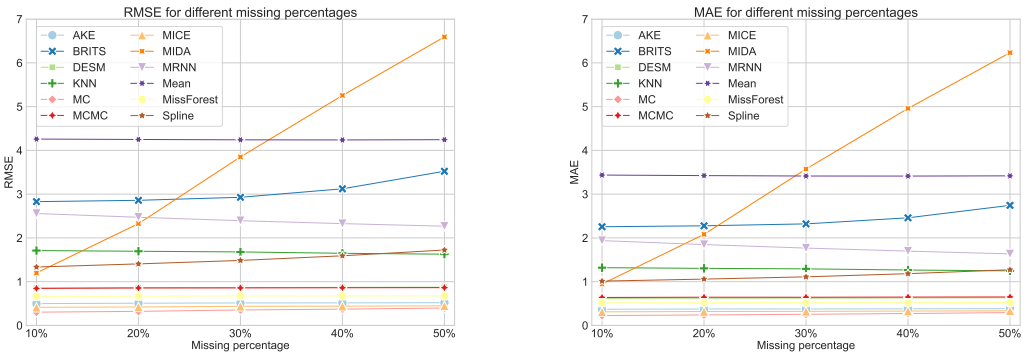
$$\text{PCE} = \frac{\text{number of successful imputations}}{\text{number of missings}} \times 100\% \tag{4.13}$$

The PCE is necessary to be able to interpret the RMSE and MAE fairly; indeed, the RMSE and MAE ignore missing data points that the model was unable to impute. For example, if a model relies on the five nearest sensors for imputations, it will be unable to provide an imputation if those sensors also have missing values. As a result, the RMSE and MAE might be underestimated for some methods that are not able to impute all samples, resulting in an overstatement of the model performance. For these cases, the PCE provides additional context.

4.3 Results and Discussion

4.3.1 Random Missings

The performance of the various imputation methods, evaluated using the metrics defined in Section 4.2.4, is depicted in Figures 4.4 and 4.5. Exact values for the RMSE, MAE, PCE, and execution times are detailed in Tables 4.3 to 4.6. As can be seen in Figure 4.4 and Tables 4.3 and 4.4 the MC imputation method achieves the best performance (smallest RMSE and MAE) for all degrees of missingness. Generally, all methods consistently outperform the baseline mean imputation, except for the MIDA method, where performance diverges for higher degrees of missingness (see Section 4.3.5). In the results, a noticeable trend is observed where methods considering the spatial features of the data generally outperform others. Methods such as MissForest and MCMC obtain a good performance and even outperform DESM, which is specifically tailored for WSNs problems. AKE, another WSN-specific method, has a very good performance and is only outperformed by MC and MICE. For nearly all methods, the performance gets worse with increasing degrees of missingness, which is expected and is also commonly observed in the literature [253, 280]. A notable exception to this rule is M-RNN, as its performance remains stable with increasing degrees of missingness. Although unexpected, we consider this result to be less significant, as the difference is relatively small, and the absolute performance of M-RNN is among the worst for our specific use case. In addition, M-RNN can efficiently handle higher degrees of missingness, explaining the result.



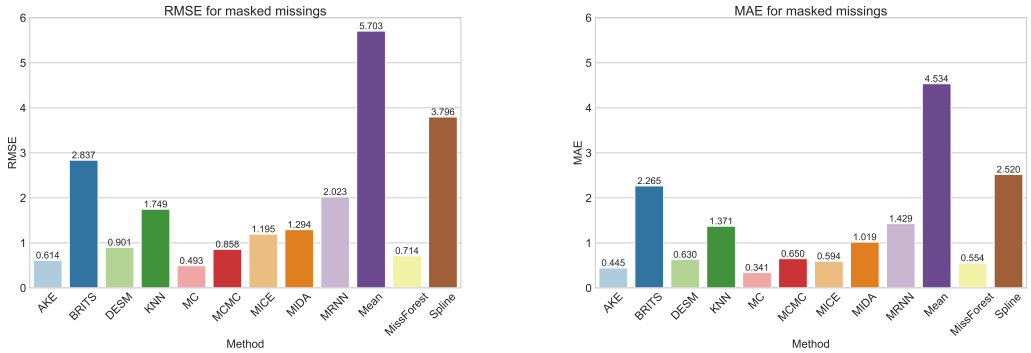
(a) (b)

Figure 4.4: (**a**) The RMSE for all models for different degrees of missingness. (**b**) The MAE for all models for different degrees of missingness.

4.3.2 Masked Missings

As random missings do not accurately represent real-life missing scenarios, we also evaluated all methods on a realistic missings dataset obtained by creating masks from real, observed missing patterns. From Figure 4.5 and Tables 4.3 and 4.4, we can conclude that the MC method performs best, as was also the case in the random missing scenario. Similarly, AKE, DESM, MCMC, and MissForest are again among the top performing

methods, indicating the better performance of spatial methods on our dataset. As expected, the performance in the masked missings scenario is generally worse than for random missings, as so-called "block" missings frequently cause a lack of "nearby" data points, which are often used to impute missing values. In particular, the performance of spline imputation and MICE is significantly worse in the masked scenario. This finding was expected for spline imputation, as it applies a linear interpolation based on the closest time points surrounding the missing value, which will often be missing itself in the case of block missings. For MICE, this can be attributed to a worse convergence of the model due to a poor initial guess. More specifically, the initial guess for the MICE algorithm was made using mean imputation. From Tables 4.3 and 4.4, we can clearly see that mean imputation performs significantly worse for masked missings when compared to its performance for random missings and thus probably did not provide an accurate initial guess.

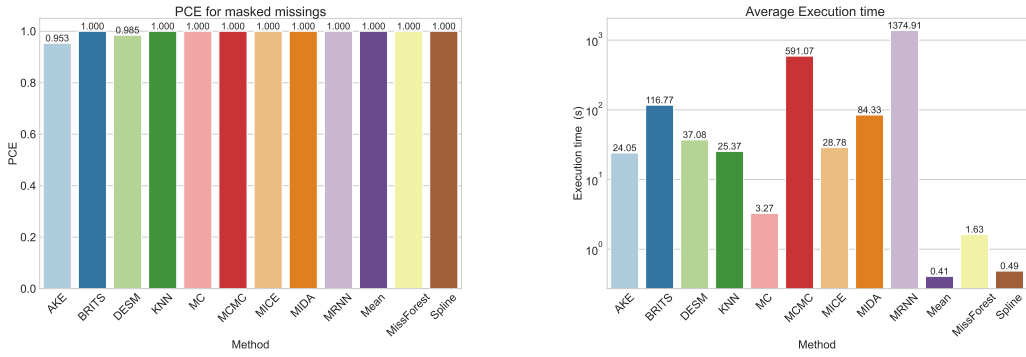


(a) (b)

Figure 4.5: (**a**) The RMSE for all models for the masked data. (**b**) The MAE for all models for the masked data.

4.3.3 Prediction Coverage Error

We should, however, interpret the results for AKE and DESM in the masked missings scenario with caution. In fact, Figure 4.6 and Table 4.5 show that these methods do not achieve a PCE of 1, meaning that they were unable to impute all missing values. Indeed, AKE and DESM were only able to impute 95.3% and 98.5% of the missing values, respectively. As AKE requires at least one sensor (in a group of k nearby sensors) to have an observed value in the given time step and DESM requires one sensor to have no missing values around the considered time step, this is likely attributable to large-scale network outages. Indeed, these outages result in spatially nearby points failing simultaneously. Although the number of unfilled missings is quite small, an end user might opt to address the remaining gaps by employing a straightforward imputation method such as mean imputation.



(a) (b)

Figure 4.6: (**a**) The PCE for all models for the masked data. (**b**) The execution time (on a logarithmic scale) for all models, averaged over all degrees of missingness, including masked missings.

4.3.4 Execution Time

A final criterion to consider when evaluating the imputation methods is the time needed to train the method (if necessary) and make the imputations, which is combined in the execution time. In Figure 4.6 and Table 4.6, the execution time averaged over all degrees of missingness and masked missings is shown. As expected, the deep learning-based methods (M-RNN, MIDA, and BRITS) were very computationally intensive to train. The MCMC method was also very computationally expensive, which was expected, as the authors state that this is the case for large datasets [250]. Finally, even though the MC method is specifically designed for large matrices, it is interesting to note that it achieved a very small execution time while also providing the best imputed values in both missing scenarios.

4.3.5 Discussion

Within our empirical evaluation, spatial methods tend to outperform others across all imputation scenarios. This can be attributed to the large number of sensors as well as the spatial proximity of the sensors. Furthermore, the evaluated time steps were limited, thus impacting the performance. In general, these results are in line with the literature. However, the deep learning-based methods exhibited poor performance on our dataset, regardless of the missing pattern. We attribute this to several factors. First, the relatively short time series considered in these data may limit the ability of certain deep learning-based methods, such as BRITS, to accurately capture temporal dependencies and patterns. Additionally, most deep learning applications tend to evaluate datasets with either a higher frequency of measurements or a higher number of measured objects or sensors. Furthermore, our dataset considers the behavior of small microclimates, in which local effects may cause temperature peaks that do not appear for other sensors, which complicates generalization to other sensors. Nevertheless, deep learning-based methods do exhibit good results for other datasets and should therefore not be ignored

as a possible imputation technique [253–255]. They also offer the added advantage of making supplementary predictions using the acquired structure [253, 255] or by training new models based on the already learned weights. For datasets where deep learning-based methods perform very well or at least comparably to other methods, this added benefit could be the deciding factor.

4.4 Conclusions

During the last decade, sensors have become increasingly important across scientific fields and industries. Unfortunately, sensor data often contain missing values, which can significantly hamper the interpretation and possible analysis of the collected data. Consequently, the importance of methods capable of imputing these missing values with accurate estimates has grown considerably. In this study, we conducted a comparison of twelve imputation methods on a unique environmental microclimate monitoring dataset collected by the CNidT citizen science project. We extend the current literature by providing an extensive comparison of different missing value imputation methods originating from different backgrounds and imputation strategies. In addition, our work considers a real WSN dataset with a large number of sensors (1500), which is uncommon in the literature. Furthermore, we extend the evaluation of the evaluated techniques from random missings to masked missings, which provides a highly realistic evaluation scenario for practical implementations.

We evaluated the imputation methods for two different missing patterns: random missings, with the degree of missingness ranging from 10% to 50%, and masked missings, which were obtained using realistic missing value patterns. For all missing patterns, the MC method outperformed all other methods. MissForest and MCMC also performed relatively well in both scenarios, while MICE only achieved good results for random missings. The methods that are designed for WSNs specifically also performed well in both scenarios; however, they were not able to provide imputations for all missing values in the masked missings scenario. Finally, the deep learning-based methods, M-RNN, MIDA, and BRITS, performed poorly for both missing patterns, which can be attributed to the characteristics of our dataset. We can conclude from the results obtained that the methods that exploit spatial correlations within the dataset tend to perform better than the other methods. This can be explained by the relatively small distance between sensors, as well as the granularity of the temporal component. Moreover, since the data encompassed the period from April to September, temperatures predominantly experienced an upward trajectory, making it challenging to discern a clear trend in the temporal aspects of the data. These results can be extrapolated to similar scenarios where the number of sensors is high and densely distributed with a comparable length of time. The success of methods such as MC, MissForest, and MCMC, particularly in capturing spatial correlations within the dataset, suggests that they would generalize well to such environments. Despite challenges posed by masked missing values, these methods still demonstrated robust performance, implying their potential applicability in scenarios with complex missing data patterns.

Future research can expand upon our study with a more detailed assessment of (other) methods on different datasets. More specifically, different numbers of sensors and temporal granularity can be evaluated to more clearly identify the impact of these dataset

specific features on the evaluated models. This can aid in the identification of a general best imputation technique across different WSNs. Furthermore, in future studies concerning missing data imputation for WSNs, additional features of the sensors or locations can be used to address missing values, such as the type of microclimate location, or other measured variables, such as the humidity in our specific use case. Also, the development of novel WSN-specific methods that efficiently exploit all structures (spatial and temporal) that are available in the data, carry significant potential. For example, a method could use an MI approach by first imputing all missing values using temporal correlations and subsequently using these imputations to obtain a more accurate spatial imputation, or vice versa. Additionally, cost-sensitive methods for missing value imputation can be evaluated, where over- or underestimations of the actual value can be penalized more heavily. Moreover, the evaluation of the temporal and spatial granularity and its impact on the imputation performance for various methods could be a valuable addition. Finally, our comparative study focuses on daily temperature values, whereas it may be interesting to evaluate it per 15-min interval or hourly and assess the imputation performance.

In conclusion, we were able to successfully impute missing values in our unique environmental monitoring dataset and provided guidelines for researchers who want to impute missing values in a similar dataset. Ultimately, we found that the best method to impute missing values is often dataset-specific and should be identified using a set of artificially induced missings, preferably both randomly generated and based on a realistic missing pattern.

4.5 Data availability statement

Publicly available datasets were analyzed in this study. The data can be found at: `https://www.vlaanderen.be/datavindplaats/catalogus/curieuzeneuzen-in-de-tuin`. Data can also be found in the SoilTemp database available at: `https://www.soiltempproject.com/` [245]. Code for the replication of our study is available on GitHub: `https://github.com/STAN-UAntwerp/Missing_Imputation_Sensors`.

4.6 Appendix

4.6.1 Hyperparamter search

Table 4.2: The hyperparameter grid search space for each evaluated model.

Method	Hyperparameter Search Space	Optimal Hyperparameter Value
AKE [244]	k: [2–25]	15
	timesteps: [5–30],	18
BRITS [255]	hidden dimensions: [5–15],	8
	learning rate: [0.0001,0.001,0.01,0.1]	0.001
KNN [247]	N neighbors: [2–25]	5
MICE [249]	N nearest features: [2–25]	24
MIDA [254]	Θ: [5,6,7,8,9]	7
	sequence length: [5–30],	5
MRNN [253]	hidden dimensions: [5–25],	7
	learning rate: [0.0001,0.001,0.01,0.1]	0.01

4.6.2 Results

Table 4.3: The RMSE scores for all imputation models, for different degrees of missingness and masked missings. The best performances are indicated in bold.

Method	10%	20%	30%	40%	50%	Masked
AKE [244]	0.499	0.508	0.514	0.516	0.520	0.614
BRITS [255]	2.829	2.859	2.928	3.122	3.523	2.837
DESM [256]	0.857	0.862	0.866	0.870	0.875	0.901
KNN [247]	1.711	1.695	1.680	1.648	1.628	1.748
MC [252]	**0.304**	**0.322**	**0.354**	**0.372**	**0.398**	**0.493**
MCMC [250]	0.847	0.856	0.858	0.861	0.864	0.858
MICE [249]	0.414	0.422	0.437	0.440	0.448	1.195
MIDA [254]	1.200	2.326	3.850	5.256	6.593	1.294
MRNN [253]	2.558	2.472	2.394	2.326	2.268	2.023
Mean Imputation	4.259	4.250	4.241	4.239	4.245	5.703
MissForest [251]	0.664	0.669	0.673	0.674	0.675	0.714
Spline [246]	1.335	1.406	1.486	1.592	1.724	3.796

Table 4.4: The MAE scores for all imputation models, for different degrees of missingness and masked missings. The best performances are indicated in bold.

Method	**10%**	**20%**	**30%**	**40%**	**50%**	**Masked**
AKE [244]	0.370	0.375	0.376	0.379	0.384	0.444
BRITS [255]	2.255	2.277	2.321	2.460	2.744	2.265
DESM [256]	0.617	0.621	0.622	0.627	0.631	0.630
KNN [247]	1.320	1.304	1.293	1.267	1.254	1.371
MC [252]	**0.226**	**0.240**	**0.255**	**0.272**	**0.294**	**0.341**
MCMC [250]	0.640	0.645	0.644	0.648	0.650	0.650
MICE [249]	0.311	0.318	0.322	0.328	0.335	0.594
MIDA [254]	0.946	2.082	3.574	4.959	6.233	1.019
MRNN [253]	1.938	1.848	1.768	1.698	1.635	1.429
Mean Imputation	3.436	3.424	3.414	3.413	3.418	4.534
MissForest [251]	0.519	0.520	0.520	0.522	0.523	0.554
Spline [246]	1.013	1.059	1.110	1.183	1.272	2.520

Table 4.5: The PCE scores for all imputation models, for different degrees of missingness and masked missings. The best performances are indicated in bold.

Method	**10%**	**20%**	**30%**	**40%**	**50%**	**Masked**
AKE [244]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.953
BRITS [255]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
DESM [256]	**1.000**	**1.000**	**1.000**	**1.000**	0.999	0.985
KNN [247]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MC [252]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MCMC [250]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MICE [249]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MIDA [254]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MRNN [253]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Mean Imputation	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MissForest [251]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Spline [246]	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 4.6: The execution times for all imputation models, for different degrees of missingness and masked missings. The lowest execution times are indicated in bold.

Method	10%	20%	30%	40%	50%	Masked
AKE [244]	13.1	18.0	28.1	26.5	46.2	12.4
BRITS [255]	52.4	81.7	103.0	123.6	303.8	36.1
DESM [256]	10.2	19.8	44.2	47.0	92.4	8.8
KNN [247]	15.3	21.7	27.6	34.2	41.6	11.8
MC [252]	1.5	2.0	3.9	3.7	6.2	2.2
MCMC [250]	537.0	552.1	572.3	567.6	881.5	435.9
MICE [249]	23.5	25.7	32.6	25.8	41.6	23.4
MIDA [254]	70.1	71.2	106.0	81.5	108.6	68.6
MRNN [253]	1180.5	1193.6	1640.9	1278.7	1811.7	1144.0
Mean Imputation	**0.3**	**0.3**	**0.5**	**0.4**	**0.6**	**0.3**
MissForest [251]	1.3	1.4	1.8	1.6	2.3	1.3
Spline [246]	0.4	0.4	0.7	0.4	0.7	0.4

4.6.3 Package Versions

Table 4.7: The versions of the Python packages used in the project.

Package	Version	Reference
python	3.11.5	[145]
scikit-learn	1.3.2	[133]
fancyimpute	0.7.0	[281]
geopy	2.4.1	[282]
keras	2.12.0	[283]
missforest	2.3.1	[251]
numpy	1.23.5	[273]
pandas	2.1.4	[284]
scipy	1.11.4	[285]
seaborn	0.13.1	[286]
tensorflow	2.12.0	[287]
tensorflow-probability	0.20.0	[287]
torch	2.1.2	[288]

Chapter 5

A global crop-specific fertilization dataset from 1961–2019

This chapter is based on:

Decorte, Thomas, Fernando Coello, Steven Mortier, Iris Janssens, Jordi Sardans, Josep Peñuelas, and Tim Verdonck. Machine Learning-Driven Global Crop-Specific Fertilization Dataset from 1961–2019. *Under Review.* 2024

5.1 Background & Summary

Inorganic fertilizers are essential for replenishing the nutrients that are removed from soils during crop harvesting. The three main nutrients provided by fertilizers, nitrogen (N), phosphorus (P) and potassium (K), play a key role in plant functions. While N and P, which are basic components of nucleotides, proteins and membrane lipids, are essential in energy metabolism [289, 290], K is essential for the transportation of water, metabolites, and nutrients across plant tissues, for defense against oxidative stresses, and for the maintenance of osmotic homeostasis [291, 292]. Although the first commercial inorganic fertilizers were developed in 1843, they were not the main anthropogenic inputs in the N, P, and K biochemical cycles until the second half of the 20th century [293]. Today, inorganic fertilizers dominate as the primary nutrient input in croplands, surpassing the second human input, manure, by over double [293], and also serve as the main N input for grasslands [294]. This substantial surge during the 20th century not only facilitated the rapid growth in human population, but also had ecological and socioeconomic ramifications, such as water eutrophication, soil degradation, climate change, and mineral resource depletion [295, 296]. In the remainder of this study, the term 'fertilizer' will refer to inorganic fertilizers, and all data and results regarding P and K will be presented in their oxidative forms (P_2O_5 and K_2O, respectively), in accordance with common references in international standards and regulations.

Given their food security, socioeconomic and environmental implications, considerable research has been conducted to discern the temporal and regional trends in the use of N, P_2O_5, and K_2O [294, 297–300]. Nevertheless, limited availability of temporal global

spatial information regarding their application across various crops have restricted these analyses to a few global and regional studies that primarily focused on N [297, 301, 302]. These studies initially estimated consumption at the country- and state-level using simple equations, based on a few crop-specific fertilization features and changes in crop surface area [297, 301, 302], or using Bayesian Markov Chain Monte Carlo modeling [303]. A global, crop-specific fertilization dataset is crucial for understanding crop nutrient management practices worldwide, identifying past trends and current gaps in fertilization, guiding agricultural policies to improve crop yields while minimizing environmental impacts, and providing input data for modeling [297]. Therefore, we aim to address this knowledge gap by providing insights into the application rates of P_2O_5 and K_2O while also seeking to improve estimates for N.

In order to accomplish this objective, we began by updating the panel datasets on cropland fertilization; enhancing the most comprehensive database developed by Ludemann *et al.* (2022) [67] by incorporating global datasets covering data from the 1970s and 1980s [304, 305], country-specific data for European countries from 2001 to 2014 [306–309]. This compilation process led to a 35% expansion of the Ludemann *et al.* (2022) database. Second, the dataset was expanded with data of various potential socioeconomic, environmental, and agricultural drivers of cropland fertilization. Third, two machine learning (ML) regression models - eXtreme Gradient Boosting or XGBoost (XGB) [190] and HistGradientBoosting (HGB) [133], both capable of handling the prevalent missing values within the dataset [133]- were applied to predict N, P_2O_5, and K_2O fertilizer application rates for the different crop classes over 60 years. Since these models are considered black-box models, feature importance was incorporated using SHapley Additive exPlanations (SHAP) [59] values to identify the global socioeconomic, agricultural, and environmental drivers of cropland fertilization and to validate the ML models. Fourth, the predictions were validated on national databases. However, since the ML models were trained on global data, which show a discrepancy with the national data, the model predictions were first adjusted to match the total annual country-level N, P_2O_5, and K_2O use in agricultural land, similar to previous studies [297–300, 310]. Crucial in this adjustment was the fraction of total country-level fertilizer use allocated to grasslands and fodder crops, as an important portion of total fertilizer use in some countries is devoted to these areas, and little previous estimates existed [293, 310–312], especially for K_2O [293]. Therefore, these fractions were estimated by reviewing scientific and technical information from 75 countries. The adjusted predictions were then validated using national databases of fertilizer application rates at the crop-level. Finally, the results were spatially allocated using crop maps of the year 2000, developed by Monfreda *et al.* (2008) [313]; the annual harvested area of each crop class in each country; and the spatial changes in cropland surface based on the Hyde v3.3. project [314].

5.2 Methods

The following section outlines the comprehensive methodology that was adopted in this study. The methodology encompasses various stages, including the collection and aggregation of different datasets and the compilation into a unified dataset, as well as any preprocessing steps that were carried out. Additionally, we introduce the ML models used in this study, as well as the respective training and evaluation procedures. Furthermore, we discuss the measures that were undertaken to explain the predictions

made by the ML models. Following this, we describe how we used the predictions to create detailed maps of global fertilizer application rates. Finally, we explain how we assessed the validity and plausibility of the dataset derived from our study.

5.2.1 Data collection

5.2.1.1 Fertilizer application rate by crops

To compile a consistent and detailed dataset of fertilizer application rates for different crops, countries, and years, 14 global datasets [67, 304–309, 315–321] were used. We discarded national databases, such as the United States of America (USA) [322] and India [323–329], to construct a homogeneous database. This approach avoids multiple year-nutrient-crop-country entries from both global and national databases, and allows us to retain external databases for validating the ML model predictions. To standardize all these datasets and minimize data loss, we classified all crop types into 13 crop groups (wheat, maize, rice, other cereals, soybean, palm fruit, other oilseeds, vegetables, fruits, roots and tubers, sugar crops, fiber crops and other crops) (Table 5.4), in alignment with the Indicative Crop Classification for agricultural census (ICC) Version 1.1 [330].

During the 80s, the International fertilizer Development Center (IFDC) published two reports [304, 305] regarding crop-specific data of fertilizer use by crop (FUBC) (hereinafter referred to as FUBC-IFDC). After the crop grouping, these publications included data for 459 country-crop-years combinations (kg ha^{-1} of N, P_2O_5, and K_2O) from 83 countries for 1973-1988. During the 90s, the Food and Agriculture Organization (FAO), in collaboration with the fertilizer industry (IFDC and International Fertilizer Association (IFA)), published five crop-specific datasets of fertilizer application rate (hereinafter referred to as FUBC-FAO). After grouping the data, these publications included data for 1693 fertilizer application rate specific to years and crops (kg ha^{-1} of N, P_2O_5, and K_2O) from 108 countries for 1984–2002, although most of the data (98%) covered 1988–2002. The data were collected using questionnaires from governmental agencies, members of industry companies, agronomists, and economic experts. In both datasets (FUBC-IFDC and FUBC-FAO), the use of fertilizer for each combination of nutrient, crop, country, and year was provided two ways: (a) as the average application rate of a fertilizer over total cropland area, and (b) as the percentage of fertilized cropland area and the application rate in that area. We transformed all data to the average application rate by multiplying the percentage of fertilized area by the application rate in that area. The data were either from a year (e.g., 1996) or a season (e.g., 1996/97). For seasonal data, we considered the starting year of the season as the year of the data in the analyses. For data for nutrient, crop, country, and year that were in more than one report, the data was selected from the most recent report. Data for crop, country, and year that were divided into crop varieties or management practices (e.g., irrigated or rain-fed rice, or soft or durum wheat) were aggregated and weighted by the area of the crop class included in the report. Data for sweet maize, or corn, were excluded, assuming that it referred to *Zea mays* var. *saccharata* and the data for silage maize, because FAOSTAT reports only the harvested area for maize grain. Values for the crop groups were derived from individual crops when either more than 90% of the harvested area (based on FAOSTAT data [331]) was dedicated to the production of a single crop, or when a combination of crops was available in the data, their weighted average was assigned to the entire group.

Since the last FAO publication, IFA has released five reports detailing the total amount of N, P_2O_5, and K_2O used for various crop classes, providing yearly or seasonal data spanning from 2006 to 2018 [67, 320, 321, 332] (hereinafter referred to as FUBC-IFA). Initially covering 11 crop types, these reports expanded to 14 types in the fourth report. They encompassed information for the European Union (EU) together as well as 27 other countries. In 2022, Ludemann published a more comprehensive dataset covering data for 66 countries, featuring EU data at the country scale, and information for 20 crop classes [67]. This report also included the FUBC-FAO data for the 1990s and prior data from IFA. However, small discrepancies between the FUBC-FAO original data and the one compiled by Ludemann *et al.* (2022) prompted us to retain the original FUBC-FAO information. To estimate the average application rate for each combination of crop, country, and year, we divided the total used amount of each fertilizer by the harvested area provided by FAOSTAT [331]. As previous research we assumed the harvested area as a proxy for the crop's annual surface on each country [297, 298]. It is worth noting that the average application rate for maize was slightly overestimated because FUBC-IFA included the amount discharged to silage maize. According to Maiz'Europ' [333], the current area of forage maize crops is 17.3 million ha (approximately 1% of the total area of maize crops in 2020) with the European Union as the most important producer of silage maize, with 6 million ha. We utilized the available raw data from Ludemann *et al.* (2022)[67], adopting FAO-IFDC datasets methods for grouping, and omitted certain countries where values were estimated based on the previous report and changes in crop surface. For the EU countries, Norway and the United Kingdom (UK), four unpublished datasets from Fertilizer Europe (FE) spanning 2001–2015 (referred to as FUBC-EFMA) [306–309] were used. These datasets offered similar information to the FUBC-FAO publications for the EU countries, the UK and Norway and allowed us to exclude the fertilizer application to silage maize, which is important in the EU[333]. However, FUBC-EFMA datasets lacked individual crop classes for rice and soybeans, resulting in missing data at the country-level for these crops since 2000 in EU countries.
The resulting dataset included data for the average fertilizer application for 3712 combinations of 13 crop classes, 114 countries, and years from 1973 to 2018. For most of the combinations of countries and crops, data were available for only a few years (on average, a country-crop combination had data for 4.1 ± 2.9 years, and 64% of the combinations had five or fewer years with available data).

In order to later validate our estimations, we compiled a series of national databases. National data was quite limited, as only a few countries conduct surveys to study fertilizer management across different crops. The two countries with most available data were the USA [322], and the UK [334], which collected long time series on cropland fertilization for the three primary nutrients. The USA dataset [322] contains fertilization information for four crops -cotton, maize, soybean, and wheat- dating back to 1964. To compare with our predictions, we converted all data to average kg ha^{-1}. Additionally, based on the same surface threshold used for global datasets, we assumed that the application rate for cotton was equivalent to that of all fiber crop classes. The UK dataset [334] provides data for four crop classes -roots and tubers, other oilseeds, sugar crops, and wheat- starting from 1998 for the three nutrients across all Great Britain. We also compiled existing information from several Asian countries, including India, the Philippines, and Pakistan [323–329, 335, 336]. The datasets from India [323–329] and Pakistan [335] did not require additional preprocessing, as they provided the data in average kg ha^{-1}. However, the dataset from Pakistan presented the information for all three nutrients combined [335].

For the dataset from the Philippines, which covers rice and maize, we converted the raw data on the regional number of 50 kg bags per hectare of different fertilizers to N and P_2O_5 using the country-specific fertilizer nutrient information [337]. Finally, we also compiled existing data from Sweden [338–341] and New Zealand [342]. The data for P_2O_5 and K_2O in the Sweden dataset, initially present in their pure nutrient form, were transformed to their oxidized forms by multiplying by the molecular weights of these elements.

5.2.1.2 Fertilizer use in other agricultural lands

An important step in the methods involves adjusting ML model predictions to national-level fertilizer use. We used the FAOSTAT database regarding fertilizer annual use at the country level for making this adjustment[343]. This database includes data on all fertilizer use for agricultural lands, covering both croplands and grasslands [343]. However, the crops included in the ML models, as well as in the FAOSTAT harvested area data[331] do not cover grasslands -whether permanent or temporary- nor fodder crops such as silage maize or fodder beet. Therefore, the primary goal of this section is to estimate the fraction of total fertilizer used for these types of agricultural lands.

Data regarding fertilizer application rate for grasslands and fodder crops is even more scarce than fertilization for other croplands. Additionally, FAOSTAT lacks information about the surface of the majority of the fodder crops [331]. Therefore, the methods used for estimation may not be as accurate as those used for other agricultural lands. Here, we reviewed technical information, such as the FUBC compiled reports [67, 304–309, 315–319, 321], and scientific information from countries where the fertilization of grasslands was considered to be higher than 1% of the total fertilizer consumption in previous research [293, 294, 310–312]. Previous research typically focused only on permanent grassland fertilization, as their goal was to distinguish agricultural fertilizer usage between arable -croplands and temporary grasslands- and non-arable land -permanent grasslands- [293, 311, 312]. However, we included in the estimation the proportion of fertilizer used for temporary grasslands and fodder crops for two main reasons: 1) our main goal was to distinguish agricultural fertilizer usage between all croplands included in the thirteen crop classes defined in the previous section and the rest of the agricultural land, 2) the majority of data available in the compiled global reports give information about all grasslands and fodder crops together [67, 304–309, 315–321]. The information estimated was the annual country proportion of N, P_2O_5, K_2O fertilizers used for agriculture for grasslands and fodder crops. Depending on the available information, we have assessed at the country- or regional-level. In total, we reviewed scientific and technical reports for 75 countries. As in previous research[293, 310–312], the methods used for estimating the share of N, P_2O_5, and K_2O usage for grasslands and fodder crops varied between countries and regions depending on the available information. Therefore, for every country, we argued the decisions taken based on the available data for providing at least as transparent as possible the estimations made. Moreover, we included a summary table (Table 5.7) with the sources used for estimating the range of values used for each country.

5.2.1.3 Potential drivers

To develop our ML models, we compiled a series of datasets that contain information on features that were identified in previous research as drivers or correlates of cropland fertilization. In this section and the next two, we clarify our rationale for the variable selection, the data sources and the methods used for estimating some of these variables. The list of all considered features can be found in Table 5.3 in the appendix. Further details about their estimations are provided below.

Environmental data Environmental variables related to climate and soil characteristics have been identified as factors that influence fertilization management in farm-level studies [344] and regional panel data [345, 346]. Therefore, we selected several potential factors, some of which have previously been shown to correlate with fertilization, such as mean annual precipitation (MAP) [345], or soil organic carbon (SOC) [346], as well as newer potential factors such as the aridity index. Data for these factors were sourced from two main databases: the CRU v.4. databases [347], for climatic factors, and the SoilGrids v.2. database [348], for soil characteristics. Obtaining values at the country-level while considering variations in climatic and soil conditions within a country can be imprecise. However, our fundamental unit of analysis is the country-level, as the FUBC values are measured on this scale. To mitigate this limitation, we used spatial information for climatic and soil characteristics along with information about the location of crops [313]. All environmental variables were estimated using Equation (5.1), but preprocessing differed across variables.

$$Env_{lic} = \frac{\sum_{g \in G}(Env_{ig} \times HArea\_M2000_{gcj})}{HArea\_M2000_{cj}} \tag{5.1}$$

Here, Env_{lic} represents the mean value of the environmental variable for country j, in year i, where crop c is located in the country; Env_{ig} is the value of the environmental variable in year i, for grid cell g ; $HArea\_M2000gc$ denotes the area of grid cell g for crop c in country j; $HArea\_M2000_{cj}$ is the total surface of crop c in country j based on Monfreda *et al.* (2008) crop maps [313]; and G denotes the set of cells where the crop is located based on Monfreda *et al.* (2008) crop maps [313]. Here, the grid cells are 5-arcminute by 5-arcminute cells on a latitude-longitude grid. For the MAP, the Env_{ig} values of Equation (5.1) are calculated by summing the precipitation from all months in the CRU v.4. dataset [347] for each grid cell g, and year i. For the mean annual temperature (MAT), the Env_{ig} values are calculated as the average of the monthly temperatures from the CRU v.4. dataset [347], weighed by the number of days of each month. The potential evapotranspiration (PET) values are derived by multiplying the daily month average from CRU v.4.[347] by the number of days in each month and summing the results. For the aridity index, we used the United Nations (UN) definition [349] of the ratio between MAP and total PET for each grid cell. As soil variables do not have temporal resolution, we simplified Equation (5.1) by removing the temporal factor. Additionally, for some soil variables like the soil cation exchange capacity (CEC), we aggregated the information by calculating the average for the first three depth layers from SoilGrid v.2. (0-5, 1-15 and 15-30 cm) [348].

Agrological data We selected agrological features that were previously identified as factors potentially related to or driving fertilizer intensification, such as holding size [350], crop area [346], or irrigation implementation [345], as well as features that should be connected to crop fertilization at the country-level, such as country fertilizer use per cropland area [343]. Most of the agrological variables used are taken directly from the sources indicated in Table 5.3. However, some required preprocessing. For holding size, we applied the methodology used by Zou *et al.* (2022) [310], which involves standardizing the information based on the average holding size according to the total agricultural area. We used holding size data from the FAOSTAT agricultural censuses [351] and previous research [352]. To estimate the annual nutrient removal for each crop class based on annual production, we used the recent compilation by FAO [293] on nutrient removal in kilograms per tonne of crop produced, along with the annual country production data from FAOSTAT [331]. Additionally, we used this compilation alone as a proxy for fertilizer recommendations, since these recommendations are generally based on the nutrient requirements of each crop [353].

Socioeconomic data Economic factors, particularly those related to the profitability of fertilizer use, have been widely studied to understand fertilizer adoption at the farming-level [354, 355]. Both input prices (fertilizers) and output prices (crops) determine profitability and can be key factors influencing fertilization decisions. However, assessing inputs at the country-level is challenging, primarily due to a lack of standardized data. The only available dataset, FAOSTAT [356], does not cover all periods and lacks standardization. To address this, we used two variables as proxies of fertilizer prices: a) global real prices for Urea, phosphate rock, and muriate of potash, as compiled by the World Bank [357]; and b) the distance from the production sites or mines, following the methodology proposed by McArthur *et al.* (2017) [358]. This methodology uses gravity models of trade, based on the premise that fertilizers are produced in a few specific countries [358]. The underlying hypothesis is that countries closer to fertilizer plants or mines are more sensitive to price variations because transport costs are a significant factor for farmers [358]. We applied a similar approach, estimating the minimum cost-adjusted distance by using the costDist function from *terra* package [359], global friction maps [360], the locations and operational years of potash [361] and phosphate mines, the locations of ammonia plants [358, 362] and the centroid of the cropland area on the country based on the Monfreda *et al.* (2008) crop maps [313]. Assessing the output prices for crops faces a similar problem: there is no standardized dataset with national-level data for the entire period. To resolve this, we used two proxies for crop prices: a) global real prices for specific commodities like wheat, maize, rice, palm oil, soybeans, sugar, and cotton, compiled by the World Bank [357], and b) standardized data from two FAOSTAT datasets [363, 364] that provide prices paid to producers at the country-level. The first dataset [363] contains information from 1990 onwards in United States dollars (USD), and local currency units (LCU), while the second dataset [364] covers 1966 to 1990, only in LCU. To standardize both datasets, we converted the older dataset into USD using annual currency exchange rates [365]. We then removed outliers independently for each crop by considering only values within 1.5 times the interquartile range. Before applying this method to the 1966-1990 dataset, we tested it on the LCU data for maize, wheat, and rice from the 1990 onwards dataset. We compared the original USD values with those obtained after converting the LCU data using exchange rates. The outlier detection method retained more than 99% of equivalent values (defined by a ratio between the original and calculated USD values of 0.99 to 1.01), while removing over 90% of non-equivalent

values. Finally, the data was converted to real prices by applying the Consumer Price Index [366].

Other socioeconomic factors, that are not directly related to the profitability of fertilizer use, have also been linked to country-level fertilizer use. These factors include the income level, reflected in the gross domestic product (GDP) per capita [367]; the population pressure, defined as the country's population divided by its agricultural land area [368]; and the farmers' knowledge about fertilizer use, as well as general education levels [354], which we measured by the percentage of total GDP spent on education. We used the sources listed in Table 5.3 to obtain data for these variables.

5.2.2 Data preprocessing

Several preprocessing steps were performed to prepare the raw data for the ML models. First, drawing from both expert domain knowledge and exploratory data analysis (e.g., using correlation matrices), the features relevant to N, P_2O_5 and K_2O fertilizer application rate were selected (Table 5.3). Since not every feature was relevant for each of the three targets, we narrowed down the dataset to data points where the average fertilizer application rate is known for all three fertilizers. This restriction ensured that the dataset comprised only labeled data points, which is crucial for supervised ML techniques. Subsequently, anomalies in the data where the fertilizer application rate was unrealistically large, i.e., greater than 5000 kg ha^{-1}, were removed. Finally, categorical features were one-hot encoded (OHE).

5.2.3 Machine learning

Previous studies within this domain typically propose linear equations to estimate the fertilizer application rate, and only consider a limited set of agricultural factors [297, 298]. However, it is well-established that natural phenomena frequently exhibit nonlinear relationships [369], rendering them unsuitable for modeling with linear methodologies. Similar studies have also employed Bayesian methods [303], with certain modeling assumptions that are not present in our study. ML has the potential to overcome these limitations. The field of ML has seen major increases in research [370] and industry [371], and, more specifically, ML has shown promising results in the field of ecology [51, 193], including agricultural research [48, 372], fertilizer consumption [47, 373] and fertilizer management [374]. For this reason, ML was used in this study to estimate the annual fertilizer application rate at the crop- and country-level. The benefit of using ML is threefold. First, ML allows us to include a larger range of variables, for example also including socioeconomic factors. Second, nonlinear ML techniques enable us to model nonlinear relationships between the variables in our dataset. Lastly, the model output can provide insights into the drivers associated with crop fertilization on a global scale, through the use of SHAP values [59] outlining the feature importance. The employed ML methods to estimate fertilizer application rate for crops differ from previous research, which typically relied solely on changes in crop area, overall fertilizer consumption, and limited data regarding fertilizer application rate at the individual crop-level [297, 298]. An advantage of our method is that it enables us to estimate values for countries where specific data is lacking by relying on other related variables. For example, the projected

data for the Union of Soviet Socialist Republics (USSR) aligns closely with national totals, even in the absence of country-specific data and without adjustments based on total consumption, as conventionally done [297, 298].

5.2.3.1 Models

In this study, two ML models based on gradient boosted regression trees were selected to predict the average annual fertilizer application at the crop- and country-level. In gradient boosting [375], an ensemble of weak learners (in our case regression trees) is trained sequentially. First, a weak learner is fitted to the original data. In the next iteration, another weak learner is fitted to the residuals, i.e., the differences between the ground truth target values and the current predictions made by the ensemble. When fitting a new weak learner to the residuals, gradient boosting adjusts its parameters in the negative gradient direction, aiming to reduce the residual error of the ensemble. This sequential learning process enables gradient boosting models to create a strong learner by combining multiple weak learners. The specific gradient boosting models employed in this study are XGB [190] and HGB [133, 191]. XGB has been shown to be a powerful tool for predictive modeling in a wide range of applications in both industry and research, including agricultural research [48] and fertilizer research [47]. It offers an optimized and scalable implementation of gradient boosting, and includes regularization techniques to prevent overfitting [190]. The HGB model is primarily based on LightGBM [191], which addresses one of the major bottlenecks in gradient boosting model training, namely the requirement to sort all samples at each node [133]. Indeed, in a traditional gradient boosting model, samples must be sorted at each node to determine the best split. This sorting process can become computationally expensive, especially when dealing with large datasets or deep trees. In HGB, the samples are first collected into a histogram, which removes the need for sorting as samples in a histogram are implicitly ordered. This optimization results in a model that is much faster to train than traditional gradient boosting models, while still achieving similar or better performance [133]. The choice for these two methods over other conventional ML approaches, such as neural networks, was primarily driven by the fact that both methods natively handle missing values. This constitutes a significant advantage, given that global fertilizer application rate data, along with the socioeconomic and agricultural variables used to predict the annual fertilizer application, are often incomplete. This also demonstrates another advantage of applying ML to this problem over the conventional approach using linear equations. Indeed, the absence of just one variable in the equation renders it impossible to compute.

5.2.3.2 Model training and evaluation

The selection of the optimal set of model hyperparameters is usually done using cross validation (CV), after which the CV error is reported as the performance of a model [376]. However, based on Stone (1974) [377], model assessment and model performance require different CV approaches. For this reason, we used nested CV, as it allowed us to find the optimal set of hyperparameters for a model and provide an unbiased estimate of the model's performance [378]. In nested CV, two levels of CV loops are used: an outer loop and an inner loop. In the outer loop, the dataset is split into training and testing sets, typically using k-fold CV. Each fold of the outer loop trains the model on

the training set and evaluates the model on the testing set. Within each fold of the outer loop, the training data is provided to an inner CV loop, in which the training data is further split into training and validation sets, also typically using k-fold CV. The inner loop is responsible for selecting the set of hyperparameters that performs best on the validation set. Finally, the performance of the selected set of hyperparameters is evaluated on the corresponding test set in the outer loop. In our study, we used a 2×5 nested CV, i.e., we had two outer loops and five inner loops. We employed a grid search that iteratively went over all possible combinations of hyperparameters, based on the explored hyperparameters as shown in Table 5.1 for both the HGB and XGB models. The performance of the models was evaluated by averaging the performances of the two models in the outer CV loop. The considered performance metrics included the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root-mean-square error (RMSE), all computed between the predicted and reported data points.

Table 5.1: Overview of the explored hyperparameters for the Histogram-based Gradient Boosting (HGB) and eXtreme Gradient Boosting (XGB) regression models.

Method	Hyperparameter	Possible values
HGB	max_depth	2, 5, 10, 20
	max_iter	25, 50, 100, 200, 500
	learning_rate	0.01, 0.1, 0.5, 1
	min_samples_leaf	5, 10, 20, 50
XGB	max_depth	2, 3, 4, 5
	n_estimators	25, 50, 100, 200, 300, 400
	colsample_by_tree	0.6, 0.7, 0.8, 0.9, 1.0
	subsample	0.6, 0.7, 0.8, 0.9, 1.0
	min_child_weight	3, 4, 5, 6, 8, 10

5.2.3.3 Model interpretability through SHAP value analysis

Unfortunately, gradient boosting methods are so-called black-box models, i.e., it is not immediately clear how certain predictions are made. However, assessing the impact of the features on the predicted fertilizer application rate in the learned models could provide us with valuable insights into the drivers of fertilizer application rate. Therefore, we resorted to explainable AI (xAI) methods to understand the predictions made by our models. More specifically, we used SHAP values [59] as they are model-agnostic, can account for interactions between features and have an intuitive interpretation. Indeed, summing the SHAP values for all features in one sample results in the prediction of the model. Additionally, like XGB and HGB, SHAP values are robust with respect to missing data by design [59]. Special attention was given to categorical values, as retrieving one SHAP value for a categorical feature that is divided into OHE features is non-trivial. However, as the SHAP values are calculated using the preprocessed input data (i.e., containing the OHE categorical features), the SHAP values for one categorical variable were obtained by adding together all SHAP values for its respective OHE features.

5.2.4 Adjustment to country totals

Previous research has always started with the same premise of allocating total fertilizer consumption at the country-level for estimating crop-level use [297, 298]. However, here we adopt a different strategy, initiating the estimation of the fertilizer consumption at the crop-level directly. Despite this shift in strategy, we still consider country-level data to be more reliable than datasets compiled from various FUBC sources. To reconcile our approach with the more dependable country-level data, we adjusted the ML predictions to align with FAOSTAT's total fertilizer consumption at the country-level [343]. As shown in Equation (5.2), we distributed the differences between the predicted total fertilizer consumption and the FAOSTAT totals equally among crops, after excluding the fraction used for grasslands and fodder crops from FAOSTAT totals.

$$Fert\_Pred_{icj} = FertML\_Pred_{icj} \times \frac{\sum_{d \in C}(FertML\_Pred_{idj} \times HArea\_FAO_{idj})}{FAOSTAT\_FERTng_{ij}} \tag{5.2}$$

Where, $Fert\_Pred_{icj}$ represents the fertilizer application rate predictions after the adjustment for year i, crop c, and country j. $FertML\_Pred_{icj}$ denotes the ML model predictions, C is the set of all crops classes included in the models, $HArea\_FAO_{idj}$ the FAOSTAT harvested area [331] of each crop class d, and $FAOSTAT\_FERTng_{ij}$ is the total FAOSTAT fertilizer consumption for the country, after removing the fraction used for grasslands and fodder crops.

5.2.5 Validation

To validate the model predictions, we compared the model predictions with national databases containing information about the average use per hectare for different fertilizers and crops. This validation is quantified using the MAE and mean absolute percentage error (MAPE) as well as with comparative plots if enough data was obtainable from the various national databases. The MAE gives an idea about the actual deviation, whilst the MAPE makes the comparison between prediction errors easier. The evaluated national databases include data obtained from for the USA [322], UK [334], India [323–329], Sweden [338–341], Philippines [336], and New Zealand [342]. For Pakistan [335], only data for the sum of fertilizer application rate is available, hence the sum of N, P_2O_5, and K_2O was used, expressed as NPK. This approach is restricted by available data in national databases for average fertilizer application rate across various crops and nutrients.

5.2.6 Gridded crop-specific application rate per fertilizer

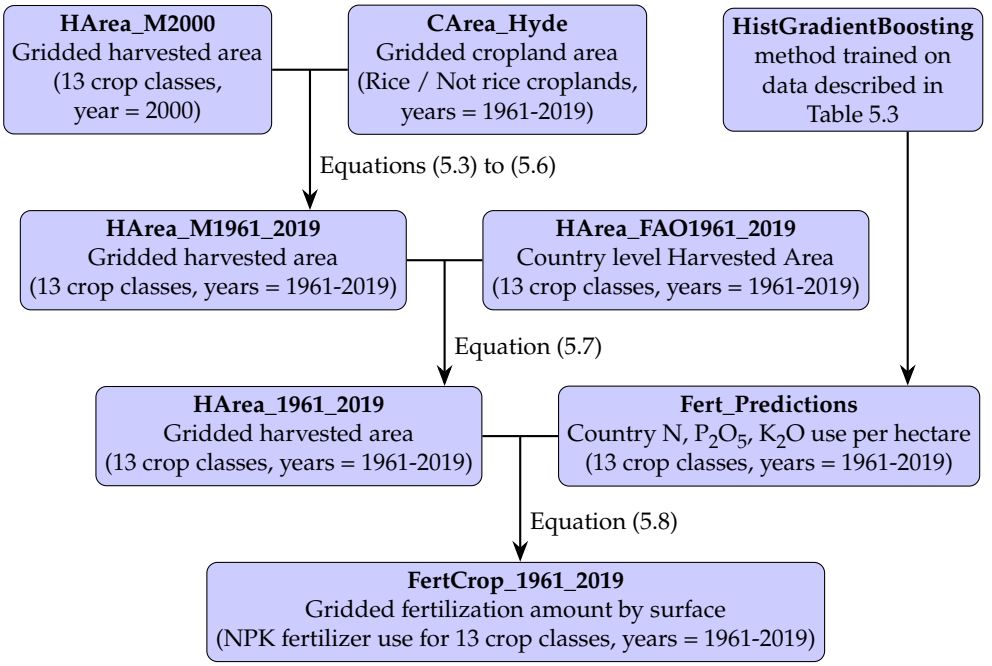


Figure 5.1: Outline of the process for generating the gridded crop-specific fertilizer dataset.

Following the generated comprehensive dataset of global fertilizer application rate, we constructed detailed 5-arcmin resolution gridded maps for each fertilizer (N, P_2O_5, and K_2O), crop class and year from 1960 to 2020. The final gridded map dataset was compiled in a three-step process, as highlighted in Figure 5.1. First, data of the gridded harvested area spanning from 1961 to 2019 for the 13 distinct crop classes (see Table 5.4) were acquired by combining data from the EARTHSTAT project of the year 2000 ($HArea\_M2000$) [313], supplemented with historical arable land and permanent crop areas per year ($CArea\_Hyde$) from the History Database of the Global Environment (HYDE version 3.3) [314]. The EARTHSTAT maps were created by combining national-, state-, and country-level census statistics with an up-to-date global dataset of croplands, organized on a 5-arcminute by 5-arcminute latitude-longitude grid. These datasets, reflecting land use around the year 2000, detail both the area harvested and the yield of 175 diverse crops worldwide [313]. Innovative maps outlining major crop groups were generated by consolidating these individual crop maps. The HYDE 3.3 project provides long time series estimates and maps for land use, including the cropland areas, based on an allocation algorithm with time-dependent weighting [314]. The elaborate information from the crop specific EARTHSTAT maps for the year 2000, in combination with the yearly changes in gridded cropland from HYDE 3.3, allowed us to make detailed gridded 5-arcmin resolution crop specific harvested areas for each of the evaluated years and crops using Equations (5.3) to (5.6):

For $CArea\_Hyde_{gi} > 0$ and crop is rice:

$$HArea\_M_{gic} = CArea\_Hyde\_R_{gi} \times \frac{HArea\_M2000_{gRice}}{CArea\_Hyde\_R_{g2000}} \tag{5.3}$$

For $CArea\_Hyde_{gi} > 0$ and crop is not rice:

$$HArea\_M_{gic} = CArea\_Hyde\_NR_{gi} \times \frac{HArea\_M2000_{gc}}{CArea\_Hyde\_NR_{g2000}} \tag{5.4}$$

For $CArea\_Hyde_{gi} > 0 \cup \sum_{c \in C} HArea\_M2000_c = 0$ and crop is rice:

$$HArea\_M_{gic} = CArea\_Hyde\_R_{gi} \times \frac{\sum_{k \in K} HArea\_M2000_{gRice}/K}{CArea\_Hyde\_R_{g2000}} \tag{5.5}$$

For $CArea\_Hyde_{gi} > 0 \cup \sum_{c \in C} HArea\_M2000_c = 0$ and crop is not rice:

$$HArea\_M_{gic} = CArea\_Hyde\_NR_{gi} \times \frac{\sum_{k \in K} HArea\_M2000_{gc}/K}{CArea\_Hyde\_NR_{g2000}} \tag{5.6}$$

Here, the indices denote the grid cell (g), the year (i), the crop (c). The harvested area ($HArea\_M_{gic}$) was generated through a series of conditional operations. These conditions stipulate that if the value of the HYDE3.3 cropland area map ($CArea\_Hyde_{gi}$) for that year i and grid cell g is larger than 0, and the crop is not rice, then the value of that grid cell for that specific crop and year is given by the HYDE3.3 cropland area ($CArea\_Hyde\_NR_{gi}$) for that grid cell/year combination. The value of the grid cell is then further adjusted by the ratio of the HYDE3.3 map of the year 2000 to the EARTHSTAT map of the year 2000 for the corresponding grid cell and crop ($\frac{HArea\_M2000_{gc}}{CArea\_Hyde\_NR_{g2000}}$). In the case of rice, the specific HYDE3.3 map for cropland area of rice was selected. Additionally, in instances where $CArea\_Hyde_{gi}$ was larger than 0 and the sum of all crops across the EARTHSTAT maps of the year 2000 is equal to 0 (e.g., when new lands are cultivated), a progressively expanding area K was evaluated to find an appropriate ratio based on the average of the k neighboring cells. The evaluated values for k were 5, 10, 25, 50, 100, 150, 200 and 250, up until a value different from zero for the ratio is found. If no value different from zero was found, the ratio value was set equal to 1. This last step made the assumption that the crop distribution in neighboring cells adequately represents the distribution in the newly cultivated area, allowing for the calculation of adjusted harvested areas. Furthermore, as the $HArea\_M2000_{gc}$ is consistently used, we assumed that the changes in crop distribution over time remain constant.

To ensure consistency with FAOSTAT data used in the model predictions, the gridded harvested area ($HArea\_M1961\_2019$) was aligned with the country-specific harvested area reported by FAOSTAT ($HArea\_FAO2000$). This adjustment, applied through Equation (5.7), provided a corrected gridded harvested area for the 13 crop classes over the 60-year period ($HArea\_1961\_2019$):

$$HArea_{gic} = HArea\_M_{gic} \times \frac{\sum_{j \in J} HArea\_FAO_{icj}}{\sum_{j \in J} HArea\_M_{icj}} \quad (5.7)$$

In this equation, $HArea\_FAO_{icj}$ represents the harvested area for year i, crop c, and country j as reported by FAOSTAT, summed over all countries (J) in grid cell g (to accommodate grid cells with multiple countries). Similarly, $HArea\_M_{icj}$ represents the estimated harvested area for the same combinations, also summed over all countries in grid cell g. The ratio of these sums adjusts the model gridded harvested area ($HArea\_M_{gic}$) to match FAOSTAT data, ensuring the resulting gridded harvested area on a country level is consistent with official statistics across the 60-year period.

Finally, the gridded harvested area ($HArea_{1961\_2019}$) was augmented with the average application rate of each predicted fertilizer (N, P_2O_5, K_2O) as per Equation (5.8):

$$FertCrop_{gic} = HArea_{gic} \times \sum_{j \in J} (Fert\_Pred_{icj} \times PercCountry_g) \quad (5.8)$$

where $Fert\_Pred_{icj}$ is the country-level prediction resulting from the HGB model after applying the adjustment, and $PercCountry_g$ refers to the percentage of grid cell g that is occupied by the country j. This process was then applied to each fertilizer separately to obtain gridded maps for each fertilizer, year, and crop combination.

5.3 Data Records

The gridded fertilizer application data for N, P_2O_5, and K_2O by crops from 1961 to 2019 are available in a Figshare repository [379]. The dataset spans from 180ºE to 180ºW and 90ºS to 90ºN at a resolution of 5 arc-min in WGS84 (EPSG: 4326). It is provided in `.tiff` format, which can be read by many tools, such as R and Python. The gridded application data by crops and fertilizers are stored in several files named "Crop_NameFertilizerYear.tiff". Here, "Crop_Name" represents each crop class listed in Table 5.4, "Fertilizer" refers to N, P_2O_5, or K_2O, and "Year" indicates any year from 1961 to 2019.

5.3.1 Crop-specific N application

On a global scale, the N application has grown for all crops (see Figure 5.5 in the appendix). For example, the average N use of the three main cereals has risen from 17.1 ± 6.1 kg ha^{-1}, 26.6 ± 7.2 kg ha^{-1}, 12.1 ± 3.9 kg ha^{-1} for wheat, maize and rice, respectively, in the 1960s, to 97.8 ± 4.2 kg ha^{-1}, 118.8 ± 4.2 kg ha^{-1}, 113.8 ± 1.9 kg ha^{-1} in the 2010s decade. Moreover, the largest increases in N application occurred in vegetable crops, with a global growth of more than 120 kg ha^{-1} between these two decades (Figure 5.5). Conversely, the lowest increases occurred in soybean, where N application rates grew by less than 20 kg ha^{-1}. At the regional scale, the intensification of N fertilizer use has shifted from higher use at the beginning of the period in the USA and Europe to being currently dominated by Asian countries such as China and India (Figure 5.5). This trend

is particularly evident for some crops like vegetables and fruits, where China now has the areas with the highest N use worldwide, whereas in the 1960s, these areas were primarily in Southern Europe and California.

5.3.2 Crop-specific P_2O_5 application

The application of P_2O_5 also experienced global increases across all crops (see Figure 5.6 in the appendix), but to a lesser extent than N. The average P_2O_5 used for the three main cereals and soybean rose from 13.8 ± 3.3 kg ha^{-1}, 13.1 ± 2.4 kg ha^{-1}, 6.3 ± 1.9 kg ha^{-1}, and 12.6 ± 2.4 kg ha^{-1} for wheat, maize, rice and soybean, respectively, in the 1960s to 35.5 ± 4.9 kg ha^{-1}, 43.0 ± 5.7 kg ha^{-1}, 39.9 ± 5.0 kg ha^{-1}, and 39.1 ± 6.6 kg ha^{-1} in the 2010s. Similar to N, the largest increases occurred in vegetable crops, where P_2O_5 application rates increased by more than 50 kg ha^{-1}. Conversely, the smallest increases were observed in the other cereal crop class, where the average P_2O_5 application rate increased by only about 2.5 kg ha^{-1} between the two decades. Regionally, a similar pattern occurred with P_2O_5 use, following the trend previously seen with N, where the hotspot shifted from Europe to Asia. This shift is particularly notable for wheat, where the hotspot of P_2O_5 intensification moved from Western Europe to northern India and northeastern China (Figure 5.6).

5.3.3 Crop-specific K_2O application

Globally, the use of K_2O has also increased across almost all crop classes (see Figure 5.7 in the appendix). For wheat, maize, rice, and soybean, the average K_2O application rates have risen from 7.2 ± 1.6, 9.8 ± 2.0, 3.4 ± 0.5, and 11.6 ± 2.6 kg ha^{-1}, respectively, to 15.4 ± 4.1, 33.1 ± 4.8, 27.3 ± 3.9, and 9.8 ± 3.2 kg ha^{-1}. Unlike N and P_2O_5, the largest difference in average K_2O application occurred for the oil palm crop, which increased from 3.7 ± 1.4 kg ha^{-1} during the 1960s to 87.6 ± 8.3 during the 2010s. Similar to P_2O_5, the other cereal class experienced the smallest change in K_2O use. In this case, the average K_2O application rate decreased from 11.7 ± 1.9 kg ha^{-1} during the 1960s to 9.8 ± 3.2 kg ha^{-1} during the 2010s. Regionally, a similar pattern emerged with K_2O, following the trend observed with N and P_2O_5, with the hotspot of K_2O fertilization shifting from Europe and the USA to Asia. However, this change was more pronounced in different crops, such as oil crops, where the use of K_2O has increased significantly in countries like Malaysia and Indonesia (Figure 5.7).

5.4 Technical Validation

This section provides a detailed discussion of the validation efforts made to confirm the validity, consistency, and plausibility of our compiled dataset and predictions. First, the performance of the ML models is evaluated. Subsequently, we use SHAP values to confirm that our models used sensible features to make their predictions, based on literature. Finally, the predictions are validated by comparing them with reported data in both national and global databases.

5.4.1 ML Model performance

The performance of the ML models predicting the fertilizer application rates for the three fertilizers is shown in Table 5.2. Both XGB and HGB significantly outperformed the naive prediction, which is uses the mean fertilizer application as its prediction. HGB consistently outperformed (or matched) XGB for all three fertilizers and performance metrics. For this reason, we will use the HGB model in the remainder of this technical validation, as well as any subsequent analyses.

Table 5.2: Performances of the eXtreme Gradient Boosting (XGB) and HistGradientBoosting (HGB) models on the test sets in a 2x5-fold nested cross validation grid search. The performance is quantified using the mean absolute error (MAE), root mean squared error (RMSE), mean squared error (MSE) and the coefficient of correlation (R^2). The naive performance of a model is defined as the performance of a model that uses the mean of all training samples as its prediction. It serves as a baseline value to compare the test performances of the models with. The best performances are indicated in boldface.

Fertilizer	Model	MAE	RMSE	MSE	R^2
N	HGB	**26.01 ± 0.94**	43.50 ± 5.13	**1905 ± 446**	**0.62 ± 0.04**
	XGB	26.67 ± 1.48	**43.35 ± 7.12**	**1905 ± 617**	**0.62 ± 0.08**
	naive	53.09 ± 0.75	70.13 ± 4.19	4927 ± 588	0.00 ± 0.00
P_2O_5	HGB	**15.19 ± 0.67**	**25.68 ± 1.18**	**660 ± 61**	**0.63 ± 0.05**
	XGB	16.83 ± 0.23	26.40 ± 0.74	697 ± 39	0.61 ± 0.04
	naive	29.97 ± 0.23	42.12 ± 1.02	1774 ± 86	0.00 ± 0.00
K_2O	HGB	**19.18 ± 0.27**	**35.74 ± 4.56**	**1287 ± 326**	**0.65 ± 0.08**
	XGB	19.99 ± 0.20	36.24 ± 4.66	1324 ± 338	0.64 ± 0.09
	naive	43.08 ± 0.76	60.25 ± 0.52	3631 ± 63	0.00 ± 0.00

5.4.2 SHAP value analysis

To examine the impact of the features on the prediction of the N, P_2O_5 and K_2O application rates, the SHAP values of the ten most important features for the three corresponding HGB models are illustrated in Figure 5.2. Agrological drivers dominated the predictions, comprising six, seven, and eight of the ten highest ranked features, respectively. The impact of the features remained consistent across all fertilizers, albeit with varying magnitudes (Figure 5.2-d,e,f). In particular, the predicted fertilizer application rates were consistently positively impacted by the country fertilizer per ha and the crop nutrient removal per ha (as red dots, i.e. high values of country fertilizer per ha and high nutrient removal per ha, corresponded with positive SHAP values), while it was negatively impacted by the crop nutrient content (red dots corresponding with negative SHAP values; Figure 5.2-d,e,f). These relationships align with the expected influence of these features on fertilization at the crop-level [380]. Across the different fertilizers, the most important socioeconomic features varied. For instance, the GDP per capita was the most important socioeconomic feature in the prediction of the P_2O_5 and K_2O application rates, while in the N prediction, the global crop price was more important. Fertilization at the country-

level is usually associated with the economic development of the country, measured by GDP [367, 381]. However, at the crop-level, this relationship only held true for the most expensive fertilizers, P_2O_5 and K_2O. For N, the most affordable nutrient [357], factors such as global crop price and N cost from production appeared to be more significant (Figure 5.2). Few environmental features seemed to be relevant for the predictions (Figure 5.2); only the soil pH, soil organic carbon stock (OCS), and aridity index appeared in the top ten for some nutrients. Although the influence of these variables appeared to be low, the direction of these relationships confirmed the findings of other authors at the farm- or regional-level for soil organic carbon content and soil pH. [344–346].

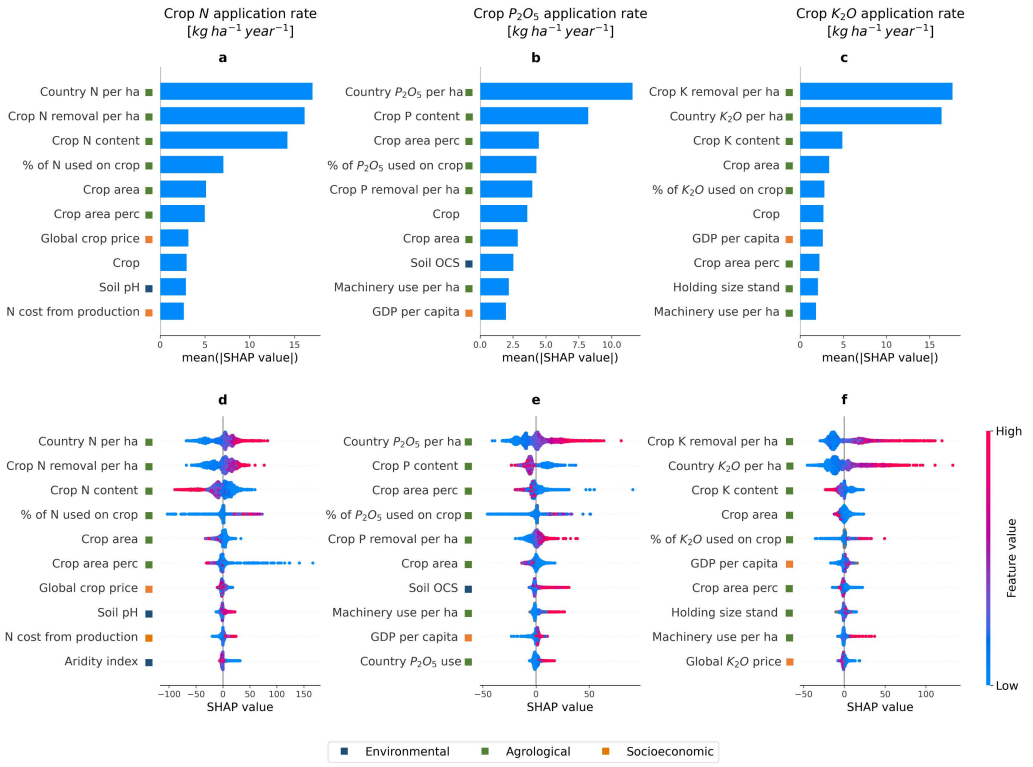


Figure 5.2: SHapley Additive eXplanation (SHAP) values of the top 10 most important features in the prediction of, respectively, the crop N (a,d), P_2O_5 (b,e) and K_2O (c,f) application rates using Histogram-based Gradient Boosted regression. (a,b,c) The top plots present the average feature importance, determined by the mean absolute SHAP value of each feature. (d,e,f) The bottom plots depict a SHAP value for each prediction and show the local feature importance and the feature effect. The color of a dot represents the value of the feature in that instance - red indicating relatively high, blue indicating relatively low values. A dot with a high SHAP value for a feature suggests a positive contribution to the prediction, whereas a very negative SHAP value leads to a lower prediction. The features are ranked in order of descending average importance and the blue, green and orange squares indicate whether the feature is an environmental, agrological or socioeconomic characteristic.

5.4.3 Validation

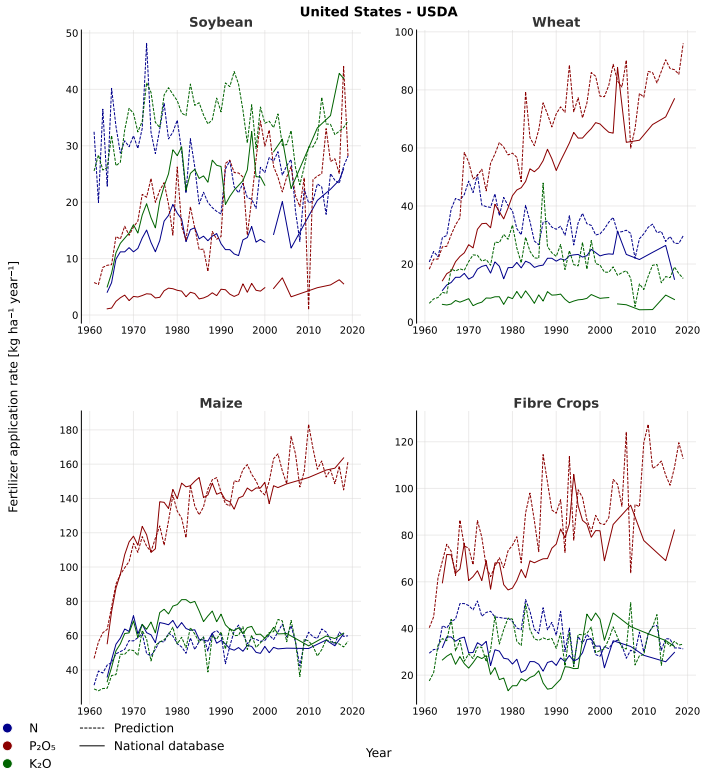


Figure 5.3: Comparison of the application rates per ha per year for various crops between our predicted model output and the data reported by the United States Department of Agriculture (USDA) for the USA.

To evaluate the validity of our results, we compare the compiled dataset based on the predictions against several national databases [322–329, 334–336, 338–342] based on the MAE and MAPE errors between both, averaged over the available years as illustrated in Table 5.5. For most country/crop combinations, the differences are within reasonable ranges, with MAE values between 5-40 kg ha^{-1} and MAPE values between 10%-50%. However, for some countries, the deviations are larger, suggesting that our models may not capture all the underlying intricacies in the data for each country or crop. This can be seen for Sweden where most results deviate between 20%-100%, or New Zealand where similar results can be found. However, it should be noted that these larger differences between our compiled dataset and the national databases cover only limited years as data was not always available for certain countries, as was the case for Sweden and New Zealand. Still, these discrepancies are slightly better than in earlier research [297]. Additionally, more detailed plots to evaluate the results per year for the USA and UK, based on the United States Department of Agriculture (USDA) and Department for Environment, Food & Rural Affairs (DEFRA) respectively, are included in Figures 5.3 and 5.4. For the USDA and DEFRA crop nutrient data, the MAPE values are less than 50% and usually less than 25%, except for USDA soybean N (Figure 5.3). Figures 5.3 and 5.4 show that our predictions follow the real observed trend for the samples and thus form

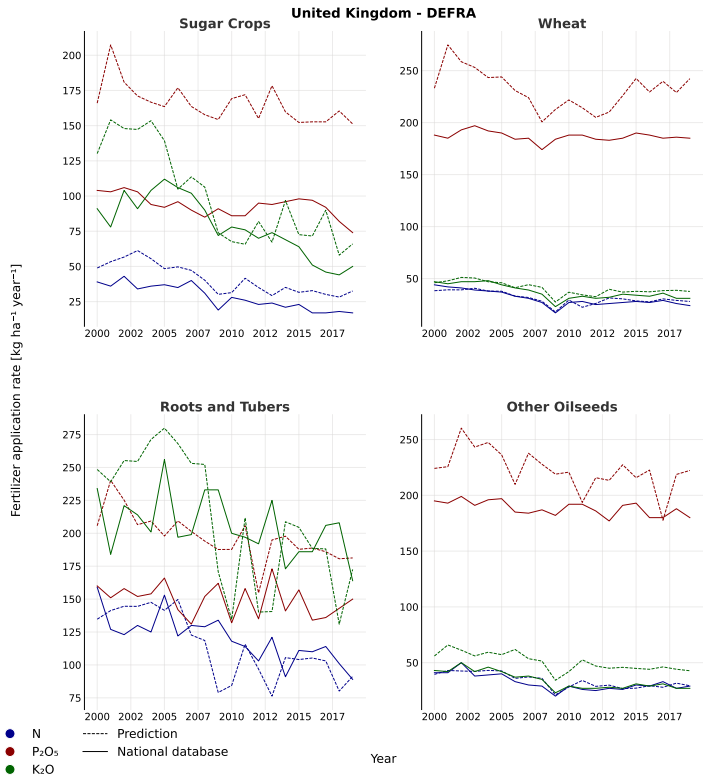


Figure 5.4: Comparison of the application rates per ha per year for various crops between our predicted model output and the data reported by the Department for Environment, Food & Rural Affairs (DEFRA) for the UK.

a reliable end source with only minimal differences. These discrepancies between the national databases and our compiled dataset can be attributed to occasional disparities between the application rates in the training data (data provided by the global dataset compilation) and the data in the national databases, e.g., the USA data for soybean N in 1998 differed by 400% between the two samples. These differences should be taken into account when comparing our results to the national databases, as our predictions are based on the global compiled dataset. As can be seen in Table 5.6, where the global databases data and the national databases are compared based on MAE and MAPE, most country/crop combination indicate an MAPE values between 10%-50%, which is similar to our resulting error in Table 5.5. Also, the lack of training samples for some country/crop combinations resulted in larger errors for these occurrences.

To conclude, the model performances and logical feature importance, derived from the SHAP values, in conjunction with the relatively minor differences between this study and regional statistics, as well as earlier literature [297], indicate that our crop-specific fertilizer application rate dataset is comparatively accurate across regions and years.

5.5 Usage Notes

In this study, we provide detailed estimates on global N, P_2O_5, and K_2O fertilizer application rate based on the HGB model output and compile a comprehensive dataset of these estimates by major crop groups between 1960 - 2020. Tabular data of the country- and crop-level predictions are made available as well as the 5-arcmin resolution gridded maps from our application, rendering an easy to use complete dataset. Subsequent analysis can be done both on the tabular data and the outputted maps, such as a trend analysis of fertilizer application rate or causal discovery to identify drivers of fertilizer application rate. Furthermore, our dataset can be leveraged as a source in other models where for example yield, ecological impact or fertilizer pricing can be seen as the output rather than use.

Our results represent an improvement and advance in efforts to evaluate historical fertilizer consumption for different crop groups and fertilizers. However, as demonstrated during the validation process, this approach still has limitations that data source users should be made aware of. The limited amount of available data for some crops, nutrients, and regions can lead to biases, particularly in regions such as Africa, during certain years, especially in the 60s, and for certain nutrients, mainly K_2O. Hence, the ML approach can be sensitive to outlying data points or noise and the limited data can make it prone to overfitting, which was mitigated as much as possible in the CV setup. In addition, our model is trained on data provided by global datasets [67, 304–309, 315–321], which means that while our predictions may align closely with them, it is essential to acknowledge that they might diverge from national data mainly due to the difference between the two data sources as highlighted by the validation. This discrepancy between global and national databases such as the USDA [322] or DEFRA [334] databases highlights the complexity of accurately capturing historical fertilizer consumption trends across different regions and crop types. Moreover, the gridded cropland data provided by the HYDE 3.3 project [314], is inconsistent with the one from satellite-derived land use (e.g., China and India [382, 383]) or data derived from a national census at regional scale (e.g., USA [384]), as stated by Adalibieke *et al.* (2023) [297]. Furthermore, utilizing neighboring cells to allocate harvested areas across different crops, as well as leveraging the EARTHSTAT map [313], implies some assumptions (see Equations (5.3) to (5.6)). The main assumption is the suggestion that the distribution pattern of a specific cell mirrors that of its neighboring cells, which constrains potential changes in cropland over cells. The consistent use of the EARTHSTAT map [313] of the year 2000 also assumed that the crop group distribution of harvested area remains constant over time between 1961-2019. Finally, it is important to recognize that there are additional uncertainties stemming from the utilization of various data sources and methodological decisions within each data source, but these lie beyond the scope of our study.

Nevertheless, our study extends the current literature by providing a more detailed historical geospatial distribution of fertilizer application rate by crop and using ML to obtain detailed predictions with high precision. The detailed description and open-source code, in combination with the limited data sources used and ability to forecast, also make the method reproducible and easy to extend to forecast fertilizer application rate. In addition, our approach does not entail any assumptions, making it more flexible and robust than precious studies. Future research can build upon our study by expanding on more detailed specific fertilizer application rate. Considering the frequency of fertilizer ap-

plication as well as the timing can be valuable for future research on the evaluation of fertilizer effectiveness and use. In addition, our study focuses on broad fertilizer applications, however, more detailed maps can be made for different types of specific fertilizers considered in our study (e.g., N fertilizer types). Furthermore, the time granularity of our maps can be improved. In addition, satellite data can be used to obtain even more fine-grained predictions, both in regions and more detailed time periods. Finally, a deeper investigation into the drivers of fertilizer application rate could enrich our understanding. While our focus has primarily been on the explainability of our model, exploring methodologies such as causal discovery or causal ML within a temporal setting could unveil the drivers of fertilizer application rate over time, potentially providing valuable insights and facilitating more detailed predictions.

5.6 Data availability statement

Our Python (3.10.3) code, encompassing the model training, prediction generation, SHAP value computation, model validation and plot creation, as well as the R (4.2.2) scripts made for map generation are made available alongside the provided data map resources [385]. Open source packages used in the code are tabulated with their respective version in Table 5.8. Access to these resources is available at the designated location [379, 385].

5.7 Appendix

5.7.1 Fertilizer usage in grasslands and fodder crops

5.7.1.1 The Americas

Argentina: In the 1960s, fertilizer application rate in Argentina was primarily directed towards sugar cane and citrus [386], with minimal application to grasslands, nearly zero in 1964 [386]. Throughout the 1970s and 1980s, the fertilizer application rate remained low, although there was a notable increase in P_2O_5 application to grasslands, reaching 28% country consumption in 1979 [304]. The substantial expansion in N and P_2O_5 fertilizer occurred during the 90s, leading to a slight rise in the share of N used for grasslands, and to a significant decrease in P_2O_5 share for grasslands [315–319]. To fill data gaps, we adopted a methodology similar to Lassaletta *et al.* (2014) [311], utilizing linear interpolation of national [387–395] and global datasets for the years lacking data, with grasslands' fertilizer share assumed as 0 in 1965 [386]. Despite potential limitations, setting the share to 0, as done in FAO nutrient budgets [293], may underestimate fertilizer application rate, particularly for P_2O_5. K_2O fertilizer application rate in Argentina remains minimal due to soil composition, with all reports except one considering it as 0 in the use for grasslands and fodder crops [304, 305, 315–319, 387–395].

Brazil: According to several sources, the use of fertilizer in Brazil's grasslands has been very low [396, 397]. The most recent values reported by IFA in 2014 and 2018 indicate that less than 1% of the fertilizer used in Brazil is used in grasslands [67, 321]. However, Lassaleta *et al.* (2014) [311] and FAO [293] considered higher percentages for N and K_2O based on regional averages [311] or previous research [293]. For P_2O_5 and K_2O, only FAO includes an estimation, considering 0 for P_2O_5, while they estimate the K_2O consumption by calculating the average between N and P_2O_5 consumption [293]. We have decided to consider 0 as the share used for grasslands and fodder crops due to the latest reported values and considering that no information is reported in previous reports [67, 304, 305, 315–319, 321].

Canada: Most of the compiled reports do not provide information about the use of fertilizers for fodder crops and grasslands [304, 305, 315–319]. The latest report, with 2018 data, indicated that 0.5% of N, 0.9% of P_2O_5, and 0.6% of K_2O fertilizers were allocated to permanent grasslands, which increased to 12%, 14.5%, and 25% respectively when considering tame hay and silage maize as well. Regarding N, FAO [293] and the 2014 estimation by Lassaleta *et al.* (2014) [311] are consistent with the 2018 estimation for all forages. However, the values for P_2O_5 and K_2O for all forages in the latest report differ significantly from those used by FAO [293] (0% for P_2O_5 and 5% for K_2O). This discrepancy in P_2O_5 may be due to FAO's reliance on Heffer *et al.* (2017) which does not consider nongrass perennial crops 0% [398], and the discrepancy for K_2O because FAO considered the average value between N and P_2O_5 [293]. We decided to utilize the percentage for all forages included in the last report [67] for the entire period. We maintained the same values throughout the period due to insufficient data to estimate any trends. Additionally, in 1974, Beaton and Berger noted that a significant share of fertilizer used in Canada was for forages, estimating 45% of total use in 1970 was for hay and grazing grasslands [399]. They suggested that their estimation might be

overestimated; however, it is unlikely that the fraction of fertilizers used for forages was 0 between 1960 and 1990.

Chile: Based on the estimations of the FAO and IFA reports, Lassaleta *et al.* (2014) [311] and FAO [293] considered a significant share of fertilizer used for grasslands. For N Lassaleta *et al.* (2014) suggested an increasing percentage from 0% in 1960 to 20% in 2005, while FAO maintained a constant percentage of 20%. For P_2O_5 and K_2O, the values used by FAO were also high, at 30% and 25% respectively. However, for Chile, using a constant value for the period overestimated the early years as the share used for grasslands for N and P_2O_5 was only 1% at the beginning of the 1960s [400]. We therefore decided to make a reconstruction similar to the one demonstrated by Lassaleta *et al.* (2014), by considering 1% as the starting share for each nutrient, and incorporating the reported values for all grasslands [67, 304, 305, 315–319, 321].

Dominican Republic: The values reported in global studies from the 1990s indicate that during this decade, the percentage of fertilizer application rate on grasslands and fodder crops ranged between 2% and 4% [315, 317, 318]. Considering these findings, Lassaleta *et al.* (2014) [311] allocated values ranging from 0% to 2% for N. We have chosen to utilize the average values from the three reports [315, 317, 318] for the period 1990-2020. This decision was influenced by the lack of available data since 1997, and by the emergence of fertilizer application rate for pasture as a new and increasing practice during the 90s [401].

Mexico: The use of fertilizer for grasslands and fodder crops appears to be nearly zero, as indicated by previous research [293, 311] and reported values [67, 315, 319, 321]. The only relevant fertilizer used for grasslands and fodder crops in Mexico appears to be related with P_2O_5 related with alfalfa production [315, 319, 402]. Due to limited available information, and the longstanding presence of alfalfa production in Mexico since the Spanish colonization, we opted to consider the average percentage (2.5%) used in the two reports with data for alfalfa [315, 319].

United States of America: According to global and national estimates from previous research, the share of N used for grasslands during the period ranged from 0% to 20% of the total [293, 301, 311]. For P_2O_5 and K_2O, the most recent estimation from FAO indicated a constant share of 0% for phosphorus and 10% for potassium [293]. To estimate the total fertilizer use for permanent and non-permanent grasslands from 1959 to 2014, we used all the available data [315, 317, 321, 399, 403]. In many sources, the information for grasslands is combined, encompassing both permanent and non-permanent grasslands. We used linear interpolation to estimate the share used for all grasslands together, replicating the method from the most recent estimation [301]. However, we included data from three additional years (1974, 1992, 1996) [315, 317, 399], and also extended the estimation to cover P_2O_5, and K_2O.

Uruguay: Grassland fertilization was actively promoted by the Uruguayan government during the 60s [404]. As early as 1963, one-third of the fertilizer used in the country was applied to pastures, with a focus on P_2O_5 due to the low P content of the Uruguayan soils [404]. These trends are reflected in the first IFDC report, which allocated 45% of the P_2O_5 used in the country for grasslands and fodder in the year 1986 [305]. However, this share decreased to 22% by 2018. In contrast, the percentage of N used for grasslands has shown an increasing trend, from almost 0% in 1986 [305] to 12% in recent years [67, 319]. K_2O is not used for these agricultural lands in the country [67, 305, 315, 319]. Given the

significant variation in percentages between decades and nutrients, we performed linear interpolation considering 33% for P_2O_5 in 1960, and 0% for N as starting points, and all the values included in the reports [67, 305, 315, 319].

Venezuela: Information regarding grassland and fodder crop fertilization in Venezuela is limited. Due to the scarcity of data and discrepancies between reported values [317, 319], FAO has considered a fertilization rate of 0% for grasslands during the specified period. Conversely, Lassaleta *et al.* (2014) [311] proposed different rates between 0% and 9% from 1960 to 2009 for N. Given the challenge of determining the most appropriate criteria, we opted to adhere to the FAO considerations. This decision is influenced by low government optimal use recommendations for grasslands compared to croplands [405], along with scientific evidence suggesting minimal fertilization for warm-climate grasslands [405, 406].

5.7.1.2 Oceania

Australia: According to Lassaletta *et al.* (2014), the share of N used for grasslands never exceeded 8.5% [311], which is similar to the 10% used by FAO in their nutrient budgets assessments of [293]. Despite an intensification in the use of N in Australian grasslands over the past three decades [407], it is noted that these grasslands were already being fertilized in the late 1950s, primarily with K_2O [408]. For instance, in 1956, 15% of the K_2O used in South Australia was directed towards pastures, a figure that rose to 42% by 1966 [408]. Therefore, we have opted to consider a constant share of 6.4% for N use since 1960 derived from the mean value of the reports [67, 305, 315–319, 321]. Regarding P_2O_5 and K_2O fertilizer, it appears that the FAO estimations [293] may have underestimated their use, particularly for K_2O. Thus, we decided to use the average value of all reports, because even with fluctuations, the variation in the reported values since 1985 is not too high, resulting in figures of $38.4 \pm 4.1\%$ for P_2O_5 and $41.6 \pm 6.9\%$ for K_2O [67, 305, 315–319, 321].

New Zealand: Previous global research presented contradictory estimates of fertilizer application rate for grasslands in New Zealand [293, 311], with figures ranging widely from 0% to 90%. However, both global and national reports consistently support the notion that the majority of the fertilizer application rate in the country is directed towards grasslands and fodder crops [67, 317, 319, 321, 342]. Therefore, we have adopted a constant percentage throughout the entire period as grasslands have been the primary type of agricultural land developed in the country since the British colonization, their fertilization has been relevant since the early 20th century [409], and the fraction used for grassland has remained constant at least in the last 30 years [67, 317, 319, 321]. The percentages selected were derived from the average of global reports [67, 317, 319, 321]: $91.1 \pm 1.4\%$ for N, $93.0 \pm 3.3\%$ for P_2O_5, and $88.8 \pm 4.4\%$ for K_2O.

5.7.1.3 Europe

Between 1980 and 2000, Europe accounted for at least half of the N fertilizer used for grasslands and fodder crops, while consuming less than one-third of the total global fertilizer consumption [294]. Consequently, the available information was broader, and the methods applied could be more comprehensive. Einarsson *et al.* (2021) provided the

most comprehensive estimation for N in most European countries [312]. They compiled and estimated the surfaces of croplands, including fodder crops, as well as temporary and permanent grasslands for the EU countries spanning from 1960 to 2019. Using their compiled data and the fertilizer application rate information from our study, we employed a similar methodology to estimate the fraction of N, P_2O_5, and K_2O used in these areas.

However, we extended the analysis to include fodder crops and all types of grasslands together, while also estimating P_2O_5, and K_2O. First, we used Equation (5.9) to estimate the ratio (R_{f-a}) between the fertilization intensity of grasslands and fodder combined, and the fertilization intensity of all agricultural land for the years with available data:

$$\frac{Q_f}{Q_a} = \frac{R_f \times A_f}{R_a \times A_a} \rightarrow \frac{Q_f \times A_a}{Q_a \times A_f} = \frac{R_f}{R_a} = R_{f-a} \tag{5.9}$$

where Q_f is the amount of fertilizer (N, P_2O_5, or K_2O) used for grasslands and fodder crops, Q_a denotes all the fertilizer of the same nutrient used in the country, A_f represent the grassland and fodder surface, and A_a represents the total agricultural land, and R_{f-a} the ratio of fertilizer application rate per area between fodder and grasslands (R_f), and all agricultural land (R_a). Therefore, R_{f-a} represents the fertilizer application relationship between fodder and grassland in comparison to all agricultural lands.

After estimating the annual R_{f-a}, we used two different procedures and equations depending on the years for which R_{f-a} data was available. If scientific literature and the observed variation in R_{f-a} indicated significant differences across the years, we performed a linear interpolation of the available values and then applied Equation (5.10). Otherwise, we applied (5.11). To assess the variation in R_{f-a} we estimated the MAE of the results derived from Equation (5.11) compared with all the reported values. When the variation of R_{f-a} occurred only in some decades within the period, we combined Equation (5.10) and Equation (5.11). Detailed explanations were provided for each country individually. For non-EU countries, we applied similar procedures as those used for the other continents. In Equations (5.10) and (5.11) presented below, $\overline{R_{f-a}}$ is the average R_{f-a} of all reports with data, and i is the year.

$$\frac{Q_{f_i}}{Q_{a_i}} = R_{f-a\,i} \times \frac{A_{f_i}}{A_{a_i}} \tag{5.10}$$

$$\frac{Q_{f_i}}{Q_{a_i}} = \overline{R_{f-a}} \times \frac{A_{f_i}}{A_{a_i}} \tag{5.11}$$

Austria: The methodology used by Einarsson *et al.* (2021) [312] results in an almost constant percentage of N used for permanent grasslands of ≈10% for the 1960-2019 period. This result led FAO to consider that 10% of fertilizer used in agricultural lands was used for permanent grasslands [293]. The intensification of grassland management began in the 1970s and 1980s [410], and the share used for grasslands was higher in the late 1970s than in the 1990s [304, 315]. For P_2O_5 and K_2O, FAO considered a constant 10% allocation for permanent grasslands [293], based on previous estimations for P_2O_5 [310] and the average value between the fraction used for N for K_2O. While historical

data suggest fluctuations in the percentage of fertilizers used for grasslands and fodder crops over time [304, 305, 315–319], the application of Equation (5.11) using constant $\overline{R_{f-a}}$ values of 0.33 for N, 0.46 for P_2O_5, and 0.32 for K_2O, and surfaces changes [312], provided an MAE of 2.33 ± 3.09%, 3.87 ± 3.47%, 3.31 ± 2.29% respectively. Only two errors higher than 10% occurred, both underestimations, namely -11.8 % for N in 1977 [304], and -10.2% for P_2O_5 for 1990 [315], suggesting higher R_{f-a} during the 1970-1990 period. Based on these results, we decided to utilize the mentioned $\overline{R_{f-a}}$ values for the period from 1991 to 2020 as well as for the period from 1961 to 1969. For the years from 1970 to 1990, we calculated the average R_{f-a} from 1977 and 1990 reports [304, 315] to minimize the errors during the period.

Belgium and Luxembourg: Belgium and Luxembourg often share statistics as a single entity in historical statistics. Consequently, we adopted the same estimation for both countries. According to Einarsson *et al.* (2021) [312], the percentage of fertilizer application rate for permanent grasslands ranged from 53% in the 1980s to 40% in the last years. They deem the N fertilization of permanent grasslands significant throughout the period based on the little available information they found [312]. The same literature confirmed the use P_2O_5, K_2O for grasslands as early as 1955, although with slightly lower applications [312] as in the actual reports. The use of constant R_{f-a} values of 1.03 for N, 0.91 for P_2O_5, and 0.81 for K_2O based on the technical reports values [67, 304, 306–309, 315–320] resulted in MAE values of 2.18 ± 1.82% for N, 5.46 ± 4.04% for P_2O_5, 3.62 ± 2.51% for K_2O. Only two instances of overestimations exceeding 10% were observed for P_2O_5 in the two last reports [67, 309]. This may be linked with the enforcement of limits on P_2O_5 application in the Flanders region since 2011 [411]. Therefore, for P_2O_5 we decided to use the average R_{f-a} for the 1960-2010 period, and use a linear interpolation of the R_{f-a} values since 2011.

Czech Republic, Slovakia, and Czechoslovakia: Information regarding grasslands and fodder crops before the disintegration of the Czechoslovak Republic is very limited [312]. Following assumptions made by Einarsson *et al.* (2021) [312], we extended the average $\overline{R_{f-a}}$ reported for the Czech Republic and Slovakia since 1993 [67, 308, 309, 319] through the period 1960-1992, considering surfaces changes, and the agricultural land of each country [312]. Potential overestimations could occur for the early years, as the fertilization of these areas compared to other croplands might have been lower than in the 1990s, like in neighboring countries such as Hungary or Germany [412, 413]. After 1993, there are four years with available data for both countries [67, 308, 309, 319]. The R_{f-a} values for all years are similar for each nutrient in each country, so we used Equation (5.11) to estimate the 1993-2020 period. This approach resulted in low deviations from the reported values for the Czech Republic (MAE = 2.08 ± 1.58% for N, 2.57 ± 1.30% for P_2O_5, 1.69 ± 1.47% for K_2O) and Slovakia (MAE = 1.49 ± 1.47% for N, 2.02± 2.87% for P_2O_5, 1.79 ± 2.13% for K_2O).

Denmark: Danish grasslands and fodder crop fertilization have a long history with N, with average rates of 45 and 17 kg ha^{-1} for temporary and permanent grasslands respectively in the early 1960s [414]. The usage of Equation (5.11) for the whole period for the three nutrients resulted in large deviations (MAE = 8.89 ± 4.40% for N, 5.36 ± 3.71% for P_2O_5, 8.42 ± 5.67% for K_2O). Therefore, as the amount of available data was large in the compiled technical reports we used Equation (5.10), and linear interpolation of all R_{f-a} values for the period 1980-2020 [67, 304, 306–309, 315–319]. For the 1960-1980 period, we utilized N data from 9 years within that timeframe [414]. Additionally, we considered

the 1980-2020 relationship between N $\overline{R_{f-a}}$ and P_2O_5 or K_2O $\overline{R_{f-a}}$, and the available N data for estimating the 1960-1980 timeframe regarding the P_2O_5 or K_2O values. We regard this assumption as the only available information for the period spanning 1960-1980 for P_2O_5 and K_2O [399] suggests a similar relationship in the application rates for all forages between N and the other nutrients, at least in the reported values since 1980 [67, 304, 306–309, 315–319].

Finland: Einarsson *et al.* (2021) [312] did not consider significant fertilization on permanent grasslands in Finland, as they mainly use arable land for forage production [415]. However, fodder crops and temporary grasslands are key parts of the agricultural production in the country [415], and they are commonly fertilized [67, 304, 306–309, 315–319]. Using Equation (5.11) for the entire period across the three nutrients resulted in minimal deviations for N and P_2O_5 (MAE = 1.57 ± 2.99%, 2.10 ± 3.34% respectively), but substantial deviations for K_2O (7.51 ± 7.38%). Given the substantial deviation for K_2O, and the large bias for R_{f-a} in 1979 [304] for N and P_2O_5, the first year with available data, we opted to use Equation (5.10), and the linear interpolation of the R_{f-a}. However, potential deviations may arise for the 1960s, as fertilizers were predominantly utilized for high-value crops during the early part of the decade [416], yet no data are available for that period.

France: Data regarding grasslands and fodder crop fertilization is less limited than in the majority of EU countries, although large differences exist between the available data. Two recent publications estimated the share of N and P_2O_5 used for permanent grasslands since 1960 [312, 417] based on country surveys at the region-level [418–421]. However, the results obtained by them differ from the FUBC-FAO and FUBC-FE reports [67, 306–309, 315–319]. For example, for 2006, Le Nöe *et al.* (2018) [417] report a share of P_2O_5 used for permanent grasslands of 27% whereas the FE reports a value for all grasslands of 20%. Considering other years with comparable data, such as 1990 or 2017, Einarsson *et al.* (2021) [312] estimate a share of 16% and 7% respectively for N used for permanent grasslands, while FAO only reports 6% for 1990, and the national survey reports 4.7% for 2017 [421]. Therefore, as it is difficult to discern the more accurate value between the two estimations, we opted to use the average between the R_{f-a} linear interpolated data from the global datasets [67, 306–309, 315–319], and from the national surveys [418–421], considering for both as 0 the share in 1955 [417] and the single estimate for the 70s [399].

Germany: The availability of data since the German reunification is substantial in global reports [306–309, 315–319]. These reports suggest a decline since 1990 in fertilizer use for all forages compared to the rest of croplands, with the drop being particularly notable for N and P_2O_5. As a result, we decided to use Equation (5.10), and interpolate the R_{f-a} values, instead of $\overline{R_{f-a}}$ for the 1990-2020 period. For the 1960-1989 period, data on grassland and fodder fertilization is scarce and primarily pertains to West Germany [399]. Most of the data available for the period are relative to N, except the 1982 IFDC-FUBC report. For the 1960-1989 period, We decided to use the linear interpolation assuming, similar to the case of France, zero fertilization of grasslands and fodder crops in 1955, as fertilization of these areas in Western Germany, where most of this agricultural land is located, was minimal before 1960 [413], using the only report with available data for the three nutrients [304]. We extrapolate the data from Western Germany for the entire country due to data availability [304, 312, 399], the prevalence of these agricultural areas in Germany [312], and because grassland fertilization in East Germany was similar to that in West Germany, at least in the late 1970s [422]. Using these approaches, we deviate by

approximately 3.9% from the N estimated data for the year 1974 [399]. Additionally, we deviated by about 10% from the N value for permanent grasslands reported by Einarsson *et al.* (2021) for 1966 (based on real data) [312]. This deviation is reasonable, considering that the average difference between Q_f/Q_a only using information for permanent grasslands or all forages for N is 7.9% [306–309, 315–319].

Greece: Fertilization has not been considered for permanent grasslands in either previous research [293, 311, 312] or technical reports [67, 306–309, 315–319]. However, since we are also considering fertilization for fodder crops, the technical reports have allocated fertilization for them, especially for alfalfa and sillage maize [67, 306–309, 315–319], which constitute the two main actual fodder crops in the country [312]. Therefore, we used Equation (5.10) and the linear interpolation of R_{f-a} because the values of the 1990s are lower than the actual ones, and we have assumed a zero level of fodder fertilization in 1960, as it was only experimental in the country [423].

Hungary: Einarsson *et al.* (2021) [312] did not consider fertilization for permanent grasslands due to the scarcity of the data and because grassland fertilization is not a common practice nowadays [312]. Reported values suggest that a significant fraction, approximately 5% of the fertilizer used since 1990 in the country was allocated to grasslands and fodder crops [67, 307–309, 315], with an even higher proportion during the 1980s [304]. Scientific information confirms that the change in the political regime in 1989 was a key driver of fertilization practices in the country, reducing the fertilizer use by five-fold in the country, and limiting fertilization of these areas to managed grasslands [424]. Furthermore, fertilization in the country commenced in the 1960s and remained stagnant during the 1980s [412]. Therefore, for the period 1960-1989, we applied Equation (5.10), and the linear interpolation of R_{f-a} from a 0 value in 1960, to the 1980 reported value [304]. For the 1990-2020 period, we used Equation (5.11), and the average $\overline{R_{f-a}}$, as there is no deviation larger than 10% from the reported values using this method.

Ireland: Ireland is likely one of the countries that use a larger proportion of fertilizers for grasslands and fodder crops [293, 311, 312], and also has more available information. Since 1972, six national surveys have been conducted, providing data for 22 years [425–430]. Moreover, the global datasets also include information from ten different years since 1987 [67, 306–309, 315–319]. For the 1986-2020 period, we used the average of the linear interpolation of the R_{f-a} values based on national surveys [425–430], and surfaces data [431–435], along with the R_{f-a} values based on the global datasets [67, 306–309, 315–319] and the Einarsson *et al.* (2021) surface compilation [312]. We excluded R_{f-a} values based on the global datasets [67, 306–309, 315–319] and the Einarsson *et al.* (2021) surface compilation [312] for the 2006-2010 period due to a change in the criteria for temporary grassland surface, which resulted in overestimations ($Q_f/Q_a > 1$). For the 1960-1985 period, we only considered the linear interpolation of the available data, all from the national surveys R_{f-a} [425–427], and surfaces [431, 432]. In cases where there was no available surface data [431] in the national databases, like 1972, we used the closest year with available data (e.g., 1970). For 2008, which has two available national surveys [429, 430], we took the average of both. We considered the share of fertilizer used for grasslands and fodder crops as zero in 1955 because almost all fertilizer was used for tillage crops in that year [436], with grassland fertilization increasing during the 1960s [437].

Italy: Einarsson *et al.* (2021) used a constant $\overline{R_{f-a}}$ for permanent grasslands for all years,

as similar values are given in various reports and scientific information [312]. When considering grasslands and fodder crops together, the R_{f-a} were also consistent for each nutrient over all years [67, 306–309, 316–319, 399], even including the 1974 data [399]. The MAE using Equation (5.11) for the entire period across the three nutrients resulted in minimal deviations comparing with the reported values [67, 306–309, 316–319, 399] (MAE = 2.24 ± 1.55% for N, 2.00 ± 1.37% for P_2O_5, and 3.21 ± 1.21% for K_2O). Therefore, we used the the $\overline{R_{f-a}}$ for the three nutrients. However, there could be potential overestimations for the 1960s decade because nearby countries like France or Germany did not use fertilizers for these agricultural lands before 1955 [417].

The Netherlands: Information regarding grassland fertilization in the country is abundant [312, 438]. However, before the development of global datasets, information regarding P_2O_5 and K_2O is very limited. For the period 1979-2019, we used Equation (5.10) considering the linear interpolation of the eleven R_{f-a} data derived from the global datasets [67, 304, 306–309, 317–319] and the agricultural surfaces changes [312]. We used the global datasets instead of the national data available because they provide information regarding the three nutrients. For the years 1960 to 1979, we used the available compilation of N application rates [438], and the total N fertilizer consumption [343] to estimate the Q_f/Q_a values for N. For P_2O_5 and K_2O, we used the ratio between the Q_f/Q_a used for N and these two nutrients for the most recent year with available data, 1979 [304], to extrapolate the results for the 1960-1979 period.

Poland: The available data in reports from the period 1988-2018 [67, 306–309, 315, 318, 319] did not show a constant R_{f-a} for any nutrient N, P_2O_5 and K_2O. Data on fertilization before 1989, during the communist government, is sparse [312, 315]. However, similar to other Eastern European countries like Hungary, it appears that fertilizer intensification in the country started during the 1960s [439], with a significant drop following the regime change [343]. As a result, we adopted the same criteria used for other Eastern European countries, setting the 1960 value to zero, and applying two distinct linear interpolations of R_{f-a}: one for the 1960-1989 period, and another for the 1990-2020 period. For the 1990-2020 period, there are seven years with available data, whereas for the 1960-1989 only 1989 has data. Despite this limited data for the earlier period, survey estimates [439] combined with FAOSTAT totals [343] suggest that the combined share of the three nutrients was between 14% and 15% in the late 1960s, which aligns with the individual nutrient shares calculated by the linear interpolation which are between 10% and 13%.

Portugal: Einarsson *et al.* (2021) did not consider fertilization of permanent grasslands, citing the relatively low surface area in the country [312]. However, recent technical reports suggest that Q_f/Q_a exceeds 20% for the three major nutrients [67, 306–309, 317–319]. We chose to apply Equation (5.10) and to interpolate the 1977-2020 data [67, 304, 306–309, 315, 316, 318, 319] because using Equation (5.10) led to discrepancies greater than 10% in some years. For the years before 1977, we retained the R_{f-a} 1977 values [304] (which resulted in $Q_f/Q_a < 2\%$) as there is no information for the earlier period.

Romania: As with other Eastern European countries, there is no available information regarding grassland and fodder crop fertilization before the political regime change in 1989. However, between 1990 and 2020, data from five years suggest that about 5% of fertilizer is used for grasslands and fodder crops [307–309, 315, 317]. For Romania, we applied Equation (5.11), using the average $\overline{R_{f-a}}$ value and the grassland and cropland surface data [312]. Potential overestimations occurred during the first decades, although

the estimated Q_f/Q_a are less than 5% for the first decades.

Spain: Previous research has not considered the fertilization of permanent grassland because this practice in the country is very uncommon [293, 312]. However, when considering temporary grasslands and fodder crops, this assumption changes, as forage crops occupy about 8% of the arable land in the country and consume nearly the same percentage of fertilizers [440]. To estimate the share of fertilizer use in these areas, we created a linear interpolation of the R_{f-a} data from the ten years with available data, ranging from 1979 to 2014, and applied Equation (5.10). Using Equation (5.11) resulted in estimations that were twice the reported values for the earlier years. Given the fraction used for these areas in 1979 was minimal (Q_f/Q_a < of 2%), potential overestimations for the first years are also likely minimal.

Sweden: In the country, fertilization of forage production areas is closely linked to the transition from natural permanent grassland to temporary grassland production on arable land that occurred during the first part of the 20th century, especially during the 1940s and 1950s [441]. Moreover, based on the available data, fertilizer intensification of these areas compared to other croplands R_{f-a} was lower during the 1970s than at the end of the century [318, 319, 399]. Therefore, we applied Equation (5.10) and performed the linear interpolation of the R_{f-a} of each nutrient of the 11 years with available data since 1974 [67, 305–309, 315, 316, 318, 319, 399]. A slight overestimation might occur for the earlier years, as the intensification of these areas was increasing before the first year with available data [441], but no data for the period was found.

United Kingdom and Northern Ireland (UK): The UK has the world's longest and most complete dataset on the fertilization of grasslands and croplands [334]. Annual time series data on fertilizer use for permanent and temporary grasslands are available for England and Wales since 1969 and for Great Britain since 1982 [334]. Northern Ireland is not included in these surveys. Additionally, there are surveys for the years 1957, 1962, and 1966 for England and Wales [442]. Two problems arise for the estimation of Q_f/Q_a from this data. The first one is that the surveys only include fertilization on permanent and temporary grassland, excluding rough grazing. The second challenge is that there is no information for Northern Ireland - which accounts for about 6% of the country's fertilizer consumption [334]-, and from 1960 to 1982, there is also no data for Scotland, who are responsible for about 14% of the country's fertilizer consumption [334]. For the period 1982-2019, we used the annual fertilizer application rates for Great Britain's tillage crops [334] and the corresponding cropland surface area [443] (excluding temporary grasslands) to estimate the total fertilizer use for croplands. We considered grassland fertilization to be the complement of the value obtained, assuming the same application rates for Northern Ireland. To include these estimations in the fraction used for fodder crops, we add the average share used for them, which is less than the 3% for all nutrients [67, 306–309, 315, 317]. For the period 1960-1981, we applied the same methodology but using the application rates [334, 442] and surfaces [444] from England and Wales, adjusted by -2.5% for N, +2.8% for P_2O_5, and +0.9% for K_2O. These adjustments are based on the observed differences between the application rates in Great Britain and those in England and Wales during the 1980s decade. Moreover, for the 1960s decade for which there are no data available for all years, we applied the linear interpolation of the years with data. We used the national databases instead of the global datasets because they provide annual information covering almost the entire period for the three nutrients, and the values between them were quite similar.

Iceland: Iceland's agriculture sector is primarily focused on livestock production, with about 90% of its agricultural land being permanent grasslands [445]. Additionally, most of the arable land is used for forage crops [445]. While grassland fertilization is a common practice in Iceland [446], there is limited information on application rates for different types of agricultural land, and no specific estimates on the proportion of fertilizer used for forage crops in the country. When we applied Equation (5.11) using the average R_{f-a} from other Nordic countries—Denmark, Sweden, and Finland, it resulted in a Q_f/Q_a ratio greater than 100%. To address this, we allocated a mid-value between 100% and the proportion of agricultural land occupied by grasslands and fodder crops, ensuring it does not exceed 100%.

Switzerland: Data on fodder crop and grassland fertilization in the country from the period 1979-1999 suggest that between 30 and 50% of the fertilizer used in the country is applied to these lands [304, 315–319]. However, whereas the data of the first two years indicate that almost 50% of N is used for grasslands and fodder crops [304, 315], only about 30% was used in 1999 [319]. Since 2000, the areas of artificial grasslands and silage maize (the two main forages that receive fertilizers [315]) have remained almost constant [447]. As there is no information available regarding grassland fertilization before 1979 or after 2000, we used the 1979 data for the period 1960-1979 and the 2000 data for the period 2000-2020. For the period from 1979 to 2000, we applied linear interpolation to the six years with available data [304, 315–319].

Norway: Fodder crops and grasslands (both temporary and permanent) play a key role in the agricultural sector of the country [448, 449]. Technical reports and scientific studies data indicate a nearly constant share of Q_f/Q_a for N, P_2O_5, and K_2O [67, 306–309, 316–319, 399]. Therefore, we used the average of all the available Q_f/Q_a data [67, 306–309, 316–319, 399], covering the period 1974-2018 for N, and from the period 1990-2018 for P_2O_5 and K_2O. The resulting values, with a share of 64.02% ± 1.76% for N, 50.02% ± 2.25% for P_2O_5, and 65.59% ± 6.07% for K_2O, were comparable to those estimated for other Scandinavian countries.

Yugoslav Socialist Federal Republic, and actual former countries: Fodder crops and grasslands played a significant role in the agricultural production of the Yugoslav Social Federal Republic (SFR) [450]. Pastures and meadows occupied 33% of the country's land, while fodder crops took up 20% of the arable land [450]. However, to the best of our knowledge, no information is available regarding fertilization for different agricultural lands before the dissolution of the country. After the dissolution, information became available in global reports for Croatia and Slovenia, but not in the other countries [67, 306–309, 316, 319]. To estimate the Q_f/Q_a values for Yugoslav SFR during the period 1961-1991, we used the weighted average by agricultural land surface [312] of the earliest R_{f-a} values from Croatia and Slovenia [306, 312, 316], given that their R_{f-a} values have changed significantly in recent years [67, 306–309, 316, 319]. We also considered the cropland, grasslands, and fodder crop surfaces of Yugoslavia SFR from the 1990s [450] to estimate the Q_f/Q_a used for the 1961-1991 period. For the period 1990-2019, for actual EU former countries, we performed the linear interpolation of the R_{f-a} values [67, 306–309, 316, 319] to estimate Q_f/Q_a considering the annual surfaces values [312]. In Serbia, the largest country, forage production is a crucial component of its agricultural sector, with about two-fifths of the agricultural land dedicated to this purpose [451]. However, as no specific information on fertilization rates has been found. We considered the average weighted R_{f-a} ratio of Croatia and Slovenia along with the 2004-2008 surfaces of agricultural

lands, grasslands, and fodder crops [451]. For smaller countries like Montenegro of North Macedonia, we assumed the average annual Q_f/Q_a values of Serbia and Croatia.

USSR and Former USSR Countries: Quantitative and qualitative information about fertilization of grassland and fodder crops before the collapse of the USSR is quite scarce [315, 452, 453]. Some publications suggest that the use of fertilizers in these areas was minimal before 1975 [452, 453]. However, data from 1990-1991, just before the collapse, from certain republics (Russia, Latvia, Estonia, or Belarus) indicate that a significant share of fertilizers was used for fodder crops and grasslands [315] (e.g., 40% for N in the Russian Federation [315]). For the period 1960-1991, we estimated the R_{f-a} for the entire USSR in 1990, weighing the value of each republic R_{f-a} [315, 454] in 1990-1991 by the total fertilizer use of each republic [315, 454]. The four republics with available data for this year (Russian Federation, Belarus, Latvia, and Estonia) account for 40% of the agricultural land of the country and 62% of its fertilizer consumption [454]. After estimating R_{f-a} for each nutrient in 1990, we used linear interpolation to estimate the annual R_{f-a} values, considering the value in 1975 as zero [452, 453]. Finally, similar to the EU countries, we considered the annual cropland, grassland, and fodder crop surfaces [454], along with the calculated R_{f-a}, to estimate the annual Q_f/Q_a. For the period from 1992 to 2020, we considered individual country information where some data was available. However, for the following actual countries, there is no information in the global reports [67, 306–309, 315–320]: Armenia, Georgia, Kazakhstan, Kyrgyzstan, Tajikistan, and Turkmenistan. For all these countries, we considered a constant Q_f/Q_a ratio during the 1992-2020 period due to the limited information. For Armenia and Georgia, we assumed the Q_f/Q_a value in 1998 for Azerbaijan, the other Caucasian country [315]. For the Central Asian countries, we used the ratio for grasslands derived from Uzbekistan's 2014 data [321], which is significantly lower than the USSR's share in 1990. This reduction seems reasonable given the significant decrease in fertilizer use, temporary grasslands, and fodder crop surfaces in the region since the USSR collapse [455].

Estonia, Latvia, Lithuania: The Baltic countries are the three former USSR countries with the most available data in global datasets [67, 306–309, 319]. Fertilizer intensification in these areas has changed significantly over the last three decades due to the abandonment of intensively managed areas [312]. This trend is reflected in the changing R_{f-a} values. Therefore, we used Equation (5.10) and the linear interpolation with the six years with available data R_{f-a} from the 1991-2018 period [67, 306–309, 319] to estimate the Q_f/Q_a values since the collapse of the USSR.

Belarus, Moldova, and Ukraine: For these three countries, limited data is available regarding fodder crops and grasslands, but some information can be found in global reports [67, 317, 321]. Thus, for each country, we used the average of the Q_f/Q_a values from the 1992-2020 period. In the case of Belarus, where two sets of data were available for grasslands and one for fodder crops [67, 321], we took the average for grasslands from both reports and the ratio that accounts for the share of grasslands and the share including fodder crops.

Russian Federation: There are three years with available data between 1992 and 2020 [316, 317, 321]. In the first two years, the data showed that an average of approximately 25% of the country's fertilizer was used on grasslands and fodder crops [316, 317]. However, in the latest report from 2014, only about 4% was attributed to these areas (excluding fodder crops not used for hay or silage) [321]. Therefore, we decide to use

the linear interpolation of the Q_f/Q_a values for the years with available data. For the late years, we likely underestimated the value because some fertilizer is used for fodder crops, like fodder beet, that are not intended for silage or hay. However, these fodder crops only accounted for about the 8% of the total fertilizer used for fodder crops and grasslands in 1990 [315].

5.7.1.4 Asia

China: Fertilization of China grasslands remains low at present [321]. Among the compiled reports, only the latest one considers a proportion of the total fertilizer application rate in China, allocating 2% for N, 4% for P_2O_5, and 3% for K_2O. Other information on grassland fertilization in China is scarce, with the few authors that provided some information describing it as sparse [456]. FAO, [293] considers this proportion as 0% for all three nutrients throughout the entire period, which differs from Lassaleta *et al.* (2014), who, based on regional averages, estimated a percentage ranging between 0 and 4.7% from 1960 to 2014. However, any global report or national more detailed information considers any fertilization. We have decided to adopt the same criteria as FAO [293], albeit potentially underestimating values for the last decades.

Iran: Fertilization of Iran's grasslands and fodder crops appears to be minimal, with few reports providing data, and only since 1990, indicating values between 2% and 6% for all three nutrients [315, 316, 321]. Other information is scarce and focused on experimental trials rather than broader country-wide applications. Considering that the first fertilization trials were developed during the 70s, and the first report with data is for 1990 [315], which reported 2% of N and 6% for P_2O_5, we considered as 0% the share for the period 1960-1990, and the average of the reports for the period 1990-2020.

Japan: Since the first report with data, in 1979, almost all reports have underscored the importance of grassland and fodder fertilization in Japan. FAO attributed a constant share of 20% for N, 0% for P_2O_5, and 10% for K_2O for the 1961-2019 period [293]. Conversely, Lassaleta *et al.* (2014) [311] suggested a growing percentage of 20% for N, starting from 0% in 1960, and increasing to 20% in 2009. Although data before 1979 is unavailable, the reported data for N use in 1979 was 15.7%, higher than the 5.2% estimated by Lassaleta *et al.* (2014) [311]. Additionally, due to the lack of data, it is challenging to determine the inception of grassland fertilization in Japan, though it appears to coincide with the transition from semi-natural grasslands to more intensive pasture during the 60s [457]. Therefore, we opted to adhere to FAO's criteria, maintaining the same percentage throughout the period, despite the potential overestimated values for the initial years. We considered the average of all available reports with data [67, 304, 315–319, 321], because FAO criteria appears to underestimate the P_2O_5, and K_2O used for grasslands, resulting in percentages of 17.3% for N, 16.9% for P_2O_5, 15.6% for K_2O.

Korea Republic: Grassland fertilization appears to be a common practice in the country nowadays [458]. However, there is no available data on the fertilization of these areas in global reports [304, 315, 317, 319], nor scientific publications. We used the same assumption as Lassaleta *et al.* (2014), which is to consider the same proportion as in Japan, the geographically and socioeconomically closest country [311]. This assumption also aligns with the observation that the sum of this percentage, and the fertilizer used for the main crops [304, 315, 317, 319] is less than the total for the country [343].

Turkey: Information about fertilization of grasslands and fodder crops in Turkey is scarce, suggesting that it is not a common practice. Lassaleta *et al.* (2014) [311] considered percentages as high as 4.8% for N in 2009, whereas FAO considered 0% for all nutrients. All the available data since 1990 except for 2014 considered some amount of fertilizer used for grasslands, and forages [67, 304, 315–319, 321]. Therefore, we used the average percentage of all reports for the period 1990 - 2020 [67, 304, 315–319, 321].

Other Asian Countries: Cambodia, Indonesia, Malaysia, The Philippines, Thailand, Vietnam, India, and Pakistan: In Asian Southeast countries, only Lassaleta *et al.* (2014) [311] considered that some fertilizer is used on grasslands, based on regional averages used for grasslands and other crops (including fruits, tea, vegetables, and forage and grasslands) [311]. However, no global report [67, 304, 305, 315–319, 321] or country-level sources [459] mentioned fertilizer application to grasslands as significant in these countries. Therefore, we have chosen to align with FAO's criteria, which assumes no fertilizer application rate for grasslands in this region [293]. We applied the same criteria for India and Pakistan, despite previous research considering a certain percentage used for grasslands [294, 311]. The data reports [67, 304, 305, 315–319, 321], the scientific literature [460, 461], and FAO [293] support the idea of non-fertilization of grassland in these two countries.

5.7.1.5 Africa

Egypt: Data regarding grassland and fodder crop fertilization in Egypt are scarce [305, 318]. As is common for many African countries, there is no fertilization of grasslands [462]. However, the few available data about the fertilization of Egyptian clover [305, 318], the main fodder crop in the country [402], suggests that a significant portion of N and P_2O_5 is utilized for fodder production, aligning with country recommendations [463]. Previous research, focused solely on grasslands, has either considered 0% allocation for the three nutrients [293] or a range between 0% and 4% for N [311]. Here, we opted to consider the average of the two reports (1986, 1997) with data [305, 318] for the entire period as Egyptian clover production has been significant since the beginning of the period [464], and the available data is not sufficient to discern any trend.

Morocco: Previous research has indicated various fractions of N fertilizer used for grasslands in the country, ranging from 0% to 11% [293, 311, 462]. With the available information, it is impossible to discern if any application for permanent grasslands occurred in the country, but not for forages such as alfalfa, Egyptian clover, or vetch [465, 466]. Additionally, due to the scarce available data in the reports, discerning any trend is challenging [315, 318, 466], although the presence of improved pastures, usually linked to fertilizer application rate, doubled during the 80s decade [465]. Here, we have opted to use the same percentage, the average of all reports, to estimate the percentage of N, P_2O_5, and K_2O, despite the potential overestimations in the first decades.

South Africa: Fertilization of grasslands and fodder crops such as alfalfa appeared to be significant throughout the study period in South Africa. Both previous scientific research [293, 311] and various technical reports [67, 315–319, 321] indicated percentages ranging 0% and 22.3% for N. For all three nutrients, the share used for grasslands and fodder crops during the 90s was higher than in the last decades [67, 315–319, 321]. This percentage appears to be higher due to larger fertilizer application rates to croplands compared to grasslands and fodder crops [67, 315], and not due to the relationship between cropland

and grassland surface [467]. While information regarding grassland fertilization prior to 1990 is limited, several factors support the hypothesis of early fertilizer application rate for grassland and fodder production. These include the fraction used for grasslands and fodder in 1990 [315], substantial research conducted on improved grasslands since 1920s [468], and the early introduction of alfalfa in 1858 [402] which is a significant consumer of P_2O_5 and K_2O in the country. Given the challenge of identifying any discernible trend and the likelihood of significant consumption at the beginning of the period, we have chosen to adopt the same percentage for the entire duration, aligning with FAO assumptions [293], despite potential slight over- and underestimations throughout the period. The average of all reports [67, 315–319, 321], resulted in percentages of 12.4% for N, 13.3% for P_2O_5, and 9.2% for K_2O.

5.7.2 Figures

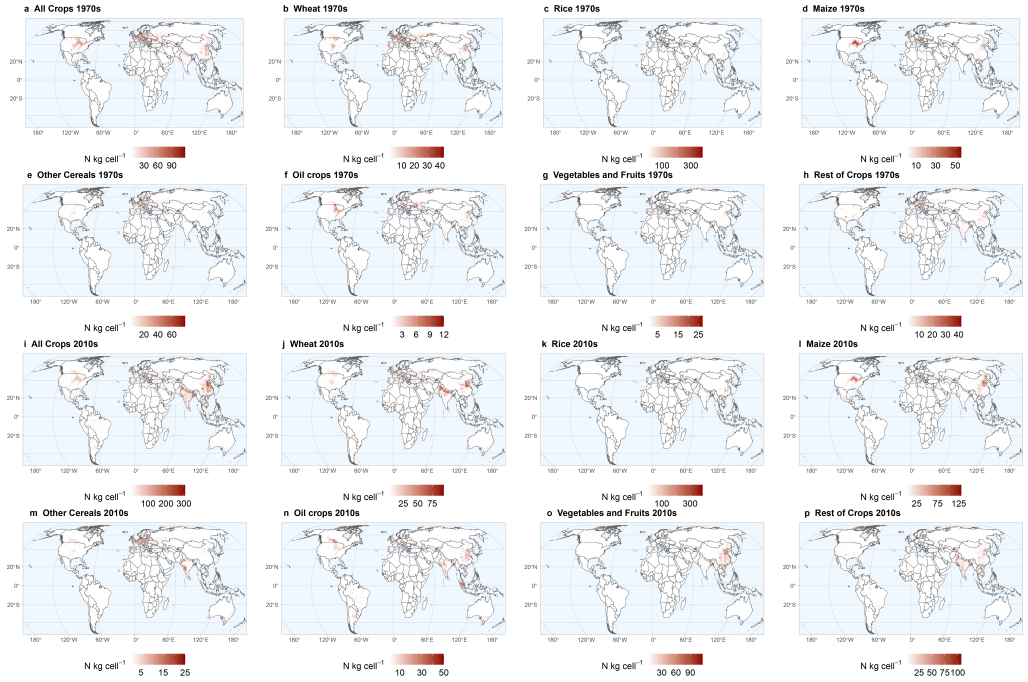


Figure 5.5: Spatial pattern of crop-specific fertilizer (N) consumed by each 0.05° grid cell for the following: a) average for the 1970s decade across all 13 crop classes, b) average for the 1970s decade for wheat, c) average for the 1970s decade for rice, d) average for the 1970s decade for maize, e) average for the 1970s decade for other cereals, f) average for the 1970s decade for all oil crops, g) average for the 1970s decade for vegetables and fruits, h) average for the 1970s decade for roots and tubers, sugar crops, fiber crops, and other crop classes, i) average for the 2010s decade across all 13 crop classes, j) average for the 2010s decade for wheat, k) average for the 2010s decade for rice, l) average for the 2010s decade for maize, m) average for the 2010s decade for other cereals, n) average for the 2010s decade for all oil crops, o) average for the 2010s decade for vegetables and fruits, p) average for the 2010s decade for roots and tubers, sugar crops, fiber crops, and other crop classes.

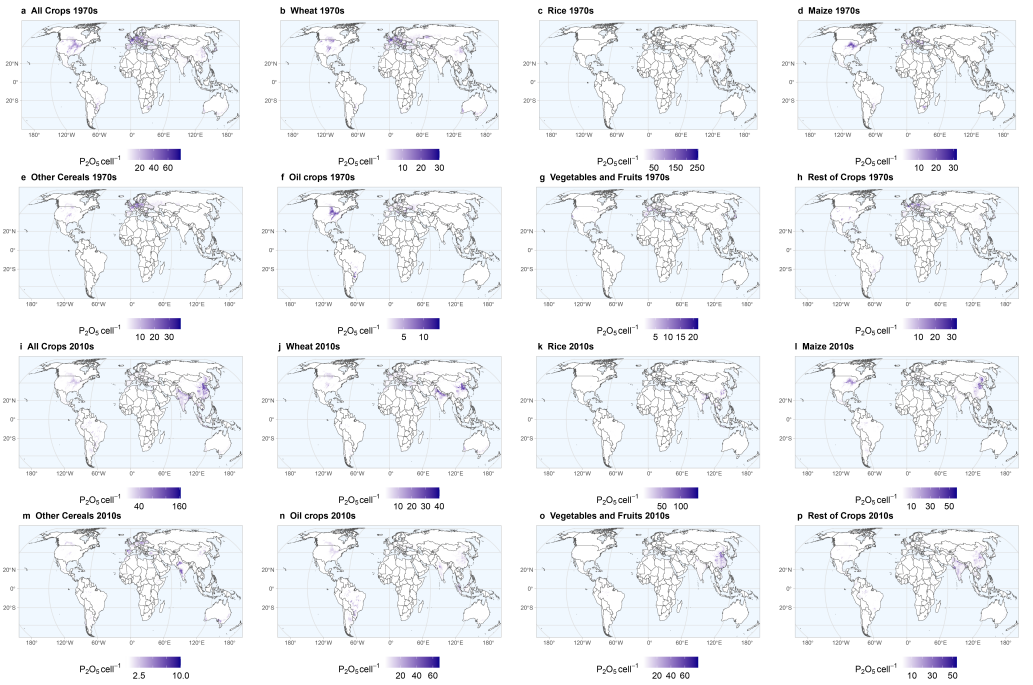


Figure 5.6: Spatial pattern of crop-specific fertilizer (P_2O_5) consumed by each 0.05° grid cell for the following: a) average for the 1970s decade across all 13 crop classes, b) average for the 1970s decade for wheat, c) average for the 1970s decade for rice, d) average for the 1970s decade for maize, e) average for the 1970s decade for other cereals, f) average for the 1970s decade for all oil crops, g) average for the 1970s decade for vegetables and fruits, h) average for the 1970s decade for roots and tubers, sugar crops, fiber crops, and other crop classes, i) average for the 2010s decade across all 13 crop classes, j) average for the 2010s decade for wheat, k) average for the 2010s decade for rice, l) average for the 2010s decade for maize, m) average for the 2010s decade for other cereals, n) average for the 2010s decade for all oil crops, o) average for the 2010s decade for vegetables and fruits, p) average for the 2010s decade for roots and tubers, sugar crops, fiber crops, and other crop classes.

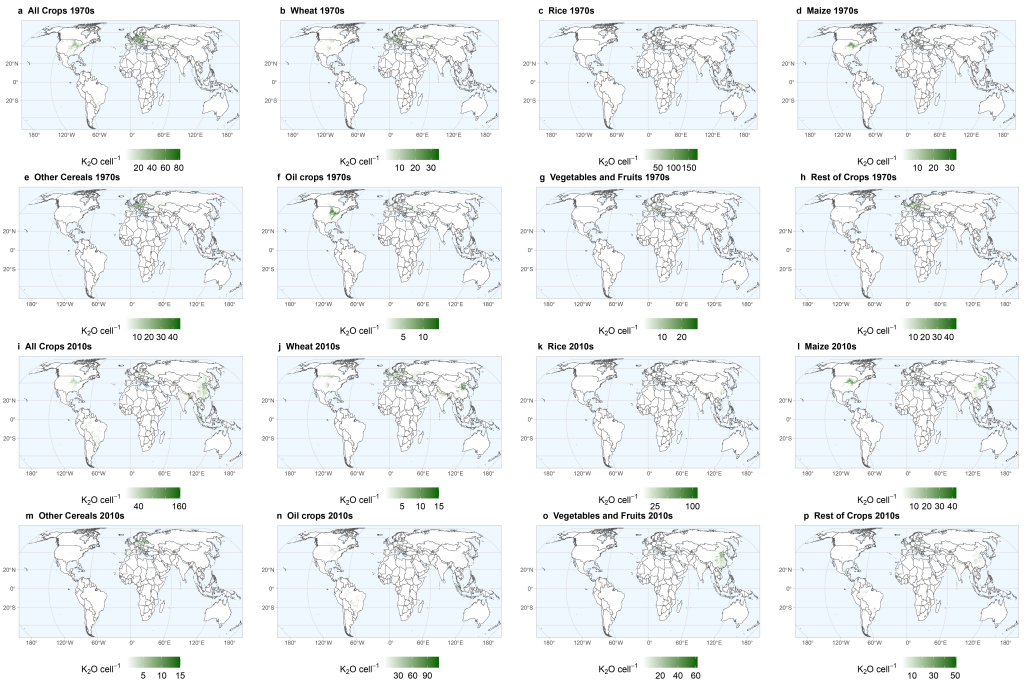


Figure 5.7: Spatial pattern of crop-specific fertilizer (K_2O) consumed by each 0.05° grid cell for the following: a) average for the 1970s decade across all 13 crop classes, b) average for the 1970s decade for wheat, c) average for the 1970s decade for rice, d) average for the 1970s decade for maize, e) average for the 1970s decade for other cereals, f) average for the 1970s decade for all oil crops, g) average for the 1970s decade for vegetables and fruits, h) average for the 1970s decade for roots and tubers, sugar crops, fiber crops, and other crop classes, i) average for the 2010s decade across all 13 crop classes, j) average for the 2010s decade for wheat, k) average for the 2010s decade for rice, l) average for the 2010s decade for maize, m) average for the 2010s decade for other cereals, n) average for the 2010s decade for all oil crops, o) average for the 2010s decade for vegetables and fruits, p) average for the 2010s decade for roots and tubers, sugar crops, fiber crops, and other crop classes.

5.7.3 Tables

Table 5.3: Environmental, agrological and socioeconomic features used in the prediction of the fertilizer application rates, accompanied by their description, unit, and data source. The *Model* column indicates whether the feature was an input for either the N, P_2O_5 or K_2O prediction, or for all 3 predictions.

	Feature	Description	Unit	Model	Source
	Year	Year of the data		All	
	Crop	Crop class		All	
	Country	Code of the country or region in FAOSTAT		All	
	Country surface	Surface of the country	km^2	All	[469]
	Region	World region		All	[470]
Environmental	PET	Annual potential evapotranspiration	mm/year	All	[347]
	MAP	Annual precipitation	mm/year	All	[347]
	TMN	Average annual temperature	° C	All	[347]
	Aridity index	Aridity index		All	[347]
	Soil N	Average soil nitrogen content at 0-30 cm depth	cg/kg	All	[348]
	Soil OCS	Average soil organic carbon stock at 0-30 cm depth	t ha^{-1}	All	[348]
	Soil sand	Average soil sand content at 0-30 cm depth	g/kg	All	[348]
	Soil silt	Average soil silt content at 0-30 cm depth	g/kg	All	[348]
	Soil clay	Average soil clay content at 0-30 cm depth	g/kg	All	[348]
	Soil pH	Average soil pH at 0-30 cm depth		All	[348]
	Soil CEC	Average soil cation exchange capacity at pH 7 at 0-30 cm depth	mmol(c)/kg	All	[348]
	Crop area	Harvested Area of the crop	ha	All	[331]
	Crop area perc	Area of the crop over the total cropland area	%	All	[331]
	Country N per ha	Amount of N fertilizer used per cropland area	t/ha	N	[343]
	Country P_2O_5 per ha	Amount of P_2O_5 fertilizer used per cropland area	t/ha	P_2O_5	[343]
	Country K_2O per ha	Amount of K_2O fertilizer used per cropland area	t/ha	K_2O	[343]
	Country N use	Amount of N fertilizer used in the country	t	N	[343]

Continued on next page

Table 5.3 – continued from previous page

	Feature	Description	Unit	Model	Source
Agrological	Country P_2O_5 use	Amount of P_2O_5 fertilizer used in the country	t	P_2O_5	[343]
	Country K_2O use	Amount of K_2O fertilizer used in the country	t	K_2O	[343]
	Holding size stand	Standardized average size of farms for each country and year	ha	All	[310]
	Crop N content	N content of the crop	kg/t	N	[293]
	Crop P content	P content of the crop	kg/t	P_2O_5	[293]
	Crop K content	K content of the crop	kg/t	K_2O	[293]
	Crop N removal per ha	Average N removal per ha for the crop, country and year	kg/ha	N	[293, 331]
	Crop P removal per ha	Average P removal per ha for the crop, country and year	kg/ha	P_2O_5	[293, 331]
	Crop K removal per ha	Average K removal per ha for the crop, country and year	kg/ha	K_2O	[293, 331]
	Irrigation implementation	Share of agricultural land irrigated in the country	%	All	[471]
	Machinery use	Number of agriculture machinery per ha of arable land for the country and year	ha^{-1}	All	[472, 473]
Socioeconomic	Global urea price	Current urea price per metric tonnes	$ current	N	[357]
	Global P-rock price	Current P price per metric tonnes	$ current	P_2O_5	[357]
	Global K_2O price	Current K_2O price per metric tonnes	$ current	K_2O	[357]
	Global crop price	Real global crop price	$ current	All	[357]
	Education	Fraction of GDP used for education	%	All	[474]
	GDP per capita	Current GDP per capita	$ current	All	[475]
	N cost from production	N fertilizer cost from production		N	[358]
	P cost from production	P_2O_5 fertilizer cost from production		P_2O_5	[358]
	K cost from production	K_2O fertilizer cost from production		K_2O	[358]
	Population pressure	Population per ha of agricultural land	persons/ha	All	[473, 476]
	National crop price	Real price paid to farmer at the country-level	$ current	All	[363, 364]

Table 5.4: Crop Classification with FAOSTAT Item Codes

Crop Class	**Crop Code**	**Description**	**Crops FAOSTAT (FAOSTAT Item Code)**
Wheat	1_1	Wheat	Wheat (15)
Maize	1_2	Maize, only for grain	Maize, corn (56)
Rice	1_3	Rice	Rice (27)
Other Cereals	1_4	Other cereals not mentioned above	Barley (44), Buckwheat (89), Canary seed (101), Fonio (94), Millet (79), Oats (75), Rye (71), Sorghum (83), Triticale (97), Quinoa (92), Cereal n.e.c (108)
Soybean	2_1	Soybean	Soya beans (236)
Palm Oil fruit	2_2	Palm oil fruit	Oil Palm fruit (254)
Other Oilseeds	2_3	Other oilseed crops not soybean and palm oil fruit	Castor oil seeds (265), Coconut, in shell (249), Jojoba seeds (277), Linseed (333), Mustard seed (292), Olives (260), Poppy seeds (296), Rape and colza seed (270), Safflower (280), Sesame seed (289), Sunflower seed (267), Tallowtree seed (305), Tung nuts (275), Other oil seeds, n.e.c (339)
Vegetables	3_1	Vegetables	Artichokes (366), Asparagus (367), Cabbages (358), Cauliflowers and broccoli (393), Chillies and peppers, green (401), Cucumber and gherkins (397), Eggplants (399), Green garlic (406), Leeks and alliaceous (407), Cantaloupes and other melons (568), Melonseed (299), Mushrooms and truffles (449), Okra (430), Onion and shallots, green (402), Onion and shallot, dry (403), Pumpkins, squash and gourds (394), Spinach (373), Tomatoes (388), Watermelons (567), Carrots and turnips (426), Lettuce and chickory (372), Cassava leaves (378), Green corn (446), Other vegetables fresh n.e.c (463)

Continued on next page

Table 5.4 – continued from previous page

Crop Class	Crop Code	Description	Crops FAOSTAT (FAOSTAT Item Code)
Fruits	3_2	Fruits	Apples (515), Apricots (526), Avocados (572), Bananas (486), Blueberries (552), Cherries (531), Sour cherries (530), Cranberries (554), Currants (550), Dates (577), Figs (569), Gooseberries (549), Pomelos and grapefruits (507), Grapes (560), Kiwi fruits (592), Lemos and limes (497), Oranges (490), Papayas (600), Peaches and nectarines (534), Pears (521), Perssimons (587), Pineapples (574), Plantains (489), Plums and sloes (536), Quinces (523), Raspberries (547), Strawberries (544), Tangerines, mandarins, clementines (495), other berries n.e.c (558), other citrus n.e.c. (512), other fruits n.e.c. (619), Other pome fruits n.e.c (542), Other stone fruits n.e.c (541), Other tropical fruits n.e.c (603)
Roots and tubers	4	Roots and tubers	Cassava (125), Potatoes (116), Sweet potatoes (122), Taro (136), Yams (137), Yautia (135), Edible roots and tubers n.e.c. (149)
Sugar crops	5	Sugar cane, sugar beet and only sugar crops	String beans (423), Sugar beet (157), Sugar cane (156), Locust beans (461), Other sugar Crops n.e.c. (161)
Fiber crops	6	Cotton and other fiber crops	Coir (813), True hemp (777), Hempseed (336), Jute (780), Kapok fruit (310), Kapok seed (311), Karite nuts (263), Abaca, manila, hemp (809), Ramie (788), Seed cotton (328), Sisal (789), Agave fibres (800), Flax (773), Kenak (782), Other fibre crops (821)

Continued on next page

Table 5.4 – continued from previous page

Crop Class	Crop Code	Description	Crops FAOSTAT (FAOSTAT Item Code)
Other crops	7	Nuts, pulses, stimulants and aromatics, natural rubber, tobacco	Almonds (221), Areca nuts (226), Cashew nuts (217), Chestnuts (220), Hazelnuts (225), Pistachios (223), Walnut (222), Brazil nuts (216), Kola nuts (224), Other nuts (234), Broad beans and horse beans, dry (181), Broad beans and horse beans, greens (420), Chick peas (191), Cow peas, dry (195), Lentils, dry (201), Lupins (210), Peas, dry (187), Peas, green (417), Pidgeon peas, dry (197), Bambara beans (203), Vetches (205), Other beans, green (414), Other pulses n.e.c (211), Coffee, green (656), Green tea (675), Tea leaves (667), Cocoa beans (661), Chickory roots (459), Mate leaves (671), Other stimulant, spice and aromatic n.e.c (723), Anise, badian, coriander, cumin, caraway, fennel and juniper (711), Cinnamon (693), Cloves (698), Ginger (720), Hop cones (677), Pepper (Piper spp.) (687), Nutmeg, mace, cardamoms (702), Vanilla (692), Chillies and peppers (689), Peppermint (748), Pyrethrum (754), Tobacco (826), Natural rubber (836), Balata, gutta-, percha-, chicle, and similar natural gums (839)

Table 5.7: Fraction of N, P_2O_5, and K_2O allocated for grasslands and fodder crops. The values given are the unique values or the range of values considered for the entire period. The mentioned sources give the information used to calculate these percentages, however, the specific country considerations are pointed throughout the Fertilizer use in other agricultural lands subsection

Country	**N share**	**P_2O_5 share**	**K_2O share**	**Sources**
Argentina	0-9.8	0-28.0	0	[304, 305, 315–319, 386–395]
Canada	12.0	14.5	25.3	[67, 398, 399]
Chile	1.2-22.9	1.5-35.0	1.2-26.9	[67, 304, 305, 315–319, 321, 400]
Dominican Republic	0-3.1	0-3.0	0-2.5	[315, 317, 318, 401]
Mexico	0	2.6	0	[315, 319]
United States of America	6.6-16.6	4.0-17.2	6.8-19.1	[315, 317, 321, 399, 403]
Uruguay	2.0-12.4	21.5-42.9	0	[67, 305, 315, 319, 404]
Australia	6.4	38.4	41.6	[67, 305, 315–319, 321]
New Zealand	91.1	93.0	88.8	[67, 317, 319, 321]
Austria	20.8-31.4	27.1-30.3	19.5-21.8	[304, 305, 312, 315–319]
Belgium and Luxembourg	52.7-66.9	35.5-62.3	41.2-52.3	[67, 304, 306–309, 312, 315–319]
Czech Republic	16.7-19.7	13.6-16.0	13.9-16.3	[67, 308, 309, 312, 319]
Slovakia	10.4-13.6	6.6-8.6	5.9-7.7	[67, 308, 309, 312, 319]
Czechoslovakia	20.8-31.4	27.1-30.3	19.5-21.8	[67, 308, 309, 312, 319]
Denmark	10.0-62.0	10.0-74.0	9.0-61.0	[67, 304, 306–309, 315–319, 414]
Finland	37.0-49.0	22.0-37.0	21.0-64.0	[67, 304, 306–309, 315–319]
France	7.0-39.0	9.0-48.0	12.0-52.0	[67, 306–309, 312, 315–319, 399, 417–421]
Germany	11.0-43.0	10.0-42.0	9.0-39.0	[304, 306–309, 312, 315–319, 399]
Greece	0-10.0	0-13.0	0-10.0	[67, 306–309, 312, 315–319]
Hungary	1.0-20.0	1.0-18.0	1.0-20.0	[67, 304, 307–309, 312, 315]
Ireland	24.0-90.0	20.0-82.0	19.0-83.0	[67, 306–309, 315–319, 425–436]
Italy	9.0-11.0	7.0-8.0	6.0-8.0	[67, 306–309, 312, 316–319, 399]
The Netherlands	52.8-77.6	9.2-58.3	10.5-26.2	[67, 304, 306–309, 312, 317–319, 438]
Poland	1.0-43.0	1.0-40.0	1.0-33.0	[67, 306–309, 312, 315, 318, 319, 439]

Continued on next page

Table 5.7 – continued from previous page

Country	N share	P_2O_5 share	K_2O share	Sources
Portugal	2.0-23.0	3.0-23.0	2.0-29.0	[67, 304, 306–309, 312, 315, 316, 318, 319]
Romania	4.3-5.6	4.2-5.4	2.2-2.8	[307–309, 312, 315, 317]
Spain	4.0-4.6	1.4-12.1	0-7.9	[67, 304, 306–309, 312, 315, 316, 318, 319, 440]
Sweden	12.7-45.1	2.1-36.7	0-7.9	[67, 305–309, 315, 316, 318, 319, 399]
United Kingdom and Northern Ireland	32.4-60.8	17.6-48.8	21.7-39.7	[67, 306–309, 315, 317, 334, 442, 444]
Iceland	97.5	97.5	97.5	[445, 446]
Switzerland	32.7-56.5	36.3-51.0	10.8-38.2	[304, 315–319]
Norway	64.0	50.0	66.0	[67, 306–309, 316–319, 399]
Yugoslav SFR	15.4	16.1	14.9	[306, 312, 312, 316, 450]
Croatia	8.8-22.4	8.7-25.9	8.8-37.8	[67, 306–309, 312, 316, 319]
Montenegro and North Macedonia	10.2-25.1	10.5-28.2	10.4-36.0	[67, 306–309, 312, 316, 319]
Serbia	11.6-27.7	12.2-31.1	12.1-37.6	[67, 306–309, 312, 316, 319, 451]
Slovenia	45.6-70.9	43.2-77.9	35.1-76.5	[67, 306–309, 312, 316, 319]
USSR	0-34.0	0-34.0	0-32.0	[315, 452–454]
Armenia, Georgia and Azerbaijan	4.0	7.0	9.0	[319]
Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan	2.0	2.0	1.5	[321]
Estonia	5.0-40.0	3.0-32.0	21.0-64.0	[67, 306–309, 319]
Latvia	7.0-81.0	6.0-60.0	6.0-65.0	[67, 306–309, 319]
Lithuania	18.0-59.0	16.0-45.0	16.0-63.0	[67, 306–309, 319]
Belarus	27.0	14.0	26.0	[67, 321]
Republic of Moldova	7.0	6.0	3.0	[317]
Ukraine	2.0	1.0	1.0	[67, 321]
Russian Federation	6.5-43.4	1.8-19.4	2.8-33.4	[316, 317, 321]
Islamic Republic of Iran	0-3.2	0-3.7	0-1.1	[315, 316, 321]
Japan and Republic of Korea	17.3	16.9	15.6	[67, 304, 315–319, 321]
Turkey	0-1.2	0-2.4	0-2.1	[67, 304, 315–319, 321]
Egypt	4.0	8.6	1.0	[305, 318]
Morocco	14.8	10.5	6.1	[315, 318, 466]
South Africa	12.4	13.3	9.2	[67, 315–319, 321]

Table 5.5: Validation of our model predictions of the average application rate per ha against national database information for certain countries and crops per fertilizer. The validation is quantified using the mean absolute error (MAE) and mean percentage error (MAPE) per fertilizer between the two data sources, expressed in the table as MAE and MAPE respectively (fertilizer). The NPK stands for the sum of all fertilizers used in the country for certain crops, this is only discussed for Pakistan as more granular data is not available. Unavailable data points are expressed as NA in the table. The sample size of the comparison per country is indicated in parentheses.

Country	**Crop**	**N**		**P_2O_5**		**K_2O**		**NPK**	
		MAE	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
United States of America	Soybean (45)	16.25	79.27	13.00	45.30	13.08	37.11	NA	NA
	Maize (45)	9.31	7.03	7.08	12.84	10.38	19.72	NA	NA
	Wheat (46)	16.20	26.38	15.23	41.01	14.21	62.69	NA	NA
	Fiber crops (43)	14.76	16.55	11.90	27.51	11.87	32.06	NA	NA
United Kingdom	Other Oilseeds (22)	37.20	16.00	3.66	10.36	16.29	32.08	NA	NA
	Wheat (22)	47.87	19.98	2.21	7.09	4.07	9.78	NA	NA
	Sugar crops (22)	72.83	43.49	11.99	29.32	24.85	22.04	NA	NA
	Roots and tubers (22)	46.62	23.20	17.95	17.03	40.51	21.44	NA	NA
India	Other Cereals (6)	11.24	30.43	6.08	34.93	1.59	38.14	NA	NA
	Rice (8)	11.43	12.70	3.54	15.40	2.86	36.29	NA	NA
	Maize (8)	5.58	12.95	5.16	35.56	3.08	59.67	NA	NA
	Wheat (8)	7.56	15.47	4.43	21.64	1.40	36.68	NA	NA
	Fiber crops (6)	15.91	15.97	12.63	49.14	4.06	91.06	NA	NA
	Sugar crops (8)	13.59	10.57	13.36	25.17	12.34	30.65	NA	NA
Sweden	Maize (4)	176.04	67.81	50.75	68.76	52.41	55.16	NA	NA
	Wheat (4)	54.80	29.05	11.15	48.54	24.17	108.43	NA	NA
	Sugar crops (4)	66.60	39.80	22.16	50.53	51.88	49.24	NA	NA
Philippines	Rice (13)	8.34	18.91	5.01	75.49	NA	NA	NA	NA
	Maize (12)	33.96	89.01	5.66	64.65	NA	NA	NA	NA
New Zealand	Fruits (5)	71.92	61.17	80.43	81.74	NA	NA	NA	NA
	Vegetables (5)	105.43	55.06	141.94	87.28	NA	NA	NA	NA
Pakistan	Rice (23)	NA	NA	NA	NA	NA	NA	66.06	41.12
	Wheat (23)	NA	NA	NA	NA	NA	NA	31.18	18.34
	Fiber crops (23)	NA	NA	NA	NA	NA	NA	71.25	29.69

Table 5.6: Comparison of the data reported by global datasets [67, 304–309, 316–321] of the average application rate per fertilizer per ha against national database information for certain countries and crops per fertilizer. The validation is quantified using the mean absolute error (MAE) and mean percentage error (MAPE) per fertilizer between the two data sources, expressed in the table as MAE and MAPE respectively (fertilizer). The NPK stands for the sum of all fertilizers used in the country for certain crops, this is only discussed for Pakistan as more granular data is not available. Unavailable data points are expressed as NA in the table.

Country	**Crop**	**N**		**P_2O_5**		**K_2O**		**NPK**	
		MAE	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
United States of America	Soybean	3.04	25.37	4.94	17.33	7.78	18.98	NA	NA
	Maize	10.45	6.38	8.41	12.36	13.69	45.07	NA	NA
	Wheat	6.04	8.96	5.34	18.20	9.65	66.89	NA	NA
	Fiber crops	21.51	20.80	8.74	24.49	11.76	32.04	NA	NA
United Kingdom	Other Oilseeds	12.42	7.31	3.41	10.55	3.01	7.85	NA	NA
	Wheat	5.80	3.04	2.00	6.67	1.20	3.32	NA	NA
	Sugar crops	3.80	3.93	3.20	9.45	9.20	10.03	NA	NA
	Roots and tubers	11.20	7.97	6.60	5.77	12.40	5.74	NA	NA
India	Other Cereals	0.30	1.03	2.49	18.89	1.01	36.54	NA	NA
	Rice	12.30	13.05	1.81	5.73	0.78	4.31	NA	NA
	Maize	19.12	43.84	13.13	124.63	0.57	19.34	NA	NA
	Wheat	3.71	3.40	1.05	2.51	0.29	4.16	NA	NA
	Fiber crops	26.52	29.83	5.70	12.94	4.28	33.55	NA	NA
	Sugar crops	14.93	9.11	2.97	5.01	26.03	46.83	NA	NA
Philippines	Rice	8.30	16.67	6.23	76.49	NA	NA	NA	NA
	Maize	31.32	69.56	6.60	78.31	NA	NA	NA	NA
Pakistan	Rice	NA	NA	NA	NA	NA	NA	44.64	27.54
	Wheat	NA	NA	NA	NA	NA	NA	29.97	18.68
	Fiber crops	NA	NA	NA	NA	NA	NA	54.30	23.83

Table 5.8: Overview of the used open source packages and respective programming language in the code for both model training, SHapley Additive exPlanations (SHAP) value computation and validation, as well as map building.

Programming language	**Package**	**Version**
Python	Python [145]	3.10.3
Python	numpy [273]	1.23.2
Python	pandas [477]	1.4.1
Python	rasterio [478]	1.3.9
Python	scikit-learn [133]	1.3.2
Python	shap [59]	0.44.0
Python	xgboost [190]	2.0.3
R	R [479]	4.2.2
R	sf [480, 481]	1.0-15
R	ncdf4 [482]	1.22
R	exactextractr [483]	0.10.0
R	readxl [484]	1.4.3
R	stringr [485]	1.5.0
R	dplyr [486]	1.1.2
R	readr [487]	2.1.4
R	ggplot2 [488]	3.4.2
R	tidyverse [489]	2.0.0
R	cshapes [490]	2.0
R	terra [359]	1.7-65

Chapter 6

Conclusions and future perspectives

In general, our findings emphasize the significant value of ML in biological sciences. Specifically, we demonstrated that ML models can assist in analyzing complex datasets by revealing patterns and relationships that were previously hidden. The rapid growth of computational power, the increasing volume of biological data, and swift advancements in ML research offer tremendous potential for its applications in this field. This final chapter summarizes the key contributions of this thesis and suggests potential directions for future research.

In Chapter 2, we introduced the WithMe dataset, and showed that it can be used to develop ML methods for detecting attention based on P3a ERPs in single-trial data. In this study, we successfully classified target and distractor stimuli using four different classification methods: xDAWN+RG, EEGNet, MiniRocket, and Rocket. EEGNet and xDAWN+RG achieved a 76% accuracy in the IS case, with EEGNet maintaining the same accuracy as in the CS case, while xDAWN+RG's accuracy fell to 73%. EEGNet's more complex structure likely enabled better generalization across subjects. Importantly, all four models can provide real-time predictions, which is essential for human-AI interaction experiments and applications. Using xAI, we found that the EEGNet model made its predictions based on EEG signals from the parietal-occipital region between 200-300ms post-stimulus, aligning with the hypotheses of domain experts. Future work could address the fraction of mislabeled data by designing an experiment solely focused on attention detection, enabling models to align labels more accurately with subjects' perceived stimuli. Additionally, it could be interesting to include a feature selection procedure, allowing models to focus on the most relevant features. Finally, as EEG data is notoriously subject-dependent, exploring other ways to achieve CS generalization could be interesting [109]. One promising strategy is the application of transfer learning [142, 143].

In Chapter 3, our research only partially confirmed the proposed hypotheses regarding the impact of soil temperature on the SOS, POS and PEAK. While we confirmed that higher soil temperatures lead to an earlier SOS, we also observed a similar shift in the POS, which was unexpected. Additionally, the PEAK increased slightly with increasing soil temperatures, which also deviated from expectations. Our analysis also confirmed that meteorological variables had the most significant impact on all three seasonal characteristics. Contrary to our hypothesis, the influence of these variables was relatively equal. Finally, annual variations in phenological characteristics were primarily driven

by soil temperature. Future research should prioritize further exploration of the relationships between soil temperature and phenology, along with the interactions between soil temperature and meteorological variables at finer temporal and spatial scales. Additionally, integrating advanced remote sensing techniques, such as satellite imagery, with ground-based measurements, has significant potential to enhance the precision of phenological studies in subarctic grasslands. From a methodological perspective, considering alternative ML model architectures like XGBoost [190] and other xAI methods like LIME [113] could be interesting, as this enables comparisons between different ML models and xAI methods.

In Chapter 4, we successfully tackled the important issue of missing values in sensor data. We started by introducing an extensive real-world dataset collected using a large-scale WSN for environmental monitoring. Current literature typically examines datasets containing a few dozen sensors, whereas our dataset comprises 1500 sensors. This feature allows us to evaluate missing value imputation methods for large-scale WSNs more reliably. Additionally, we expanded upon the existing literature by introducing "masked missings", which provide a far more realistic evaluation scenario compared to the common practice of using random missings. After thoroughly evaluating both random and masked missing scenarios, we found that the MC method consistently outperformed the others. In general, methods that specifically account for spatial correlations tended to perform better. Following this work, there are many directions for future research. From an evaluation perspective, future research can build upon our study by exploring additional datasets with varying numbers of sensors and temporal granularities, which would enhance our understanding of how dataset characteristics influence method performance. On a methodological note, we expect great value in the development of WSN-specific imputation methods. For example, one could consider a two-step approach, where the first step is focused on temporal correlations and the second step on spatial correlations. More specifically, one could first apply the iterative imputing network as proposed by [491]. The benefit of the iterative component is twofold: it enables the method to deal with sparse data, which (trivially) is often the case in a missing value problem. Additionally, iterative imputation methods are more capable of dealing with clustered missings (so-called missing blocks). In the next step, one could exploit spatial correlations by taking inspiration from methods such as [244], which assumes that the relationship between two spatially close sensors can be estimated by a linear relationship, and accordingly fits a linear regression model between two sensors, which is ultimately used to make the imputations. A possible improvement to this proposed method could be using more complex ML models such as XGBoost [190] or MLPs to model the relationships between (multiple) nearby sensors. Finally, creating a specialized open-source Python package for missing value imputation methods would be highly valuable. The leading package for general missing value imputation, fancyimpute [281], has been in bare maintenance mode for three years. Meanwhile, Autoimpute [492] primarily focuses on simpler imputation techniques, and ImputeBench [259] is dedicated solely to time series imputation.

Finally, in Chapter 5, we compiled a comprehensive dataset of global N, P_2O_5, and K_2O application rates by major crop groups from 1961 to 2019. We used ML, specifically gradient boosted decision trees, to estimate fertilizer application rates. These estimates were made based on a collection of crop class related data, historical fertilizer application rates, as well as various socioeconomic, environmental and agricultural variables that were identified as possible drivers of cropland fertilization. We corrected the estimates by considering the proportion of fertilizer use allocated to grasslands and fodder crops,

ensuring they more accurately reflect the total annual country-level N, P_2O_5, and K_2O application rates use in agricultural land. Ultimately, we provide a tabular dataset of the country- and crop-level estimates, as well as gridded maps with a resolution of 5 arc minutes, both of which are published as open-source resources. Our results present a clear improvement in efforts to evaluate historical fertilizer consumption for different crop groups, countries, and fertilizers. Future research can build upon our study by for example considering the frequency of fertilizer application, as well as the associated timings of the application within a year. Additionally, while our study focuses on the application of broad fertilizer classes, more detailed estimates can be made for the different types of fertilizers, e.g., different N fertilizer types. The data published in this study also opens up a lot of possible directions for future research. For example, a thorough analysis of the drivers of fertilizer application rate could deepen our understanding, and also inform future decisions made by policy-makers. While we use xAI to get insights regarding these drivers, exploring methodologies such as causal structure discovery could provide us with a more comprehensive and causally accurate understanding of the main drivers of fertilizer application rate.

Bibliography

[1] Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. *Journal of Global Health*, 8:20303, 2018. 1

[2] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021. 1

[3] Vasant Dhar. Data science and prediction. *Commun. ACM*, 56(12):64–73, dec 2013. 1

[4] Kirill Dubovikov. *Managing Data Science: Effective strategies to manage data science projects and build a sustainable team*. Packt Publishing Ltd, 2019. xv, 2

[5] Alan M. Turing. Computing machinery and intelligence. 1950. 2

[6] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. 2

[7] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 11 1958. 2

[8] Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 479(480):104, 1969. 2

[9] Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160, 1976. 2

[10] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Classification and regression trees. *Classification and Regression Trees*, pages 1–358, 1 1984. 2

[11] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. Support-vector networks. *Machine Learning 1995 20:3*, 20:273–297, 9 1995. 2

[12] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 4 1980. 2

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[14] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature 2016 529:7587*, 529:484–489, 1 2016. 2

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017. 3

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 3

[17] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei Openai. Language models are few-shot learners. 2020. 3

[18] OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed on 09-08-2024. 3

[19] Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, Emidio Capriotti, Rita Casadio, Salvador Capella-Gutierrez, Davide Cirillo, Alessio Del Conte, Alexandros C. Dimopoulos, Victoria Dominguez Del Angel, Joaquin Dopazo, Piero Fariselli, José Maria Fernández, Florian Huber, Anna Kreshuk, Tom Lenaerts, Pier Luigi Martelli, Arcadi Navarro, Pilib Broin, Janet Piñero, Damiano Piovesan, Martin Reczko, Francesco Ronzano, Venkata Satagopam, Castrense Savojardo, Vojtech Spiwok, Marco Antonio Tangaro, Giacomo Tartari, David Salgado, Alfonso Valencia, Federico Zambelli, Jennifer Harrow, Fotis E. Psomopoulos, and Silvio C.E. Tosatto. Dome: recommendations for supervised machine learning validation in biology. *Nature Methods 2021 18:10*, 18:1122–1127, 7 2021. xv, 3

[20] Tanzila Saba. Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. *Journal of Infection and Public Health*, 13(9):1274–1289, 2020. 4

[21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature 2021 596:7873*, 596:583–589, 7 2021. 4

[22] Rizwan Qureshi, Muhammad Irfan, Taimoor Muzaffar Gondal, Sheheryar Khan, Jia Wu, Muhammad Usman Hadi, John Heymach, Xiuning Le, Hong Yan, and Tanvir Alam. Ai in drug discovery and its clinical relevance. *Heliyon*, 2023. 4

[23] Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of cheminformatics*, 10:1–24, 2018. 4

[24] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024. 4

[25] Larry Squire, Darwin Berg, Floyd E Bloom, Sascha Du Lac, Anirvan Ghosh, and Nicholas C Spitzer. *Fundamental neuroscience*. Academic press, 2012. 4

[26] U Rajendra Acharya, Yuki Hagiwara, and Hojjat Adeli. Automated seizure prediction. *Epilepsy & Behavior*, 88:251–261, 2018. 4

[27] Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, and Dagan Feng. Early diagnosis of alzheimer's disease with deep learning. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 1015–1018. IEEE, 2014. 4

[28] Muhammad Tanveer, Bharat Richhariya, Riyaj Uddin Khan, Ashraf Haroon Rashid, Pritee Khanna, Mukesh Prasad, and Chin-Teng Lin. Machine learning techniques for the diagnosis of alzheimer's disease: A review. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–35, 2020. 4

[29] Dhruv Chandra Lohani and Bharti Rana. Adhd diagnosis using structural brain mri and personal characteristic data with machine learning framework. *Psychiatry Research: Neuroimaging*, 334:111689, 2023. 4

[30] Ritu Gautam and Manik Sharma. Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis. *Journal of medical systems*, 44(2):49, 2020. 4

[31] Violaine Guy, Marie Hélène Soriani, Mariane Bruno, Théodore Papadopoulo, Claude Desnuelle, and Maureen Clerc. Brain computer interface with the P300 speller: Usability for disabled people with amyotrophic lateral sclerosis. *Annals of Physical and Rehabilitation Medicine*, 61(1):5–11, 1 2018. 4, 12

[32] Matthias Kaper, Peter Meinicke, Ulf Grossekathoefer, Thomas Lingner, and Helge Ritter. BCI competition 2003 - Data set IIb: Support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, 51(6):1073–1076, 6 2004.

[33] Yoji Okahara, Kouji Takano, Tetsuo Komori, Masahiro Nagao, Yasuo Iwadate, and Kenji Kansaku. Operation of a P300-based brain-computer interface by patients with spinocerebellar ataxia. *Clinical Neurophysiology Practice*, 2:147–153, 1 2017. 4, 12

[34] Eda Akman Aydin, Omer Faruk Bay, and Inan Guler. P300-based asynchronous brain computer interface for environmental control system. *IEEE Journal of Biomedical and Health Informatics*, 22(3):653–663, 5 2018. 4, 12

[35] Usman Masud, Muhammad Iram Baig, Faraz Akram, and Tae Seong Kim. A P300 brain computer interface based intelligent home control system using a random forest classifier. *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, 2018-January:1–5, 2 2018. 4, 12

[36] Shiro Ikegami, Kouji Takano, Kiyohiko Kondo, Naokatsu Saeki, and Kenji Kansaku. A region-based two-step P300-based brain–computer interface for patients with amyotrophic lateral sclerosis. *Clinical Neurophysiology*, 125(11):2305–2312, 11 2014. 4, 12

[37] Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a P300-based brain–computer interface (BCI)? *Neuroscience Letters*, 462(1):94–98, 9 2009. 4, 12, 21

[38] Jana Wäldchen and Patrick Mäder. Plant species identification using computer vision techniques: A systematic literature review. *Archives of computational methods in engineering*, 25:507–543, 2018. 4

[39] Jana Wäldchen and Patrick Mäder. Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11):2216–2225, 11 2018. 29

[40] Amiza Amir, Nik Adilah Hanin Zahri, Naimah Yaakob, and R Badlishah Ahmad. Image classification for snake species using machine learning techniques. In *Computational Intelligence in Information Systems: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2016)*, pages 52–59. Springer, 2017. 4

[41] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55:96, 2 2022. 4

[42] Amir Mosavi, Mohsen Salimi, Sina Faizollahzadeh Ardabili, Timon Rabczuk, Shahaboddin Shamshirband, and Annamaria R Varkonyi-Koczy. State of the art of machine learning models in energy systems, a systematic review. *Energies*, 12(7):1301, 2019. 4

[43] Yongliang Yan, Tohid N. Borhani, Sai Gokul Subraveti, Kasturi Nagesh Pai, Vinay Prasad, Arvind Rajendran, Paula Nkulikiyinka, Jude Odianosen Asibor, Zhien Zhang, Ding Shao, Lijuan Wang, Wenbiao Zhang, Yong Yan, William Ampomah, Junyu You, Meihong Wang, Edward J. Anthony, Vasilije Manovic, and Peter T. Clough. Harnessing the power of machine learning for carbon capture, utilisation, and storage (ccus) – a state-of-the-art review. *Energy & Environmental Science*, 14:6122–6157, 12 2021. 4

[44] David John Gagne, Amy McGovern, Sue Ellen Haupt, Ryan A Sobash, John K Williams, and Ming Xue. Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and forecasting*, 32(5):1819–1840, 2017. 4

[45] Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, Ashesh Chattopadhyay, A Singh, et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093, 2021. 4

[46] Abhinav Sharma, Arpit Jain, Prateek Gupta, and Vinay Chowdary. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873, 2020. 4

[47] Max Grell, Giandrin Barandun, Tarek Asfour, Michael Kasimatis, Alex Silva Pinto Collins, Jieni Wang, and Firat Güder. Point-of-use sensors and machine learning enable low-cost determination of soil nitrogen. *Nature Food 2022 2:12*, 2(12):981–989, 12 2021. 4, 78, 79

[48] Liujun Xiao, Guocheng Wang, Enli Wang, Shengli Liu, Jinfeng Chang, Ping Zhang, Hangxin Zhou, Yuchen Wei, Haoyu Zhang, Yan Zhu, Zhou Shi, and Zhongkui Luo. Spatiotemporal co-optimization of agricultural management practices towards climate-smart crop production. *Nature Food 2024 5:1*, 5(1):59–71, 1 2024. 4, 78, 79

[49] Karl Friederichs. A definition of ecology and some thoughts about basic concepts. *Ecology*, 39(1):154–159, 1958. 4

[50] Di Chen, Yexiang Xue, Daniel Fink, Shuo Chen, and Carla P Gomes. Deep multispecies embedding. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3639–3646, 2017. 4

[51] Sylvain Christin, Eric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 10 2019. 4, 29, 78

[52] YoonKyung Cha, Jihoon Shin, ByeongGeon Go, Dae-Seong Lee, YoungWoo Kim, TaeHo Kim, and Young-Seuk Park. An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *Journal of Environmental Management*, 291:112719, 2021. 4

[53] Pramaditya Wicaksono, Prama Ardha Aryaguna, and Wahyu Lazuardi. Benthic habitat mapping model and cross validation using machine-learning classification algorithms. *Remote Sensing*, 11(11):1279, 2019. 4

[54] Simon Vollert, Martin Atzmueller, and Andreas Theissler. Interpretable machine learning: A brief survey from the predictive maintenance perspective. In *2021 26th IEEE international conference on emerging technologies and factory automation (ETFA)*, pages 01–08. IEEE, 2021. 5

[55] Pawan Kumar and Manmohan Sharma. Data, machine learning, and human domain experts: None is better than their collaboration. *International Journal of Human–Computer Interaction*, 38(14):1307–1320, 2022. 5

[56] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019. 5

[57] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020. 5

[58] Laura Moss, David Corsar, Martin Shaw, Ian Piper, and Christopher Hawthorne. Demystifying the black box: the importance of interpretability of predictive models in neurocritical care. *Neurocritical care*, 37(Suppl 2):185–191, 2022. 5

[59] Scott M Lundberg, Paul G Allen, and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017. 5, 29, 36, 72, 78, 80, 118

[60] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022. 5

[61] A Saranya and R Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, page 100230, 2023. 5

[62] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. 6

[63] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666, 2022. 6

[64] Steven Mortier, Renata Turkeš, Jorg De Winne, Wannes Van Ransbeeck, Dick Botteldooren, Paul Devos, Steven Latré, Marc Leman, and Tim Verdonck. Classification of targets and distractors in an audiovisual attention task based on electroencephalography. *Sensors 2023, Vol. 23, Page 9588*, 23:9588, 12 2023. 7

[65] Steven Mortier, Amir Hamedpour, Bart Bussmann, Ruth Phoebe Tchana Wandji, Steven Latré, Bjarni D. Sigurdsson, Tom De Schepper, and Tim Verdonck. Inferring the relationship between soil temperature and the normalized difference vegetation index with machine learning. *Ecological Informatics*, 82:102730, 2024. 7

[66] Thomas Decorte, Steven Mortier, Jonas J Lembrechts, Filip J R Meysman, Steven Latré, Erik Mannens, and Tim Verdonck. Missing value imputation of wireless sensor data for environmental monitoring. *Sensors 2024, Vol. 24, Page 2416*, 24:2416, 4 2024. 8

[67] Cameron I. Ludemann, Armelle Gruere, Patrick Heffer, and Achim Dobermann. Global data on fertilizer use by crop and by country. *Scientific Data*, 9(1):502, 2022. xxi, 8, 72, 73, 74, 75, 90, 92, 93, 94, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 114, 115, 117

[68] Fernando Coello, Thomas Decorte, Iris Janssens, Steven Mortier, Jordi Sardans, Josep Peñuelas, and Tim Verdonck. Global crop-specific fertilization dataset from 1961-2019. *arXiv preprint arXiv:2406.10001*, 2024. 8

[69] FET-BAM. `https://www.fet-bam.eu/`. [ONLINE]. Accessed on 06-05-2024. 9

[70] Tullia Calogiuri, Mathilde Hagens, Jan Willem Van Groenigen, Thomas Corbett, Jens Hartmann, Rick Hendriksen, Iris Janssens, Ivan A. Janssens, Guillermo Ledesma Dominguez, Grant Loescher, Steven Mortier, Anna Neubeck, Harun Niron, Reinaldy P. Poetra, Lukas Rieder, Eric Struyf, Michiel Van Tendeloo, Tom De Schepper, Tim Verdonck, Siegfried E. Vlaeminck, Sara Vicca, and Alix Vidal. Design and construction of an experimental setup to enhance mineral weathering through the activity of soil organisms. *JoVE (Journal of Visualized Experiments)*, 2023:e65563, 11 2023. 9

[71] The PRINCE Out-of-Distribution Challenge | by Thibaud Rahier | Criteo R&D Blog | Medium. `https://medium.com/criteo-engineering/the-prince-out-of-distribution-challenge-1aad96cbecd3`. [ONLINE]. Accessed on 06-05-2024. 10

[72] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 10

[73] Jorg De Winne, Paul Devos, Marc Leman, and Dick Botteldooren. With No Attention Specifically Directed to It, Rhythmic Sound Does Not Automatically Facilitate Visual Task Performance. *Frontiers in Psychology*, 13:3221, 6 2022. 11, 14, 23, 26

[74] Erik Van der Burg, Christian N.L. Olivers, Adelbert W. Bronkhorst, and Jan Theeuwes. Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053–1065, 10 2008. 11

[75] Steven J. Luck, Geoffrey F. Woodman, and Edward K. Vogel. Event-related potential studies of attention. *Trends in Cognitive Sciences*, 4(11):432–440, 11 2000. 12

[76] Geoffrey F. Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics 2010 72:8*, 72(8):2031–2046, 2010. 12

[77] R Näätänen. Selective attention and evoked potentials in humans — a critical review. *Biological Psychology*, 2(4):237–307, 1975. 12

[78] R Näätänen, AW Gaillard, and S Mäntysalo. Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, 42(4):313–329, 1978.

[79] Risto Näätänen. The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behavioral and brain sciences*, 13(2):201–233, 1990.

[80] Connie C Duncan, Robert J Barry, John F Connolly, Catherine Fischer, Patricia T Michie, Risto Näätänen, John Polich, Ivar Reinvang, and Cyma Van Petten. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, p300, and n400. *Clinical Neurophysiology*, 120(11):1883–1908, 2009. 12

[81] R Johnson. The amplitude of the P300 component of the event-related potential: Review and synthesis. *Advances in psychophysiology*, 3(April):69–137, 1988.

[82] Heather M. Gray, Nalini Ambady, William T. Lowenthal, and Patricia Deldin. P300 as an index of attention to self-relevant stimuli. *Journal of Experimental Social Psychology*, 40(2):216–224, 3 2004.

[83] Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7(NOV):732, 11 2013. 12

[84] Christian Scharinger, Alexander Soutschek, Torsten Schubert, and Peter Gerjets. Comparison of the working memory load in N-back and working memory span tasks by means of EEG frequency band power and P300 amplitude. *Frontiers in Human Neuroscience*, 11:212526, 1 2017. 12

[85] Terence W Picton and others. The P300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9:456, 1992.

[86] Samuel Sutton, Magery Braren, Joseph Zubin, and E. R. John. Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, 150(3700):1187–1188, 11 1965. 12

[87] John Polich. Updating P300: An Integrative Theory of P3a and P3b. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 118(10):2128, 10 2007. 12

[88] John Polich. *Detection of change: event-related potential and fMRI findings*. Springer, 2003. 12, 16

[89] E. Donchin. Surprise!... Surprise? *Psychophysiology*, 18(5):493–513, 9 1981. 12

[90] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain Computer Interfaces, a Review. *Sensors 2012, Vol. 12, Pages 1211-1279*, 12(2):1211–1279, 1 2012. 12

[91] Joseph N. Mak and Jonathan R. Wolpaw. Clinical Applications of Brain—Computer Interfaces: Current State and Future Prospects. *IEEE Reviews in Biomedical Engineering*, 2:187–199, 2009. 12

[92] J. N. Mak, Y. Arbel, J. W. Minett, L. M. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, and D. Erdogmus. Optimizing the P300-based brain–computer interface: current status, limitations and future directions. *Journal of Neural Engineering*, 8(2):025003, 3 2011. 12

[93] Swati Aggarwal and Nupur Chugh. Review of Machine Learning Techniques for EEG Based Brain Computer Interface. *Archives of Computational Methods in Engineering*, 1:1–20, 1 2022. 12

[94] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain-Computer Interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009. 12, 17, 18

[95] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *http://dx.doi.org/10.1080/2326263X.2017.1297192*, 4(3):155–174, 7 2017. 17

[96] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 4 2012. 17

[97] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, 112:172–178, 2013. 12, 17

[98] Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):1–30, 2018. 12, 13, 18

[99] Arnaldo E Pereira, Dereck Padden, Jay J Jantz, Kate Lin, and Ramses E Alcaide-Aguirre. Cross-Subject EEG Event-Related Potential Classification for Brain-Computer Interfaces Using Residual Networks. *Preprint*, 9 2018.

[100] Hongfei Zhang, Zehui Wang, Yinhu Yu, Haojun Yin, Chuangquan Chen, and Hongtao Wang. An improved EEGNet for single-trial EEG classification in rapid serial visual presentation task. *Brain Science Advances*, 2022(2):111–126, 2022. 12

[101] Foroogh Shamsi, Ali Haddad, Laleh Najafizadeh, Boyu Zang, Yanfei Lin, Zhiwen Liu, al , Bashar Awwad Shiekh Hasan, and John Q Gan. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001, 4 2019. 12

[102] J. P. Kulasingham, V. Vibujithan, and A. C. De Silva. Deep belief networks and stacked autoencoders for the P300 Guilty Knowledge Test. *IECBES 2016 - IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 127–132, 2016.

[103] Zhengqing Miao, Meirong Zhao, Xin Zhang, and Dong Ming. LMDA-Net:A lightweight multi-dimensional attention network for general EEG-based brain-computer interfaces and interpretability. *NeuroImage*, 276:120209, 8 2023. 12

[104] Angus Dempster, François Petitjean, and Geoffrey I. Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 9 2020. 12, 18, 19

[105] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 248–257, 12 2020. 18, 19

[106] Kusumika Krori Dutta. Multi-class time series classification of EEG signals with recurrent neural networks. *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019*, pages 337–341, 1 2019. 12

[107] Stephen M. Gordon, Matthew Jaswa, Amelia J. Solon, and Vernon J. Lawhern. Real world BCI: Cross-domain learning and practical applications. *BCIforReal 2017 - Proceedings of the 2017 ACM Workshop on An Application-Oriented Approach to BCI Out of the Laboratory, co-located with IUI 2017*, pages 25–28, 3 2017. 12

[108] Dongrui Wu. Online and Offline Domain Adaptation for Reducing BCI Calibration Effort. *IEEE Transactions on Human-Machine Systems*, 47(4):550–563, 8 2017. 12

[109] Bo Qun Ma, He Li, Wei Long Zheng, and Bao Liang Lu. Reducing the subject variability of eeg signals with adversarial domain generalization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11953 LNCS:30–42, 2019. 12, 119

[110] Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2020. 12

[111] Hiroshi Morioka, Atsunori Kanemura, Jun ichiro Hirayama, Manabu Shikauchi, Takeshi Ogawa, Shigeyuki Ikeda, Motoaki Kawanabe, and Shin Ishii. Learning a common dictionary for subject-transfer decoding with resting calibration. *NeuroImage*, 111:167–178, 5 2015. 12

[112] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021, Vol. 23, Page 18*, 23(1):18, 12 2020. 13

[113] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:1135–1144, 8 2016. 13, 29, 46, 120

[114] Mohammed Saidul Islam, Iqram Hussain, Md Mezbaur Rahman, Se Jin Park, and Md Azam Hossain. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. *Sensors 2022, Vol. 22, Page 9859*, 22(24):9859, 12 2022. 13

[115] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, 7 2017. 13

[116] Valentin Gabeff, Tomas Teijeiro, Marina Zapater, Leila Cammoun, Sylvain Rheims, Philippe Ryvlin, and David Atienza. Interpreting deep learning models for epileptic seizure detection on EEG signals. *Artificial Intelligence in Medicine*, 117:102084, 7 2021. 13

[117] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 12 2013. 13, 23

[118] Amr Farahat, Christoph Reichert, Catherine M. Sweeney-Reed, and Hermann Hinrichs. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of Neural Engineering*, 16(6):066010, 10 2019. 18, 23

[119] Zülfikar Aslan and Mehmet Akin. A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals. *Physical and Engineering Sciences in Medicine*, 45(1):83–96, 3 2022. 13

[120] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 0(7 DEC):267, 2013. xv, 15, 26

[121] Mike X Cohen. *Analyzing neural time series data: theory and practice*. MIT press, 2014. 15, 16

[122] Salil H Patel and Pierre N Azzam. Characterization of N200 and P300: Selected Studies of the Event-Related Potential. *International Journal of Medical Sciences*, 2(4):147–154, 2005. 16

[123] Tamer Demiralp, Ahmet Ademoglu, Yorgo Istefanopulos, Canan Başar-Eroglu, and Erol Başar. Wavelet analysis of oddball P300. *International Journal of Psychophysiology*, 39(2-3):221–227, 1 2001. 16

[124] Joseph Geraghty and George Schoettle. Single-Subject vs. Cross-Subject Motor Imagery Models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13519 LNCS:442–452, 2022. 17

[125] O. Yeon Kwon, Min Ho Lee, Cuntai Guan, and Seong Whan Lee. Subject-Independent Brain-Computer Interfaces Based on Deep Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):3839–3852, 10 2020. 17

[126] Parisa Ghane, Narges Zarnaghinaghsh, and Ulisses Braga-Neto. Comparison of Classification Algorithms Towards Subject-Specific and Subject-Independent BCI. *9th IEEE International Winter Conference on Brain-Computer Interface, BCI 2021*, 2 2021.

[127] Siamac Fazli, Cristian Grozea, Márton Danóczy, Florin Popescu, Benjamin Blankertz, and Klaus-Robert Müller. Subject independent EEG-based BCI decoding. *Advances in Neural Information Processing Systems*, 22, 2009. 17

[128] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 8 1987. 17

[129] Pooja, S. K. Pahuja, and Karan Veer. Recent Approaches on Classification and Feature Extraction of EEG Signal: A Review. *Robotica*, 40(1):77–101, 1 2022. 17

[130] Alexandre Barachant and Marco Congedo. A plug&play P300 BCI using information geometry. *arXiv preprint arXiv:1409.0107*, 2014. 17

[131] Marco Congedo, Alexandre Barachant, and Anton Andreev. A new generation of brain-computer interface based on riemannian geometry. *arXiv preprint arXiv:1310.8115*, 2013. 17

[132] Alexandre Barachant, Quentin Barthélemy, Jean-Rémi King, Alexandre Gramfort, Sylvain Chevallier, Pedro L C Rodrigues, Emanuele Olivetti, Vladislav Goncharenko, Gabriel Wagner vom Berg, Ghiles Reguig, Arthur Lebeurrier, Erik Bjäreholt, Maria Sayu Yamamoto, Pierre Clisson, and Marie-Constance Corsi. pyRiemann/pyRiemann: v0.3, 7 2022. 17, 26

[133] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. 17, 19, 26, 35, 70, 72, 79, 118

[134] Alberto Zancanaro, Giulia Cisotto, João Ruivo Paulo, Gabriel Pires, and Urbano J. Nunes. CNN-based Approaches For Cross-Subject Classification in Motor Imagery: From the state-of-the-art to DynamicNet. *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2021*, 2021. 18

[135] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Annh Ratanamahatana, and Eamonn Keogh. The UCR Time Series Archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 10 2018. 18

[136] Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872*, 2019. 19, 26

[137] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019. 19

[138] Robert M.G. Reinhart, Daniel H. Mathalon, Brian J. Roach, and Judith M. Ford. Relationships between pre-stimulus gamma power and subsequent P300 and reaction time breakdown in schizophrenia. *International Journal of Psychophysiology*, 79(1):16–24, 1 2011. 21

[139] Fali Li, Qin Tao, Wenjing Peng, Tao Zhang, Yajing Si, Yangsong Zhang, Chanlin Yi, Bharat Biswal, Dezhong Yao, and Peng Xu. Inter-subject P300 variability relates to the efficiency of brain networks reconfigured from resting- to task-state: Evidence from a simultaneous event-related EEG-fMRI study. *NeuroImage*, 205:116285, 1 2020. 21

[140] Abeer Al-Nafjan and Mashael Aldayel. Predict Students' Attention in Online Learning Using EEG Data. *Sustainability 2022, Vol. 14, Page 6553*, 14(11):6553, 5 2022. 26

[141] Bin Hu, Xiaowei Li, Shuting Sun, and Martyn Ratcliffe. Attention Recognition in EEG-Based Affective Learning Research Using CFS+KNN Algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1):38–45, 1 2018. 26

[142] Feng Li, Yi Xia, Fei Wang, Dengyong Zhang, Xiaoyu Li, and Fan He. Transfer Learning Algorithm of P300-EEG Signal Based on XDAWN Spatial Filter and Riemannian Geometry Classifier. *Applied Sciences 2020, Vol. 10, Page 1804*, 10(5):1804, 3 2020. 26, 119

[143] Nathalie TH Gayraud, Alain Rakotomamonjy, and Maureen Clerc. Optimal transport applied to transfer learning for p300 detection. In *BCI 2017-7th Graz Brain-Computer Interface Conference*, page 6, 2017. 26, 119

[144] Steven Mortier, Jorg De Winne, Pengfei Sun, Wannes Vanransbeeck, Renata Turkes, Zhongju Yuan, Tim Verdonck, Marc Leman, Paul Devos, and Dick Botteldooren. WithMe preprocessed dataset, 10 2023. 26

[145] Guido Van Rossum and Fred L Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. 26, 70, 118

[146] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 26

[147] Anna Zmarz, Mirosław Rodzewicz, Maciej Dąbski, Izabela Karsznia, Małgorzata Korczak-Abshire, and Katarzyna J Chwedorzewska. Application of UAV BVLOS remote sensing data for multi-faceted analysis of Antarctic ecosystem. *Remote Sensing of Environment*, 217:375–388, 2018. 27

[148] Sha Huang, Lina Tang, Joseph P Hupy, Yang Wang, and Guofan Shao. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, 2021. 27, 44

[149] Yu Zhao, Qi Feng, and Aigang Lu. Spatiotemporal variation in vegetation coverage and its driving factors in the guanzhong basin, nw china. *Ecological Informatics*, 64, 9 2021. 27

[150] Jae Hyun Ryu, Dohyeok Oh, and Jaeil Cho. Simple method for extracting the seasonal signals of photochemical reflectance index and normalized difference vegetation index measured using a spectral reflectance sensor. *Journal of Integrative Agriculture*, 20(7):1969–1986, 2021. 27

[151] Manuela Balzarolo, Karen Anderson, Caroline Nichol, Micol Rossini, Loris Vescovo, Nicola Arriga, Georg Wohlfahrt, Jean-Christophe Calvet, Arnaud Carrara, Sofia Cerasoli, Sergio Cogliati, Fabrice Daumard, Lars Eklundh, Jan A. Elbers, Fatih Evrendilek, Rebecca N. Handcock, Jörg Kaduk, Katja Klumpp, Bernard Longdoz, Giorgio Matteucci, Michele Meroni, Lenoardo Montagnani, Jean-Marc Ourcival, Enrique P. Sánchez-Cañete, Jean-Yves Pontailler, Radoslaw Juszczak, Bob Scholes, and M. Pilar Martín. Ground-Based Optical Measurements at European Flux Sites: A Review of Methods, Instruments and Current Controversies. *Sensors*, 11(8):7954–7981, 2011. 27

[152] Rossana Monica Ferrara, Costanza Fiorentino, Nicola Martinelli, Pasquale Garofalo, and Gianfranco Rana. Comparison of different ground-based NDVI measurement methodologies to evaluate crop biophysical properties. *Italian Journal of Agronomy*, 5(2):145–154, 2010. 27

[153] Manel Rhif, Ali Ben Abbes, Beatriz Martinez, Rogier de Jong, Yanfang Sang, and Imed Riadh Farah. Detection of trend and seasonal changes in non-stationary remote sensing data: Case study of tunisia vegetation dynamics. *Ecological Informatics*, 69:101596, 2022. 27

[154] Pieter S.A. Beck, Clement Atzberger, Kjell Arild Høgda, Bernt Johansen, and Andrew K. Skidmore. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *Remote Sensing of Environment*, 100(3):321–334, 2006. 27

[155] Feng Li, Guo Song, Zhu Liujun, Zhou Yanan, and Lu Di. Urban vegetation phenology analysis using high spatio-temporal ndvi time series. *Urban Forestry & Urban Greening*, 25:43–57, 2017. 27

[156] Xue-Qing Ma, Pei Leng, Qian-Yu Liao, Yun-Jing Geng, Xia Zhang, Guo-Fei Shang, Xiaoning Song, Qian Song, and Zhao-Liang Li. Prediction of vegetation phenology with atmospheric reanalysis over semiarid grasslands in inner mongolia. *Science of The Total Environment*, 812:152462, 2022. 27

[157] Add Odland, Olavi Junttila, and Jade Nilsen. Growth responses of matteuccia struthiopteris plants from northern and southern norway exposed to different temperature and photoperiod treatments. *Nordic Journal of Botany*, 23(2):237–246, 2003. 27

[158] Jørgen A. B. Mølmann, Sigridur Dalmannsdottir, Anne Linn Hykkerud, Timo Hytönen, Amos Samkumar, and Laura Jaakola. Influence of arctic light conditions on crop production and quality. *Physiologia Plantarum*, 172(4):1931–1940, 2021. 27

[159] Howard E. Epstein, Isla Myers-Smith, and Donald A. Walker. Recent dynamics of arctic and sub-arctic vegetation. *Environmental Research Letters*, 8(1), 2013. 27, 28

[160] IPCC. Technical Summary. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. In V Masson-Delmotte, P Zhai, A Pirani, S L Connors, C Péan, S Berger, N Caud, Y Chen, L Goldfarb, M I Gomis, M Huang, K Leitzell, E Lonnoy, J B R Matthews, T K Maycock, T Waterfield, O Yelekçi, R Yu, and B Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 33–144. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. 27, 28

[161] Philipp R. Semenchuk, Mark A.K. Gillespie, Sabine B. Rumpf, Nanna Baggesen, Bo Elberling, and Elisabeth J. Cooper. High Arctic plant phenology is determined by snowmelt patterns but duration of phenological periods is fixed: An example of periodicity. *Environmental Research Letters*, 11(12), 2016. 27, 28

[162] Christopher Potter and Olivia Alexander. Changes in Vegetation Phenology and Productivity in Alaska Over the Past Two Decades. *Remote Sensing*, 12(10), 2020. 28, 43

[163] Kyle A. Arndt, Maria J. Santos, Susan Ustin, Scott J. Davidson, Doug Stow, Walter C. Oechel, Thao T.P. Tran, Brian Graybill, and Donatella Zona. Arctic greening associated with lengthening growing seasons in Northern Alaska. *Environmental Research Letters*, 14(12):125018, 2019. 28, 43

[164] Alexander Thomas Merrington. A Time Series Analysis of Vegetation Succession on Lava Flow Fields at Hekla Volcano: Assessing the Utility of Landsat Data, 2019. 28

[165] Pieter S.A. Beck and Scott J. Goetz. Satellite observations of high northern latitude vegetation productivity changes between 1982 and 2008: ecological variability and regional differences. *Environmental Research Letters*, 6(4):045501, 10 2011. 28

[166] Michael M. Loranty and Scott J. Goetz. Shrub expansion and climate feedbacks in Arctic tundra. *Environmental Research Letters*, 7(1), 2012. 28

[167] D. A. Walker, H. E. Epstein, M. K. Raynolds, P. Kuss, M. A. Kopecky, G. V. Frost, F. J.A. Danils, M. O. Leibman, N. G. Moskalenko, G. V. Matyshak, O. V. Khitun,

A. V. Khomutov, B. C. Forbes, U. S. Bhatt, A. N. Kade, C. M. Vonlanthen, and L. Tichý. Environment, vegetation and greenness (NDVI) along the North America and Eurasia Arctic transects. *Environmental Research Letters*, 7(1), 2012. 28

[168] Howard E. Epstein, Martha K. Raynolds, Donald A. Walker, Uma S. Bhatt, Compton J. Tucker, and Jorge E. Pinzon. Dynamics of aboveground phytomass of the circumpolar Arctic tundra during the past three decades. *Environmental Research Letters*, 7(1), 2012. 28

[169] Uma S. Bhatt, Donald A. Walker, Martha K. Raynolds, Peter A. Bieniek, Howard E. Epstein, Josefino C. Comiso, Jorge E. Pinzon, Compton J. Tucker, and Igor V. Polyakov. Recent Declines in Warming and Vegetation Greening Trends over Pan-Arctic Tundra. *Remote Sensing 2013, Vol. 5, Pages 4229-4254*, 5(9):4229–4254, 8 2013. 28

[170] Isla H. Myers-Smith, Jeffrey T. Kerby, Gareth K. Phoenix, Jarle W. Bjerke, Howard E. Epstein, Jakob J. Assmann, Christian John, Laia Andreu-Hayles, Sandra Angers-Blondin, Pieter S.A. Beck, Logan T. Berner, Uma S. Bhatt, Anne D. Bjorkman, Daan Blok, Anders Bryn, Casper T. Christiansen, J. Hans C. Cornelissen, Andrew M. Cunliffe, Sarah C. Elmendorf, Bruce C. Forbes, Scott J. Goetz, Robert D. Hollister, Rogier de Jong, Michael M. Loranty, Marc Macias-Fauria, Kadmiel Maseyk, Signe Normand, Johan Olofsson, Thomas C. Parker, Frans Jan W. Parmentier, Eric Post, Gabriela Schaepman-Strub, Frode Stordal, Patrick F. Sullivan, Haydn J.D. Thomas, Hans Tømmervik, Rachael Treharne, Craig E. Tweedie, Donald A. Walker, Martin Wilmking, and Sonja Wipf. Complexity revealed in the greening of the Arctic. *Nature Climate Change 2020 10:2*, 10(2):106–117, 1 2020. 28

[171] Martha Raynolds, Borgthór Magnússon, Sigmar Metúsalemsson, and Sigurdur H. Magnússon. Warming, sheep and volcanoes: Land cover changes in Iceland evident in satellite NDVI trends. *Remote Sensing*, 7(8):9492–9506, 2015. 28

[172] Halldór Björnsson, Bjarni D. Sigurðsson, Brynhildur Davíðsdóttir, Jón Ólafsson, Ólafur S. Ástþórssson, Snjólaug Ólafsdóttir, Trausti Baldursson, and Trausti Jónsson. *Loftslagsbreytingar og áhrif þeirra á Íslandi: Skýrsla vísindanefndar um loftlagsbreytingar 2018*. 2007. 28

[173] Haraldur Olafsson and Iman Rousta. Influence of atmospheric patterns and North Atlantic Oscillation (NAO) on vegetation dynamics in Iceland using Remote Sensing. *European Journal of Remote Sensing*, 54(1):351–363, 2021. 28

[174] Dennis Metze, Jörg Schnecker, Coline Le Noir de Carlan, Biplabi Bhattarai, Erik Verbruggen, Ivika Ostonen, Ivan A. Janssens, Bjarni D. Sigurdsson, Bela Hausmann, Christina Kaiser, and Andreas Richter. Soil warming increases the number of growing bacterial taxa but not their growth rates. *Science Advances*, 10(8):eadk6295, 2024. 28, 45

[175] Chao Fang, Niel Verbrigghe, Bjarni D. Sigurdsson, Ivika Ostonen, Niki I. W. Leblans, Sara Marañón-Jiménez, Lucia Fuchslueger, Páll Sigurðsson, Kathiravan Meeran, Miguel Portillo-Estrada, Erik Verbruggen, Andreas Richter, Jordi Sardans, Josep Peñuelas, Michael Bahn, Sara Vicca, and Ivan A. Janssens. Decadal soil warming decreased vascular plant above and belowground production in a subarctic grassland by inducing nitrogen limitation. *New Phytologist*, 240(2):565–576, 2023. 45

[176] Niel Verbrigghe, Kathiravan Meeran, Michael Bahn, Alberto Canarini, Erik Fransen, Lucia Fuchslueger, Johannes Ingrisch, Ivan A. Janssens, Andreas Richter, Bjarni D. Sigurdsson, Jennifer L. Soong, and Sara Vicca. Long-term warming reduced microbial biomass but increased recent plant-derived c in microbes of a subarctic grassland. *Soil Biology and Biochemistry*, 167:108590, 2022. 28

[177] Biplabi Bhattarai, Bjarni D Sigurdsson, Páll Sigurdsson, Niki Leblans, Ivan Janssens, Wendelien Meynzer, Arun Kumar Devarajan, Jaak Truu, Marika Truu, and Ivika Ostonen. Soil warming duration and magnitude affect the dynamics of fine roots and rhizomes and associated C and N pools in subarctic grasslands. *Annals of Botany*, 132(2):269–279, 07 2023. 28, 45

[178] Angel Chen, Trevor C. Lantz, Txomin Hermosilla, and Michael A. Wulder. Biophysical controls of increased tundra productivity in the western Canadian Arctic. *Remote Sensing of Environment*, 258:112358, 2021. 28

[179] Lorna E. Street and S. Caldararu. Why are Arctic shrubs becoming more nitrogen limited? *New Phytologist*, 233(2):585–587, 1 2022. 28

[180] René Van Der Wal and Audun Stien. High-arctic plants like it hot: a long-term investigation of between-year variability in plant biomass. *Ecology*, 95(12):3414–3427, 12 2014. 28

[181] Anne D. Bjorkman, Mariana García Criado, Isla H. Myers-Smith, Virve Ravolainen, Ingibjörg Svala Jónsdóttir, Kristine Bakke Westergaard, James P. Lawler, Mora Aronsson, Bruce Bennett, Hans Gardfjell, Starri Heiðmarsson, Laerke Stewart, and Signe Normand. Status and trends in Arctic vegetation: Evidence from experimental warming and long-term monitoring. *Ambio*, 49(3):678–692, 2020. 28

[182] Niki I.W. Leblans, Bjarni D. Sigurdsson, Sara Vicca, Yongshuo Fu, Josep Penuelas, and Ivan A. Janssens. Phenological responses of Icelandic subarctic grasslands to short-term and long-term natural soil warming. *Global Change Biology*, 23(11):4932–4945, 2017. 28, 29, 30, 32, 43, 44

[183] Kaleem Mehmood, Shoaib Ahmad Anees, Akhtar Rehman, Shao"an Pan, Aqil Tariq, Muhammad Zubair, Qijing Liu, Fazli Rabbi, Khalid Ali Khan, and Mi Luo. Exploring spatiotemporal dynamics of ndvi and climate-driven responses in ecosystems: Insights for sustainable management and climate resilience. *Ecological Informatics*, 80:102532, 2024.

[184] J. S. Kimball A. S. Hope and D. A. Stow. The relationship between tussock tundra spectral reflectance properties and biomass and vegetation composition. *International Journal of Remote Sensing*, 14(10):1861–1874, 1993.

[185] DA Walker, HE Epstein, MK Raynolds, P Kuss, MA Kopecky, GV Frost, FJA Daniëls, MO Leibman, NG Moskalenko, GV Matyshak, et al. Environment, vegetation and greenness (ndvi) along the north america and eurasia arctic transects. *Environmental Research Letters*, 7(1):015504, 2012.

[186] E Herrera Estrella, A Stoeth, NY Krakauer, and N Devineni. Quantifying vegetation response to environmental changes on the galapagos islands, ecuador using the normalized difference vegetation index (ndvi). *Environmental Research Communications*, 3(6):065003, 2021.

[187] Hao Wang, Huiying Liu, Ni Huang, Jian Bi, Xuanlong Ma, Zhiyuan Ma, Zijian Shangguan, Hongfang Zhao, Qisheng Feng, Tiangang Liang, Guangmin Cao, Bernhard Schmid, and Jin-Sheng He. Satellite-derived ndvi underestimates the advancement of alpine vegetation growth over the past three decades. *Ecology*, 102(12):e03518, 2021. 28, 43

[188] Lieven Michielsen. Plant communities and global change: adaptation by changes in present species composition or adaptation in plant traits. a case study in iceland. Master's thesis, Universiteit Antwerpen, 2014. 28

[189] M A Hearst, S T Dumais, E Osuna, J Platt, and B Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 7 1998. 29

[190] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:785–794, 3 2016. 29, 72, 79, 118, 120

[191] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30, 2017. 29, 79

[192] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 12 1943. 29

[193] Anne E. Thessen. Adoption of Machine Learning Techniques in Ecology and Earth Science. *One Ecosystem 1: e8621*, 1:e8621–, 2016. 29, 78

[194] Pierre Barré, Ben C. Stöver, Kai F. Müller, and Volker Steinhage. LeafNet: A computer vision system for automatic plant species identification. *Ecological Informatics*, 40:50–56, 7 2017. 29

[195] Xing Chen, Jun Zhao, Yan hua Chen, Wei Zhou, and Alice C. Hughes. Automatic standardized processing and identification of tropical bat calls using deep learning approaches. *Biological Conservation*, 241:108269, 1 2020.

[196] Deepti Barhate, Sunil Pathak, and Ashutosh Kumar Dubey. Hyperparameter-tuned batch-updated stochastic gradient descent: Plant species identification by using hybrid deep learning. *Ecological Informatics*, 75:102094, 2023. 29

[197] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5(9), 9 2019. 29

[198] Melanie Clapham, Ed Miller, Mary Nguyen, and Chris T. Darimont. Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. *Ecology and Evolution*, 10(23):12883–12892, 12 2020. 29

[199] Lin Ye and Qinghua Cai. Forecasting Daily Chlorophyll a Concentration during the Spring Phytoplankton Bloom Period in Xiangxi Bay of the Three-Gorges Reservoir by Means of a Recurrent Artificial Neural Network. *Journal of Freshwater Ecology*, 24(4):609–617, 2011. 29

[200] Hee Sun Cho, Kwang Hee Choi, Sang Don Lee, and Young Seuk Park. Characterizing habitat preference of Eurasian river otter (Lutra lutra) in streams using a self-organizing map. *Limnology*, 10(3):203–213, 6 2009.

[201] Tanya Strydom, Michael D. Catchen, Francis Banville, Dominique Caron, Gabriel Dansereau, Philippe Desjardins-Proulx, Norma R. Forero-Muñoz, Gracielle Higino, Benjamin Mercier, Andrew Gonzalez, Dominique Gravel, Laura Pollock, and Timothée Poisot. A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1837):20210063, 11 2021. 29

[202] Wang Li, Robert Buitenwerf, Michael Munk, Peder Klith Bøcher, and Jens Christian Svenning. Deep-learning based high-resolution mapping shows woody vegetation densification in greater Maasai Mara ecosystem. *Remote Sensing of Environment*, 247:111953, 9 2020. 29

[203] Qinghua Guo, Shichao Jin, Min Li, Qiuli Yang, Kexin Xu, Yuanzhen Ju, Jing Zhang, Jing Xuan, Jin Liu, Yanjun Su, Qiang Xu, and Yu Liu. Application of deep learning in ecological resource research: Theories, methods, and challenges. *Science China Earth Sciences*, 63(10):1457–1474, 10 2020. 29

[204] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys (CSUR)*, 55(2):96, 2 2022. 29

[205] Paul A. O'Gorman and John G. Dwyer. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 10 2018.

[206] Sachin Kumar. A novel hybrid machine learning model for prediction of co2 using socio-economic and energy attributes for climate change monitoring and mitigation policies. *Ecological Informatics*, 77:102253, 2023. 29

[207] Yuanyi Gao, Xiaobao Xue, Guoqing Qin, Kai Li, Jiahao Liu, Yulong Zhang, and Xinjiang Li. Application of machine learning in automatic image identification of insects - a review. *Ecological Informatics*, 80:102539, 2024. 29

[208] Marta Jemeļjanova, Alexander Kmoch, and Evelyn Uuemaa. Adapting machine learning for environmental spatial data - a review. *Ecological Informatics*, 81:102634, 2024. 29

[209] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 12 2021. 29

[210] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *Computer Vision–ECCV 2014*, 8689(PART 1):818–833, 2014. 29

[211] Yoshifumi Masago and Maychee Lian. Estimating the first flowering and full blossom dates of Yoshino cherry (Cerasus × yedoensis 'Somei-yoshino') in Japan using machine learning algorithms. *Ecological Informatics*, 71:101835, 11 2022. 29

[212] Bohao He, Yanghe Zhao, Wei Mao, and Robert J. Griffin-Nolanb. Explainable artificial intelligence reveals environmental constraints in seagrass distribution. *Ecological Indicators*, 144:109523, 11 2022. 29

[213] Jungsu Park, Woo Hyoung Lee, Keug Tae Kim, Cheol Young Park, Sanghun Lee, and Tae Young Heo. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Science of The Total Environment*, 832:155070, 8 2022. 29

[214] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. 29

[215] Bjarni D. Sigurdsson, Niki I.W. Leblans, Steven Dauwe, Elín Gudmundsdóttir, Per Gundersen, Gunnhildur E. Gunnarsdóttir, Martin Holmstrup, Krassimira Ilieva-Makulec, Thomas Kätterer, Bryndís Marteinsdóttir, Marja Maljanen, Edda S. Oddsdóttir, Ivika Ostonen, Josep Peñuelas, Christopher Poeplau, Andreas Richter, Páll Sigurdsson, Peter Van Bodegom, Håkan Wallander, James Weedon, and Ivan Janssens. Geothermal ecosystems as natural climate change experiments: The ForHot research site in Iceland as a case study. *Icelandic Agricultural Sciences*, 29(1):53–71, 2016. 29, 30, 31, 33, 45

[216] Xiaoyang Zhang, Mark A. Friedl, Crystal B. Schaaf, Alan H. Strahler, John C.F. Hodges, Feng Gao, Bradley C. Reed, and Alfredo Huete. Monitoring vegetation phenology using MODIS. *Remote Sensing of Environment*, 84(3):471–475, 3 2003. 29, 33

[217] Venja M. Roeber, Thomas Schmülling, and Anne Cortleven. The photoperiod: Handling and causing stress in plants, 1 2022. 30, 44

[218] S. R. Adams and F. A. Langton. Photoperiod and plant growth: A review, 2005. 30, 44

[219] Bjarni D Sigurdsson. Elevated [CO2] and nutrient status modified leaf phenology and growth rhythm of young Populus trichocarpa trees in a 3-year field study. *Trees*, 15(7):403–413, 2001. 30

[220] Niel Verbrigghe, Niki I.W. Leblans, Bjarni D. Sigurdsson, Sara Vicca, Chao Fang, Lucia Fuchslueger, Jennifer L. Soong, James T. Weedon, Christopher Poeplau, Cristina Ariza-Carricondo, Michael Bahn, Bertrand Guenet, Per Gundersen, Gunnhildur E. Gunnarsdóttir, Thomas Kätterer, Zhanfeng Liu, Marja Maljanen, Sara Marañón-Jiménez, Kathiravan Meeran, Edda S. Oddsdóttir, Ivika Ostonen, Josep Peñuelas, Andreas Richter, Jordi Sardans, Páll Sigurðsson, Margaret S. Torn, Peter M. Van Bodegom, Erik Verbruggen, Tom W.N. Walker, Håkan Wallander, and Ivan A. Janssens. Soil carbon loss in warmed subarctic grasslands is rapid and restricted to topsoil. *Biogeosciences*, 19(14):3381–3393, 2022. 30, 44, 45

[221] C Beer, P Porada, A Ekici, and M Brakebusch. Effects of short-term variability of meteorological variables on soil temperature in permafrost regions. *The Cryosphere*, 12(2):741–757, 2018. 30, 45

[222] Xiaoqing Tan, Siqiong Luo, Hongmei Li, Xiaohua Hao, Jingyuan Wang, Qingxue Dong, and Zihang Chen. Investigating the Effects of Snow Cover and Vegetation on Soil Temperature Using Remote Sensing Indicators in the Three River Source Region, China. *Remote Sensing*, 14(16), 2022. 45

[223] Jing Xie, Fabia Hüsler, Rogier de Jong, Barbara Chimani, Sarah Asam, Yeran Sun, Michael E Schaepman, and Mathias Kneubühler. Spring Temperature and Snow Cover Climatology Drive the Advanced Springtime Phenology (1991–2014) in the European Alps. *Journal of Geophysical Research: Biogeosciences*, 126(3):e2020JG006150, 2021. 30

[224] Wenjuan Hou, Jiangbo Gao, Shaohong Wu, and Erfu Dai. Interannual variations in growing-season NDVI and its correlation with climate variables in the southwestern karst region of China. *Remote Sensing*, 7(9):11105–11124, 2015. 30, 45

[225] Andrew R Conn, Nicholas I M Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000. 34

[226] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 35

[227] P.J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990. 35

[228] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 7 2019. 35

[229] Stein Rune Karlsen, Arve Elvebakk, Kjell Arild Høgda, and Tom Grydeland. Spatial and Temporal Variability in the Onset of the Growing Season on Svalbard, Arctic Norway — Measured by MODIS-NDVI Satellite Data. *Remote Sensing*, 6(9):8088–8106, 2014. 43

[230] Courtney G Collins, Sarah C Elmendorf, Robert D Hollister, Greg H R Henry, Karin Clark, Anne D Bjorkman, Isla H Myers-Smith, Janet S Prevéy, Isabel W Ashton, Jakob J Assmann, Juha M Alatalo, Michele Carbognani, Chelsea Chisholm, Elisabeth J Cooper, Chiara Forrester, Ingibjörg Svala Jónsdóttir, Kari Klanderud, Christopher W Kopp, Carolyn Livensperger, Marguerite Mauritz, Jeremy L May, Ulf Molau, Steven F Oberbauer, Emily Ogburn, Zoe A Panchen, Alessandro Petraglia, Eric Post, Christian Rixen, Heidi Rodenhizer, Edward A G Schuur, Philipp Semenchuk, Jane G Smith, Heidi Steltzer, Ørjan Totland, Marilyn D Walker, Jeffrey M Welker, and Katharine N Suding. Experimental warming differentially affects vegetative and reproductive phenology of tundra plants. *Nature Communications*, 12:3442, 2021. 44

[231] Andrey V. Malyshev, Hugh A.L. Henry, and Juergen Kreyling. Relative effects of temperature vs. photoperiod on growth and cold acclimation of northern and southern ecotypes of the grass arrhenatherum elatius. *Environmental and Experimental Botany*, 106:189–196, 2014. 44

[232] Quang Thanh Bui, Quang Tuan Pham, Van Manh Pham, Van Thuy Tran, Dinh Hung Nguyen, Quoc Huy Nguyen, Huu Duy Nguyen, Nhung Thi Do, and Van Manh Vu. Hybrid machine learning models for aboveground biomass estimations. *Ecological Informatics*, 79, 3 2024. 44

[233] Binghua Zhang, Li Zhang, Dong Xie, Xiaoli Yin, Chunjing Liu, and Guang Liu. Application of synthetic ndvi time series blended from landsat and modis data for grassland biomass estimation. *Remote Sensing*, 8, 2016.

[234] Maria Lumbierres, Pablo F Méndez, Javier Bustamante, Ramón Soriguer, and Luis Santamaría. Modeling biomass production in seasonal wetlands using modis ndvi land surface phenology. *Remote Sensing*, 9, 2017.

[235] Eileen Perry, Kathryn Sheffield, Doug Crawford, Stephen Akpa, Alex Clancy, and Robert Clark. Spatial and temporal biomass and growth for grain crops using ndvi time series. *Remote Sensing*, 14, 2022. 44

[236] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer, 2020. 45

[237] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020. 45

[238] Steven Mortier, Amir Hamedpour, Bart Bussmann, Phoebe Ruth Tchana Wandji, Bjarni Diðrik Sigurdsson, Tim Verdonck, and Tom De Schepper. Forhot dataset 2014-2019, 2023. 47

[239] Muhammad Ali Jamshed, Kamran Ali, Qammer H. Abbasi, Muhammad Ali Imran, and Masood Ur-Rehman. Challenges, Applications, and Future of Wireless Sensors in Internet of Things: A Review. *IEEE Sensors Journal*, 22(6):5482–5494, 3 2022. 49

[240] N. Achyutha Prasad, H. V. Chaitra, G. Manjula, Mohammad Shabaz, Ana Beatriz Martinez-Valencia, K. B. Vikhyath, Shrawani Verma, and José Luis Arias-Gonzáles. Delay optimization and energy balancing algorithm for improving network lifetime in fixed wireless sensor networks. *Physical Communication*, 58:102038, 6 2023. 49

[241] S. R. Jino Ramson and D. Jackuline Moni. Applications of wireless sensor networks — A survey. In *International Conference on Innovations in Electrical, Electronics, Instrumentation and Media Technology (ICEEIMT 2017)*, pages 325–329. IEEE, 11 2017. 49

[242] Dionisis Kandris, Christos Nakas, Dimitrios Vomvas, and Grigorios Koulouras. Applications of Wireless Sensor Networks: An Up-to-Date Survey. *Applied System Innovation 2020, Vol. 3, Page 14*, 3(1):14, 2 2020. 49

[243] Linghe Kong, Mingyuan Xia, Xiao Yang Liu, Guangshuo Chen, Yu Gu, Min You Wu, and Xue Liu. Data loss and reconstruction in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 25(11):2818–2828, 11 2014. 49

[244] Liqiang Pan and Jianzhong Li. K-Nearest Neighbor Based Missing Data Estimation Algorithm in Wireless Sensor Networks. *Wireless Sensor Network*, 02(02):115–122, 3 2010. 50, 51, 60, 68, 69, 70, 120

[245] Jonas J Lembrechts, Juha Aalto, Michael B Ashcroft, Pieter De Frenne, Martin Kopeckỳ, Jonathan Lenoir, Miska Luoto, Ilya MD Maclean, Olivier Roupsard, Eduardo Fuentes-Lillo, et al. Soiltemp: A global database of near-surface temperature. *Global change biology*, 26(11):6616–6629, 2020. 50, 67

[246] David M Kreindler and Charles J Lumsden. The effects of the irregular sample and missing data in time series analysis. In *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data*, pages 149–172. CRC Press, 2016. 50, 51, 57, 68, 69, 70

[247] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays . *Bioinformatics*, 17(6):520–525, 1 2001. 51, 58, 68, 69, 70

[248] Donald B Rubin. Multiple imputation. In *Flexible Imputation of Missing Data, Second Edition*, pages 29–62. Chapman and Hall/CRC, 2018. 51, 58

[249] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2 2011. 51, 58, 68, 69, 70

[250] Daniel Schunk. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. *AStA Advances in Statistical Analysis*, 92(1):101–114, 2 2008. 51, 58, 59, 65, 68, 69, 70

[251] Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 1 2012. 51, 61, 68, 69, 70

[252] Rahul Mazumder, Trevor Hastie, Hastie@stanford Edu, Robert Tibshirani, Tibs@stanford Edu, and Tommi Jaakkola. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010. 51, 59, 68, 69, 70

[253] Jinsung Yoon, William R. Zame, and Mihaela Van Der Schaar. Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, 5 2019. 51, 55, 62, 63, 66, 68, 69, 70

[254] Lovedeep Gondara and Ke Wang. MIDA: Multiple imputation using denoising autoencoders. In *PAKDD 2018: Advances in Knowledge Discovery and Data Mining*, volume 10939 LNAI, pages 260–272. Springer Verlag, 2018. 51, 61, 68, 69, 70

[255] Wei Cao, Dong Wang, Jian Li, Hao Zhou Bytedance, A I Lab, Yitan Li, Bytedance Ai Lab, and Lei Li. BRITS: Bidirectional Recurrent Imputation for Time Series. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 51, 61, 66, 68, 69, 70

[256] Yingshu Li, Chunyu Ai, Wiwek P Deshmukh, and Yiwei Wu. Data Estimation in Sensor Networks Using Physical and Statistical Methodologies. In *2008 The 28th International Conference on Distributed Computing Systems*, pages 538–545, 2008. 51, 59, 68, 69, 70

[257] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019. 51

[258] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in Big Data*, 4:693674, 7 2021. 51, 55

[259] Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. Mind the gap: An experimental evaluation of imputation of missing values techniques in time series. In *Proceedings of the VLDB Endowment*, volume 13, 2020. 51, 120

[260] Ceylan Yozgatligil, Sipan Aslan, Cem Iyigun, and Inci Batmaz. Comparison of missing value imputation methods in time series: the case of turkish meteorological data. *Theoretical and applied climatology*, 112:143–167, 2013. 51

[261] JJ Lembrechts, S Van de Vondel, C Allonsius, I Nijs, and F Meysman. Curieuzeneuzen in de tuin: eerste ervaringen met de microklimaatnetwerken van de toekomst. *Bodem: tijdschrift voor informatie-uitwisseling en discussie over duurzaam bodembeheer.-Alphen aan den Rijn, 1991, currens*, 33, 2022. 52

[262] Curieuzeneuzen in de tuin (curious noses in the garden), 2021. `https://curieuzeneuzen.be/home-en/` (accessed on 7 February 2024). 52

[263] Jonas J Lembrechts, Johan van den Hoogen, Juha Aalto, Michael B Ashcroft, Pieter De Frenne, Julia Kemppinen, Martin Kopeckỳ, Miska Luoto, Ilya MD Maclean, Thomas W Crowther, et al. Global maps of soil temperature. *Global Change Biology*, 28(9):3110–3144, 2022. 52

[264] Jonas J Lembrechts, Jonathan Lenoir, Brett R Scheffers, and Pieter De Frenne. Designing countrywide and regional microclimate networks. *Global Ecology and Biogeography*, 30(6):1168–1174, 2021. 52

[265] Jan Wild, Martin Kopeckỳ, Martin Macek, Martin Šanda, Jakub Jankovec, and Tomáš Haase. Climate at ecologically relevant scales: A new temperature and soil moisture logger for long-term microclimate measurement. *Agricultural and Forest Meteorology*, 268:40–47, 2019. 53, 55

[266] Georgios A Florides and Soteris A Kalogirou. Annual ground temperature measurements at various depths. In *CLIMA 2005*, 2005. 54

[267] Nitin R Chopde and Mangesh Nichat. Landmark based shortest path detection by using a* and haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):298–302, 2013. 54

[268] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. 55

[269] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[270] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997. 55

[271] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018. 55

[272] Ravindra S Lokupitiya, Erandathie Lokupitiya, and Keith Paustian. Comparison of missing value imputation methods for crop yield data. *Environmetrics: The official journal of the International Environmetrics Society*, 17(4):339–349, 2006. 55

[273] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 55, 70, 118

[274] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 3 2011. 58

[275] Stef van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242, 6 2007.

[276] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 12 2011. 58

[277] Augustine Kong, Jun S Liu, and Wing Hung Wong. Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288, 1994. 59

[278] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984. 59

[279] Mihail Halatchev Le Gruenwald. Estimating missing values in related sensor data streams. In *COMAD*, pages 83–94, 2005. 62

[280] Yulong Deng, Chong Han, Jian Guo, and Lijuan Sun. Temporal and Spatial Nearest Neighbor Values Based Missing Data Imputation in Wireless Sensor Networks. *Sensors*, 21(5), 2021. 62, 63

[281] Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python. `https://github.com/iskandr/fancyimpute`. 70, 120

[282] P Lopez Gonzalez-Nieto, M Gomez Flechoso, MA Arribas Mocoroa, A Muñoz Martin, ML Garcia Lorenzo, G Cabrera Gomez, JA Alvarez Gomez, A Caso Fraile, JM Orosco Dagan, R Merinero Palomares, et al. Design and development of a virtual laboratory in python for the teaching of data analysis and mathematics in geology: Geopy. In *INTED2020 Proceedings*, pages 2236–2242. IATED, 2020. 70

[283] François Chollet et al. Keras. `https://keras.io`, 2015. 70

[284] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. 70

[285] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde,

Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 70

[286] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. 70

[287] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from `https://www.tensorflow.org/`. 70

[288] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 70

[289] Guohua Xu, Xiaorong Fan, and Anthony J. Miller. Plant nitrogen assimilation and use efficiency. *Annual Review of Plant Biology*, 63:153–182, 2012. 71

[290] Jianbo Shen, Lixing Yuan, Junling Zhang, Haigang Li, Zhaohai Bai, Xinping Chen, Weifeng Zhang, and Fusuo Zhang. Phosphorus dynamics: From soil to plant. *Plant Physiology*, 156:997–1005, 2011. 71

[291] Jordi Sardans and Josep Peñuelas. Potassium control of plant functions: Ecological and agricultural implications. *Plants*, 10, 2021. 71

[292] Jordi Sardans and Josep Peñuelas. Potassium: A neglected nutrient in global change. *Global Ecology and Biogeography*, 24:261–275, 2015. 71

[293] C. I. Ludemann, N. Wanner, P. Chivenge, A. Dobermann, R. Einarsson, P. Grassini, A. Gruere, K. Jackson, L. Lassaletta, F. Maggi, G. Obli-Laryea, M. K. van Ittersum, S. Vishwakarma, X. Zhang, and F. N. Tubiello. A global FAOSTAT reference database of cropland nutrient budgets and nutrient use efficiency (1961–2020): nitrogen, phosphorus and potassium. *Earth System Science Data*, 16(1):525–541, 2024. 71, 72, 75, 77, 92, 93, 94, 95, 98, 100, 103, 104, 105, 110

[294] R. Xu, H. Tian, S. Pan, S. R. S. Dangal, J. Chen, J. Chang, Y. Lu, U. M. Skiba, F. N. Tubiello, and B. Zhang. Increased nitrogen enrichment and shifted patterns in the world's grassland: 1860-2016. *Earth System Science Data*, 11(1):175–187, 2019. 71, 75, 94, 104

[295] MA Sutton, Bleeker A, Bekunda M, Grizzetti B, de Vries W, van Grinsven HJM, Abrol YP, Adhya TK, Billen G, Davidson EA, Datta A, Diaz R, Erisman JW, Liu XJ, Oenema O, Palm C, Raghuram N, Reis S, Scholz RW, Sims T, Yan XY, and Zhang Y. *Our Nutrient World: The challenge to produce more food and energy with less pollution*. Centre for Ecology and Hydrology (CEH), Edinburgh UK on behalf of the Global Partnership on Nutrient Management and International Nitrogen Initiative., 2013. 71

[296] Josep Penuelas, Ivan A. Janssens, Philippe Ciais, Michael Obersteiner, and Jordi Sardans. Anthropogenic global shifts in biospheric n and p concentrations and ratios and their impacts on biodiversity, ecosystem productivity, food security, and human health. *Global Change Biology*, 26:1962–1985, 2020. 71

[297] Wulahati Adalibieke, Xiaoqing Cui, Hongwei Cai, Liangzhi You, and Feng Zhou. Global crop-specific nitrogen fertilization dataset in 1961-2020. *Scientific data*, 10(1):617, 12 2023. 71, 72, 74, 78, 79, 81, 88, 89, 90

[298] Chaoqun Lu and Hanqin Tian. Global nitrogen and phosphorus fertilizer use for agriculture production in the past half century: Shifted hot spots and nutrient imbalance. *Earth System Science Data*, 9(1):181–192, 2017. 74, 78, 79, 81

[299] Kazuya Nishina, Akihiko Ito, Naota Hanasaki, and Seiji Hayashi. Reconstruction of spatially detailed global map of NH_4^+ and NO_3^- application in synthetic nitrogen fertilizer. *Earth System Science Data*, 9:149–162, 2017.

[300] Graham K. MacDonald, Elena M. Bennett, Philip A. Potter, and Navin Ramankutty. Agronomic phosphorus imbalances across the world's croplands. *Proceedings of the National Academy of Sciences of the United States of America*, 108:3086–3091, 2011. 71, 72

[301] Peiyu Cao, Chaoqun Lu, and Zhen Yu. Historical nitrogen fertilizer use in agricultural ecosystems of the contiguous united states during 1850-2015: Application rate, timing, and fertilizer types. *Earth System Science Data*, 10, 2018. 72, 93

[302] Zhen Yu, Jing Liu, and Giri Kattel. Historical nitrogen fertilizer use in china from 1952 to 2018. *Earth System Science Data*, 14:5179–5194, 2022. 72

[303] Richard T Conant, Aaron B Berdanier, and Peter R Grace. Patterns and trends in nitrogen use and nitrogen recovery efficiency in world agriculture. *Global Biogeochemical Cycles*, 27(2):558–566, 2013. 72, 78

[304] Adolfo Martinez and Ray B Diamond. Fertilizer Use Statistics In Crop Production. Technical Report T-24, International Fertilizer Developemnt Center, 9 1982. xxi, 72, 73, 75, 90, 92, 93, 95, 96, 97, 98, 99, 101, 103, 104, 114, 115, 117

[305] Adolfo Martinez. Fertilizer use statistics and crop yields. Technical Report T-37, International Fertilizer Development Center, Muscle Shoals, Alabama, 1 1990. 72, 73, 92, 93, 94, 96, 100, 104, 114, 115

[306] EFMA. Fertilizer application by crop in EU countries 2001/02. Technical report, EFMA, unpublished. 72, 74, 96, 97, 98, 99, 100, 101, 102, 114, 115

[307] EFMA. Fertilizer application by crop in EU countries 2006/07. Technical report, EFMA, unpublished. 98, 99, 114, 115

[308] Fertilizer Europe. Fertilizer application by crop in EU countries 2011/12. Technical report, Fertilizer Europe, unpublished. 96, 114

[309] Fertilizer Europe. Fertilizer application by crop in EU countries 2014/15. Technical report, Fertilizer Europe, unpublished. xxi, 72, 73, 74, 75, 90, 96, 97, 98, 99, 100, 101, 102, 114, 115, 117

[310] T. Zou, X. Zhang, and E. A. Davidson. Global trends of cropland phosphorus use and sustainability challenges. *Nature*, 611:81–87, 2022. 72, 75, 77, 95, 110

[311] Luis Lassaletta, Gilles Billen, Bruna Grizzetti, Juliette Anglade, and Josette Garnier. 50 year trends in nitrogen use efficiency of world cropping systems: The relationship between yield and nitrogen input to cropland. *Environmental Research Letters*, 9:105011, 2014. 75, 92, 93, 94, 98, 103, 104

[312] Rasmus Einarsson, Alberto Sanz-Cobena, Eduardo Aguilera, Gilles Billen, Josette Garnier, Hans J.M. van Grinsven, and Luis Lassaletta. Crop production and nitrogen use in european cropland and grassland 1961–2019. *Scientific Data*, 8:288, 2021. 72, 75, 95, 96, 97, 98, 99, 100, 101, 102, 114, 115

[313] Chad Monfreda, Navin Ramankutty, and Jonathan A Foley. Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles*, 22(1), 2008. 72, 76, 77, 82, 90

[314] Kees Klein Goldewijk, Arthur Beusen, Jonathan Doelman, and Elke Stehfest. Anthropogenic land use estimates for the holocene–hyde 3.2. *Earth System Science Data*, 9(2):927–953, 2017. 72, 82, 90

[315] FAO, IFA, and IFDC. Fertilizer use by crop 1. Technical Report ESS/MISC/1992/3, FAO, Viale delle Termi di Caracalla, 00100, Rome, 10 1992. 73, 75, 90, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 114, 115

[316] FAO, IFA, and IFDC. Fertilizer use by crop 2. Technical Report ESS/MISC/1994/4, FAO, Viale delle Termi di Caracalla, 00100, Rome, 1994. xxi, 99, 100, 101, 102, 103, 114, 115, 117

[317] FAO, IFA, and IFDC. Fertilizer use by crop 3. Technical Report ESS/MISC/1996/1, FAO, Viale delle Termi di Caracalla, 00100, Rome, 1996. 93, 94, 99, 100, 102, 103, 114, 115

[318] FAO, IFA, and IFDC. Fertilizer use by crop 4. Technical report, FAO, Viale delle Termi di Caracalla, 00100, Rome, 1999. 93, 99, 100, 104, 114, 115

[319] FAO, IFA, and IFDC. Fertilizer use by crop 5. Technical report, FAO, Viale delle Termi di Caracalla, 00100, Rome, 2002. 75, 92, 93, 94, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 114, 115

[320] Pattrick Heffer. Asessment of Fertilizer Use by Crop at the Global Level 2006/07 - 2007/08. Technical Report A/09/55, International Fertilizer Industry Association, rue Marbeuf - 75008, Paris, April 2009. 74, 96, 102

[321] Patrick Heffer, Armelle Gruère, and Terry Roberts. Assessment of fertilizer use by crop at the global level 2014-2014/15. Technical Report A/17/134, International Fertilizer Industry Association & International Plant Nutrition Insitute, November 2017. xxi, 73, 74, 75, 90, 92, 93, 94, 102, 103, 104, 105, 114, 115, 117

[322] USDA. Fertilizer use and price, 2019. https://www.ers.usda.gov/data-products/fertilizer-use-and-price/. 73, 74, 81, 88, 90

[323] Agricultural Census Division. All india bulletin on input survey 1986-87. Technical report, Goverment of India, 9 1992. 73, 74, 81

[324] Agricultural Census Division. All india report on input survey 1991-1992. Technical Report AGRI/2000-1A, Goverment of India, 1 2000.

[325] Agricultural Census Division. All india report on input survey 1996-1997. Technical report, Goverment of India, 2007.

[326] Agriculture Census Division. All india report on input survey 2001-02. Technical report, Ministry of Agriculture, 8 2008.

[327] Agriculture Census Division. All india report on input survey 2006-07. Technical report, Ministry of Agriculture, 3 2012.

[328] Agriculture Census Division. All india report on input survey 2011-12. Technical report, Minister of Agriculture & Farmers Welfare, 4 2016.

[329] Agriculture Census Division. All india report on input survey 2016-17. Technical report, Minister of Agriculture, Cooperation & Farmers Welfare, 2 2021. 73, 74, 81, 88

[330] FAO. World programme for the census of agriculture 2020: Programme, concepts and definitions, 2015. 73

[331] FAOSTAT. Crops and livestock products, 2023. https://www.fao.org/faostat/en/#data/QCL. 73, 74, 75, 77, 81, 109, 110

[332] Pattrick Heffer. Asessment of Fertilizer Use by Crop at the Global Level 2010 - 2010/11. Technical Report A/13/111, International Fertilizer Industry Association, rue Marbeuf - 75008, Paris, August 2013. 74

[333] Maiz'Europ'. Figures, 2024. https://www.maizeurop.com/en/structure/cepm/figures/. 74

[334] DEFRA. British survey of fertiliser practice dataset, 7 2023. https://www.gov.uk/government/statistical-data-sets/british-survey-of-fertiliser-practice-dataset. 74, 81, 88, 90, 100, 115

[335] NFDC. Statistics national fertilizer development centre, 2023. http://www.nfdc.gov.pk/stat.html. 74, 81

[336] Philippine Statistics Authority. Estimated inorganic fertilizer use by geolocation, grade and year, by area harvested and year, 2023. https://openstat.psa.gov.ph/. 74, 81, 88

[337] Roehlano M Briones. The fertilizer industry and philippine agriculture: Policies, problems, and priorities. *Philippine Journal of Development*, 43, 2017. 75

[338] Official Statistics of Sweden. Use of fertilisers and animal manure in agriculture in 2010/11. Technical Report SM 1203, Official Statistics of Sweden, 6 2014. 75, 81, 88

[339] Official Statistics of Sweden. Use of fertilisers and animal manure and cultivation measures in agriculture 2012/13. Technical Report SM 1402, Official Statistics of Sweden, 6 2014.

[340] Official Statistics of Sweden. Use of fertilisers and animal manure and cultivation measures in agriculture 2015/16. Technical Report SM 1702, Official Statistics of Sweden, 6 2017.

[341] Official Statistics of Sweden. Use of fertilisers and animal manure and cultivation measures in agriculture 2018/19. Technical Report SM 2002, Official Statistics of Sweden, 6 2020. 75, 81

[342] Stats NZ. Agricultural production survey, 2021. `https://www.stats.govt.nz/indicators/fertilisers-nitrogen-and-phosphorus`. 75, 81, 88, 94

[343] FAOSTAT. Fertilizers by nutrient, 2023. `https://www.fao.org/faostat/en/#data/RFN`. 75, 77, 81, 99, 103, 109, 110

[344] Paswel P. Marenya and Christopher B. Barrett. Soil quality and fertilizer use rates among smallholder farmers in western Kenya. *Agricultural Economics*, 40(5):561–572, 2009. 76, 87

[345] Kaushik Bora. Rainfall shocks and fertilizer use: A district level study of India. *Environment and Development Economics*, 27:556–577, 2022. 76, 77

[346] Christian Levers, Van Butsic, Peter H. Verburg, Daniel Müller, and Tobias Kuemmerle. Drivers of changes in agricultural intensity in Europe. *Land Use Policy*, 58:380–393, 2016. 76, 77, 87

[347] Ian Harris, Timothy J. Osborn, Phil Jones, and David Lister. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1):190, 2020. 76, 109

[348] Laura Poggio, Luis M. De Sousa, Niels H. Batjes, Gerard B.M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, 2021. 76, 109

[349] N. Middleton and D. S.G. Thomas. World atlas of desertification, 1992. 76

[350] Xiaotang Ju, Baojing Gu, Yiyun Wu, and James N. Galloway. Reducing China's fertilizer use by increasing farm size. *Global Environmental Change*, 41:26–32, 2016. 77

[351] FAOSTAT. Structural data from agricultural censuses, 2024. `https://www.fao.org/faostat/en/#data`. 77

[352] Sarah K. Lowder, Jakob Skoet, and Terri Raney. The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, 87:16–29, 2016. 77

[353] L. Jordan-Meille, G. H. Rubæk, P. A.I. Ehlert, V. Genot, G. Hofman, K. Goulding, J. Recknagel, G. Provolo, and P. Barraclough. An overview of fertilizer-P recommendations in europe: Soil testing, calibration and fertilizer recommendations. *Soil Use and Management*, 28:419–435, 2012. 77

[354] G. Feder, R. E. Just, and D. Zilberman. Adoption of agricultural innovations in developing countries: a survey. *Economic Development & Cultural Change*, 33:255–298, 1985. 77, 78

[355] M. Hossain and V. P. Singh. Fertilizer use in Asian agriculture: Implications for sustaining food security and the environment. *Nutrient Cycling in Agroecosystems*, 57:155–169, 2000. 77

[356] FAOSTAT. Fertilizers Archive, 2020. `https://www.fao.org/faostat/en/#data/RA`. 77

[357] World Bank. World Bank Commodity Price Data (Pink Sheet). Technical report, 2024. `https://www.worldbank.org/en/research/commodity-markets`. 77, 87, 110

[358] John W. McArthur and Gordon C. McCord. Fertilizing growth: Agricultural inputs and their effects in economic development. *Journal of Development Economics*, 127:133–152, 2017. 77, 110

[359] Robert J Hijmans, Roger Bivand, Karl Forner, Jeroen Ooms, Edzer Pebesma, and Michael D Sumner. Package 'terra'. *Maintainer: Vienna, Austria*, 2022. 77, 118

[360] Global maps of travel time to healthcare facilities. *Nature Medicine*, 26:1835–1838, 2020. 77

[361] Ulrich Kleine-Kleffmann. The discovery of the first potash mine and the development of the potash industry since 1861. *Journal of Plant Nutrition and Soil Science*, 186:615–622, 12 2023. 77

[362] Lieven Clarisse, Martin Van Damme, Cathy Clerbaux, and Pierre François Coheur. Tracking down global NH3 point sources with wind-adjusted superresolution. *Atmospheric Measurement Techniques*, 12:5457–5473, 2019. 77

[363] FAOSTAT. Producer prices, 2023. `https://www.fao.org/faostat/en/#data/PP`. 77, 110

[364] FAOSTAT. Producer prices (old series), 2023. `https://www.fao.org/faostat/en/#data/PA`. 77, 110

[365] World Bank. Official exchange rate (LCU per US $, period average), 2023. `https://data.worldbank.org/indicator/PA.NUS.FCRF`. 77

[366] Jongrim Ha, M. Ayhan Kose, and Franziska Ohnsorge. One-stop source: A global database of inflation. *Journal of International Money and Finance*, 137, 2023. 78

[367] D. Tilman, J. Fargione, B. Wolff, C. D'Antonio, A. Dobson, R. Howarth, D. Schindler, W. H. Schlesinger, D. Simberloff, and D. Swackhamer. Forecasting agriculturally driven global environmental change. *Science*, 292(5515):281–284, 2001. 78, 87

[368] Tao Xiang, Tariq H. Malik, and Klaus Nielsen. The impact of population pressure on global fertiliser use intensity, 1970–2011: An analysis of policy-induced mediation. *Technological Forecasting and Social Change*, 152:1–12, 2020. 78

[369] Nils Chr Stenseth and Atle Mysterud. Climate, changing phenology, and other life history traits: Nonlinearity and match-mismatch to the environment. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13379–13381, 10 2002. 78

[370] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scientific machine learning benchmarks. *Nature Reviews Physics 2022 4:6*, 4(6):413–420, 4 2022. 78

[371] Massimo Bertolini, Davide Mezzogori, Mattia Neroni, and Francesco Zammori. Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, 175:114820, 8 2021. 78

[372] Devdatta A Bondre and Santosh Mahagaonkar. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, 4:371–376, 2019. 78

[373] Carla Pacheco, Mario Guimaraes, Eduardo Bezerra, Dacy Lobosco, Jorge Soares, Pedro Henrique González, Adalberto Andrade, Cristina Gomes De Souza, and Eduardo Ogasawara. Exploring Data Preprocessing and Machine Learning Methods for Forecasting Worldwide Fertilizers Consumption. *Proceedings of the International Joint Conference on Neural Networks*, 2022-July, 2022. 78

[374] Peng Xu, Geng Li, Yi Zheng, Jimmy C.H. Fung, Anping Chen, Zhenzhong Zeng, Huizhong Shen, Min Hu, Jiafu Mao, Yan Zheng, Xiaoqing Cui, Zhilin Guo, Yilin Chen, Lian Feng, Shaokun He, Xuguo Zhang, Alexis K.H. Lau, Shu Tao, and Benjamin Z. Houlton. Fertilizer management for global ammonia emission reduction. *Nature 2024 626:8000*, 626(8000):792–798, 1 2024. 78

[375] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *https://doi.org/10.1214/aos/1016218223*, 28(2):337–407, 4 2000. 79

[376] Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1):1–15, 3 2014. 79

[377] M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1 1974. 79

[378] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):1–8, 2 2006. 79

[379] Fernando Coello, Steven Mortier, Thomas Decorte, Iris Janssens, Tim Verdonck, Jordi Sardans, and Josep Peñuelas. Fertilizer application rate maps per crop and year. *figshare* `https://doi.org/10.6084/m9.figshare.25435432`, 2024. 84, 91

[380] Yongbo Liu, Xubin Pan, and Junsheng Li. A 1961–2010 record of fertilizer use, pesticide application and cereal yields: a review. *Agronomy for Sustainable Development*, 35:83–93, 2015. 86

[381] Stefano Longo and Richard York. Agricultural exports and the environment: A cross-national study of fertilizer and pesticide consumption. *Rural Sociology*, 73:82–104, 2008. 87

[382] Mingliang Liu and Hanqin Tian. China's land cover and land use change from 1700 to 2005: Estimations from high-resolution satellite data and historical archives. *Global Biogeochemical Cycles*, 24(3), 2010. 90

[383] Hanqin Tian, Kamaljit Banger, Tao Bo, and Vinay K Dadhwal. History of land use in india during 1880-2010: Large-scale land transformations reconstructed from satellite data and historical archives. *Global and Planetary Change*, 121:78–88, 2014. 90

[384] Zhen Yu and Chaoqun Lu. Historical cropland expansion and abandonment in the continental US during 1850 to 2016. *Global Ecology and Biogeography*, 27(3):322–333, 2018. 90

[385] Iris Janssens, Fernando Coello, Steven Mortier, Thomas Decorte, Jordi Sardans, Josep Peñuelas, and Tim Verdonck. Code for "machine learning-driven global crop-specific fertilization dataset from 1960–2020". *figshare* `https://doi.org/10.6084/m9.figshare.25435594`, 2024. 91

[386] CEPAL, FAO, and BID. El uso de fertilizantes en Argentina. Technical Report E/CN.12/741, 3 1966. 92, 114

[387] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2006, 2006. `https://fertilizar.org.ar/estadisticas/`. 92

[388] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2011, 2011. `https://fertilizar.org.ar/estadisticas/`.

[389] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2012, 2012. `https://fertilizar.org.ar/estadisticas/`.

[390] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2013, 2013. `https://fertilizar.org.ar/estadisticas/`.

[391] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2014, 2014. `https://fertilizar.org.ar/estadisticas/`.

[392] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2015, 2015. `https://fertilizar.org.ar/estadisticas/`.

[393] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2016, 2016. `https://fertilizar.org.ar/estadisticas/`.

[394] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2017, 2017. `https://fertilizar.org.ar/estadisticas/`.

[395] Fertilizar Asociación Civil. Consumo de fertilizantes por cultivo 2018, 2018. `https://fertilizar.org.ar/estadisticas/`. 92, 114

[396] FAO. Fertilizer use by crop in Brazil. Technical Report TC/D/Y5376E/1/05.04/300, FAO, Viale delle Terme di Caracalla 00100, Rome, 2004. 92

[397] Robert M. Boddey, Deise F. Xavier, Bruno J.R. Alves, and Segundo Urquiaga. Brazilian agriculture: The transition to sustainability. *Journal of Crop Production*, 9:593–621, 2003. 92

[398] Sarah J. Pogue, Nicole Bamber, Hugues Imbeault-Tétreault, Adeline Baudoin, Aaron L. McPherson, Steven Kega, Madavine Tom, Roland Kröbel, Kim H. Ominski, Karen A. Beauchemin, Brenna Grant, and Tim A. McAllister. Regionalized life cycle inventory data collection and calculation for perennial forage production in canada: methodological best practices and limitations. *International Journal of Life Cycle Assessment*, 2023. 92, 114

[399] James D. Beaton and Joseph Berger. *Present and Potential Use of Fertilizer for Forage Production in Temperate Zones*, chapter 2, pages 17–37. American Society of Agronomy, 1974. 92, 93, 97, 98, 99, 100, 101, 114, 115

[400] CEPAL, FAO, and BID. El uso de fertilizantes en Chile. Technical Report E/CN.12/757, 6 1966. 93, 114

[401] Rubén D. Nuñez. Estudio sobre el mercado de fertilizantes en la República Dominicana. Technical Report IICA/E72/9, Instituto Inteamericano de Cooperación para la Agricultural, United States Agency for International Development, Secretariado Técnico de la Presidencia Oficina Nacional de Planificación ONAPLAN, 9 1999. 93, 114

[402] Réal Michaud, W. F. Lehman, and M. D. Rumbaugh. *World distribution and historical development*, chapter 2. Agronomy Monographs. American Society of Agronomy, 1988. 93, 104, 105

[403] L. B Nelson, M. H Mcvickar, R. D Munson, L. F Seatz, S. L Tisdale, and W. C. White. *Changing patterns in fertilizer use*. Soil Science Society of America, inc, 1968. 93, 114

[404] Darrell A. Russel, W. Joe Free, and Donald L. McCune. *Potential for Fertilizer Use on Tropical Forages*, chapter 3, pages 39–65. American Society of Agronomy, 1974. 93, 114

[405] Eduardo Casanova. Problemática de los fertilizantes en Venezuela. *Venesuelos*, 12:5–16, 2004. 94

[406] J. C.B. Dubeux, L. E. Sollenberger, B. W. Mathews, J. M. Scholberg, and H. Q. Santos. Nutrient cycling in warm-climate grasslands. *Crop Science*, 47:915–928, 2007. 94

[407] R. P. Rawnsley, A. P. Smith, K. M. Christie, M. T. Harrison, and R. J. Eckard. Current and future direction of nitrogen fertiliser use in australian grazing systems. *Crop and Pasture Science*, 70:1034–1043, 2019. 94

[408] P. M. Barrow. Potassium fertilizer use in south australia. *Agronomy branch report*, 1:1–11, 1968. 94

[409] P. H. Williams and R. J. Haynes. Influence of improved pastures and grazing animals on nutrient cycling within new zealand soils. *New Zealand Journal of Ecology*, 14:49–57, 1990. 94

[410] Karl Buchgraber, Andreas Schaumberger, and Erich M Pötsch. Grassland farming in Austria - status quo and future prospective. *Grassland Science in Europe*, 16:13–24, 2011. 95

[411] Stany Vandermoere. *Improving agricultural phosphorus use efficiency and reducing soil phosphorus losses at the field scale.* PhD thesis, 2020. 96

[412] T. Németh. *Past, Present and Future Status of N-Fertilization Policies in Hungary*, pages 243–252. Springer Netherlands, 2000. 96, 98

[413] Philip M. Raup. Postwar recovery of Western German agriculture. *Journal of Farm Economics*, 32:1–14, 1950. 96, 97

[414] H. J. Styhr Petersen. Forecasting Danish nitrogen fertilizer consumption. *Industrial Marketing Management*, 6:211–221, 1977. 96, 114

[415] P Virkajärvi, M Rinne, J Mononen, O Niskanen, K Järvenranta, and A Sairananen. Dairy production systems in Finland. *18th Symposium of the European Grassland Federation, Grassland and forages in high output dairy farming systems*, 20:51–66, 2015. 97

[416] Geniana R Edwards and Geraldine W Abbott. *The Agricultural Economy of Finland*, volume 169 of *ERS foreign*. US Department of Agriculture, Economic Research Service, 1966. 97

[417] J. Le Noë, G. Billen, F. Esculier, and J. Garnier. Long-term socioecological trajectories of agro-food systems revealed by n and p flows in french regions from 1852 to 2014. *Agriculture, Ecosystems and Environment*, 265:132–143, 2018. 97, 99, 114

[418] Ministère de l'Agriculture. *Les Praires en 1982*, volume 233 of *Collections de statistique agricole*. 1984. 97

[419] Véronique Rabaud and Michelle Cesses. Enquête sur les practiques culturales - 2001. Technical Report 159, Ministère de l'Agriculture (SSP), 6 2004.

[420] Ministère de l'Agriculture (SSP). Pratiques culturales - 2006. `https://agreste.agriculture.gouv.fr/agreste-web/disaron/Dos8/detail/`, 2006.

[421] Ministère de l'Agriculture (SSP). Pratiques culturales sur les grandes cultures - 2017. `https://agreste.agriculture.gouv.fr/agreste-web/disaron/Chd2009/detail/`, 2020. 97, 114

[422] H. A. Pearson, C. H Herbel, and D. T Pendleton. A tour of East German agriculture. *Rangelands*, 1:9–11, 1979. 97

[423] D.A Panos, S. Sotiriadis, and E Fikas. Grassland's progress in Greece. *Der Z'uchter*, 31:37–47, 1961. 98

[424] Yong Hou, Lin Ma, Katalin Sárdi, István Sisák, and Wenqi Ma. Nitrogen flows in the food production chain of hungary over the period 1961–2010. *Nutrient Cycling in Agroecosystems*, 102:335–346, 2015. 98

[425] W. E. Murphy, W. F. O'Keeffe, and Foras Taluntais. Fertiliser use surveys, 1972, 1974 & 1975. Technical Report 14, Fertiliser Association of Ireland, 1978. 98, 114

[426] W. E. Murphy and W. F. O'Keefe. Fertiliser use survey 1981-82. Technical Report 24, The Fertiliser Association of Ireland, 11 1983.

[427] W. E. Murphy and O'Keefem W. F. Fertiliser use survey. Technical Report 27, The Fertiliser Association of Ireland, 11 1987. 98

[428] B. S. Coulter, W. E. Murphy, N. Culleton, G. Quinlan, and L. Connolly. A survey of fertilizer use from 2001-2003 for grassland and arable crops. Technical report, Teagasc, 7 2005.

[429] S.T.J Lalor, B. S. Coulter, G. Quinlan, and L. Connolly. A survey of fertilizer use in Ireland from 2004-2008 for grasslands and arable crops. Technical report, Teagasc, 3 2010. 98

[430] Emma Dillon, Cathal Buckley, Brian Moran, John Lennon, and David Wall. Teagasc national farm survey, fertiliser use survey 2005-2015. Technical report, Teagasc, 5 2018. 98

[431] Ireland CSO. Table ACEN2: Area used by rural district, farm land utilisation and year (1926–1980), 2007. `https://data.cso.ie/`. 98

[432] Ireland CSO. Table AQA02: Farm land utilisation in june by type of land use, year and region (1980–1999), 2020. `https://data.cso.ie/`. 98

[433] Ireland CSO. Table AQA01: Area farmed in june by type of land use, year and region (1991–2007), 2021. `https://data.cso.ie/`.

[434] Ireland CSO. Table AQA05: Area farmed in june by type of land use, year and region (2008–2012), 2020. `https://data.cso.ie/`.

[435] Ireland CSO. Table AQA06: Area farmed in june by type of land use, year and region (2013–2023), 2024. `https://data.cso.ie/`. 98

[436] T. Walsh, P. F. Ryan, and J. Kilroy. A half century of fertiliser and lime use in Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland*, 19:104–136, 1957. 98, 114

[437] John. F. Heavy. The economic optimum use of fertilisers in Ireland. Technical Report 2, Fertiliser Association of Ireland, 10 1969. 98

[438] W. H Prins. *Limits to nitrogen fertilizer on grassland*. PhD thesis, 3 1983. 99, 114

[439] Eliza Kurek. The fertilizing of main crops in peasant farming and its effectiveness. *Zagadnienia Ekonomiki Rolnej*, 3:59–71, 1971. 99, 114

[440] Pilar García-Serrano Jiménez, Juan José Lucena Marotta, Sebastián Ruano Criado, and Mariano Nogales García. *Guía Práctica de la Fertilización Racional de los Cultivos en España. Parte 1.* Ministerio de Medio Ambiente y Medio Rural y Marino, 2010. 100, 115

[441] Ewert Åberg. Recent changes in swedish crop production. *Advances in Agronomy*, 7:39–74, 1955. 100

[442] B. M. Church and J. Webber. Fertiliser practice in england and wales: A new series of surveys. *Journal of the Science of Food and Agriculture*, 22:1–7, 1971. 100, 115

[443] DEFRA. United Kingdom land areas, livestock numbers and agricultural workforce on agricultural holdings on 1 June, 2024. `https://www.gov.uk/government/statistical-data-sets/structure-of-the-agricultural-industry-in-england-and-the-uk-at-june`. 100

[444] DEFRA. Crops areas and livestock numbers in England from the June Census of Agriculture: 1900-2010, 2024. https://www.gov.uk/government/statistical-data-sets/structure-of-the-agricultural-industry-in-england-and-the-uk-at-june. 100, 115

[445] Torfi Jóhannesson. *Agriculture in Iceland Conditions and Characteristics*. The Agricultural University of Iceland, 2010. 101, 115

[446] A. Helgadóttir, E. Eythórsdóttir, and T. Jóhannesson. Agriculture in iceland - a grassland based production. *Grassland Science in Europe*, 18:30–43, 2013. 101, 115

[447] Office fédéral de la statistique. Surface agricole utile sans les alpages, 1 2024. https://www.bfs.admin.ch/bfs/fr/home/statistiques/agriculture-sylviculture/agriculture.assetdetail.30245951.html. 101

[448] Asbjørn Nordgård. Orientation and intensity of Norwegian agriculture. *Norsk Geografisk Tidsskrift*, 29:169–220, 1975. 101

[449] H Steinshamn, L Nesheim, and A K Bakken. Grassland production in Norway. *Grassland Science in Europe*, 21:15–25, 2016. 101

[450] The Federal Administration for Plant Protection and Veterinarian Medicine. Yugoslavia: Country report to the fao international technical conference on plan genetic resources (leipzig, 1996). Technical report, 11 1995. 101, 115

[451] Z. Lugić, D. Lazarević, P. Erić, V. Mihajlović, and S. Vučković. The state of forage crops production in Serbia. *Biotechnology in Animal Husbandry*, 26:29–47, 2010. 101, 102, 115

[452] G. Loza and I. Kurtsev. The growth of productive forces in agriculture in the tenth five-year plan. *Problems in Economics*, 19:3–22, 1977. 102, 115

[453] W. Klatt. Reflections on the 1975 soviet harvest. *Soviet Studies*, 28:485–498, 1976. 102

[454] Jacklyn Y. Shend. *Agricultural statistics of the former USSR republics and the Baltic States*. Number 863 in Statistical bulletin (United States Department of Agriculture). U.S. Department of Agriculture, 1993. 102, 115

[455] Mekhlis Suleimenov and Peter Oram. Trends in feed, livestock production, and rangelands during the transition period in three Central Asian countries. *Food Policy*, 25:681–700, 2000. 102

[456] F. Zhang, J. Qi, F. M. Li, C. S. Li, and C. B. Li. Quantifying nitrous oxide emissions from chinese grasslands with a process-based model. *Biogeosciences*, 7:2039–2050, 2010. 103

[457] Atushi Ushimaru, Kei Uchida, and Takeshi Suka. *Grassland Biodiversity in Japan: Threats, Management and Conservation*. Taylor & Francis Group, 6 2017. 103

[458] Bae Hun Lee, Ji Yung Kim, Kyung Il Sung, and Byong Wan Kim. Investigation on the actual state of grassland in Republic of Korea. *Journal of The Korean Society of Grassland and Forage Science*, 39, 2019. 103

[459] FAO. Fertilizer use by crop in Indonesia. Technical Report TC/D/Y7063E/1/05.05/300, FAO, Viale delle Terme di Caracalla 00100 Rome, 2005. 104

[460] P. K. Ghosh, S. K. Mahanta, and S. N. Ram. Nitrogen dynamics in grasslands. In Yash P. Abrol, Tapan K. Adhya, Viney P. Aneja, Nandula Raghuram, Himanshu Pathak, Umesh Kulshrestha, Chhemendra Sharma, and Bijay Singh, editors, *The Indian Nitrogen Assessment: Sources of Reactive Nitrogen, Environmental and Climate Effects, Management Options, and Policies*, pages 187–205. Elsevier, 2017. 104

[461] Muhammad Irfan and Nighat Hasnain. Nitrogen emissions from agriculture sector in Pakistan: context, pathways, impacts and future projections. In Tariq Aziz, Abdul Wakeel, Muhammad Arif Watto, Muhammad Sanaullah, Muhammad Aamer Maqsood, and Aysha Kiran, editors, *Nitrogen Assessment: Pakistan as a Case-Study*, chapter 6, pages 99–125. 2022. 104

[462] Ahmed S. Elrys, Mohamed K. Abdel-Fattah, Sajjad Raza, Zhujun Chen, and Jianbin Zhou. Spatial trends in the nitrogen budget of the african agro-food system over the past five decades. *Environmental Research Letters*, 14, 2019. 104

[463] FAO. Fertilizer use by crop in Egypt. Technical Report TC/D/Y5863E/1/01.05/300, FAO, Viale delle Terme di Caracalla, 00100 Rome, Italy, 2005. 104

[464] Hadi S. Esfahani. Aggregate trends in four main agricultural regions in Egypt, 1964–1979. *International Journal of Middle East Studies*, 20:135–164, 1988. 104

[465] M Bounejmate. *The Role of Legumes in the Farming Systems of the Mediterranean Areas: proceedings of a workshop on the role of legumes in the farming systems of the Mediterranean areas, UNDPI ICARDA*, volume 38 of *Developments in plant and soil science*, chapter The role of legumes in the farming systems of Morocco, pages 85–93. Kluwer Academic Publishers, 1 edition, 1990. 104

[466] FAO. Utilisation des engrais par culture au maroc. Technical Report TC/D/A710F/1/10.06/300, FAO, Viale delle Terme di Caracalla, 00100 Rome, Italy, 2006. 104, 115

[467] Maria Niedertscheider, Simone Gingrich, and Karl Heinz Erb. Changes in land use in South Africa between 1961 and 2006: An integrated socio-ecological analysis based on the human appropriation of net primary production framework. *Regional Environmental Change*, 12:715–727, 2012. 105

[468] A. Smith and J. M.L.C. Rhind. Eight decades of pasture plant improvement in South Africa. *Journal of the Grassland Society of Southern Africa*, 1:25–28, 1984. 105

[469] United Nations. Land area, 2024. `https://data.un.org/Data.aspx?d=FAO&f=itemCode:6601&c=2,4,5,6,7&s=countryName:asc,elementCode:asc,year:desc&v=1`. 109

[470] FAOSTAT. Definitions and standards. country group 2022, 2022. `https://data.apps.fao.org/catalog/iso/457196b9-1d93-410d-8ad3-a57aadf09a1a`. 109

[471] World Bank. Agricultural irrigated land (% of total agricultural land), 2023. `https://data.worldbank.org/indicator/AG.LND.IRIG.AG.ZS`. 110

[472] World Bank. Agricultural machinery, tractors, 2023. https://data.worldbank.org/indicator/AG.AGR.TRAC.NO. 110

[473] FAOSTAT. Land use, 2022. https://www.fao.org/faostat/en/#data/RL. 110

[474] World Bank. Goverment expenditure in education, total (% of GDP), 10 2023. https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS. 110

[475] United Nations. Per capita gdp at current prices - us dollars, 2023. https://data.un.org/Data.aspx?d=SNAAMA&f=grID:101;currID:USD;pcFlag:1&c=2,3,5,6&s=_crEngNameOrderBy:asc,yr:desc&v=1. 110

[476] FAOSTAT. Annual population, 2023. https://www.fao.org/faostat/en/#data/OA. 110

[477] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. 118

[478] Sean Gillies et al. Rasterio: geospatial raster i/o for Python programmers, 2013–. "https://github.com/mapbox/rasterio". 118

[479] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/. 118

[480] Edzer Pebesma and Roger Bivand. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC, 2023. 10.1201/9780429459016. 118

[481] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. 118

[482] David Pierce and Maintainer David Pierce. Package 'ncdf4'. 2019. https://www.vps.fmvz.usp.br/CRAN/web/packages/ncdf4/ncdf4.pdf. 118

[483] Daniel Baston. *exactextractr: Fast Extraction from Raster Datasets using Polygons*, 2020. https://CRAN.R-project.org/package=exactextractr. 118

[484] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*, 2023. https://readxl.tidyverse.org,https://github.com/tidyverse/readxl. 118

[485] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2023. https://github.com/tidyverse/stringr. 118

[486] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023. https://github.com/tidyverse/dplyr. 118

[487] Hadley Wickham, Jim Hester, and Jennifer Bryan. *readr: Read Rectangular Text Data*, 2024. https://readr.tidyverse.org. 118

[488] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. https://ggplot2.tidyverse.org. 118

[489] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. 118

[490] Nils B. Weidmann, Doreen Kuse, and Kristian Skrede Gleditsch. The geography of the international system: The cshapes dataset. *International Interactions*, 36(1):86–106, 2010. 118

[491] Jingguang Zhou and Zili Huang. Recover missing sensor data with iterative imputing network. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 120

[492] Joseph Kearney, Shahid Barkat, and Arnab Bose. kearnz/autoimpute: Python package for imputation methods. 120