Domain-Informed Neural Networks for Earth Observation

Thesis approved by
the Department of Computer Science
RPTU University Kaiserslautern-Landau
for the award of the Doctoral Degree

Doctor of Engineering (Dr.-Ing.)

to

Miro Benjamin Miranda Lorenz

Date of Defense:	February 20, 2026
Dean of Dept.:	Prof. Dr. Christoph Garth
Reviewers:	Prof. Dr. Prof. h.c. Andreas Dengel
	Prof. Dr. Matias Valdenegro Toro
Head of Committee:	Prof. Dr. Jörg Dörr

DE-386

Miro Benjamin Miranda Lorenz
Domain-Informed Neural Networks for Earth Observation

Contact: mmlorenz@rptu.de

RPTU University Kaiserslautern-Landau
Department of Computer Science
Gottlieb-Daimler-Straße 47
67663 Kaiserslautern

„Da steh ich nun, ich armer Tor!
Und bin so klug als wie zuvor.“

- Faust. *Der Tragödie erster Teil*

Für meine Liebsten.

ABSTRACT

Driven by the increased availability of Earth Observation data, Deep Learning has been increasingly adopted in Earth Observation applications. At the same time, we observe a tendency towards larger Deep Learning models, powered by ever larger datasets. Such purely data-driven models are exceptionally powerful. However, scenarios exist where purely data-driven methods reach limits. For instance, when data is insufficient, when physical principles must be considered, or when information about the uncertainty of a prediction is required. All these aspects nourish skepticism and continue to hinder the success of Deep Learning approaches in Earth Observation applications.

This thesis explores the inclusion of prior knowledge into a learning system, defined as *Domain-Informed Learning*. Here, prior knowledge is a source of information that exists independently of the model. This work focuses on agricultural applications and time series analysis, and develops methods that are extendable to a variety of Earth Observation tasks. This work introduces a novel, large-scale dataset for crop yield prediction using Earth Observation data. Following, three techniques of integrating prior knowledge are explored, namely 1) *data space enrichment*, 2) *conditional learning*, and 3) *uncertainty estimation*. The first technique analyzes data sources and time series representations for crop yield prediction. Furthermore, novel data fusion methods are presented. In the following, this work analyzes conditional learning with prior knowledge. A novel physics-guided approach for drought stress estimation is proposed. The last part emphasizes the inherent variability in Earth Observation tasks. Therefore, the concept of uncertainty estimation is introduced by focusing on missing data and distribution shifts. We present a novel method for uncertainty estimation, inspired by naturally occurring missing time steps. Finally, we overcome the performance collapse under distribution shift by coupling Bayesian inference with prior knowledge.

In conclusion, this thesis contributes to the field of research by making models more reliable, easier to understand, and more trustworthy. It offers new perspectives on Earth Observation and emphasizes the importance of understanding how confident our predictions are and that they remain consistent with real-world physical laws.

ACKNOWLEDGEMENTS

Today, I feel a deep sense of happiness and gratitude. I am thankful for the opportunity to write this thesis and for every person who will take the time to read this work.

First, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Andreas Dengel, for providing an academic environment characterized by freedom, trust, and opportunities for growth. I am grateful for the many discussions and his valuable feedback, which helped me stay focused and move forward with clarity. Throughout this journey, I always felt supported, and I deeply appreciate the inspiration that kept me motivated along the way.

I would also like to thank my second supervisor, Prof. Dr. Matias Valdenegro-Toro, for his guidance, expertise, and the continuous discussions that shaped this work in many positive ways. I am especially grateful for the opportunity to spend a research stay at the University of Groningen and for the warm welcome I received from his research group. The inspiring atmosphere and collaborative spirit made this experience truly memorable, and I will always remember these moments.

My heartfelt thanks go to Dr. Marcela Charfuelan, who supported me in every possible way. I am deeply grateful for her constant availability, encouragement, and the many insightful discussions from which I learned so much. I truly appreciate the opportunity to work in an environment defined by patience, respect, and equality.

Furthermore, I would like to thank all my colleagues from the Yield Consortium and the ESA-Lab team for the past years filled with discussions, shared memories, and enjoyable conference trips.

LIST OF PUBLICATIONS

Parts of the research presented in this thesis, including figures and tables, have already been published in the following peer-reviewed publications:

Publications Related to the Thesis

- **M. Miranda**, D. Pathak, P. Helber, B. Bischke, H. Najjar, F. Mena, C. Sanchez, M. Valdenegro-Toro, M. Charfuelan, M. Nuske, and A. Dengel. *YieldSAT: A Multimodal Benchmark Dataset for High-Resolution Crop Yield Prediction.* Accepted in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026.
- **M. Miranda**, M. Charfuelan, M. Valdenegro-Toro, and A. Dengel. *Informed Learning for Estimating Drought Stress at Fine-Scale Resolution Enables Accurate Yield Prediction.* European Conference on Artificial Intelligence (ECAI), pp. 5384–5391. IOS Press, 2025.
- **M. Miranda**, A. Dinesh, D. N. Lesmes-Leon, F. Mena, M. Charfuelan, and A. Dengel. *regDiff: Regression Diffusion for Earth Observation.* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2025.
- **M. Miranda**, F. Mena, M. Charfuelan, and A. Dengel. *Informed Learning for Efficient Crop Yield Prediction.* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2025.
- **M. Miranda**, F. Mena, and A. Dengel. *An Analysis of Temporal Dropout in Earth Observation Time Series for Regression Tasks.* International Symposium on Intelligent Data Analysis (IDA), pp. 389–402. Springer, 2025.
- A. Münzberg* and **M. Miranda***. *Vorhersage von Landwirtschaftlichen Erträgen und Wachstum.* In: Hybride KI mit Machine Learning und

*Shared first authorship.

Knowledge Graphs: Innovative Lösungen aus der Praxis, pp. 153–167. Springer, 2025.

- **M. Miranda**, M. Charfuelan, and A. Dengel. *Exploring Physics-Informed Neural Networks for Crop Yield Loss Forecasting.* NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning, 2024.

- **M. Miranda***, D. Pathak*, M. Nuske, and A. Dengel. *Multi-Modal Fusion Methods with Local Neighborhood Information for Crop Yield Prediction at Field and Subfield Levels.* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4307–4311. IEEE, 2024.

- D. Pathak*, **M. Miranda***, F. Mena, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, H. Najjar, J. Siddamsetty, D. Arenas, M. Vollmer, M. Charfuelan, M. Nuske, and A. Dengel. *Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level.* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2767–2770. IEEE, 2023.

- **M. Miranda**, L. Drees, and R. Roscher. *Controlled Multi-Modal Image Generation for Plant Growth Modeling.* 26th International Conference on Pattern Recognition (ICPR), pp. 5118–5124. IEEE, 2022.

Publications with Partial Contribution

- H. Najjar, **M. Miranda**, M. Nuske, R. Roscher, and A. Dengel. *Explainability of Sub-Field Level Crop Yield Prediction using Remote Sensing.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025.

- F. Mena, D. Arenas, **M. Miranda**, and A. Dengel. *On What Depends the Robustness of Multi-Source Models to Missing Data in Earth Observation?* IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2025.

- F. Mena, D. Pathak, H. Najjar, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, **M. Miranda**, J. Siddamsetty, M. Nuske, M. Charfuelan, D. Arenas, M. Vollmer, and A. Dengel. *Adaptive Fusion of Multi-Modal Remote Sensing Data for Optimal Sub-Field Crop Yield Prediction.* Remote Sensing of Environment, 318:114547, 2025.

- P. Helber, B. Bischke, P. Habelitz, C. Sanchez, D. Pathak, **M. Miranda**, H. Najjar, F. Mena, J. Siddamsetty, D. Arenas, M. Vollmer, M. Charfuelan, M. Nuske, and A. Dengel. *Crop Yield Prediction: An Operational Approach to Crop Yield Modeling on Field and Subfield Level with Machine Learning Models*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2763–2766. IEEE, 2023.
- C. Sanchez, D. Pathak, **M. Miranda**, M. Charfuelan, P. Helber, M. Nuske, B. Bischke, P. Habelitz, N. Rahman, F. Mena, H. Najjar, J. Siddamsetty, D. Arenas, M. Vollmer, and A. Dengel. *Influence of Data Cleaning Techniques on Sub-Field Yield Predictions*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4852–4855. IEEE, 2023.

CONTENTS

ACRONYMS

ADM Additional Data Modality

BNN Bayesian Neural Network

BBB Bayes By Backprop

CNN Convolutional Neural Network

cGAN conditional Generative Adversarial Network

CV Cross-Validation

contGAN controllable GAN

ConvLSTM Convolutional LSTM

DE Deep Ensemble

DEM Digital Elevation Map

DL Deep Learning

DIL Domain-Informed Learning

DM Diffusion Model

ECE Expected Calibration Error

EO Earth Observation

ESA European Space Agency

ET Evapotranspiration

FAO Food and Agriculture Organization

FOV Factors of Variation

FF Feature Fusion

FCL Fully Connected Layer

FID Fréchet Inception Distance

GAN Generative Adversarial Network

GCF Ground Cover Fraction

IF Input Fusion

KL Kullback-Leibler

LFMC Life Fuel Moisture Content

LORO Leave-One-Region-Out

LOYO Leave-One-Year-Out

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

PICP Prediction Interval Coverage Probability

MC Monte Carlo

MSE Mean Squared Error

ML Machine Learning

MLP Multilayer Perceptron

MPIW Mean Prediction Interval Width

MC-TD Monte Carlo Temporal Dropout

MC-cTD Monte Carlo Concrete Temporal Dropout

NLL Negative Log-Likelihood

NN Neural Network

NDVI Normalized Difference Vegetation Index

NDWI Normalized Difference Water Index

PM2.5 Particle Matter 2.5

PICP Prediction Interval Coverage Probability

PG Physics-Guided

pp percentage points

PU Predictive Uncertainty

QICE Quantile Interval Coverage Error

RF Random Forest

RegDiff Regression Diffusion

RNN Recurrent Neural Network

RMSE Root-Mean-Square Error

RRMSE Relative Root-Mean-Square Error

R^2 Coefficient of Determination

RS Remote Sensing

SAR Synthetic Aperture Radar

S1 Copernicus Sentinel-1

S2 Copernicus Sentinel-2

SDG Sustainable Development Goal

SCL Scene Classification Layer

TD Temporal Dropout

t-SNE t-Distributed Stochastic Neighbor Embedding

VAE Variational Autoencoder

VI Vegetation Index

LIST OF FIGURES

LIST OF TABLES

PART I

INTRODUCTION & BACKGROUND

1 | INTRODUCTION

1.1 Motivation

The growing number of conflicts, intensified competition for resources, and changing climate conditions are combined in ways that are both far-reaching and destructive [1]. The so-called *polycrisis* emerges when rapidly developing events coincide with slowly unfolding stresses, shifting a system from equilibrium into a volatile state of disequilibrium [1].

Polycrisis

Agriculture, for example, exists within a field of global tension, being significantly affected by polycrisis. First, conflicts and wars. The war in Ukraine, for instance, has had an outsized impact on the global food system, causing spikes in food and energy prices due to the major contributions of the involved parties to fuel, fertilizer, and essential food commodity exports [2]. Immediately after the onset of the war, global food markets experienced a significant surge in prices, as reflected in the Food and Agriculture Organization (FAO) *Food Price Index (FFPI)*[1] reached the highest level ever recorded [3, 4]. Second, the frequency of natural disasters such as floods extreme temperatures, wildfires, and droughts is a major concern that is suspected to increase economic damages [5, 6]. Linked to slowly changing climate conditions, agricultural productivity is under pressure, resulting in severe economic and production losses every year [7]. Third, by around 2050, the world population is expected to reach approximately 10 billion people [8]. Simultaneously, more than 60 % of the world population is projected to live in cities by 2050 [9]. Resulting populous cities (megacities) already face severe challenges, such as air pollution (exposure to fine particulate matter), which is suspected to negatively impact health and life expectancy [10]. Moreover, with fewer people living in rural areas, the availability of the workforce for food production is constantly declining, leading to the fourth development: a decline in the workforce. Already today, the number of farms, farmers, and people employed in the agricultural sector in the *European Union (EU)* has declined to 4.2 % of the total employment, with an annual decline rate of 2.5% [11, 12]. At the same time, the average farm size is growing continuously [11]. Ultimately, fewer individuals must manage

Conflicts & wars

Climate change

Population growth

Decline in workforce

[1] https://www.fao.org/worldfoodsituation/foodpricesindex/en/ [accessed: April 17, 2026]

Requirements

more land by increasing productivity with increased sustainability to achieve a world without hunger and malnutrition, a core objective of the Sustainable Development Goals (SDGs) (SDG 2- zero hunger, SDG 13- climate action) [13]. A polycrisis, however, is a serious threat to this goal and affects different stakeholders at different levels. Politicians require novel tools to quickly adapt to emerging socioeconomic challenges, manage disaster response, and support long-term planning. At the same time, industry stakeholders, such as insurance companies, need new techniques to estimate yield losses at scale and at low cost. Finally, farmers require innovative tools to support decision-making and management practices, to compensate for the declining workforce, and to adapt to changing climate conditions.

Earth observation & Machine Learning

Many of the above-mentioned stressors can be captured from space using Earth Observation (EO) technologies. Earth-orbiting satellites, ground-based sensors, and Remote Sensing (RS) technologies continuously deliver information about planet Earth at a global scale with high spatial resolution and high temporal frequency. For instance, satellite programs like the *Copernicus Sentinel missions* [2], already published petabytes of openly accessible data that can be used for many scientific disciplines such as land monitoring, disaster control, climate science, and agriculture [14]. However, the large volumes of EO data are highly complex with varying temporal, spatial, and spectral resolution. To process such data, dedicated and scalable methods are required, such as data-driven Machine Learning (ML) and Deep Learning (DL) technologies. Through the ability to extract patterns from complex, high-dimensional data, ML is expected to enhance informed decision-making in EO applications [15]. Nevertheless, many EO applications go beyond the development of actionable end-to-end models. Instead, the main goal is to develop hypotheses and theories from learned patterns and relationships to advance scientific knowledge and discovery [16].

Despite their potential, especially DL methods are often criticized for their *black-box* nature, with little interpretability and knowledge conformity. In fact, there are scenarios where purely data-driven methods are limited. For example, when insufficient data is available, when physical principles must be considered, or when information about the uncertainty of a prediction is required [17, 18]. Moreover, there is a high demand to make predictions more explainable [19]. A main limitation of common DL models is that they tend to produce solutions that might be inconsistent with existing knowledge and their

[2] https://sentinels.copernicus.eu/ [accessed: April 17, 2026]

inability to provide sufficient justification of the discovered patterns [20]. This characteristic can have severe consequences, leading to skepticism that limits the success of DL approaches. On the other side, various scientific disciplines, such as agriculture, have a long-standing tradition of research and knowledge discovery. For instance, the cause-effect relationship between water availability and crop yield can be approximated by multiple differential equations. However, this knowledge is often overlooked by purely data-driven methods, limiting their knowledge-awareness and potentially even their accuracy. Consequently, in the light of the growing dominance of data-driven models, a question arises: *To what extent is prior knowledge relevant to overcome the limitations of purely data-driven models in EO tasks today?*

This thesis explores the importance of prior knowledge in ML for EO applications. This thesis considers prior knowledge as a source of information that exists independently of the learning algorithm. This paradigm, defined as *Domain-Informed Learning (DIL)*, exploits the richness of existing knowledge to improve the effectiveness and ultimately to increase trust in ML approaches. Only limited research has systematically evaluated the benefits of integrating prior knowledge into data-driven methods in EO applications. A major problem arises from the different knowledge sources and types of knowledge representations that formulate a cause-effect mechanism between physical variables. Likewise, different knowledge integration practices exist within the learning models, ranging from the data space and model architecture to model regularization and the final hypothesis. It is an open question which knowledge sources and integration practices are useful in DIL for EO applications.

Prior knowledge

Research Gap

Finally, the EO domain has very specific requirements and challenges arising from the characteristics and inhomogeneity of the satellite-derived data. Unlike other domains, EO data is multimodal and high-dimensional with varying temporal, spatial, and spectral resolutions. This requires highly specialized model architectures and data fusion techniques. This uniqueness further increases the difficulty of integrating prior knowledge.

1.2 Research Questions & Goals

The main objective of this thesis is to analyze the importance of prior knowledge for EO applications by extending existing research. The core research question of this thesis is :

> **Core Research Question**
> Can the integration of prior knowledge increase the effectiveness of ML in EO applications?

To answer this question, this thesis systematically evaluates different knowledge sources, representations, and integration practices. This thesis focuses on agricultural applications, including yield prediction, plant growth modeling, and drought stress estimation. Additionally, this thesis explores further applications of climate science, including wildfire risk forecasting and air pollution prediction. Specifically, this thesis focuses on the three sub-questions:

1. **Research Question (RQ1):** How can data space enrichment be used for large-scale crop yield prediction at the field and subfield level?
   Goal: Investigating various data modalities and data representations while considering different temporal and spatial resolutions with novel data fusion schemes. Simultaneously, the various input time series representations will be analyzed with respect to important physiological properties.

2. **Research Question (RQ2)**: How can prior knowledge be used as conditions to increase the knowledge conformity and physical consistency?
   Goal: Studying different conditions (e.g., natural laws) to address the *black-box* nature of DL models to explicitly increase the explainability and consistency with physical conditions. Explicitly exploring the integration of prior knowledge into the learning algorithm for regularization.

3. **Research Question (RQ3):** How can uncertainty estimation quantify the unknown and shape the final hypothesis to increase trust in data-driven predictions?
   Goal: Studying different uncertainty estimation methods while proposing novel approaches that build upon important characteristics of the EO domain. Investigate the impact of distribution shifts to quantify the effectiveness of DL methods.

To answer the research questions and quantify the success of the studied and presented methods, the thesis particularly focuses on:

1. **Superior performance**: The inclusion of domain knowledge should result in an improvement in the quantitative or qualitative evaluation metrics. Moreover, the integration should increase the robustness of the model to limited training data, distribution shifts, and missing data.

2. **Increased explainability & Trustworthiness**: A domain-informed model should be more explainable and trustworthy compared to its data-driven counterpart. This includes increased knowledge conformity, and physical consistency with the modeled process. A model should tell the user which predictions to trust and which not, or be constrained to a physically plausible solution space.

3. **Novel perspectives**: The proposed method should provide new insights into the application that were not available previously and that result in new perspectives for future research or knowledge discovery.

1.3 Thesis Contribution

Parts of the research results that are presented in this work, including figures and tables, have already been published in peer-reviewed conferences, journal articles, and book chapters, or are currently under review. In detail, the thesis presents the following contributions for which I have been the main contributor:

1. **Contribution for RQ1:** The thesis presents an analysis of input modalities and models for crop yield prediction. The thesis presents a novel attention-based feature fusion method that incorporates local spatial context to improve crop yield prediction. Furthermore, the thesis investigates time series representations and proposes a novel sampling method based on physiological properties. This part is based on the publications:

   - **M. Miranda**, F. Mena, M. Charfuelan, and A. Dengel. Informed Learning for Efficient Crop Yield Prediction. In *IEEE IGARSS International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2025.
   - A. Münzberg* and **M. Miranda***. Vorhersage von Landwirtschaftlichen Erträgen und Wachstum. In *Hybride KI mit Machine Learning und Knowledge Graphs: Innovative Lösungen aus der Praxis*, pages 153–167. Springer, 2025.
   - H. Najjar, **M. Miranda**, M. Nuske, R. Roscher, and A. Dengel. Explainability of Sub-Field Level Crop Yield Prediction using Remote Sensing. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

*Shared first authorship.

- **M. Miranda***, D. Pathak*, M. Nuske, and A. Dengel. Multimodal Fusion Methods with Local Neighborhood Information for Crop Yield Prediction at Field and Subfield Levels. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4307–4311. IEEE, 2024.

- D. Pathak*, **M. Miranda***, F. Mena, C. Sanchez, P. Helber, B. Bischke, P. Habelitz, H. Najjar, J. Siddamsetty, D. Arenas, M. Vollmer, M. Charfuelan, M. Nuske, and A. Dengel. Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 2767–2770. IEEE, 2023.

2. **Contribution for RQ2:** The thesis presents novel methods for integrating prior conditions into the learning algorithm, explicitly for regularization. A novel method for estimating drought stress and yield loss is proposed, leveraging fundamental natural laws. This part discusses a conditional Generative Adversarial Network (GAN) for multimodal image generation in plant growth modeling, utilizing a novel loss function. Further, the part presents a novel conditional Diffusion Model (DM) for time series regression. This part is based on the publications:

   - **M. Miranda**, M. Charfuelan, M. Valdenegro-Toro, and A. Dengel. Informed Learning for Estimating Drought Stress at Fine-Scale Resolution Enables Accurate Yield Prediction. In *European Conference on Artificial Intelligence (ECAI)*, pages 5384–5391. IOS Press, 2025.

   - **M. Miranda**, A. Dinesh, D. N. Lesmes-Leon, F. Mena, M. Charfuelan, and A. Dengel. regDiff: Regression Diffusion for Earth Observation. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2025.

   - **M. Miranda**, M. Charfuelan, and A. Dengel. Exploring Physics-Informed Neural Networks for Crop Yield Loss Forecasting. In *NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024.

   - **M. Miranda**, L. Drees, and R. Roscher. Controlled Multi-Modal Image Generation for Plant Growth Modeling. In *26th International Conference on Pattern Recognition (ICPR)*, pages 5118–5124. IEEE, 2022.

3. **Contribution for RQ3:** This part investigates different uncertainty estimation methods and proposes a novel method that leverages missing time steps in EO time series data. Furthermore, the thesis explores the impact of distribution shifts and proposes potential mitigation strategies based on prior knowledge. This part is based on, or currently under review in, the following publications:

    - **M. Miranda**, D. Pathak, P. Helber, B. Bischke, H. Najjar, F. Mena, C. Sanchez, M. Valdenegro-Toro, M. Charfuelan, M. Nuske, and A. Dengel. YieldSAT: A Multimodal Benchmark Dataset for High-Resolution Crop Yield Prediction. Accepted in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

    - **M. Miranda**, F. Mena, and A. Dengel. An Analysis of Temporal Dropout in Earth Observation Time Series for Regression Tasks. In *International Symposium on Intelligent Data Analysis (IDA)*, pages 389–402. Springer, 2025

4. **Further Contribution:** The thesis introduces a novel publicly available dataset for high-resolution crop yield prediction with EO data. The dataset contains information about multiple countries, crop types, and years. This work is under review in:

    - **M. Miranda**, D. Pathak, P. Helber, B. Bischke, H. Najjar, F. Mena, C. Sanchez, M. Valdenegro-Toro, M. Charfuelan, M. Nuske, and A. Dengel. YieldSAT: A Multimodal Benchmark Dataset for High-Resolution Crop Yield Prediction. Accepted in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Several contributions are published as open source to foster research in this direction. The following contributions are available open-source or are under review and will be released after acceptance:

- **YieldSAT**, Dataset, presented in Section 2.2
  `https://yieldsat.github.io/`

- **Physics-Guided LSTM**, Code, presented in Chapter 6
  `https://github.com/mmiranda-l/Yield-Loss`

- **MC-TD**, Code, presented in Chapter 9
  `https://github.com/mmiranda-l/Temporal-Dropout`

I was further involved in collaborations that led to peer-reviewed publications and that contribute to the overall storyline of this thesis. Each contribution will be acknowledged in the respective chapters and is listed in List of Publications.

> **Writing style**
> For the rest of the thesis and for the ease of reading, I will use *"we"* to refer to the reader, myself, and the co-authors of the articles I authored.

1.4 Overview of the Thesis

The remaining content of this thesis is divided into five parts, structured along the three research questions outlined earlier. Additionally, preliminary and background information is provided to follow the content of this work. Finally, this work concludes with a summary and an outlook on future research. An overview of the thesis is given in Figure 1.1.

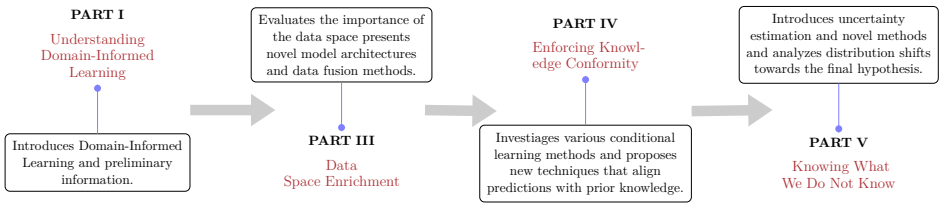


Figure 1.1: Overview of the thesis

PART I – INTRODUCTION & BACKGROUND

Chapter 2 provides an overview of EO data sources and applications. Presents the *YieldSAT* datasets in the light of the current literature.

Chapter 3 introduces the notation that is used throughout this work and introduces important concepts of ML and DL. Furthermore, the concepts of DIL are discussed in the light of the current literature.

PART II – DATA SPACE ENRICHMENT

Chapter 4 presents an extensive analysis of input modalities and models for crop yield prediction and proposes a novel fusion method that leverages local neighborhood information.

Chapter 5 explores different time series sampling methods for crop yield prediction and proposes a novel method based on physiological processes.

PART III – ENFORCING KNOWLEDGE CONFORMITY

Chapter 6 presents a novel approach for temporal crop drought stress estimation that enables explainable and physically consistent yield predictions.

Chapter 7 discusses a generative method for controlled image generation for temporal plant growth modeling by integrating image priors. Furthermore, a conditional diffusion approach is presented for time series regression.

PART IV – KNOWING WHAT WE DO NOT KNOW

Chapter 8 provides a general overview of uncertainty estimation for EO regression tasks.

Chapter 9 presents a novel method for uncertainty estimation, leveraging naturally occurring missing time steps in EO data.

Chapter 10 explores different uncertainty estimation methods for crop yield prediction. Furthermore, the impact of distribution shifts is assessed by finding explanations and potential mitigation strategies.

PART V – CONCLUSION

Chapter 11 summarizes the important results and findings of this thesis and outlines potential future research directions.

2 EARTH OBSERVATION & APPLICATIONS

Chapter Highlights:

1. A general overview of EO data and remote sensing technologies.
2. EO data is multimodal and high-dimensional, with varying temporal, spatial, and spectral resolutions.
3. YieldSAT is a novel dataset for crop yield prediction with EO data at scale and high resolution.

In this chapter, an overview of popular EO data sources and satellite missions is given. Additionally, the applications and datasets that are used in the thesis are presented and discussed. Specifically, we present the *YieldSAT* dataset, a novel satellite benchmark for crop yield prediction.

2.1 Satellite Missions

Countless satellite missions are currently operating and continuously deliver information about the Earth's surface from various perspectives. The amount of available EO is growing exponentially and is often freely accessible. For instance, satellite programs like the *Copernicus Program*[1] have reported that the *Sentinel missions*[2] alone have already published approximately 51 petabytes (PB) since the start of the program [14].
Together with other satellite missions, the *New Space Era* enables new perspectives on Earth [21]. A detailed description of satellite technologies is beyond the scope of this thesis; still, we briefly describe the satellite missions that are relevant to this work. The development of accessible EO data and its estimated future growth are depicted in Figure 2.1.

[1] https://www.copernicus.eu/en [accessed: April 17, 2026]
[2] https://sentinels.copernicus.eu/ [accessed: April 17, 2026]

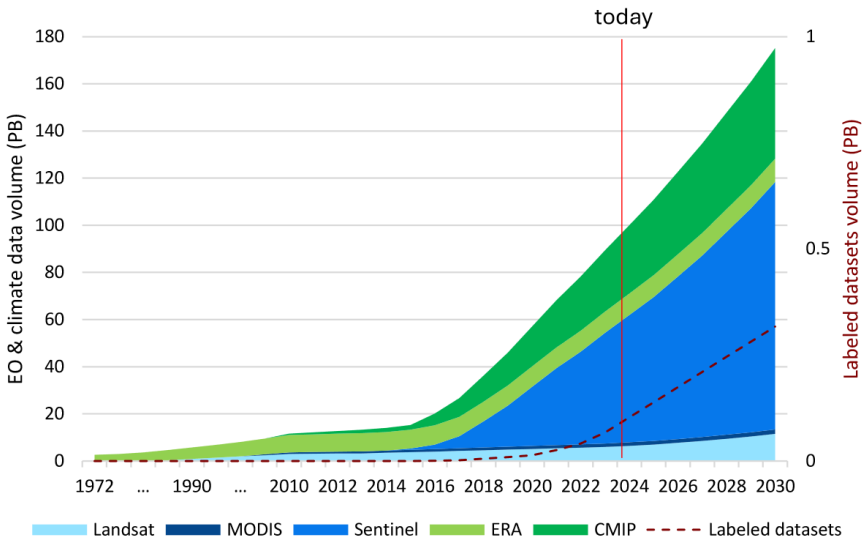


Figure 2.1: Development and estimated future volume of accessible EO data in petabytes (PB). Further, the volume of labeled datasets (dashed line) is shown. (Figure taken from [22])

2.1.1 Optical & Multispectral Imagery

Several satellite missions exist that capture optical and multispectral data. Optical and multispectral sensors record light reflected from specific parts of the electromagnetic spectrum at the Earth's surface. Optical data refers to visible light, which spans roughly 400 to 700 nm and is usually represented by three bands: red, green, and blue (RGB). Multispectral data covers additional wavelengths beyond human vision, such as near-infrared (NIR). Key differences between datasets include spectral, spatial, and temporal resolution. Satellite missions such as the Landsat-8 and the Copernicus Sentinel-2 program provide these types of data.

Sentinel-2 Satellite Imagery

Sentinel-2 Data

The Copernicus Sentinel-2[3] satellite mission is a European initiative using two earth-orbiting satellites, *Sentinel-2A* and *Sentinel-2B*, for wide-swath, high-resolution, multispectral imaging. Launched in July 2015, this mission provides data across many wavelengths with a high revisit time of about five days at the Equator. The data is made available systematically and free of charge for public, scientific, and commercial use. Copernicus Sentinel-2 (S2) satellites capture the Earth surface at up to 10 m × 10 m spatial resolution.

[3] https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-data/sentinel-2 [accessed: April 17, 2026]

Table 2.1: Band information of the Sentinel-2 L2A satellite imagery

Band Name	Spatial Resolution	Central Wavelength (nm)
B01 - Coastal Aerosol	60 m × 60 m	443
B02 - Blue	10 m × 10 m	490
B03 - Green	10 m × 10 m	560
B04 - Red	10 m × 10 m	665
B05 - Vegetation Red Edge	20 m × 20 m	705
B06 - Vegetation Red Edge	20 m × 20 m	740
B07 - Vegetation Red Edge	20 m × 20 m	783
B08 - Near-Infrared (NIR)	10 m × 10 m	842
B8A - Vegetation Red Edge	20 m × 20 m	865
B09 - Water Vapor	60 m × 60 m	945
B10 - Short-Wave Infrared (SWIR) Cirrus	60 m × 60 m	1375
B11 - Short-Wave Infrared (SWIR)	20 m × 20 m	1610
B12 - Short-Wave Infrared (SWIR)	20 m × 20 m	2190

(a) Annual Crop (b) Permant Crop (c) Pasture (d) Forest

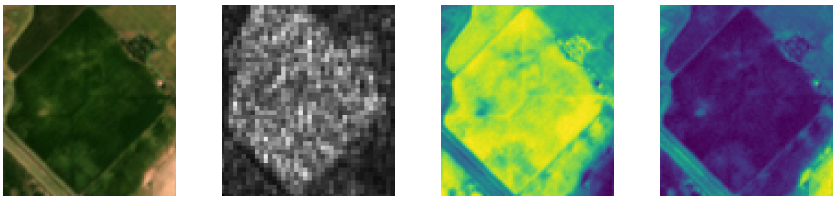
Figure 2.2: Example Sentinel-2 images with labels from the EuroSAT dataset. (Source: [23])

In total, thirteen bands are available, each covering different wavelengths. Details for these bands are listed in Table 2.1. Sample S2 images from the EuroSAT dataset illustrate different land cover classes. The twin satellites largely contributed to the recent success in EO by contributing to services and applications of agriculture, land monitoring, climate change, risk mapping, and many more[4]. Moreover, satellite data is increasingly used to quantify sustainable development, particularly in agriculture [21], as part of the SDGs.

2.1.2 Synthetic Aperture Radar

Synthetic Aperture Radar (SAR) is another data modality that is frequently used in EO applications. Unlike optical and multispectral data, SAR data is collected by an active sensor that measures backscattered microwave beams from the Earth's surface. Noteworthy, SAR sensors operate in the longer wavelength. This has several advantages over optical and multispectral data. First, SAR sensors can operate in all weather conditions by penetrating clouds

[4] https://sentiwiki.copernicus.eu/web/s2-applications [accessed: April 17, 2026]



(a) Optical Image (RGB) (b) SAR Image (c) NDVI (d) NDWI

Figure 2.3: Comparison between an optical Image in RGB, SAR, NDVI, and NDWI. Both images capture the same geographic location at the same time.

and other atmospheric disturbances. Second, they can operate at night. The Copernicus Sentinel-1 (S1) mission[5] provides SAR data in 20 m × 20 m spatial resolution and an approximate revisit time of 6 days. A comparison between a SAR (S1) image and an optical (S2) image is given in Figure 2.3. Both images display the exact same geographic location at a similar time point.

2.1.3 Derivatives of Satellite Data

Derivatives (indices) from satellite data describe relationships between spectral information and specific applications. For example, a Vegetation Index (VI) provides information related to crop growth and vegetation. These derivatives provide information that is often human-interpretable. An important example is the Normalized Difference Vegetation Index (NDVI), which provides information on the vegetation density, biomass, and crop health and is derived from multispectral data [24]. The NDVI is calculated as:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)}. \tag{2.1}$$

The NDVI ranges between -1 and 1, where very low values indicate areas of barren rock, water, or settlement, and very high values indicate dense vegetation or crops at their peak of growth. Another vegetation index is the Normalized Difference Water Index (NDWI), which is used to remotely sense vegetation liquid water and is less sensitive to atmospheric effects [25]. An example image for optical and SAR and respective derivatives is depicted in Figure 2.3. An overview of commonly used VIs for crop yield modeling is shown in Table A.2.

[5] https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1 [accessed: April 17, 2026]

2.1.4 Additional Data Modalities

Further EO data sources provide important context for many downstream tasks and complement the previous described data modalities. This data is defined as auxiliary data or Additional Data Modality (ADM) in this thesis.
An important modality is meteorological (weather) information, including temperature, precipitation, solar radiation, or wind speed. The *European Center for Medium Range Weather Forecasts (ECMWF)*[6] is an archive for weather information with the ERA5 [26], providing atmospheric, climate, land, and oceanic variables. Meteorological information is often used to quantify extreme weather events such as droughts and flooding.

Weather

Further sensed data sources include *topographic* information, such as a Digital Elevation Map (DEM), which provides information about the Earth's surface elevation. This is particularly interesting in the context of water movements. For instance, in an agricultural field, terrain topography affects the direction of water flow, which can result in waterlogging. Commonly, terrain information is provided through active radar sensors. A common library for terrain data is provided by the *NASA's Shuttle Radar Topography Mission (SRTM)* [27] in $30\,\mathrm{m} \times 30\,\mathrm{m}$ resolution.

Topography

Soil information is another data modality that is frequently used in this thesis. Soil data provides valuable information on soil characteristics, including texture, nutrient content, and water availability. This is particularly important for modeling plant development, since plant growth depends on soil properties. The *SoilGrids* library provides global soil data of diverse soil properties [28].

Soil

2.2 Datasets & Applications

Although unlimited EO data is provided daily, the number of labeled datasets is highly limited. Zhu et al. [22] pointed out that only 0.1% of the total available data volume is labeled datasets (see Figure 2.1). Consequently, and in contrast to expectations, EO remains in a low-data regime, leading many deployed models to suffer from limitations. In particular, regression tasks remain largely underexplored in EO due to the lack of dedicated benchmark datasets.
In this chapter, we present all the datasets used throughout the thesis. Following

[6] https://www.ecmwf.int/ [accessed: April 17, 2026]

the overview of existing datasets, we describe the YieldSAT dataset, a novel EO benchmark for crop yield prediction.

2.2.1 Crop Yield Prediction with YieldSAT

Early and accurate crop yield prediction at a large scale and high spatial resolution is critical for ensuring global food security. Yield prediction becomes increasingly important for assessing food security, increasing productivity, and ultimately reducing hunger. The FAO reported that in 2024, between 638 and 720 million people still faced hunger, while almost 2.60 billion people were unable to afford a healthy diet [31]. On the other hand, agriculture is vulnerable to climate change and extreme weather events that decrease productivity, result in significant yield losses, and can result in severe economic damages [32, 33, 34, 35, 7]. To mitigate this, early assessment of the expected yields is increasingly demanded for applications such as farm management, risk insurance, and policymaking. Nevertheless robust, generalized models require large and diverse datasets. While EO data is openly accessible without limit, the acquisition of ground truth yield data is often hindered by high acquisition costs, inconsistent quality, data privacy regulations, and reluctance to share data. Consequently, public yield datasets are scarce, noisy, or incomplete. Therefore, many studies are refined to national or regional yield prediction based on reported statistics (such as from the USDA National Agricultural Statistics Service (NASS)[7])[36, 37]. However, such data can have errors and uncertainties [38]. Additionally, such data is inadequate to model in-field variability, optimize management strategies, support decision-making, and ultimately increase productivity. Consequently, public yield data with high spatial resolution is scarce, restricted to very small areas and crop types, and often not ready to train ML models. There is a severe lack of open datasets, limiting the advancement of Agriculture and EO research [39].

Food security

To fill this gap, we created a novel dataset for large-scale, high-resolution crop yield prediction: the *YieldSAT* dataset.

Acknowledgement

The dataset was collected as part of a large collaborative project on agricultural yield prediction. This project was partially funded by the European Space

Parts of this chapter, including figures and tables, have been published already in [29] and [30].

[7] https://www.nass.usda.gov/ [accessed: April 17, 2026]

Table 2.2: Comparison of the YieldSAT dataset with other crop yield prediction datasets.

Dataset	Countries	Crops	Years	Fields	Pixel-Level	Resolution		Features	Curated
						Spatial	Temporal (optical)		
SwissYield [40]	1	2	2017–2021	73	✓	10 m × 10 m	∼5 days	14	✗
CropNet [36]	1	4	2017-2022	0	✗	9 km × 9 km	∼14 days	13	✗
YieldSAT	**4**	**4**	**2016-2024**	**2176**	✓	10 m × 10 m	**∼5 days**	**>70**	✓

Table 2.3: Overview of the YieldSAT dataset.

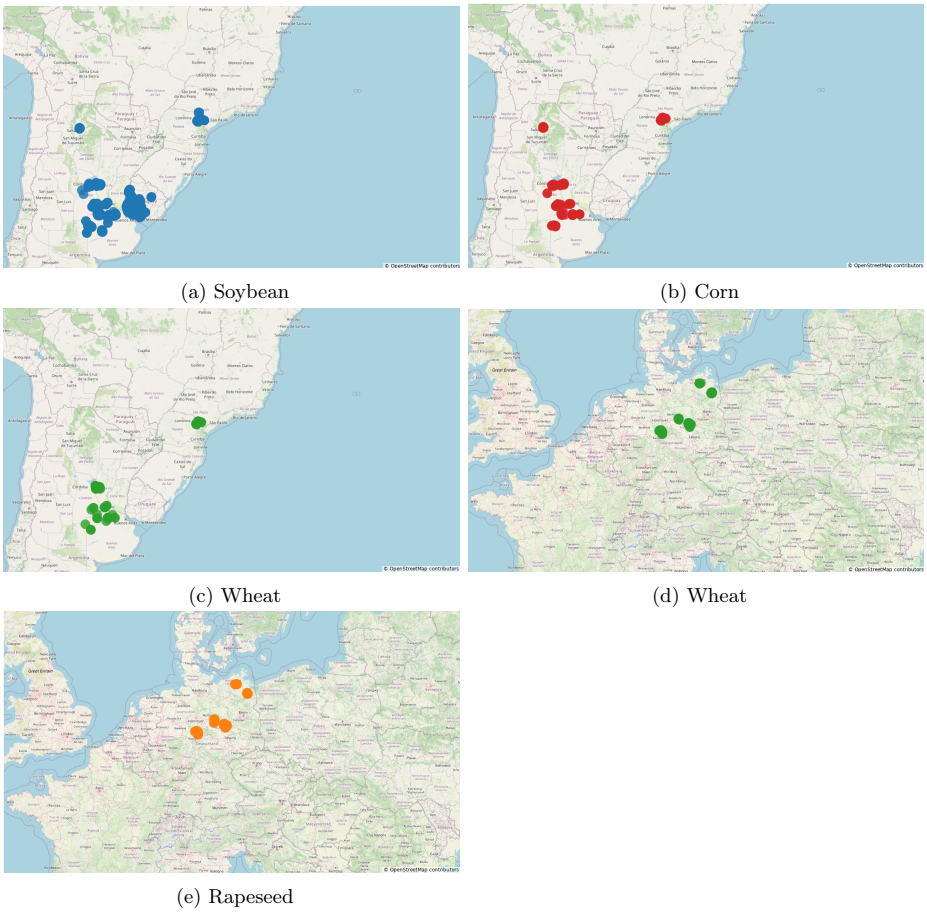
YieldSAT	Fields					Pixels	Area (ha)		Years	S2 Images
	Corn	Rapeseed	Soybean	Wheat	Total		Total	Average		
Argentina	185	✗	443	126	754	∼5.3 M	∼56,050	74.3	2017-2024	42,996
Brazil	118	✗	293	140	551	∼4.2 M	∼43,125	78.2	2017–2024	19,308
Uruguay	✗	✗	572	✗	572	∼2.1 M	∼32,804	57.3	2018–2022	24,318
Germany	✗	111	✗	188	299	∼0.6 M	∼6,460	21.6	2016–2022	26,998
Summary	**303**	**111**	**1,308**	**454**	**2,176**	**∼12.2 M**	**∼138,439**	**57.8**	**2016-2024**	**113,630**

Agency (ESA) InCubed Programme [8] as part of the project *AI4EO Solution Factory* [9].

Overview & Comparison

The yieldSAT dataset is a multimodal dataset, containing high-resolution (10 m × 10 m) yield data, paired with multimodal EO data, such as optical and multispectral time series imagery, weather data, soil data, and topographic information. The dataset covers a wide range of geographic locations across multiple countries and crop types. Specifically, it includes information from Argentina, Brazil, Uruguay, and Germany with approximately 138 439 ha (1384.39 km^2). In Figure 2.4, the geographic locations of the available fields are displayed, colored by crop type. Moreover, the dataset contains several crop types, including soybean (*Glycine max L.*), corn (*Zea mays*), rapeseed (*Brassica napus L. ssp. napus*), and wheat (*Triticum aestivum*). The dataset spans eight years, from 2016 to 2024, and contains a total of 2176 labeled fields. Soybean is the most represented crop, with 1308 fields, while rapeseed is the least represented, with 111 fields. Notable differences in the field size exist between countries. For example, Brazil has the largest average field size, with 78.8 ha. In contrast, Germany has smaller field sizes with an average of 21.6 ha. The average field size is 57.8 ha across the entire dataset. Altogether, the dataset includes approximately 12.2 million ground truth yield samples (pixels with 10 m × 10 m resolution). A detailed description of the available ground truth data is depicted in Table 2.3. Each field and pixel is coupled with multimodal data sources, including multispectral satellite imagery from the S2 mission, weather, soil, and elevation data. In total, more than 70 features are available for each sample. Additionally, more than 113,630 S2 images are

[8] https://incubed.esa.int/ [accessed: April 17, 2026]
[9] https://www.ai4eo-solution-factory.de/ [accessed: April 17, 2026]



(a) Soybean (b) Corn

(c) Wheat (d) Wheat

(e) Rapeseed

Figure 2.4: Yield Data spatial distribution for each region and crop type. The yield data is colored by crop type. Map data copyrighted by OpenStreetMap contributors.

available, covering the growing period from seeding to harvesting in each field.

Comparison to literature

The yieldSAT dataset is unique across the literature. A comparison to other popular datasets for crop yield modeling is provided in Table 2.2. For instance, the data presented by Perich et al. [40] cover only 73 cereal crop fields from one country and use only a subset of S2 bands. In addition, the CropNet [36] dataset offers only regional-level yield data (from the USDA NASS[10]) for the US only, with low temporal and spatial resolution. The data is coupled only to selected bands and derivatives from S2 and to meteorological variables.
In particular, the yieldSAT dataset is the first multimodal dataset for predicting crop yields at the field and subfield (pixel) levels, spanning multiple countries, crop types, and years, and containing a large number of EO data modalities with high temporal and spatial resolution.

[10] https://www.nass.usda.gov/ [accessed: April 17, 2026]

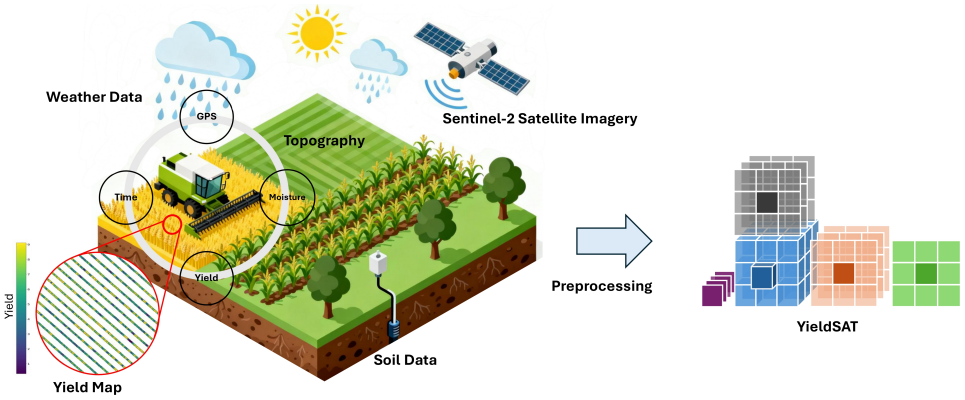


Figure 2.5: Shematic overview of the data collection and preprocessing. At harvest, a combine collects point data containing various information (yield, geolocation, timestamp, moisture). After cleaning, the point data gets rasterized to 10×10m resolution using a rasterization grid. For each yield map, EO data is collected. Finally, the target and input data are preprocessed into a ML-ready data format for pixel-wise yield prediction. (Image partly generated with [41])

Data Collection

The data collection consists of (1) the ground truth yield data acquisition and preprocessing, (2) EO data collection, and (3) creation of ML-ready datasets. The ground truth data collection process was done in collaboration with farmers, research institutions, and industry. The process is summarized in Figure 2.5.

Yield Data Collection: The YieldSAT dataset contains subfield-level yield data that was collected by combine harvesters. At harvest, a combine harvester equipped with yield monitors drives through the field and collects georeferenced, equidistant data points as point vector data. Each data point contains information such as the geographic coordinates (latitude and longitude) of each measurement, the amount of wet yield, and the moisture content. Additionally, metadata is collected, such as the timestamp of each measurement, speed, working width, moving direction, and elevation. Together, all data points form a yield map, a collection of data point vector data for a single field. Yield map data provides valuable information about the spatial variability and productivity of a single field and enables farmers to identify productivity zones, estimate yield quality, quantify damage and losses, and serves as a foundation for future activities. Nevertheless, yield map data is highly inhomogeneous and therefore requires careful data preprocessing.

Georeferenced vector data

Yield map

Yield Data Preprocessing: Raw yield data is commonly stored as georeferenced vector data in the *shapefile* format. Still, many other data formats exist that can store geospatial yield data. Therefore, if the data is not provided in the shapefile format, a format conversion is performed. Moreover, yield

The shapefile format is a geospatial vector data format that stores geographical information along with its features.

maps are manually inspected and curated. The curation is stored as metadata for each yield map. Only yield data that appeared realistic to agricultural and *Geographic Information System (GIS)* experts were considered for further processing.

Combine harvester yield data is highly inhomogeneous across farmers, regions, and countries. Differences arise from the use of different machines, languages, units, and management practices. This makes the processing of combine harvester data from various locations challenging and requires systematic data preprocessing. Therefore, a standardized preprocessing pipeline harmonizes the data, using semi-automatic and automatic detection and translation of feature names across various languages, as well as conversion to the metric system. Additionally, automatic transformation from the geographic *World Geodetic System 84 (WGS84)* to a projected *Universal Transverse Mercator (UTM) Coordinate Reference System (CRS)* is done. Following, yield measurements are cleaned, since the reliability of the yield sensors is affected by several parameters, including speed, grain flow, and GPS accuracy. Combine harvester data is often miscalibrated and is associated with sensor errors, positioning inconsistencies, missing information, delays between grain collection and measurement, and focus during turns [42].

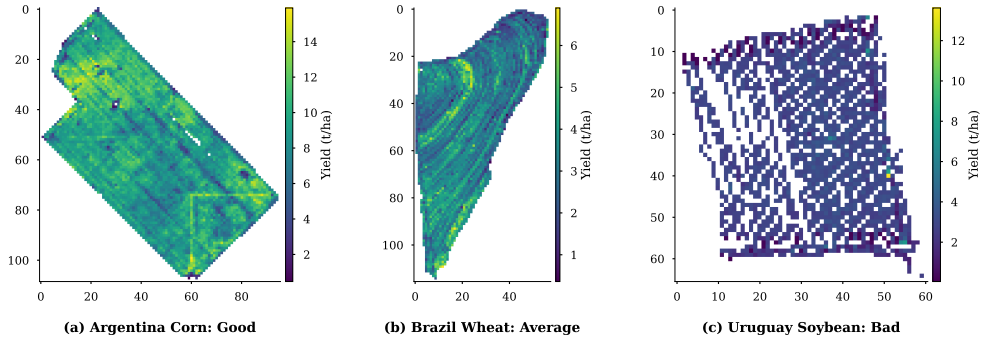
Consequently, improving data quality is necessary by removing erroneous values for yield, moisture, timestamp, and deactivated harvesters [43]. Nevertheless, we aimed to minimize preprocessing and outlier removal to preserve the data in its original form. For this, zero-yield points and biologically infeasible points were removed based on expert rules. This includes a crop-specific maximum yield value shown in Table 2.4. Yield values above that threshold are considered sensor errors and are subsequently removed. The data was further filtered using statistical thresholds. Yield measurements must fall within three standard deviations ($\pm 3\sigma$). This *three-sigma-rule* ensures that 99.7% of data remains, with the rest classified as outliers. This method is a common and simple outlier detection method that is also applied in the context of yield data preprocessing [40, 30]. Finally, the scaled yield (dry yield) is calculated based on the measured wet yield, adjusted to a fixed standard moisture as:

$$y_s = y_w * \frac{1 - m_m}{1 - m_s}, \tag{2.2}$$

with y_s as the scaled yield (dry yield), y_w the wet yield, m_m as the measured moisture, and m_s the standard moisture. The standard moisture for every crop type is given in Table 2.4. The scaled yield is calculated because it is less affected by measurement noise (weather and time of harvest). Additionally,

Table 2.4: Maximum accepted yield values in t/ha for every crop type and the standard moisture.

Standard Values	Wheat	Rapeseed	Soybean	Corn
Max. Yield (t/ha)	20	10	15	45
Standard Moisture (%)	15	9	15	16



(a) Argentina Corn: Good (b) Brazil Wheat: Average (c) Uruguay Soybean: Bad

Figure 2.6: Example yield data together with the curation.

the scaled (dry) yield is the true indicator of the grain output used to estimate revenue potential for farmers, traders, and crop insurance. After this, a georeferenced boundary polygon is created for each yield map that is used as a georeference for EO data acquisition, which is required for training ML models.

At the end of the data preprocessing, each yield map is assigned a quality score. This score combines visual expert decisions and automatic curation guidelines. Figure 2.6 shows a set of randomly selected fields. It highlights the varying quality levels within the dataset. Low-quality yield maps often exhibit sparse sampling, spatial misalignment, or erroneous measurements, such as unrealistically high yield values. Artifacts may be present as well. These can include patterns caused by harvester turns or delays in the measurement process.

Earth Observation Data Collection: All collected EO features used to train yield prediction models were selected based on stringent selection criteria: (1) demonstrated or theoretically established influence on crop development and yield, (2) open and freely accessible, (3) global coverage, and (4) high spatial- and temporal resolution. An overview of the resulting data modalities and detailed information about the collected data are provided in Table 2.5.

Optical & Multispectral Satellite Data

For each field, a set of EO data modalities was acquired that are known to be a predictor of crop yield. Specifically, S2 Level-2A optical and multispectral time series imagery, containing all 13 spectral bands, was acquired for the period

Table 2.5: Overview of all available data modalities in the yieldSAT dataset and their characteristics.

Modality	Source	Product	Spatial Resolution	Temporal Resolution
Multispectral	Sentinel-2 L2A	B01 - Coastal Aerosol B02 - Blue B03 - Green B04 - Red B05 - Red Edge 1 B06 - Red Edge 2 B07 - Red Edge 3 B08 - NIR B8A - Narrow NIR B09 - Water vapour B11 - SWIR 1 B12 - SWIR 2 Scene Classification Layer (SCL)	10 m × 10 m	∼ 5 days
Weather	Era5 [26]	Max Temperature Mean Temperature Min Temperature Total Precipitation	30 km × 30 km	Daily
Soil	SoilGrids [28]	Soil Organic Carbon Nitrogen Cation Exchange Capacity Clay Silt Sand pH Volumetric fraction of course fragments	250 m × 250 m	Static
Topography	SRTM [27]	Elevation (DEM) Slope Curvature Topographic wetness index (twi) Aspect	30 m × 30 m	

between seeding and harvesting. All available S2 bands and time steps within the corresponding growing season were collected, resulting in a time series with a temporal resolution of approximately one image every five days. In Table 2.1, information about each band of the S2 product is given. Additionally, the Scene Classification Layer (SCL) is available for each time step, providing per-pixel class information for the S2 product at a 20 m × 20 m resolution. In total, the SCL layer provides 12 class labels, such as "vegetated," "non-vegetated," or "clouded." Following, spectral bands not originally available at 10 m × 10 m resolution were upsampled to 10 m × 10 m to ensure a unified spatial resolution across all bands. The S2 bands were preserved in their original form, and no specific indices were calculated (e.g., NDVI, NDWI), such as in [36]. This increases the flexibility of the dataset for downstream tasks.

The SCL provides a classification layer for the S2 product.

Additional Data Modalities

While optical and multispectral satellite data are known to be good predictors of crop yields [40], additional data modalities can be integrated to build yield prediction models. For example, optical data can be affected by cloud occlusion, rendering certain time steps uninformed [44]. Therefore, to compensate for the shortcomings of S2 data, ADMs were acquired for every field. Specifically, other data modalities are known to correlate with crop yield, including weather, soil, and topographic information, which further fulfilled the selection criteria. Weather data for each field is acquired from the ERA5 program [26] within

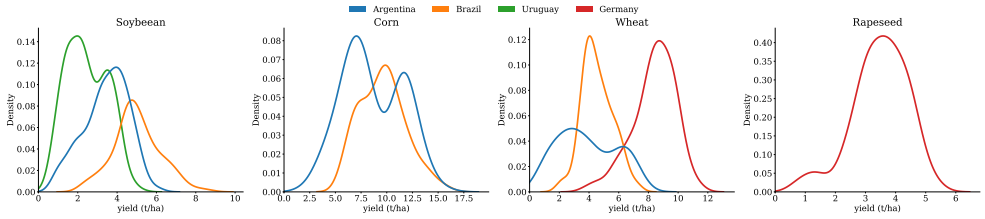


Figure 2.7: Yield data distribution plots for each crop type and country.

Table 2.6: Kruskal–Wallis H-test between yield distributions grouped by countries and by crop type.

Evaluation	Country	Crop
p-value	4.15×10^{-178}	7.07×10^{-192}

the growing period. Soil data was acquired from *SoilGrids* in 250 m × 250 m resolution [28]. Topography data was acquired from the SRTM mission [27] in 30 m × 30 m resolution. For soil and topography data, raster images are created for each feature and upsampled to 10 m × 10 m resolution using a cubic spline interpolation to match the resolution of the S2 images. For soil, all soil properties are sampled at depths of 0-5, 5-15, 15-30, 30-60, 60-100, and 100-200cm. For the topography data, the RichDEM [45] library was used for feature engineering to derive additional features, including curvature, slope, and aspect. Moreover, the Topographic Wetness Index (TWI) is calculated following [46].

Rasterization & Pixel-Wise Yield Mapping: To train yield prediction models using supervised learning, the data alignment criteria must be fulfilled. To align the S2 imagery and the ADM with the yield data, the yield maps are rasterized so that each S2 product pixel spatially aligns with the corresponding yield data pixel across the entire time series. For this purpose, a rasterization grid derived from the S2 imagery with the spatial resolution of 10 m × 10 m was placed over the yield map. All yield points within a given pixel are averaged, yielding a rasterized yield map with the spatial resolution of the S2 data. Each pixel in the S2 and ADMs is spatially aligned with a pixel in the target yield map. For spatio-temporal data (S2), this applies for the entire time series.

Rasterization

Data Quality & Data Analysis: The YieldSAT dataset is diverse in terms of ground truth yield data and data modalities, such as S2 reflectance patterns. We highlight the unique characteristics of each crop type and country. For example, we see distinct yield distributions across crop types and countries. Corn and wheat tend to have the highest yields, with a wide spread. Soybeans and rapeseed generally exhibit lower yield values than other crops. Most yield

Table 2.7: Pairwise post-hoc comparisons of the yield distributions between individual **countries**. Each cell displays the statistical significance level of the difference between two countries based on Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Comparision	Argentina	Brazil	Germany	Uruguay
Argentina	-	****	****	****
Brazil	****	-	ns	****
Germany	****	ns	-	****
Uruguay	****	****	****	-

Table 2.8: Pairwise post-hoc comparisons of the yield distributions between individual **crops**. Each cell displays the statistical significance level of the difference between two crops based on Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Comparision	Corn	Rapeseed	Soybean	Wheat
Corn	-	****	****	****
Rapeseed	****	-	ns	****
Soybean	****	ns	-	****
Wheat	****	****	****	-

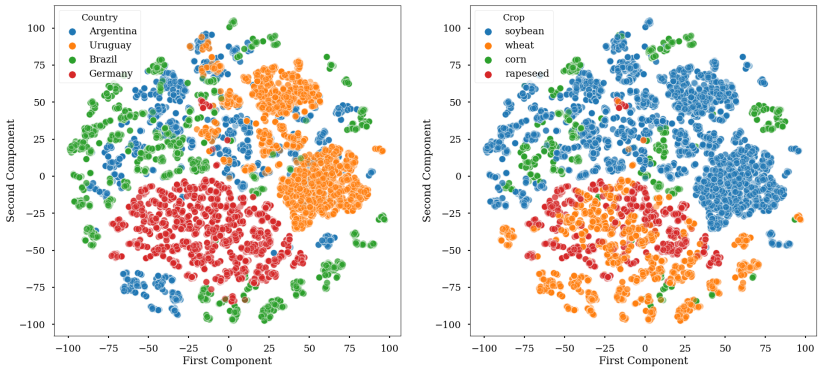


Figure 2.8: t-SNE plot of the surface reflectance of the S2 time series, grouped by countries (left) and crops (right).

Distribution shifts

distributions do not follow a normal distribution. This is shown in Figure 2.7. To test whether the data distributions differ significantly across crop types and countries, we perform a Kruskal-Wallis test. It compares the distributions between these groups. Both tests reveal significantly different distributions between countries and crop types. Next, we perform a post-hoc analysis to compare pairwise differences within each group. Dunn's test with a Holm-Bonferroni correction is used to account for multiple comparisons. Pairwise comparison results between each country are shown in Table 2.7. Notably, each group has a different yield distribution, with most p-values ≤ 0.0001. No significance is present except between Germany and Brazil. For crop types,

another Dunn's test shows most distributions are significantly different, with p-values ≤ 0.0001 (Table 2.8). Only rapeseed and soybean show no significant difference in yield distributions. Additionally, we observe high diversity in the patterns of the data modalities (e.g., surface reflectance of the S2 time series) between countries and crop types. In Figure 2.8, a t-Distributed Stochastic Neighbor Embedding (t-SNE) [47] of the surface reflectance of the S2 time series is shown, colored by countries (left) and crops (right). The plot shows that the surface reflectance differs between countries and crops.

2.2.2 Further Yield Prediction Datasets

The *SwissYield*[11] dataset was published by Perich et al. [40]. Similar to the YieldSAT dataset, this dataset is a regression dataset containing subfield-level yield measurements coupled with EO data modalities. The dataset contains 73 yield maps of cereal crops collected between 2017 and 2021 in Switzerland. The yield maps are rasterized to $10\,\text{m} \times 10\,\text{m}$ resolution using S2 data as the reference. For each pixel, multispectral optical sensor data from S2 (10 bands) is collected. Moreover, weather data was acquired including mean temperature and total precipitation from the ERA5 program. The time series spans the period from seeding to harvesting, with an approximate 5-day sampling rate. The dataset has a size of 54098 yield pixels.

Food security

2.2.3 Life Fuel Moisture Content Estimation

The Life Fuel Moisture Content (LFMC) dataset is a regression dataset for moisture content estimation, published by Rao et al. [48]. LFMC estimation estimates the mass of water per unit dry biomass in vegetation and is particularly important for wildfire risk assessment. This is a regression task in which the vegetation water (moisture) per dry biomass is predicted for a specific location.

Wildfire

$$\text{LFMC}(\%) = \frac{\text{fresh mass} - \text{dry mass}}{\text{dry mass}} \times 100. \tag{2.3}$$

The data was collected between 2015 and 2019 in 12 states of the western US. In total, 2578 samples were collected over an area of $3.7\,\text{km}^2$ containing several vegetation types. The area was selected for its increased fire activity and diverse climatological, topographical, and ecological characteristics. The

[11] https://www.research-collection.ethz.ch/handle/20.500.11850/581023 [accessed: April 17, 2026]

Table 2.9: Overview of the input features from the LFMC dataset. (Source: Rao et al. [48])

Radar (Sentinel-1)	Optical (Landsat 8)	Mixed (dB)	Static
VV	Red	$\frac{VV}{Red}$, $\frac{VH}{Red}$	Canopy height
VH	Green	$\frac{VV}{Green}$, $\frac{VH}{Green}$	Elevation
$VH - VV$	Blue	$\frac{VV}{Blue}$, $\frac{VH}{Blue}$	Terrain Slope
	NIR	$\frac{VV}{NIR}$, $\frac{VH}{NIR}$	Silt fraction
	SWIR	$\frac{VV}{SWIR}$, $\frac{VH}{SWIR}$	Sand fraction
	NDVI	$\frac{VV}{NDVI}$, $\frac{VH}{NDVI}$	Clay fraction
	NIRv	$\frac{VV}{NIRv}$, $\frac{VH}{NIRv}$	Land cover
	NDWI	$\frac{VV}{NDWI}$, $\frac{VH}{NDWI}$	

temporal features are multispectral optical sensor data from Landsat-8 [12], containing 5 bands. From the optical data, the NDVI, NDWI, and the near-infrared vegetation index (NIR_v) are derived, which are known to correlate with the LFMC. Additionally, SAR sensor data from S1 at C-band (5.4 GHz) backscatter with 3 bands is provided. Specifically, the vertical-vertical (VV), vertical-horizontal (VH) polarization, and their product are used. Moreover, mixed features are calculated containing ratios of optical and SAR data to follow a Physics-Guided (PG) approach. These features sampled over 4-month before the moisture measurement, resulting in a signal of 4 time steps. ADMs are static sensors for topographic information from the National Elevation Dataset [49], canopy height from GLAS-based lidar [50], soil properties from the Unified North American Soil Map [51], and land-cover class from GLOBCOVER [13]. The available features are provided in Table 2.9 and were interpolated to a pixel resolution of 250 m × 250 m.

2.2.4 Air Pollution Forecasting

Air pollution

The Particle Matter 2.5 (PM2.5) [14] is a multivariate time-series regression dataset containing information about the air particles of 2.5 microns or fewer in diameter (in $\mu g/m^3$) in the cities of Beijing, Shanghai, Guangzhou, Chengdu, and Shenyang. The dataset was published by Chen [52]. The dataset allows the estimation of air pollution in major cities and was collected between 2010 and 2015. As the input data, atmospheric conditions are used such as dew

[12] https://landsat.gsfc.nasa.gov/satellites/landsat-8/ [accessed: April 17, 2026]
[13] https://due.esrin.esa.int/page_globcover.php [accessed: April 17, 2026]
[14] https://archive.ics.uci.edu/dataset/394/pm2+5+data+of+five+chinese+cities [accessed: April 17, 2026]

point, temperature, humidity. Moreover, atmospheric dynamics such as the atmospheric pressure, combined wind direction, cumulative wind speed, and the season. Finally the hourly and cumulative precipitation are provided. These are captured at hourly resolution, where we consider a three-day window for prediction. In total, 52854 measurement instances are contained in the dataset.

2.2.5 Plant Growth Modeling

The *MixedCrop* dataset[15], is an image dataset that contains temporal RGB images of field patches, collected by a mixed-cropping experiment at the site Campus Klein-Altendorf in 2020, Germany in the context of the PhenoRob programe[16]. The dataset contains images of monocultures of spring wheat (*Triticum aestivum*, *SW*) and of faba bean (*Vicia faba L.*, *FB*), and as mixtures (*MX*) of both. All plots were planted under different conditions and management practices. For instance, under different seeding densities and different treatments. Alongside, various field data were collected, including nutrient supply, soil information, weather data, and manual height and biomass measurements [53]. In total, 110 features are available. Additionally, information was collected at 11 acquisition times during the growing period, including plant mean height and georeferenced orthophotos, captured with a *Unmanned Aerial Vehicle (UAV)* at a ground resolution of 3 mm in RGB space for each field plot. From this, patches are extracted that are aligned in space. An example is depicted in Figure 2.9.

Sustainable agriculture

[15] https://www.phenorob.de/data-2/index.html [accessed: April 17, 2026]
[16] https://www.phenorob.de/index.html [accessed: April 17, 2026]

Faba Bean (FB) Spring Wheat (SW) Mixture (MX)

Figure 2.9: Example orthophotos of two aligned image pairs for different field plots, namely for faba bean (FB), spring wheat (SW), and mixtures (MX). The left image is taken at a mid-growth stage (8 weeks), and the right image is taken 4 weeks after plant emergence.

3 PRELIMINARIES

This chapter presents all the background and preliminary information that is required to follow the content of this thesis. We provide an overview of the notation used in this thesis, including a taxonomy and basic techniques of ML, as well as an overview of DIL.

3.1 Notation & Basic Techniques

The notation used throughout this thesis is defined by:

- P, p: Probability measure, density function
- $\mathcal{X}, x$: Input space, instance of the input space
- $\mathcal{Y}, y$: Target space, instance of the target space
- $\mathcal{D}$: Dataset of input and target instance pairs
- $\mathcal{H}, h$: Hypothesis space, hypothesis
- $\mathcal{F}, f$: Function space, instance of a function
- $\mathcal{Z}, z$: Latent space, instance of the latent space
- $\mathbb{E}[\cdot]$: Expected value of a random variable
- $V(\cdot)$: Variance of a random variable
- T, t: Time domain, time step
- $\{\cdot\}$: Set

Further, we refer to a prediction of a target instance as $\hat{y}$. Additionally, this thesis distinguishes between different data types of the input and target instances:

- $\boldsymbol{x} \in \mathbb{R}^{1\times1\times B}$: Pixel or vector instance of the input space with bands B.
- $\boldsymbol{X} \in \mathbb{R}^{W\times H\times B}$: Image instance of the input space with width W, height H, and bands B.

- $\boldsymbol{y} \in \mathbb{R}^{1\times1\times B}$: Pixel or vector instance of the target space with B bands.
- $Y \in \mathbb{R}^{W\times H\times B}$: Image instance of the target space with width W, height H, and bands B.
- **Pixel dataset**: This dataset contains spatially aligned pixel pairs $\{\boldsymbol{x}, \boldsymbol{y}\}$. Both instances are aligned in space, i.e., each pixel shows the same geographical location.
- **Time Series Pixel Dataset**: This dataset consists of paired and spatially aligned pixel pairs $\{\boldsymbol{x}, \boldsymbol{y}\}$, however, the input instance contains T time steps $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T]$. Likewise, the target remains unchanged.
- **Image Dataset**: This dataset consists of aligned image pairs of the input space and target space $\{X, Y\}$. Similar to the pixel datasets, both instances are aligned in space, i.e., each pixel corresponds to the same geographic location.
- **Time Series Image Dataset**: This dataset contains spatially aligned image pairs $\{X, Y\}$. Similar to the time series pixel dataset, the input instance contains T time steps with $X = [X_1, X_2, ..., X_T]$. Likewise, the target instance remains unchanged.

We can commonly derive a time series pixel dataset from an image dataset if the image alignment is satisfied. For each study, deviating notations may be used. Changes will be defined subsequently.

3.2 A Reflection on Machine Learning and Knowledge

Machine Learning & Deep Learning

In general, ML is a broad family of algorithms that are designed to learna patterns from data without being explicitly programmed for it [54]. The simplest example might be a *linear regression* model that learn the importance of input features for a prediction. Suppose you want to predict the biomass of a plant based on its height. For this, you collect data on the plant height and learn to adjust a linear line so that the prediction matches the observed data. After the learning process, the model can predict new values for plants it has not seen before. Traditional ML algorithms like *linear regression* or *decision trees* are commonly explainable, and a decision can be justified by providing meaningful explanations of the underlying logic. For example, a *decision tree* follows a graphical tree structure and decisions are linearized based on *if-then* rules [55]. Another category of ML algorithms is DL. DL commonly relies

on multiple processing units, designed to learn several levels of abstraction, commonly using Neural Networks (NNs) [56]. Such models, commonly known as *back-box* models, are highly powerful and can process high-dimensional data but at the cost of being less explainable [55].

More formally, the process of ML is defined as learning a mapping function $\mathcal{F}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ from the input space $\mathcal{X}$ to the target space $\mathcal{Y}$ by optimizing over the model parameters θ. For example, in a supervised learning setting, the training data consists of instance pairs, e.g., $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ that define the relations between the instances in the training set. Following, the model tries to extract rules by learning parameters that minimize an objective (loss) function:

Supervised learning

$$\mathcal{F}_\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, \hat{y}_i), \tag{3.1}$$

with $\mathcal{L}(y_i, \hat{y}_i)$ as the objective function, $\hat{y}_i$ as the prediction, and y_i as the ground truth. An important loss function for regression applications is the squared error (see Equation 3.3), which measures the squared difference between a predicted and a target value [57]. Other objective functions exist that will be introduced throughout this thesis.

More general, ML algorithms share the property of extracting patterns from data by performing *induction* [54].

> **Induction**
> Induction is a reasoning process that involves summarizing observations (e.g., data) to generalized concepts or rules.

The induction or learning process creates a hypothesis that reflects the underlying data. However, several hypotheses are commonly consistent with the observed data. ML can be considered as the search over the entire hypothesis space to find the model that is most consistent with the observed data. In practice, many solutions exist that satisfy the observed data, and there is no clear superior hypothesis. Moreover, different models may predict new, unseen samples differently. Yet all remain similarly consistent with the observed data during the learning process [54]. Consequently, a learning system must decide which hypothesis it becomes. The bias for a specific solution is called *Inductive bias*. Importantly, every learning algorithm has its own inductive bias, that guides the model to a unique solution. Otherwise, a prediction would be ambiguous and model would make different predictions for the same input. Interestingly, the inductive bias that caused the algorithm to become

Hypothesis space

Inductive bias

a unique hypothesis leads to a fundamental problem in ML itself. **That is, what is the best model?** This brings us to the *No Free Lunch Theorem (NFL)* [58]. The NFL theorem states that across every possible problem, no single algorithm exists that performs better than all other algorithms. In fact, any algorithm that excels in one specific task fails in another task. This has significant consequences for the field of ML as it implies that a universal superior model does not exist. Consequently, ML can only be viewed in the context of a particular problem, and the success of a model depends on tailoring the model to a particular task. This requires careful model selection, as algorithmic choices and underlying assumptions must align with the data and the objectives. Therefore, practitioners should evaluate not only the theoretical capabilities of a model but also its suitability for the context in which it will be deployed.

No Free Lunch Theorem

In fact, what helps the model produce a hypothesis is passed on to the creator of the learning model itself. Consequently, to achieve superior performance in a specific task, prior knowledge (domain expertise) is required that defines a hypothesis over the hypothesis space.

3.2.1 Deep Learning Building Blocks

Table 3.1: Common building blocks in deep learning with different inductive biases and assumptions. (Source: [59])

Layer Components	Entity	Relation	Inductive Bias	Invariance
Fully Connected Layer	Neurons	All-to-all	Weak	-
Convolutional Layers	Grid elements (image)	Local	Locality	Spatial translation
Recurrent Layers	Timesteps	Sequential	Sequentiality	Time translation

In fact, ML without prior knowledge does not and cannot exist as it is inherently structured by assumptions about the world. Traditionally, ML is viewed as an abstract function approximator that discovers relationships between entities in order to define general rules or concepts. Regardless, it is commonly overlooked that every ML method embeds substantial prior knowledge, as well as mathematical, structural, and conceptual properties. A historical example lies in the very building blocks of ML. In his fundamental work, Rosenblatt [60] explored the nature of machine thinking, raising foundational questions about the properties of perception, generalization, and memory. He introduced the *Perceptron*, the direct ancestor of modern NNs. Perceptrons (or neurons) form interconnected systems and pathways shaped by stimuli and experience, analogously to the biological nervous system. Rosenblatt

An entity is an element characterized by attributes.

Perceptron

Parts of this section are guided by [54]

envisioned a model of general intelligence *"without becoming deeply enmeshed in the special."* He further discussed crucial properties of perceptrons, such as random initialization, arguing that a network (brain) begins in a largely random state at birth, except for some genetic components. Only through learning and neural plasticity, meaningful connections and functional organization emerge from experience. Likewise, all building blocks of ML carry task-specific assumptions and inductive bias [59].

Fully Connected Layer

A Fully Connected Layer (FCL) [61] is one of the fundamental building blocks of modern NNs architectures that build on the previously discussed perceptrons. It connects every neuron in the previous layer to every neuron in the current layer. Each neuron generates a weighted sum of the inputs and adds a bias term. Finally, an activation function is applied. Importantly, every neuron in the previous layer is connected to every neuron in the current layer, allowing each input to influence every output by defining any output value. This high degree of connectivity reduces the model's inductive bias by imposing minimal structural assumptions about which inputs are relevant. Consequently, the network gains the flexibility to model complex, global relationships and to perform general tasks such as classification and regression. By stacking multiple FCLs, one obtains a Multilayer Perceptron (MLP), a core architecture capable of approximating arbitrary continuous functions. In this way, such architectures emulate general aspects of intelligence through specific, learnable inductive biases encoded in their structure and training process.

Convolutional Layer

Convolutional layers [62] are commonly used in image processing. These layers apply a convolution operation to an input, typically an image, using learned filters or kernels to create a feature map. This feature map highlights specific properties of the input. Compared to the FCL, a convolutional layer learns sparse relations that focus on locality and translational invariance to provide a stronger inductive bias. Locality assumes that neighboring entities (e.g., pixels) have high covariance that diminishes with distance. Translational invariance assumes that the rule of locality can be applied across the input. By sharing weights and using local connections, convolutional layers reduce computational complexity and parameter count compared to a FCL, making them more efficient for many computer vision tasks.

Recurrent Layers

Recurrent layers [63] are another fundamental building block of NN, designed to capture structure in sequential information, such as in time. These layers include recurrent connections that process inputs sequentially, with the output (or hidden state) from the previous time step serving as input to the current state. Through this mechanism, recurrent layers learn relationships across sequential inputs and hidden states, enabling the network to model temporal dependencies and dynamics over time. Analogous to how convolutional layers exploit spatial locality, recurrent layers assume temporal locality, that is, dependencies between neighboring time steps that are approximately invariant in time, often modeled under a Markovian assumption.

In Section A.1 we provide a detailed overview of advanced model architectures that are relevant for this work.

3.2.2 Domain-Informed Learning

Although ML building blocks already integrate prior knowledge, the explicit integration of prior knowledge into the learning algorithm is an intensively studied field of research, aiming to increase the model's effectiveness, robustness, explainability, or knowledge conformity.

inductive bias & prior knowledge

> **Domain-Informed Learning**
> *Learning from prior knowledge and data.*

prior knowledge

In this thesis, we refer to prior knowledge as a source of validated *scientific information* that exists independently of the learning algorithm. Scientific knowledge is a validated source of information (e.g., natural laws) that can be represented as algebraic equations, differential equations, simulation results, data, or validated concepts. Several definitions for DIL exist that describe the explicit integration of knowledge into a learning system, such as *Theory-Guided Data Science* [16], *Informed Machine Learning* [17], *Knowledge-Guided Machine Learning* [20], or *Physics-Informed Machine Learning* [64]. Table 3.2 provides an overview of existing definitions of learning with data and knowledge. All methods share the common goal of addressing the nature of knowledge discovery in scientific disciplines that go beyond the development of actionable end-to-end models, as in classical ML. Instead, the main goal is to develop hypotheses and theories from learned patterns and relationships to achieve

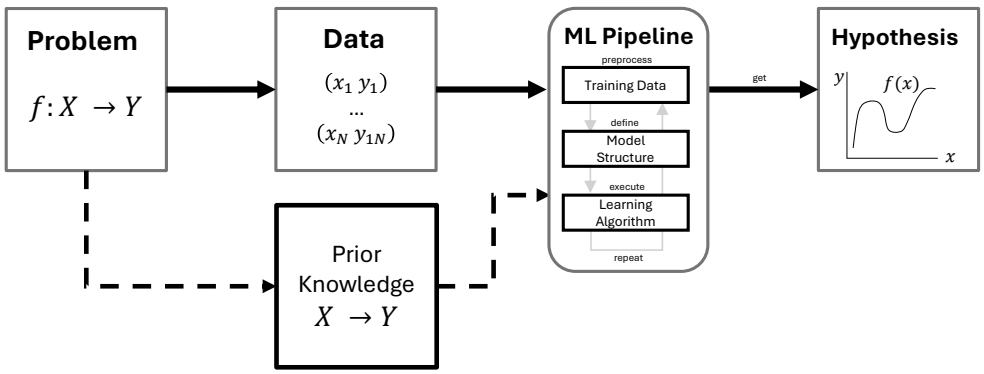


Figure 3.1: Schematic overview of domain-informed learning. The ML pipeline receives data and prior knowledge to learn a hypothesis for a specific problem. The dashed line indicates the integration of prior knowledge. (Figure derived from [17])

Table 3.2: Domain-Informed Learning literature definitions.

Theory-Guided Data Science	*"seamlessly blending scientific knowledge in data science models."* [16]
Informed Machine Learning	*"learning from a hybrid information source that consists of data and prior knowledge."* [17]
Knowledge-Guided Machine Learning	*"aims to use both scientific knowledge and data in ML frameworks."* [20]
Physics-Informed Machine Learning	*"integrates seamlessly data and mathematical physics models."* [64]

scientific advancements.

Definitions proposed by Karniadakis et al. [64] focus on the learning of physically consistent solutions, e.g., by integrating formalized mathematical knowledge such as differential equations. Such approaches explicitly address the lack of representative training samples, arguing that data scarcity can be compensated for by incorporating prior knowledge. In fact, one reason for the limited success of purely data-driven models in scientific disciplines is the limited access to sufficient and representative training data, compared to mainstream problems such as those in natural language processing or object detection. Therefore, models are often under-constrained. Moreover, physical variables are often complex and dynamically change over time. Insufficient training data, therefore, cannot capture the true nature of the physical variables, resulting in poorly performing models and even misleading conclusions [16].

In contrast, Von Rueden et al. [17] provides a more general approach by considering different knowledge representations and integration practices. In general, three main questions define DIL:

Table 3.3: Overview of knowledge integration strategies. The referenced studies are related to this work.

Integration	Motivation	Goal	Challenges	Studies
Data Space	Less data	Data augmentation (e.g., simulations)	Noisy data, mismatch	[65, 66, 67, 68, 69, 70]
Model Structure	Effectiveness (e.g., performance, explainability)	Knowledge-based architectures (e.g., time invariances), probabilistic relations	Costs, feasibility	[71, 69, 72, 73, 74, 75]
Learning Algorithm	Knowledge conformity, less data	knowledge-based regularization (e.g., loss terms)	Weighting knowledge vs. Data, robustness	[37, 76, 70, 77, 78]

1. How is the knowledge represented?
2. How is the knowledge integrated into the model pipeline?

Several scientific knowledge representations can be used to enrich ML algorithms, as discussed earlier. For example, algebraic equations in the form of a VI can formalize a relationship between the reflectance of an image of a plant and its biomass. Integrating such formalized concepts can facilitate the learning process. Considering knowledge integration, prior knowledge can be integrated into the (1) training data, (2) model structure, and (3) learning process (e.g., through regularization or conditions). Integrating prior knowledge at different steps in the ML-pipeline serves distinct purposes and has distinct advantages and limitations, as shown in Table 3.3.

3.2.3 Limitations in the Current Literature

Many studies related to DIL for EO rely on custom datasets. For example, the lack of widely adopted benchmarks for DIL in agriculture limits direct comparisons between methods and reduces reproducibility. Although expert knowledge can enhance the performance of ML models, as shown in Table 3.3, knowledge sources are difficult to represent and embed in ML pipelines. Most studies in EO applications focus on enriching the data space using simulation results or augmented data, but commonly lack a comparison with a baseline model trained without expert knowledge. Only a few papers incorporate prior knowledge directly into the learning algorithm (e.g., the loss function) to increase the conformity with physical laws. Additionally, only a few papers follow

a probabilistic approach by incorporating expert knowledge using Bayesian Networks.
No study systematically evaluates different knowledge sources and integration practices using a standardized approach. Consequently, it is an open question which knowledge sources and integration practices are truly effective, and where purely data-driven methods excel. In the following parts of this work, we will analyze different knowledge sources and integration strategies to systematically answer RQ1, RQ2, and RQ3.

3.3 Evaluation

This thesis uses different model evaluation techniques that will be described here. The objective of training a ML model is to foster generalization and avoid overfitting.

> **Generalization & Overfitting**
> Generalization refers to the ability to accurately predict new, unseen data after being trained on a training dataset. Unlike generalization, overfitting occurs when a model only fits the training data and performs poorly on new, unseen data.

To assess the generalization performance of a ML model, evaluation strategies are required. Therefore, we briefly summarize key evaluation and training techniques. Training a ML model typically involves splitting the data into a *training* set and an independent *test* set. Often, another independent *validation* set is used specifically for hyperparameter optimization of the model. Additionally, K-Fold Cross-Validation (CV) is a common technique that divides the data into K non-overlapping folds, where one fold is held out consecutively during training and used solely for testing. The overall performance is subsequently reported over all folds. CV is commonly used to estimate the generalization performance of a model by detecting overfitting. Further, it is recommended to include prior information about the data that counteracts overfitting. *Stratification* involves arranging the data so that each fold has the same class distribution. Furthermore, *grouping* ensures that samples that belong to the same group only appear in either the training or testing set. Moreover, defining domain-specific CV scenarios helps to assess the generalization power of a model. For example, EO data from distinct regions or years often exhibit distinct distributions. Therefore, it is desirable to assess the temporal or

Hyperparameter

Stratification & Grouping

Transferability

spatial generalization (transferability) by using Leave-One-Year-Out (LOYO) or Leave-One-Region-Out (LORO) CV scenarios. Here, a specific group (e.g., a region) is held out during training and used only for inference.

3.3.1 Quantitative Evaluation

Regression

This thesis focuses on regression applications. Regression is a predictive modeling task that involves predicting a numeric value. Unlike classification, regression predicts a continuous quantity and requires specific evaluation measures to assess the predictive power of the model. For quantitative evaluation in a regression setting over N samples, the Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root-Mean-Square Error (RMSE), Relative Root-Mean-Square Error (RRMSE), BIAS, and the Coefficient of Determination (R^2) are calculated as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{3.2}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{3.3}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{3.4}$$

$$\text{RRMSE} = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} y_i}}, \tag{3.5}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \tag{3.6}$$

$$BIAS = \frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i), \tag{3.7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3.8}$$

Image generation

We also evaluate the image generation power of certain models that require specific evaluation methods. The Fréchet Inception Distance (FID) score [79] is a prominent metric used to assess the quality of generated images. The FID score compares the distributions of generated and real images, rather than comparing individual images, by passing the samples to a pretrained InceptionNet [80] trained on the ImageNet dataset [81]. The generated and

real distributions are compared at the feature level by computing the Fréchet distance between their multivariate Gaussian feature representations extracted from a specific layer of the InceptionNet.

$$\mathrm{FID}(\mathcal{N}_r, \mathcal{N}_g) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right), \tag{3.9}$$

with μ_r and Σ_r as the mean and covariance of the real image features and μ_g and Σ_g for the generated images, respectively. $\mathrm{Tr}(\cdot)$ is the trace operator (sum of diagonal elements). The FID is a robust and efficient measure of image quality that is considered to correlate well with human judgment regarding image diversity and quality [82]. However, it also has disadvantages. For a finite set of samples, the FID is inaccurate and deviates from the true score. The FID is characterized by a high bias [83] and depends on the model being evaluated [84]. Therefore, comparisons between different models are unreliable.

3.3.2 Qualitative Evaluation

Besides quantitative evaluation, qualitative evaluation based on human perception is an important aspect for assessing a model's performance. This is done on individual examples using expert judgement. Since qualitative evaluation can be biased, this work uses stringent guidelines for qualitative evaluation:

- Realistic appearance: predicted samples should look realistic, sharp, artifact-free, and exhibit clear and domain-realistic detail.
- Consistency of conditions: predictions should be consistent with the input and other injected conditions while producing low prediction errors.
- Variability and generalization: the model should generalize to unseen data and produce diverse but realistic results.
- Distribution match: prediction and target distribution must be close to each other.

PART II

DATA SPACE ENRICHMENT

4 MULTIMODAL LEARNING

Chapter Highlights:

1. S2 multispectral satellite imagery is an important predictor of crop yield using ML. The data space can be further enriched with ADMs.
2. The importance of a ADMs depends on country and crop type, but can be mitigated by selecting advanced data fusion methods.
3. Spatial locality information improves pixel-based yield prediction.

In this chapter, we focus on the data space for large-scale crop yield prediction to answer RQ1. We aim to enrich the data space (model input) by evaluating different data modalities and their representations to improve the effectiveness of the model with respect to the evaluation criteria, defined in Section 1.2. Since we are working with complex and high-dimensional input data that require systematic data preprocessing and data fusion, we will further answer the following subquestion:

Data space enrichment

- **RQ1.1** Which model architecture performs best for time series regression in EO application?
- **RQ1.2** Which temporal, spectral, and spatial data modalities consistently contribute to crop yield prediction performance?
- **RQ1.3** Which data fusion method effectively combines data modalities with different temporal, spectral, and spatial resolutions?

To answer the research questions, we analyze two data fusion methods. In Section 4.4, we analyze multimodal data sources and assess their individual importance using a simple Input Fusion (IF) method. In Section 4.5, we examine a Feature Fusion (FF) method that integrates multimodal data with varying temporal and spatial resolutions. We further explore the integration of neighborhood information, VIs, and simulations in pixel-based yield prediction.

Parts of this chapter, including figures and tables, have been published already in [85], [86], [87], and [88].

4.1 Data Fusion & Machine Learning

As discussed earlier, the availability of data modalities in EO has increased rapidly over the past decades, offering deeper insights from multiple perspectives. Exploiting the richness and diversity of different data modalities can support a better understanding of the task [89]. Consequently, we could assume that adding more data to an ML model will continuously improve performance. However, studies demonstrate that this is not always the case, and the opposite is observed instead [90]. Training with multiple modalities, known as *multimodal learning*, raises several challenges, stemming from the nature of the data modalities and the ML model itself. A common challenge is the difference between spatial, temporal, and spectral resolution, which increases the difficulty of information extraction. Additionally, sensors and data collection are commonly affected by different noise levels. Therefore, combining data modalities in a meaningful way is necessary to leverage the full information from multiple sensors. This process is known as data fusion. Several definitions of data fusion are available in the literature [91]. In the following, we define data fusion as:

Multimodal learning

> **Data Fusion**
> Data fusion is the integration of complementary data sources with different sensor characteristics and acquisition conditions, using automated methods to achieve more robust, accurate, and domain-adapted analysis.

Mena et al. [92] pointed out that three main aspects must be considered during data fusion: (1) *what to fuse*?, (2) *where to fuse?*, and (3) *how to fuse?*
In the following, two approaches for crop yield prediction are described, capitalizing on IF and FF methods.

4.2 Introduction to Crop Yield Prediction

Today, agriculture faces numerous challenges from population growth and shifting environmental factors. The ever-growing population, combined with changing environmental conditions, requires increased productivity, food security, and sustainability [93]. In response to these challenges, digital agriculture has emerged as an indispensable strategy [94]. For example, large-scale crop monitoring enables farmers, insurers, and policymakers to identify gaps between actual and potential yields, better assess the impact of environmental

stress, and improve farming practices [95]. Building on this, recent advances in ML and EO have substantially enhanced the accuracy of yield prediction by introducing robust, scalable models and providing large training datasets, respectively [96, 97]. The S2 satellite mission exemplifies this progress by providing global multispectral data with high temporal frequency and resolutions up to $10\,\mathrm{m} \times 10\,\mathrm{m}$. This data enables thorough monitoring of crop development from seeding to harvesting at subfield resolution by capturing distinct crop features, including soil characteristics, chlorophyll, nitrogen, and water content. Furthermore, in the context of these technological improvements, models such as Transformers [98] and Recurrent Neural Networks (RNNs) are highly scalable and can process long time series, which are common across various EO tasks. Nevertheless, the amount of available input modalities introduces unique challenges. For instance, processing the entire time series introduces significant computational and operational challenges due to the volume and dimensionality of the data. Moreover, atmospheric conditions and sensor errors can lead to missing information in optical RS data, reducing the usability of affected instances [99, 100]. Likewise, adding more data modalities can further increase the computational complexity. Furthermore, working with multimodal data originating from sensors with varying temporal and spatial resolutions requires identifying an appropriate data fusion scheme [92].

4.3 Related Work

Yield prediction using ML and EO has gained widespread interest [101], yet remains an ongoing challenge [96]. The rise of EO and ML technologies enables yield prediction at scale, based on the globally available data sources and powerful ML models [102, 103]. One can broadly classify studies by ground-truth yield data type, region of interest, crop type, and ML method. As ground truth data, regional-level [104], field-level [105], and subfield-level data [40] are common choices. Nevertheless, most studies focus on specific regions, crop types, and individual years [106, 107, 104, 108, 40]. This can significantly increase the risk of regional or temporal overfitting [40]. In a field setting, a primary challenge in yield prediction arises from the wide range of field sizes, which leads to infrequent processing of the entire field and renders numerous vision techniques unsuitable. Consequently, many methods operate at the pixel level and disregard local neighborhood effects, such as infield dynamics, terrain, and field boundary effects. Individual studies, as evidenced in [107], address this challenge by introducing a convolutional Long Short-Term

Memory (LSTM) approach that incorporates spatial information [109]. Other common ML methods are Random Forests (RFs), MLPs, NNs, Convolutional Neural Networks (CNNs), and RNNs [110, 107, 111]. Lately, the Transformer architecture has emerged as a key component that is capable of processing long sequence data [112].

As previously mentioned, yield prediction typically involves processing time series data to capture crop development from seeding to harvest. Therefore, a key component is the selection of relevant time steps, as improper choices can even lead to economic risks or performance reduction [113]. This particularly occurs with very long time series. Expanding further on data choices, models are commonly trained on different RS data modalities such as satellite imagery, DEM, soil, and weather data [114, 103, 115]. Although most data modalities impact crop yield, only a subset is included in most studies. Consequently, it still must be determined whether the inclusion of multiple modalities continuously improves model performance across different crop types and climate zones. Regarding data fusion, concatenating the input features of multiple RS observations, known as IF, is a common choice [116]. However, certain approaches involve combining hidden features, known as FF. For instance, Maimaitijiang et al. [117] demonstrated a FF method for soybean yield prediction. Considering fusion strategies, the attention mechanism is used to aggregate time series RS data [118, 119], or to highlight input features [120].

4.4 An Analysis of Input Modalities

In this section, we provide an in-depth analysis of various data modalities and models for crop yield prediction. We set up baseline methods and, moreover, present a simple but effective fusion method that combines multiple modalities with different temporal, spatial, and spectral resolutions at the input level.

Baseline

Often, models are trained on diverse sets of remotely sensed input modalities. These include satellite imagery, weather, soil, and DEM data [114, 103, 115]. All the mentioned modalities are important factors for yield formation. However, only a subset is commonly evaluated in most studies. It must still be determined whether including multiple modalities is beneficial for model performance in all cases. To bridge this gap, we extensively analyze the contribution of various input modalities for crop yield prediction using different ML models. We propose an IF approach that combines all data modalities at the input level. This study aims for generalizability by evaluating the proposed approach on

the YieldSAT dataset (see Subsection 2.2.1) that covers a large geographic area with multiple crop types, regions, and years. This method provides a lightweight yet powerful approach for crop yield prediction, serving as a baseline for future studies. Results are evaluated at the field and subfield levels.
We demonstrate that the contribution of each ADM depends on the region and crop type, highlighting the necessity for careful and systematic modality selection based on these factors. Similarly, the selection of the ML model should be guided by the specific country and crop characteristics, We consider a model that was trained on S2 data only as the data-driven counterpart and evaluate the integration of domain expertise by adding auxiliary data modalities. Finally, we study the integration of VIs and simulations that encode prior knowledge.

Sentinel-2 only as baseline

4.4.1 Methodology

Data

Table 4.1: Yield maps data per country and crop type for different years.

Country	Crop	Years	#Fields	#Samples
Germany	Rapeseed	2016 – 2022	111	∼ 0.3M
Germany	Wheat	2016 – 2023	188	∼ 0.3M
Argentina	Soybean	2017 – 2022	190	∼ 1.4M
Uruguay	Soybean	2018 – 2022	572	∼ 1.7M

In this section, the *time series pixel dataset* is used (see notation in Section 3.1), consisting of an input $\boldsymbol{x}$ and target $\boldsymbol{y}$. As target data, a subset of the YieldSAT dataset is used (see Section 2.2). In detail, we use data from Germany, Argentina, and Uruguay for wheat, rapeseed, and soybean crops. An overview of the data is given in Table 4.1. For most experiments S2 images are used for model training with all spectral bands, as provided in Table 2.1. In addition, we include weather, soil, and DEM data. Detailed information on all available features is given in Table 2.5.

Time series pixel dataset

Sentinel-2 Data:

Additional Data Modalities:

For each sample, the input time series is represented by 24 time steps, where each time step corresponds to a single month within the growing period. For this, S2 images are used as reference data by selecting the best cloud-free S2 image among all images within each time interval. This time series representation is based on [121] and will be further analyzed in Subsection 5.2.1.
Once the reference images are selected, all ADMs are concatenated with the S2 product at each time step. For this, daily weather data is aggregated between each time interval based on the timestamp of the each S2 image. We also

24 months

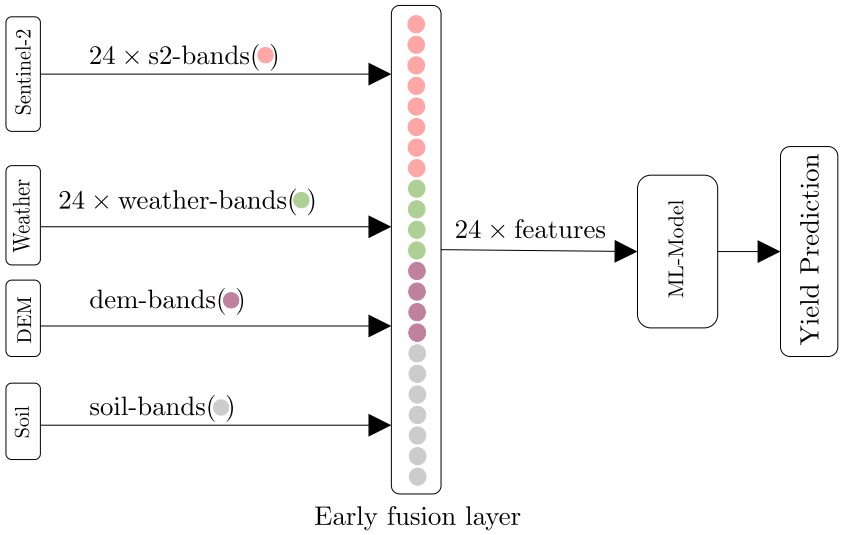


Figure 4.1: Overview of the IF method for crop yield predictions. Various data modalities with differen temporal, spatial, and resolutions are fused at the input level. A machine learning model is then trained at the pixel level to generate yield predictions at a spatial resolution of 10 m × 10 m.

tested the average but found that aggregation yields the best result. The Soil and DEM features are vectorized and repeated over each time step. This IF technique results in a multivariate time series, where each sample represents a pixel of the yield map. Each sample contains different numbers of input features, depending on the selected ADM. For soil, all eight available soil features at depths of 0-200 cm are used as provided in Table 2.5. Importantly, the IF can be processed by any ML and therefore offers high variability.

Architecture, Training & Evaluation

LightGBM & LSTM baseline

This section sets up baseline models. Therefore, only well-known ML and DL models are used to predict the crop yield, namely the LightGBM model [122] and the LSTM model. Details on the model implementation and training are provided in Section A.1.

Following the IF method, a multivariate time series is created as the model input $\boldsymbol{x}$, where each time step corresponds to a concatenated feature. The time series is then passed to an ML model for a regression task, where each sample represents a pixel with a spatial resolution of 10 m × 10 m derived from the S2 images. An overview of the proposed framework is illustrated in Figure 4.1.

For evaluation, the stratified group K-fold CV is used, as described in Section 3.3, with 10 independent folds. The performance was evaluated quantitatively and qualitatively using the evaluation metrics and guidelines described in Section 3.3.

Table 4.2: Contribution of different modalities in soybean yield prediction for Argentina using the LSTM model

Modalities	Field-Level		Subfield-Level	
	MAPE (↓) %	R^2 (↑) -	MAPE (↓) %	R^2 (↑) -
S2-Weather-Soil-DEM	0.11	0.76	0.24	0.63
S2-DEM	**0.09**	**0.82**	**0.24**	**0.65**
S2-Soil	0.10	0.76	0.25	0.61
S2-Weather	0.11	0.78	0.25	0.63
S2	0.11	0.74	0.25	0.61

Table 4.3: Results show the best-performing combination of different modalities and ML methods for distinct crops and countries at the field and subfield level crop yield prediction. ARG = Argentina, URG = Uruguay, GER = Germany. S = soybean, R = rapeseed, W = wheat.

Evaluation			Field-Level		Subfield-Level	
Model	Modalities	Dataset	MAPE (↓) %	R^2 (↑) -	MAPE (↓) %	R^2 (↑) -
LSTM	S2-DEM	ARG-S	0.09	0.82	0.24	0.65
LightGBM	S2-Weather-Soil-DEM	URG-S	0.20	0.77	1.02	0.42
LSTM	S2-Soil	GER-R	0.15	0.78	0.39	0.45
LightGBM	S2-Weather-Soil-DEM	GER-W	0.09	0.68	0.29	0.37

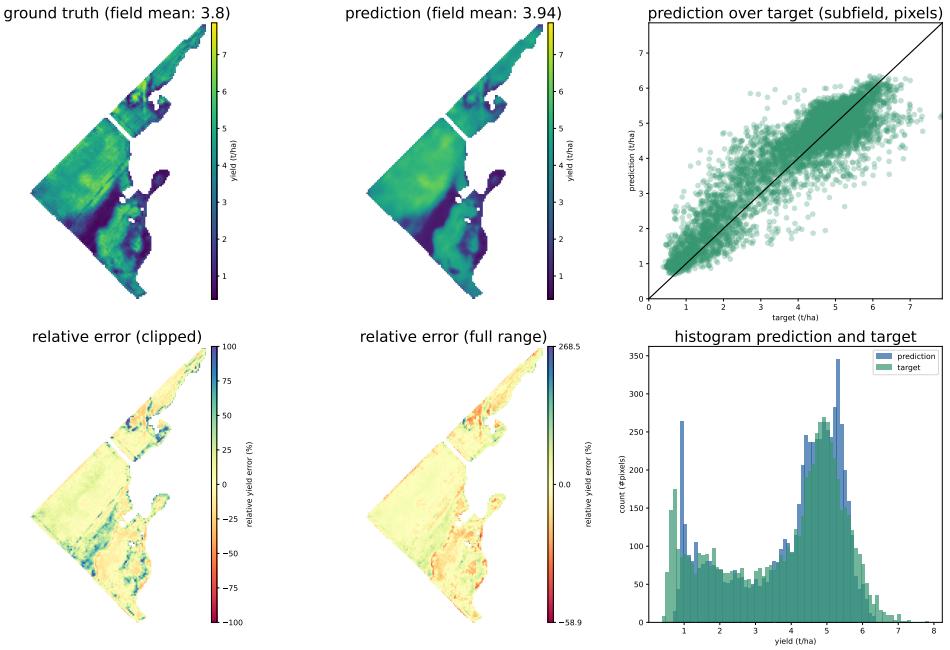


Figure 4.2: Qualitative evaluation plots for a single field from Argentina. The model was trained on S2 and DEM data. Upper left: ground truth yield map, followed by the predicted yield map, the scatterplot between predicted, and ground truth data. Lower left: relative prediction clipped at 100%, followed by the relative prediction error in full range, and the distribution plot of predictions against the target.

4.4.2 Results

Table 4.4: Contribution of the individual auxiliary modalities in soybean yield prediction for Argentina using the LightGBM model

Modalities	Field-Level		Subfield-level	
	MAPE (↓) %	R^2 (↑) -	MAPE (↓) %	R^2 (↑) -
Weather	0.17	0.54	0.17	0.54
DEM	0.24	0.13	0.49	0.06
Soil	0.24	0.10	0.50	0.03
S2	0.11	0.74	0.26	0.59

Data Modalities

In Table 4.2, an example of combining ADMs with S2 for soybean in Argentina is shown using the LSTM architecture. We show each individual data modality combined with S2. Additionally, the model trained on all data modalities is displayed. Finally, the model trained only on S2 is shown as the baseline. First, we show that including ADMs improves performance in all cases, independent of the ADM used. However, we observe varying levels of improvement. For instance, S2 combined with DEM results in the strongest improvement, evidenced by an R^2 of 0.82 at the field level, representing a significant improvement of 8 percentage points (pp) over the model trained on S2 data only. Additionally, including weather data improves the model by 4 pp. In contrast, including soil data results in an improvement of only 2 pp at the field level and none at the subfield level. Surprisingly, including all data modalities results in only a 2 pp improvement at the field and subfield levels. In Table 4.3, similar results are reported for all other crops and regions. We present results for the best-performing combination of data modalities and models at the field and subfield levels. Foremost, we observe that the combination of data modalities and models differs across countries and crop types. For instance, in Uruguay for soybean and in Germany for wheat, the combination of all data modalities with the LightGBM model performed best. In contrast, for rapeseed in Germany, S2 combined with soil data using the LSTM model performed best. Additionally, we observe significant differences in model performance in absolute numbers. For instance, the best performance is achieved in Argentina with an R^2 of 0.82 and a MAPE of 0.09, while wheat in Germany performs the worst with an R^2 of only 0.68 and a MAPE of 0.09.

Based on the qualitative evaluation, we observe good performance across countries, crops, and years for all models. We note that the presented framework captures in-field variability with high accuracy. In addition, low prediction errors and good distributional matches are observed across many examples. An example is depicted in Figure 4.2 for a soybean field in Argentina. The example

Table 4.5: Comparison of different knowledge sources to enrich the data space. Different VIs are used. SD = simulated drought stress. The results are displayed for soybean in Argentina with the LSTM model.

Modalities	Field-Level		Subfield-Level	
	MAPE (↓) %	R^2 (↑) -	MAPE (↓) %	R^2 (↑) -
VI	0.13	0.64	0.27	0.55
S2+SD	**0.12**	**0.74**	**0.25**	**0.62**
S2	0.11	**0.74**	**0.25**	0.61

was generated with the LSTM model. All figures support the exceptionally high performance for this example. The infield variability is well-preserved. However, there are regions with high pixel-wise error, especially in areas of lower yield. Nevertheless, the distribution of predicted and target yield values exhibits high consistency. Finally, in Table 4.4 we present the potential of the ADMs for crop yield prediction using the LightGBM model. These results focus on soybeans in Argentina. As expected, the individual ADMs performs worse when used alone, especially compared to the S2 data. For example, the soil variable shows only an R^2 of 0.1 at the field level. Surprisingly, the weather data alone achieves a noticeable R^2 score of 0.54 at the field level. Regardless, despite this improvement at the field level, the performance remains the same at the subfield level. This pattern is due to the lack of spatial resolution in weather data, which is repeated across every pixel within a single field.

Further Data Space Enrichment

In Table 4.5, we exchange the S2 bands with commonly used VIs for crop yield prediction. The overview of the used VIs is given in Table A.2. Additionally, we include simulation results from a process-based model that simulates temporal crop drought stress using the FAO-56 method [123], implemented in Python [124]. More details about process-based drought stress simulations will be provided in Chapter 6. Interestingly, adding VIs indices does not further improve model performance compared to the baseline model trained only on S2. In contrast, adding drought stress indices into the data space only slightly improves the MAPE score at the field-level and in R^2 at the subfield-level. However, the improvements are negligible.

4.4.3 Summary

ML models are well-suited for yield predictions over countries, crops, and years. Surprisingly, we observe that the choice of data modalities and models

depends on region and crop type, underscoring the importance of selecting input features for ML-based crop yield prediction. However, this process can be time-consuming and computationally expensive. Nevertheless, models trained on multimodal data consistently outperform those trained only on S2 data. Adding ADMs increases field-level performance and, moreover, improves subfield-level performance. In contrast, further enriching the data space, e.g., with VIs, does not improve the performance. Additionally, adding simulation results on drought stress from process-based crop models only slightly improves model performance, likely due to a gap between synthetic and real-world data that increases input noise. Additionally, generating simulation results and VIs is expensive, reducing the effectiveness of such methods for data space enrichment. In this study, we specifically focused on evaluating IF methods and their performance in crop yield prediction. It is still an open question whether alternative fusion approaches could more effectively capture yield-driving features and and avoid the expensive modality selection.

4.5 Multimodal Fusion with Neighborhood Information

In the previous section, we observed that IF requires careful selection of modality, depending on the country and crop type. Unfortunately, this is an expensive and time-consuming process. To overcome this, this section evaluates a FF method that handles each modality independently, using a modality-specific encoder to better account for each modality's individual characteristics. Moreover, the encoded modalities are combined using an attention mechanism [98] to avoid explicit modality selection. In addition, previous research has primarily focused on pixel-based yield mapping, processing each pixel independently. However, this overlooks local neighborhood effects within the field. This includes terrain and soil characteristics, as well as the spread of pests and diseases within the field. Treating each pixel as an independent sample may not capture such dynamic patterns within the field. Therefore, including local neighborhood information might provide a more comprehensive understanding of spatial relationships, patterns, and interactions within the field environment. In Chapter 3, we discussed that convolutional layers capture locality with strong inductive bias. To explicitly account for local neighborhood effects, a CNN architecture is employed, along with each sample's geographical coordinates. We highlight that incorporating neighborhood information improves over the baselines. Moreover, we highlight that the FF method based on attention mechanisms better captures the non-linear nature of yield formation.

Locality with convolutional layers

We consider a model trained solely on S2 data, without spatial context, as the data-driven baseline and evaluate the integration of domain expertise by adding ADMs and spatial context.

No locality with S2 as baseline

4.5.1 Methodology

Data

For this study, the same subset of the yieldSAT dataset as described in Subsubsection 4.4.1 is used to ensure comparability. Additionally, the same input modalities are used, including all available bands of the S2 product and the ADMs (weather, soil, and DEM). Furthermore, to account for the geographical context and neighborhood information, the sample coordinates (coord) are used as latitude (φ) and longitude (λ). All coordinates of a respective sample are projected into a three-dimensional space as:

$$(x, y, z) = (\cos\varphi\cos\lambda,\ \cos\varphi\sin\lambda,\ \sin\varphi).$$

This projection maps the latitude and longitude on the unit sphere to avoid wraparounds and singularities and eliminates the artificial discontinuity at $\pm 180°$.

Architecture, Training & Evaluation

To capture and model local neighborhood information, each pixel is processed together with its surroundings. For this, the *time series image dataset* is used (see notations in Section 3.1), consisting of an input image X and a target Y. For each sample, a window of size $H \times W$ is spanned around each pixel, with the center pixel being the pixel of interest, and the surrounding pixels provide additional information. For all experiments, a window size of 5×5 is used.

Local neighborhood information

For evaluation, the stratified group K-fold CV is used, as described in Section 3.3, with 10 independent folds. The model performance was evaluated quantitatively and qualitatively using the evaluation metrics and guidelines described in Section 3.3. To assess the impact of local neighborhood information and the data fusion method, we present results from the baseline model defined in the previous section. The baseline model is defined by the IF model trained on S2 and all combined ADM, as described in Section 4.4.

3D-LSTM: For the IF model, the input changes insignificantly. In the context of labeled training data, where the input is represented as $X \in \mathcal{X}$ and the corresponding target as $Y \in \mathcal{Y}$. The dimensions of the input data are now

denoted as $X \in \mathbb{R}^{N \times B \times T \times H \times W}$, and for the target, $Y \in \mathbb{R}^{N}$. In this representation, N refers to the total number of samples, B the number of bands, and T the number of time steps. Additionally, $H \times W$ provides the spatial context. The input modalities are fused at the input level as described in Section 4.4 using the 24-month time series sampling method. Building on the input structure, we use a conv3d block that applies 3D convolution across multiple input planes to capture the locality within the field. Further, the output is passed through a LSTM cell with 2 layers and 64 hidden units, as previously described in Section 4.4. Implementation and training details are given in Subsubsection A.1.2. We refer to this model as 3D-LSTM.

24 months

3D-LSTM

Multimodal Attention Fusion (MMAF): In the following, the FF architecture is described. For this, we utilize independent modality encoders. We distinguish between temporal features and static features. For temporal modalities, including S2 and weather, the dense time series is used. Since each field has varying time intervals between seeding and harvesting, padding is required to ensure a common time series length of the input. In contrast to the IF pipeline, static features, such as soil and DEM, are not repeated over each time step but are instead processed only once using specific modality encoders for each modality. Hence, the input vector is a set of individual inputs $\{X_{S2}, X_{DEM}, X_{soil}, \boldsymbol{x}_{weather}, X_{coord}\}$. The main components of this architecture include a conv3d, a conv2d, an FCL, and an LSTM block. Additionally, a *scaled dot-product attention* [98] is used to fuse the feature representations across each data modality. The spatio-temporal features (S2) are processed by the 3D-LSTM, as described earlier, to map the input to a feature representation. Spatial features, including DEM, soil, and coordinates, have no temporal dimension and are therefore handled by a conv2d block. In parallel, temporal features (weather) have no spatial information and are processed using a LSTM block. To fuse the features of each modality, we use a scaled dot-product attention mechanism [98]. This mechanism enables attention pooling, in which a learnable query interacts with each modality via cross-attention. This generates attention weights for each data modality. This allows the model to focus on important data modalities while ignoring less important ones. Importantly, we can use the attention weights to assess the importance of each individual data modality. The architecture of the FF model is illustrated in Figure 4.3. We refer to this model as MMAF. Details on the model implementation and training are provided in Section A.1.

Multimodal Attention Fusion

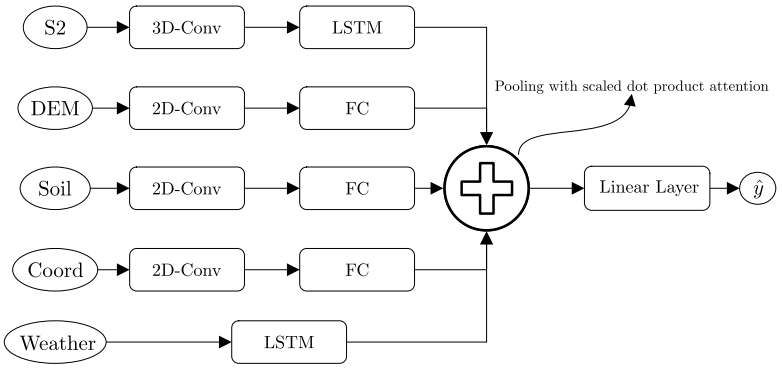


Figure 4.3: Overview of the MMAF architecture. Each modality is separately encoded. Subsequently, the modalities are fused using scaled dot-product attention pooling and fed to a linear layer.

4.5.2 Results

The quantitative results are shown in Table 4.6 for both data fusion strategies and the baseline model. We highlight that including spatial information results in performance either equal to or better than that of the baseline model. For the 3D-LSTM, an improvement of 4 pp in R^2 at the field level is reported on the wheat dataset in Germany. However, we also observe cases with no improvement, as evidenced by the soybean dataset in Uruguay. Here, a reduction of 1 pp in the R^2 is shown. In contrast, the MMAF consistently outperforms both the baseline and the 3D-LSTM model across all datasets. More specifically, on the soybean dataset in Argentina, an R^2 of 0.86 at the field level is reported, representing an improvement of 10 pp over the baseline model and 8 pp over the 3D-LSTM model. Additionally, a reduction in RMSE of 0.13 t/ha is reported at the field level and 0.8 t/ha at the subfield. Moreover, on the wheat dataset in Germany, an improvement of 15 pp R^2 at the field level is achieved over the baseline model.

Attention Scores: We further examine the attention weights for each data modality in the MMAF model to assess the relative importance of each modality. Figure 4.4 shows the average attention weight for each data modality. The average attention weights are calculated across all samples in the validation split over all CV folds. The figure highlights how the model learns to assign different levels of attention to the individual data modalities. However, the S2 modality consistently shows the highest attention scores. Nevertheless, we also notice that attention weights vary slightly across different countries and crop types. For instance, in the soybean dataset in Uruguay, the ADMs receives the least attention across all samples, with most attention being concentrated on

Table 4.6: Overview of the performance for the 3D-LSTM and MMAF fusion method with local neighborhood information. Results are shown for different crop types and countries in the yieldSAT dataset. The best scores are highlighted. ARG = Argentina, URG = Uruguay, GER = Germany. S = soybean, R = rapeseed, W = wheat.

Evaluation		Field-Level				Subfield-Level			
Dataset	Model	MAE (↓) t/ha	RMSE (↓) t/ha	MAPE (↓) %	R^2 (↑) -	MAE (↓) t/ha	RMSE (↓) t/ha	MAPE (↓) %	R^2 (↑) -
ARG-S	3D-LSTM	0.37	0.49	0.10	0.78	0.67	0.90	0.25	0.62
	MMAF	**0.27**	**0.39**	**0.08**	**0.86**	**0.60**	**0.81**	**0.23**	**0.70**
	Baseline	0.40	0.52	0.11	0.76	0.66	0.89	0.24	0.63
URG-S	3D-LSTM	0.37	0.52	0.18	0.76	0.81	1.22	**0.91**	0.40
	Feature Fusion	**0.32**	**0.46**	**0.17**	**0.81**	**0.78**	**1.19**	**0.91**	**0.43**
	Baseline	0.35	0.51	0.20	0.77	**0.78**	1.22	1.02	0.42
GER-R	3D-LSTM	0.49	0.64	**0.14**	0.77	0.90	1.22	0.36	0.46
	MMAF	**0.44**	**0.60**	**0.14**	**0.80**	**0.87**	**1.20**	**0.35**	**0.49**
	Baseline	0.49	0.65	0.15	0.77	0.93	1.23	0.38	0.46
GER-W	3D-LSTM	0.80	1.05	0.09	0.70	1.67	2.30	0.27	0.39
	MMAF	**0.61**	**0.83**	**0.07**	**0.81**	**1.51**	**2.13**	**0.25**	**0.48**
	Baseline	0.84	1.11	0.09	0.66	1.71	2.37	0.29	0.35

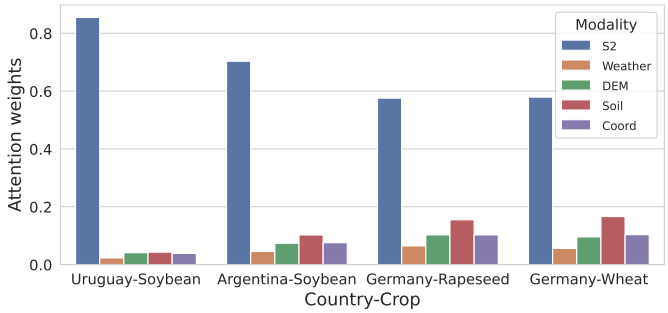


Figure 4.4: Bar plot of the aggregated attention weights derived from the MMAF model for all data modalities and different countries and crop types.

the S2 imagery. In contrast, Germany shows higher attention to ADMs, with soil exhibiting the highest values over wheat and rapeseed crops.

4.5.3 Summary

We introduced two methods that incorporate spatial context by using convolutional layers and geographical coordinates, namely the 3D-LSTM and the MMAF model. The results show that including neighborhood information improves crop yield prediction. Furthermore, the MMAF method accounts for the varying temporal and spatial resolutions of the multimodal input data by treating each modality separately using an attention-based FF architecture. We demonstrated that attention weights can be used to fuse data modalities, thereby avoiding the need for expensive modality selection. Additionally, we demonstrated that S2 receives the highest attention values, compared to the ADMs.

4.6 Discussion

In this chapter, the potential of multimodal learning for crop yield prediction was explored to answer RQ1. For this, various data modalities with different temporal, spatial, and spectral resolutions were analyzed under two different data fusion schemes. We observed distinct performance levels across all models and datasets (crops) (see Table 4.3 and Table 4.6). These differences may result from variations in data quality, dataset size, and data complexity. For example, the German dataset is much smaller than those from Argentina or Uruguay. Additionally, data quality in Germany and Uruguay is lower than in Argentina. Data quality is a serious concern in yield-mapping and directly affects model performance and reliability. It is important to dedicate greater attention to data collection, calibration, and preprocessing. In answering RQ1.1 we showed that the LSTM architecture provides consistently good results and can be extended with additional components like 3D convolution (3D-LSTM) or advanced fusion methods (MMAF).

In Section 4.4 to answer RQ1.2, we thoroughly analyzed input modalities using a simple yet effective IF approach. Surprisingly, each dataset had a different combination of data modalities that performed best (Table 4.3). Nevertheless, including ADMs always improves performance, whereas adding VIs or simulations does not. However, this method requires expensive modality selection. To address this limitation and to answer RQ1.2 and RQ1.3, we introduced the MMAF architecture, which delivers improved model performance while continuously utilizing all data modalities. This is based on the attention mechanism. Consequently, the model learns to concentrate on more important features. Interestingly, this mechanism enables the evaluation of which data modalities contribute to the model performance in each agro-ecological environment (Figure 4.4). Consistently, S2 was found to have the highest importance to the model prediction. Nevertheless, the contribution of the ADMs should not be underestimated for individual environments. This is complemented by further studies [125, 87]. Nevertheless, ADMs alone are inefficient for crop yield prediction, and therefore, always require the integration of satellite imagery (see Table 4.4). Multispectral satellite data captures distinct characteristics of crop development, including biochemical properties, vegetation density, and differences in the crop cycle [74], which are not captured by the auxiliary variables. Additionally, further limitations of previous studies are addressed by integrating spatial locality to better account for in-field dynamics. The results demonstrate that including neighborhood information enhances model performance in crop yield prediction, especially for the MMAF model. In

contrast, the 3D-LSTM model showed limited benefit from including local neighborhood information. We argue that this is due to the concatenation of features with differing spatial resolutions at the input level. This may cause redundancies. For instance, weather data is repeated across the spatial dimension. Likewise, soil and DEM data are repeated over time. This may introduce redundancy and could lead to confusion within the model. Consequently, the spatial context may not be leveraged effectively. Wang et al. [90] underlined the difficulty of training multimodal models and emphasized that although a multimodal network receives more information when integrating auxiliary data, only slight or no improvements can often be observed in practice. This may be due to overparameterization or overfitting. It is, therefore, important to acknowledge the specific characteristics of each input stream.

Context & Limitation

Although this chapter investigated different data modalities and data fusion methods, additional modalities and fusion methods remain to be investigated [92]. For instance, Mena et al. [125] proposed an adaptive gated fusion method for crop yield prediction by capitalizing on a Gated Unit that allows the computation of fusion weights for each data modality. Notably, they support our results by underlining S2 as the most important modality for yield prediction. Moreover, while we investigated the contribution of adding auxiliary to our models and analyzed the attention weights for each modality, we did not explicitly assess feature importance. This was investigated in [126, 87]. For instance, Najjar et al. [87] extensively evaluated various feature attribution methods for crop yield prediction to assess the contribution of each data modality. The results underline that S2 is an important predictor for crop yield. Nevertheless, the attribution depends on individual S2 bands and specific auxiliary features. More importantly, this study highlights the importance of each time step for yield prediction, identifying individual time steps and growth stages that influence yield prediction. This will be investigated in the next chapter.

4.7 Conclusion

This chapter demonstrates the potential of EO-based yield prediction using multimodal data and ML. However, selecting an appropriate data fusion method remains challenging. Our results show that IF is a simple and efficient approach, though it struggles with differing spatial and temporal resolutions and requires costly modality selection. To overcome these limitations, we introduced an attention-based FF method, MMAF.

5 DOMAIN-INFORMED TIME SERIES ANALYSIS

Chapter Highlights

- The number of processed time steps impacts the computational efficiency. The representation of time series data is therefore of significant importance for efficient, large-scale crop yield prediction.
- The monthly sampling and growth stage sampling are two informed time series representations that improve computational efficiency while maintaining high performance.
- The self-attention mechanism of Transformer models can be leveraged for intrinsic explainability of important time steps.

In the previous chapter, we demonstrated the importance of the input features. We highlighted challenges in working with multimodal data that is characterized by different temporal, spatial, and spectral resolutions. In this chapter, we continue with RQ1 and the subquestion RQ1.2, focusing specifically on the temporal dimension of the input data. We analyze the importance of time series data for yield prediction. We investigate different representations of time series data and demonstrate the advantage of including expert knowledge.

5.1 Time Series Analysis in Crop Yield Prediction

We have already demonstrated that S2 satellite imagery, coupled with ML, enables large-scale yield prediction for different crop types with global scalability. However, when using the entire available time series, the abundance of input data can introduce significant computational costs. Besides, atmospheric conditions and sensor malfunctions often lead to missing information in optical EO data, as discussed earlier. This may reduce the information content in affected instances [99, 100]. While ML models can, in principle, learn to ignore uninformative signals, the computational burden remains high [74]. Additionally, several studies indicate a positive relationship between model

Parts of this chapter, including figures and tables, have been published already in [127], and [87].

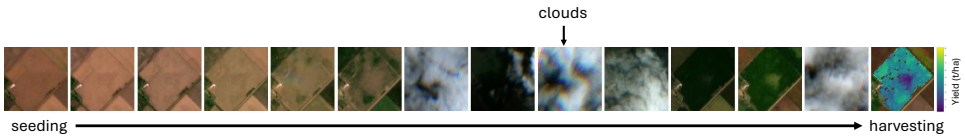


Figure 5.1: Example S2 time series of a field from seeding (left) to harvesting (right). The last image shows the yield collected at harvest. Note that some images are corrupted by clouds.

performance and the availability of informative (cloud-free) image instances [128, 125]. Nevertheless, Najjar et al. [87] emphasized that individual time steps are more important for crop yield prediction. This raises the question of whether we can reduce the time series length while maintaining high performance and simultaneously reducing the computational complexity.

In this section, we investigate different time series modeling techniques in terms of performance, efficiency, and explaiability. In detail, we analyze the entire time series and the monthly sampling used in the previous chapter. Moreover, we present an approach that selects time steps based on the crop growth stages.

Growth stages

More specifically, a single time step per growth stage is selected to account for the underlying crop phenology. This method achieves comparable or superior performance to the previous baseline methods while reducing training time. For all experiments, the Transformer architecture [98] is used due to the intrinsic explainability of the self-attention mechanism [129]. We leverage the attention mechanism to enhance explainability by highlighting critical time steps for yield prediction.

5.2 Methodology

5.2.1 Deriving Domain-Informed Time Series

Modeling the plant growing period requires a proper definition of the available time series to capture essential crop characteristics [74, 113]. Hence, the objective should be to acquire sufficient time steps such that the model can learn important patterns that correlate with the measured yield. At the same time, the computational costs should be minimized. This section compares the different time series modeling techniques and evaluates their performance, applicability, and computational efficiency.

Cost Function: A major factor that must be considered when selecting a time series representation is the computational complexity. For instance,

Computational complexity

Transformers perform pairwise computations over all time steps, making them computationally expensive for long sequences. In a Transformer, the self-attention is computed by pairwise interactions across all time steps T. At each time step, a query (Q), a key (K), and a value (V) are produced. The attention weights are then computed as:

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{5.1}$$

with the matrix QK^T, representing the pairwise comparison of the time steps, with dimension $T \times T$. The cost function is therefore defined as $\mathcal{O}(T^2 \cdot d)$. Here, d defines the embedding size of the input data. Given this, the sequence length T significantly affects the model's overall computational complexity. Besides computational complexity, it is also important to consider the implementation effort, generalizability, and explainability of a time series representation. Often, these aspects are at odds with one another. For instance, reducing computational complexity often requires more domain knowledge to handcraft features or simplify representation. For non-experts, designing effective time series representations is challenging and can even lead to significant errors, as critical time steps in crop development may be overlooked or misrepresented.

Crop type and region agnostic

In the following, we describe three time series representations that require increasing amounts of domain knowledge while decreasing computational complexity. Each method must be agnostic to every crop type and country.

Dense Time Series

The dense time series is the complete time series of all available time steps, including noisy measurements such as cloud-corrupted instances. An example of such a time series is given in Figure 5.1 For this, minimal preprocessing is required, making it easy to implement. Moreover, the model can learn from sufficient data and therefore often outperforms other preprocessing methods [40]. However, dense time series often contain higher levels of artifacts, such as sensor errors, missing time steps, or cloud-corrupted images. Although cloud-occluded images can, in theory, be ignored by a NN, cloud contamination remains a significant challenge when working with optical EO data. For example, by degrading model performance or increasing predictive uncertainty [74, 130, 44]. Furthermore, uninformative time steps do not contribute to a successful algorithm but increase the computational demand. Consequently, the computational costs of a dense time series are high. Moreover, a dense time series is often restricted to powerful DL models, as they are more suited

to learning from large and noisy datasets [74]. However, classical ML models, including LightGBM, are still very popular in the agricultural domain, as they do not require heavy resources and are often much more interpretable than a black-box NN.

Monthly Sampling

The monthly sampling selects a single image per month, as proposed in [121]. This time-series representation defines a time series spanning two calendar years (24 months). For this, the seeding year (SY) and harvesting year (HY) for each yield sample are defined. Subsequently, a period from 1.1.SY to 31.12.HY is defined (SY-1 if the seeding year is equal to the harvesting year) and split into 24 equally spaced time intervals. Subsequently, each time interval represents a single month. Lastly, a single satellite image is placed into each time interval. To reduce the number of cloud-corrupted instances, the first and best cloud-free image is selected each month using the SCL Layer. More specifically, only pixels with the SCL class 4 (”vegetated”) and 5 (”not vegetated”) are considered. The remaining instances, i.e., those before seeding and after harvesting and cloud-corrupted instances, are masked to ensure a homogeneous representation of non-informative measurements. This framework requires a higher implementation effort but significantly reduces the time series length compared to the dense time series. Moreover, the sequence consists of equally spaced intervals and is crop type agnostic. Still, it results in redundancies, since most crop types have a shorter growing period than 2 years (e.g., the average growing period for soybeans in Argentina in our dataset is 156 days). This results in the models being fed with non-informative data, unnecessarily increasing computational cost. Moreover, the monthly sampling method may not provide sufficient information to capture the complex, non-linear behavior of plant development, since crop development is divided into specific growth stages with variable lengths [131] rather than months.

Growth Stage Sampling

The growth stage sampling further reduces the number of processed time steps while aiming for high information content. This method assumes that a single data point per growth stage is sufficient to accurately model the crop growing cycle. Here, crop growth stages refer to the distinct developmental stages a plant undergoes during its lifespan. However, Vallentin et al. [113] pointed out that the relationship between yield and RS information is not consistent over the distinct growth stages. Thus, the acquisition time is the most important

Algorithm 1 Assign Cloud-Free Sentinel-2 Images to Growth Stages

Require: Yield data D with N samples
Require: Sentinel-2 images I with SCL mask Q from seeding to harvesting
Require: Growth stages definitions $G = \{g_1, g_2, \ldots, g_t\}$
Require: Time bounds for growth stages $T = \{(t_1^s, t_1^e), \ldots, (t_t^s, t_t^e)\}$

```
for each sample d in D do
    Initialize S_d = {}                          ▷ Storage for time series of cloud-free images
    for each growth stage g_i ∈ G do
        images ← {I_j | I_j ∈ I, t_j ∈ [t_i^s, t_i^e], Q ∈ {4, 5}}          ▷ Filter images
        if images ≠ ∅ then
            S_d[g_i] ← images[1]                 ▷ Assign first image to current stage
        else
            S_d[g_i] ← null                      ▷ No valid image for this stage
    Save S_d as the time series for sample d
return Time series {S_d | d ∈ D}
```

Figure 5.2: Algorithm for the crop growth stage time series sampling method. For each distinct growth stage, an input signal is sampled and placed into each time interval.

factor for yield prediction. As plants show different correlations between growth stages and yield [113], the question arises: *which growth stage should be focused on?* Najjar et al. [87] noted that the importance of time steps varies between crop types in crop yield prediction with ML. Nevertheless, a crop-agnostic approach should provide at least a single image per stage. Therefore, a single best cloud-free S2 image per growth stage is selected. Therefore, for each sample, the input data is represented as a sequence of T time steps corresponding to the distinct crop growth stages. We focus on the BBCH system [131] that divides the growing period into 10 major growth stages, from *germination* (0) to *senescence* (9). An overview of the growth stages for different crop types is given in Table A.1. For each growth stage, the start and end points must be available. Next, we place the first, best, and cloud-free S2 image in each time interval, using the SCL layer to select only pixels from classes 4 (vegetated) and 5 (non-vegetated). A detailed description of the algorithm is provided as pseudocode in Figure 5.2. Additionally, Figure 5.3 provides a comparison between the monthly sampling method and the growth stage sampling method. Note that growth stages vary in length, leading to unevenly spaced intervals. Moreover, the monthly sampling methods include time steps that exceed the growing period.

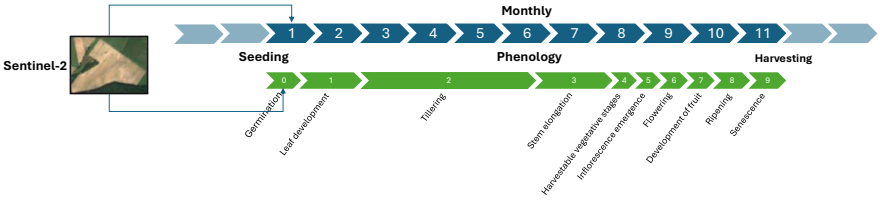


Figure 5.3: Visualization and comparison of the monthly sampling and growth stage sampling method. For each time interval, a single signal is sampled.

Table 5.1: S2 statistics for each country and crop type.

Country	Crop	Years	Avg. # S2 Images	Cloud Coverage (%))
Germany	Rapeseed	2016 - 2022	80	41.58
Germany	Wheat	2016 - 2023	96	48.15
Argentina	Soybean	2017 - 2022	43	8.34
Uruguay	Soybean	2018 - 2022	40	6.90

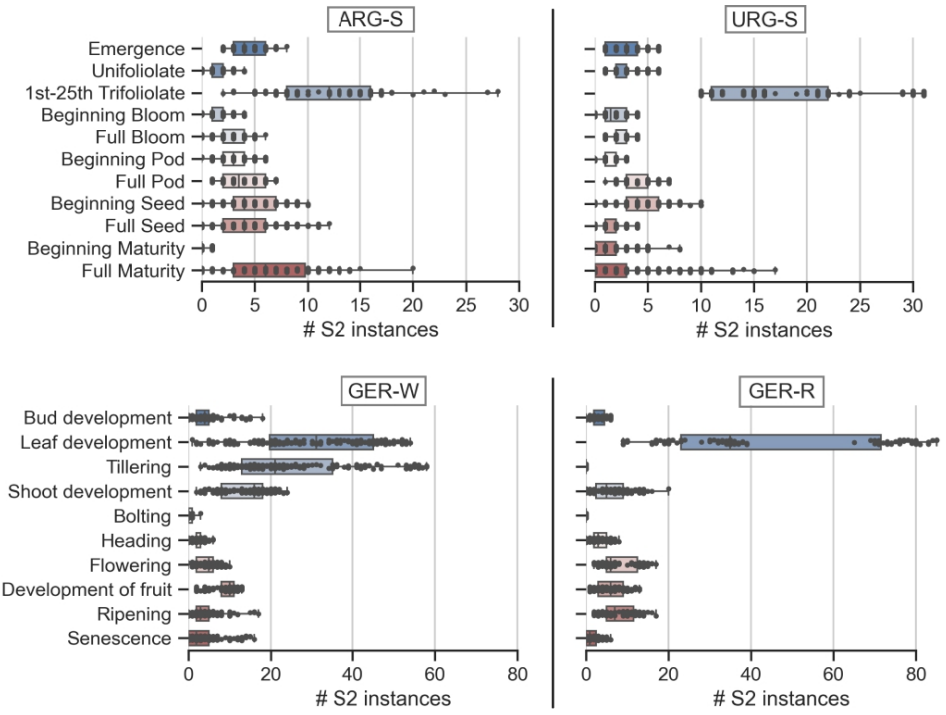


Figure 5.4: Number of S2 instances for all crop datasets over each growth stage. A translation of the growth stages for each crop type is given in Section A.2. ARG = Argentina, URG = Uruguay, GER = Germany. S = soybean, R = rapeseed, W = wheat. (Source: [87])

5.2.2 Data

We use the same subset of the YieldSAT dataset as in the previous experiments. Details on the target data are given in Table 4.1. As model input, only S2 data

Table 5.2: Average time per forward pass for different time series representations in milliseconds. The models were trained on a V100-16GB GPU with a batch size of 2048.

Evaluation	Monthly	Growth Stage	Dense
Time	33.16 ms	18.09 ms	269.11 ms

is used with all spectral bands. We use only S2 data for model training, since we demonstrated earlier that S2 alone is a good yield predictor that meets all the requirements of this study while reducing computational requirements. Following, we explore the availability of cloud-free S2 instances. Table 5.1 shows the number of available S2 instances for each country and crop type. Notably, for Germany, significantly more S2 images are available. This is due to the longer growing periods in winter crops. However, cloud coverage is also higher in Germany, which is commonly attributed to the higher cloudiness during the winter period [132]. Moreover, we provide the availability of S2 instances for each growth stage and dataset in Figure 5.4. Notably, not every growth stage has S2 data, either because no high-quality image is available or because of the crop physiology. For instance, rapeseed lacks distinct growth stages (e.g., tillering and booting), so no images are available. The same applies to soybeans. A detailed overview of the growth stages for each crop type and the generation of growth stage data is provided in Section A.2.

5.2.3 Architecture, Training & Evaluation

All experiments are conducted using the Transformer encoder architecture (see Subsection 3.2.1). Implementation and training details are given in Subsubsection A.1.2. The model is trained and evaluated on countries and crop types independently, using the stratified-grouped K-Fold CV (see Section 3.3) and the evaluation criteria as described in Section 3.3. Additionally, we evaluate the training times for each time series representation. Finally, the attention scores for the growth stage sampling are evaluated.

5.3 Results

Table 5.2 shows the average time per forward pass for each time representation. All models were trained on a V100-16GB GPU. The growth stage sampling method requires only 18.09 ms per forward pass, compared to 33.16 ms for the monthly sampling and 269.11 ms for the dense time series. The total training

Table 5.3: Total training times for different time series representations (hh:mm). Models were trained on a V100-16GB GPU. The fastest method is highlighted.

Evaluation	Germany		Argentina	Uruguay
	Rapeseed	Wheat	Soybean	Soybean
Monthly	00:55	03:11	09:51	07:06
Growth Stage	**00:43**	**01:13**	**06:01**	**06:27**
Dense	01:11	02:58	09:15	11:25

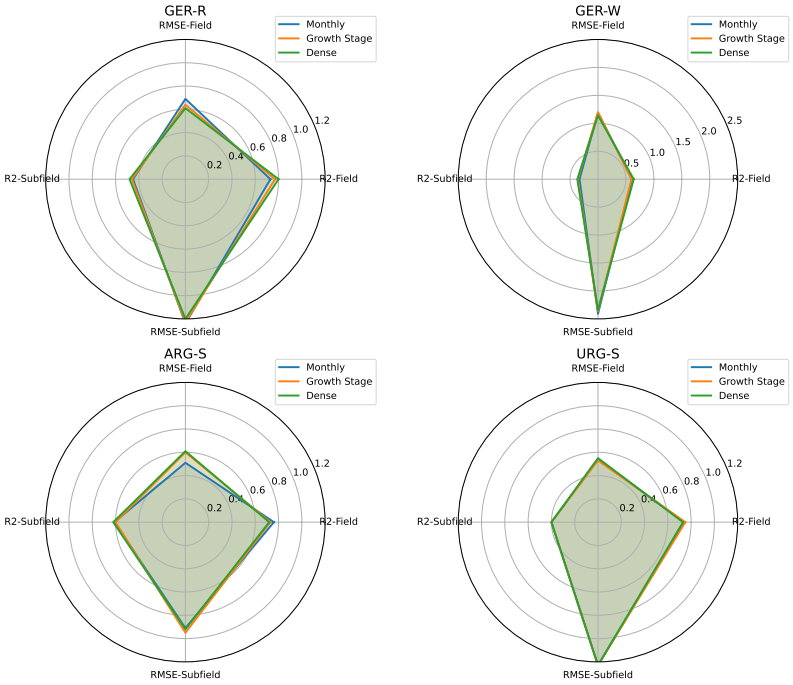


Figure 5.5: Radial error plot per dataset and time series sampling method for the R^2 and RMSE at the field and subfield level. ARG = Argentina, URG = Uruguay, GER = Germany. S = soybean, R = rapeseed, W = wheat.

times also differ significantly. Table 5.3 presents the total training times for all time representations and datasets. As expected, the dense sampling method requires the longest training times, with a maximum of 11:25 h for the soybean in Uruguay. In contrast, the growth stage time series requires significantly less time across the datasets. Specifically, only 06:27 h is required for soybeans in Uruguay. Likewise, a reduction of approximately 4 hours has been observed in soybeans in Argentina. However, for Germany (wheat), the difference is less significant.

Figure 5.5 shows the R^2 and RMSE metrics for all datasets at the field and subfield level as a radial plot. This plot facilitates the comparison between the datasets and the time series modeling technique. Across all methods, we observe accurate performance across countries and crops for all time-series representations, with only minor but inconsistent variations. The plot shows

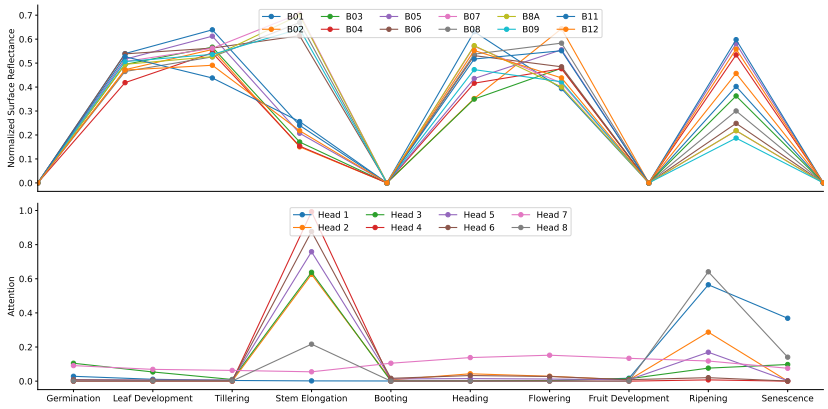


Figure 5.6: Example time series for a randomly selected wheat field. Top: S2 reflectance for each band and growth stage. Bottom: self-attention scores for each attention head and growth stage.

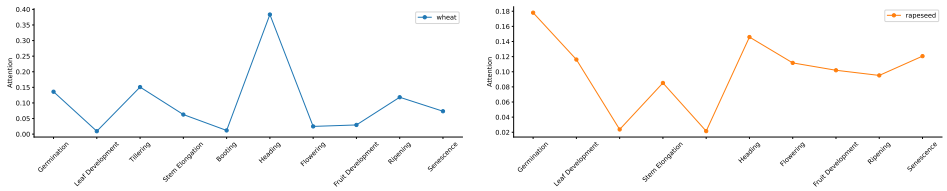


Figure 5.7: Temporal attention for the growth stage time series sampling method for wheat (left) and rapeseed (right). The attention weights are averaged over a random selection of 2048 samples.

that all time series sampling methods perform comparably well across datasets and metrics. Only small differences are visible (e.g., on the ARG-S dataset). We even observe improvements in R^2 and RMSE at the field and subfield levels for the growth stage sampling compared to monthly sampling.

Self-attention for explainability

We continue analyzing the self-attention mechanism for intrinsic explainability. The model leverages multi-head attention, defined as an attention matrix. The attention scores define the relevance of one time step to another in the sequence and are used to weight each input for the final output. Figure 5.6 visualizes the S2 surface reflectance over each time step. Zero surface reflectance means no data is available. The input time series shows that only specific growth stages are missing. Below, the self-attention of all attention heads is displayed for every growth stage. The model exhibits a notable pattern in its attention allocation between growth stages. For instance, high attention scores are assigned to *shoot development* and *ripening*. In contrast, less attention is assigned to missing, early, and mid-stages. Nevertheless, individual heads also allocate attention to the mid and early stages. In Figure 5.7, we display the temporal attention averaged over the entire dataset for wheat (left) and rapeseed (right).

For wheat, we observe that different growth stages receive high attention, including *germination*, *tillering*, *shoot development*, and *ripening*. In contrast, for rapeseed, the early growth stage (*germination*) and the later growth stages (*heading*, *flowering*, *fruit development*, *ripening*, and *senescence*) play a more important role. Importantly, no attention is allocated to missing growth stages (e.g., *tillering* and *bolting*).

5.4 Discussion

In this chapter, the importance of time series representation of the input data for model accuracy, efficiency, and explainability was examined to answer RQ1 and RQ1.2. We identified critical aspects that must be considered when selecting a time series representation. First, the availability of time series data and time series sampling method is essential for effective yield prediction [113]. However, we showed that various time series sampling methods perform comparably well across countries and crop types (Figure 5.5). Nevertheless, although all evaluated time series representations are crop type- and region-agnostic, each method has its own advantages and limitations. While the dense time series requires no preprocessing and only little domain knowledge, the computational burdens are exceptionally high (see Table 5.3). This raises the concern that, with greater data availability, training DL models might be accessible to only a small fraction of the community. Additionally, it causes a bigger carbon footprint. Unfortunately, research is still paying little attention to computational efficiency by introducing ever-growing models trained on ever-growing and massive datasets [133].

Consequently, as energy demands grow, DL models will finally contribute to global carbon emissions. Assessing the climate impact of DL systems is essential for mitigating the negative climate effects of future DL research [134]. Reducing computational burdens can contribute to a more inclusive ML research and toward achieving the SDGs, especially SDG 13 (climate action).

Moreover, the dense time series sampling is less interpretable. To mitigate this, additional post-processing steps are required, e.g., attention aggregation [87], which incurs additional overhead. Here, the monthly sampling and the growth stage sampling can overcome this limitation by significantly reducing computational complexity while increasing the interpretability of a time series. This particularly holds for the growth stage sampling, since each time step marks a specific stage in the life cycle of each crop. The results support the hypothesis that a single image per growth stage is sufficient to capture the entire growing

cycle. This is underlined by the analyzed attention score. We demonstrated that less attention is paid to time steps with cloud-corrupted or missing data. Nevertheless, the growth stage sampling requires more preprocessing effort but significantly less computational burden (Table 5.3).
Additionally, we demonstrated high intrinsic explainability by analyzing the attention scores (see Figure 5.7, Figure 5.6) and demonstrated consistent alignment with agronomic principles. For instance, for wheat crops, *tillering*, *heading*, and *ripening* are associated with high attention scores. *Tillering* is critical for canopy formation and is also associated with yield-limiting factors [135, 136]. *Ripening* is similarly important, as grain filling is crucial for high yields. Especially, environmental stress can cause a significant reduction in yield during these stages. This distribution of attention indicates the model's alignment with agronomic principles, emphasizing the importance of both early vegetative growth and later reproductive processes in predicting crop yield.

Context & Limitations

Further research is required to determine whether attention scores can reliably contribute to more explainable models [129]. Moreover, additional methods are required to carefully assess the importance of each time step. Nevertheless, we observe high consistency between attention scores, attribution methods, and agronomic principles. For instance, similar to our results, Najjar et al. [87] demonstrated that *tillering* was highly important using explicit feature attribution methods. This is underlined by agronomic principles that the *tillering* stage is the most important phase during canopy formation. Consequently, a model must place attention on this stage. Nevertheless, collecting growth stage information is expensive and requires expertise and specialized models. Facilitating growth stage acquisition can significantly improve the efficiency of ML with high impact. Therefore, the results suggest that this method is particularly useful for domain experts who focus on downstream interpretation and efficient training.

5.5 Conclusion

Time series analysis remains a challenge in EO. We demonstrated that time series patterns significantly influence the computational complexity and explainability. However, those aspects are antagonistic, and selecting a time series depends on a user's needs and must be done carefully.

PART III

ENFORCING KNOWLEDGE CONFORMITY

6 PHYSICS-GUIDED LEARNING

Chapter Highlights

- Integrating prior knowledge into the learning algorithm and the loss function, defined as PG, increases physical consistency, explainability, and trustworthiness.
- A NN can approximate the relationship between crop stress and crop yield by solving the yield response to water function.
- Prior knowledge can be used to upsample simulation data of crop stress from low spatial resolution to high spatial resolution.

In the previous chapter, we studied the integration of domain knowledge into the data space. However, this does not guarantee scientific consistency. Instead, the model remains a black-box and might not follow the underlying physical principles of plant growth [17]. In this chapter, we aim to overcome this limitation and dive into PG learning by constraining the learning algorithm to a physically plausible solution. Constraining the model with expert knowledge enforces scientific (physical) consistency, increases model explainability, and supports the trustworthiness of predictions.

> **Physical Consistency**
> Physical consistency refers to the ability of a model to respect the known laws, constraints, and symmetries of the modeled system. It ensures that these principles are explicitly enforced during learning or inference, such that physically impossible or inconsistent predictions cannot be made.

The scientific consistency of a prediction is a critical aspect in natural sciences that supports a solution space that follows underlying physical principles. *Physics-Informed Machine Learning* [64] enforces physical consistency through regularization (e.g., *Physics-Informed Neural Networks (PINNs)* [139, 140]), as discussed in Chapter 3. Karpatne et al. [16] even outlined that physical

Parts of this chapter, including figures and tables, have been published already in [137] and [138].

consistency must be considered as an evaluation metric, especially in safety-critical applications. This chapter focuses on answering RQ2 by raising the following subquestions:

- **RQ2.1** Which knowledge sources exist that can be used to condition a learning algorithm to physically plausible solutions in EO applications?
- **RQ2.2** How can those knowledge sources be integrated into the learning algorithm to increase the physical consistency?

6.1 Natural Disasters and Food Security

The frequency and intensity of disasters increased steadily over the last decades [7]. Data show that the rising frequency of disasters is a major concern today, resulting in many lives lost, destroyed livelihoods, and increased food losses. The global development of disasters is depicted in Figure 6.1. Agriculture is significantly more vulnerable to disaster and extreme weather events than any other productive sector, threatening global food security and increasing the risk of hunger and malnutrition. A direct impact of natural disasters on the agricultural sector is through reduced productivity and resulting economic losses every year [32, 33, 34]. Importantly, *Least developed countries (LDCs)* and *Low to Middle Income Countries (LMICs)* are disproportionately affected with both short and long-term impacts. Extreme weather conditions, including droughts and floods, are major concerns for agricultural productivity, with 34% of crop and livestock losses traced to drought and 19% to floods [7]. Only in 2022, more than USD 200 billion in economic losses occurred worldwide due to disasters [7]. Consequently, closing the gap between potential and actual yields by adapting to extreme weather conditions is an urgent task to sustain global food security [141]. Nevertheless, the frequency of severe droughts and floodings is expected to increase, causing either water scarcity or water abundance [142, 32].

Evapotranspiration: a proxy for crop stress

6.1.1 The Relationship Between Drought Stress & Yield

Water shortage impact crop development. Therefore, research has intensively investigated the modeling of water requirements to enable timely interventions in the event of a water scarcity. In an agricultural context, this commonly involves the estimation of the crop *Evapotranspiration (ET)*.

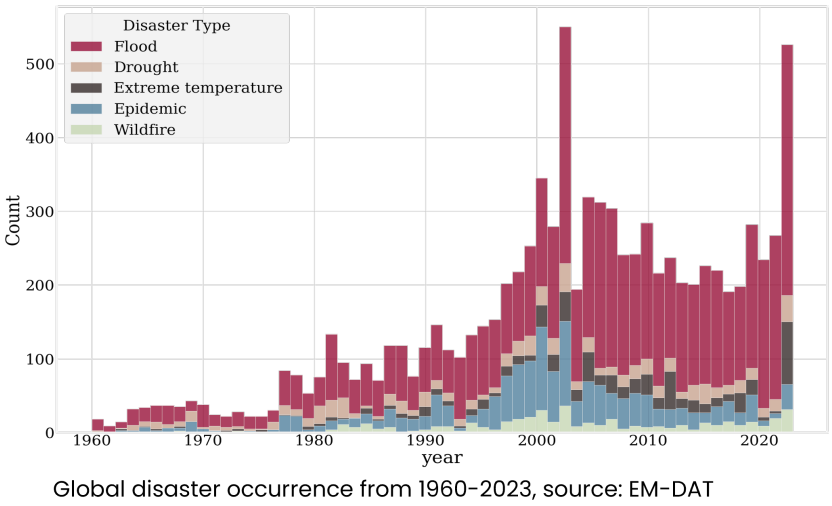


Figure 6.1: Global development of reported disasters by different categories from 1960 to 2024. (Data from: [6])

Evapotranspiration (ET)

The ET describes all biophysical processes in which liquid water is converted to water vapor from the topsoil (evaporation) and vegetation (transpiration) and is an indicator of crop stress and crop health [143, 144]. Moreover, the ET is closely related with crop growth and crop yield [145, 146]. This makes the ET is a core component of crop management, such as irrigation management and crop water deficit estimation, and is explicitly important for adapting to extreme weather conditions. The ET is defined by different components (e.g., temperature, solar radiation, soil, terrain elevation, and crop properties). The FAO (FAO-56 method [123]) provides a detailed description of the involved biophysical processes, which is based on the Penman-Monteith equation to estimate ET in mm per day.

Maximum and actual evapotranspiration

For a specific crop type, Allen et al. [123] differentiate between the maximum ET (ET_x) and the actual ET (ET_a). The maximum ET (ET_x) is defined under standard, non-limiting environmental conditions and depends only on climate conditions and crop parameters. Standard conditions imply full productivity, absence of disease, adequate fertilization, and optimal soil water availability. The maximum ET is defined by:

$$ET_x = K_c \cdot ET_o, \tag{6.1}$$

In contrary, the ET_a represents the actual ET under limiting conditions that are caused by low water potential that result in water stress and a reduction in ET, causing a reduction in productivity. Various factors cause limiting conditions such as soil salinity, limited soil water, soil infertility, diseases, and poor management.

Earlier, the FAO described the relationship between ET and the relative yield loss [147], stating that relative reduction in yiel is defined by the relative reduction in ET. This relationship is defined by the *yield response to water function*:

Evapotranspiration and yield loss

$$y_l = \left(1 - \frac{y_a}{Y_x}\right) = K_y \left(1 - \frac{ET_a}{ET_x}\right), \tag{6.2}$$

where:

y_l: relative yield loss [%]

y_a: actual yield

Y_x: potential maximum yield

K_y: yield response factor

Consequently, the reduction in yield is defined by the reduction in ET. The relationship is further defined by the dimensionless yield response factor K_y. This factor genetic component captures the impact of the reduction in ET and the reduction in yield for each genotype. Specifically, $K_y > 1$ indicates a higher sensitivity to water scarcity with a proportionally larger yield reduction, and $K_y < 1$ indicates higher resilience to water deficits. Different studies empirically estimated different K_y coefficients for different crop types. However, reports often differ, making the equation difficult to solve in practice. Furthermore, K_y values change over time, as many crops exhibit variable susceptibility to water scarcity. This further increases the difficulty of estimating the yield response factor. An overview for various crop types is given in [148]. More information is provided in [138, 137].

As mentioned earlier, droughts are primarily responsible for reducing agricultural productivity. Consequently, the response of crop yields to water scarcity has been a central focus of research for decades, serving as an important parameter for assessing crop resilience under extreme weather conditions [149]. Commonly process-based models (simulation models), have been developed to model the relationship between weather conditions and crop productivity. Process-based models are defined by biological and physical principles, solving many differential equations. Therefore, such models offer high explainability. However, process-based models often struggle with large volume of

high-dimensional data, are characterized by high computational costs, and require extensive calibration. Therefore, applying process-based models to large areas and high spatial resolution is limited. Finally, process-based models are commonly simplified representations of reality and rely on approximations to maintain computational efficiency [142], often resulting in inaccurate performance [150, 20]. Process-based models for daily ET estimation often idealize the crop stress component and therefore may not accurately estimate the actual ET. Consequently Equation 6.2 is difficult to solve in practice.
As we have seen earlier, ML models can handle complex, large, and high-dimensional data efficiently [151] and are therefore increasingly utilized for EO applications. In Chapter 4, we have demonstrated scalability and accuracy, even at fine-scale resolution for crop yield prediction. Nevertheless, DL models are often criticized for their black-box nature, that limit their explainability and trustworthiness [152, 19]. More importantly, ML rarely follow the physical principles of the modeled process [37], that can cause invalid outcomes. Increasing the physical consistency is an essential part for reducing the black-box nature of DL models. There is a growing interest to combine the advantages of data-driven approaches with the interpretability of process-based models [153, 142, 19, 17].

This chapter presents a method for PG learning by coupling interpretable process-based models and DL models. For this, crop yield is formulated as a function of temporal water scarcity by sequentially learning the actual ET, and the crop susceptibility to water scarcity. This is used to derive the expected yield loss by sequentially solving the crop yield response to water function [147]. This is done at high spatial resolution by using multispectral S2 satellite imagery from the and coarse weather data. Physical consistency is enforced by using a novel PG loss function.

6.2 Related Work

Several studies have been conducted that integrate prior knowledge to enhance crop modeling under extreme weather conditions by integrating crop drought information. However, most studies focus on enriching the data space with domain knowledge. For instance, Shuai and Basso [65] focused on integrating a crop drought index into the data space of the ML model and demonstrated a significantly improved performance. Especially, an increased robustness to

extreme weather conditions for maize yield predictions is reported. Likewise, Shahhosseini et al. [154] integrated simulated hydrological features into the data space and argued that weather information alone is insufficient for accurate yield estimation. More importantly, Jahromi et al. [144] integrated ET into the data space that was calculated using an energy balance concept and satellite data [143]. Ultimately, the study showed superior performance when integrating ET information.

Although several studies demonstrate the importance of including information on prior knowledge into the data space, enriching or expanding the data space does not guarantee physical consistency [17]. Only a few studies exist that particularly enforce scientific consistency through model regularization. He et al. [37] demonstrated a PG model for crop yield prediction by acknowledging key components of the carbon cycle, specifically the mass conservation into the loss function. Interestingly, this method accounts for spatial fairness. Additionally, He et al. [76] presents a methodology that extracts physical features from simulation data to estimate crop yield while preserving the physical features.

6.3 Modeling Drought Stress with Physics-Guided Learning

While ET_x can be sufficiently estimated by using process-based models, estimating the ET_a is difficult and depends on many unknown parameters. Consequently, Equation 6.2 is difficult to solve in practice. This is commonly done using field trials that are time-consuming, expensive and not feasible over large areas [155]. Additionally, process-based models are often restricted to low spatial resolution due to computational costs and the low resolution of the input data (weather).

Instead, this section proposes to estimate the actual ET (ET_a) and the susceptibility to water scarcity (K_y) using an NN approach. This serves a two purposes. First, estimating ET_a and K_y using a NN addresses the limitations the noisy process-based models. Therefore, the ET can be approximated with high precision and at a higher spatial resolution. Here, the S2 satellite data and the ground truth yield data support a type of super-resolution for predicting ET at the pixel level. Allen et al. [143] demonstrated that satellite data can approximate ET_a at high resolution. More importantly, including Equation 6.2 enables yield predictions that are based on physical principles (i.e., the reduction in yield is defined by the reduction in ET).

We use the notation for the time series pixel dataset as introduced in Section 3.1, with an input $\boldsymbol{x} \in \mathcal{X}$ defined as a sample from a time series. However, we refer to the target as $y_a \in \mathcal{Y}_a$ since we are dealing with *actual* and *potential* targets. Here, the actual target reflects the actual crop yield, harvested at a time t. This data is provided from field measurements, such as described in Subsection 2.2.1. In contrast, the potential yield is the maximum yield that could be achieved under non-limiting conditions.

Potential yield

The goal is to learn a function with two output heads predicting ET_a and K_y, $f_\theta(x) = [ET_a, K_y]$, by optimizing over the model's parameters such that the relative reduction yield is equal to the relative reduction in ET, as defined by the relationship derived in Equation 6.2. Importantly, both $ET_a = (ET_a^t)_{t=0}^T$ and $K_y = (K_y^t)_{t=0}^T$ are estimated over the entire growing period. Finally, the cumulative yield loss (Y_l) at the end of the time series (T) is given by the integral:

$$Y_l = \int_0^T K_y^t \left(1 - \frac{ET_a^t}{ET_x^t}\right) dt \approx \sum_{t=0}^{T} K_y^t \left(1 - \frac{ET_a^t}{ET_x^t}\right). \tag{6.3}$$

The reduction of ET over each time step must be equal to the reduction from the potential yield to the actual yield. Following, the final prediction of the actual yield is given by:

$$\hat{y}_a = Y_l \cdot Y_x. \tag{6.4}$$

Forcing the model to approximate $\hat{y}_a$ improves the estimations of ET_a. Moreover, Y_x is defined in Subsection 6.3.2. To satisfy the relationship between a reduction in crop yield and a reduction in ET, a novel PG loss term is integrated, consisting of a data-dependent part $\mathcal{L}_l$ and the physical constraint part $\mathcal{L}_{phys}$. For this, it is assumed that ET_x values are sufficiently accurate and are provided by a simulation model, as reported in [156]. This allows the model to focus on the prediction of ET_a and K_y to achieve an accurate solution.

To guide the learning of ET_a values, physical constraints are integrated. First,

ET_a values must always be below or equal to ETx over the entire time series. The loss term is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_l + \lambda_2 \mathcal{L}_{\text{phys}} \tag{6.5}$$

$$\mathcal{L}_l = \mathbb{E}\left[(\hat{y_a} - y_a)^2\right] \tag{6.6}$$

$$\begin{aligned}\mathcal{L}_{\text{phys}} = \mathbb{E}\Bigg[& \underbrace{1_{\{ET_a<0\}} \cdot (ET_a)^2}_{\text{lower bound penalty}} + \\ & \underbrace{1_{\{ET_a>ET_x\}} \cdot (ET_a - ET_x)^2}_{\text{upper bound penalty}} + \\ & \underbrace{1_{\{0\leq ET_a\leq ET_x\}} \cdot (ET_a - ET_x)^2}_{\text{within bounds MSE}}\Bigg]\end{aligned} \tag{6.7}$$

Here, $1\{\cdot\}$ is an indicator function that is equals 1 if the condition inside the braces is true and 0 otherwise. The data-dependent part forces the model to estimate ET_a so that, using Equation 6.4, the predicted actual yield ($\hat{y_a}$) matches the target yield (y_a). The second component forces the model to maintain ET_a values bounded between [0, ET_x] while sufficiently close to ET_x such that is satisfies Equation 6.2 [137]. Finally, λ_1, λ_2 are hyperparameters that control the weighting of both terms.

6.3.1 Physics-Guided LSTM

For the implementation, a LSTM backbone with 2 layers is employed, where each hidden state is passed to a sequential layer with 128 hidden units, incorporating a linear layer, batch normalization, and dropout of 0.2. Finally, two linear layers are incorporated with a single output channel each, predicting K_y^t and ET_a^t, respectively. We refer to this network as *PG-LSTM*. A LSTM network is selected because of the good performance in previous studies for crop yield prediction. However, since LSTM models can struggle with very long and multidimensional sequences, we additionally evaluate the inclusion of an attention mechanism, as proposed in [37]. More specifically, we employ the scaled dot-product attention that was previously used for generating channel attention. We refer to this model as PG-LSTM^{attn}. A schematic overview of the training and inference scheme of the proposed method is given in Figure 6.2. Moreover, simulated ET_x values are used in the optimization loss to guide the training. At each time step, the model produces an estimation of ET_a and, K_y, which is then used to calculate the yield through the yield response to water function.

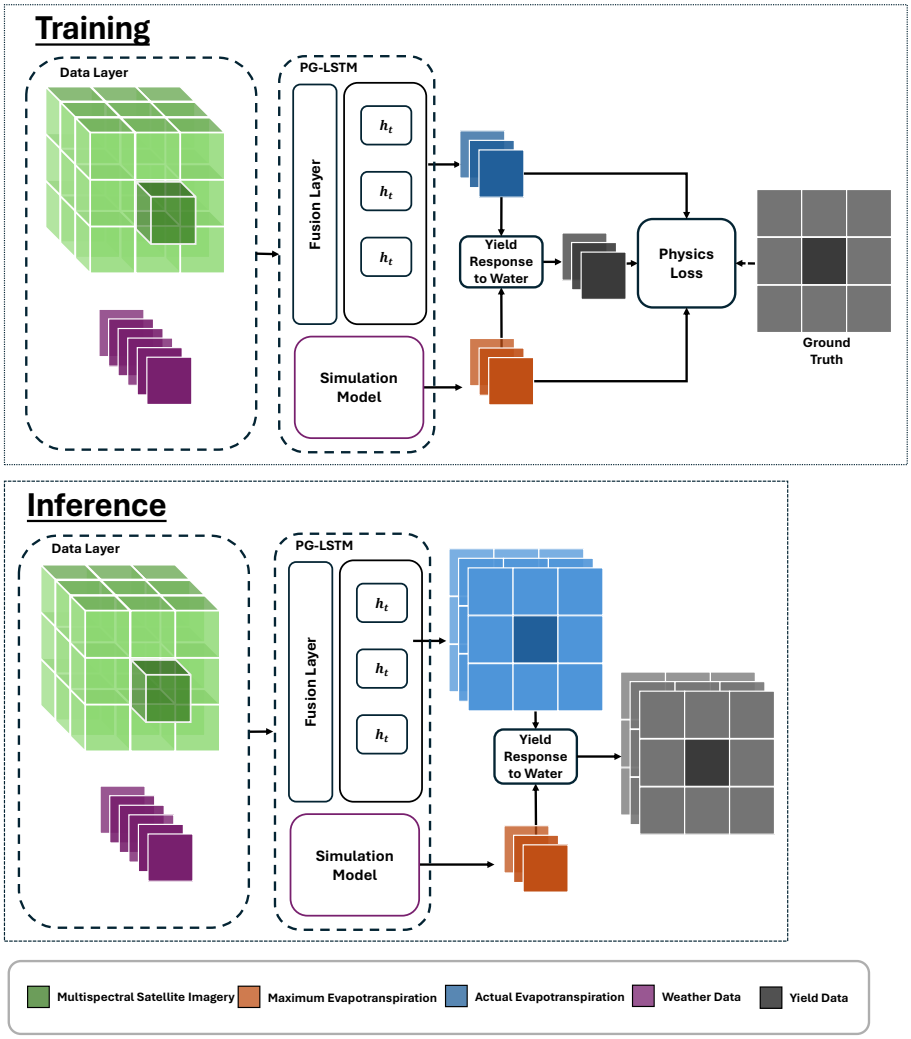


Figure 6.2: Overview of the PG-LSTM architecture for drought stress estimation and crop yield prediction. The training (top) and the inference (bottom) are shown. The data is modeled at the pixel-level and approximates the crop yield using the yield response to water function, which enhances the estimation of actual ET. In contrast, the simulated maximum ET lacks spatial context.

6.3.2 Experimental Setup

Data & Simulations

In this study, the *SwissYield* dataset is used as described in Subsection 2.2.2. The dataset provides ground truth yield measurements at 10 m × 10 m resolution for cereal crops. Since no information about the potential maximum yield is available, the maximum yield sample is defined as the maximum yield across the entire dataset:

$$Y_x = \max(\mathcal{Y}_a). \tag{6.8}$$

This assumes that at least a single sample in the entire dataset was cultivated under non-limiting conditions.

As the model input, S2 data and weather data are used. Weather data is included to better account for extreme environmental conditions. More specifically, total precipitation and the minimum and maximum, and average temperatures are used. Data modalities are fused at the input level using the dense time series of S2 images as described in Section 4.4. More information about the dataset is given in Subsection 2.2.2.

Simulating evapotranspiration

For each field, ET is simulated over time using a process-based crop model. More precisely, only the maximum ET (ET_x). For this, the FAO paper-56 [123] is employed that simulates ET_x over time using a Python [124]. The required weather features are obtained from the ERA5 global reanalysis [26]. This data is usually given at 10 m above the Earth's surface. However, to calculate ET, some features must be available at 2 m above the surface. Therefore, features that are not available at 10 m height are adjusted following [123]. Relevant soil data is collected from the *SoilGrids* [28] and *Hihydrosoil* [157] projects for every sample. A detailed overview of the data modalities that are used in the NN and the simulation model is shown in Table 6.1. Crop parameters are taken from [123, 155]. An example time series of S2 satellite imagery, the

Table 6.1: Overview of the data modalities that are used for each model type.

Modality	Source	Product	Neural Network	Simulation Model
Multispectral	Sentinel- 2 L2A	B02 - Blue	✓	✗
		B03 - Green	✓	✗
		B04 - Red	✓	✗
		B05 - Red Edge 1	✓	✗
		B06 - Red Edge 2	✓	✗
		B07 - Red Edge 3	✓	✗
		B08 - NIR	✓	✗
		B8A - Narrow NIR	✓	✗
		B11 - SWIR 1	✓	✗
		B12 - SWIR 2	✓	✗
Weather	ERA5	Max Temperature	✓	✓
		Mean Temperature	✓	✓
		Min Temperature	✓	✓
		Total Precipitation	✓	✓
		maximum relative humidity	✗	✓
		minimum relative humidity	✗	✓
		average wind speed at 2m	✗	✓
		Incoming solar radiation	✗	✓
		average dew point temperature	✗	✓
Soil	SoilGrids	Silt	✗	✓
		Clay	✗	✓
		Sand	✗	✓
	HihyrdroSoil	Field Capacity (pF2)	✗	✓
		Permanent Wilting Point (pF4.2)	✗	✓
Terrain	SRTM	Elevation (DEM)	✗	✓

corresponding weather, and simulation data are shown in Figure 6.3. All data modalities cover the entire growing period from seeding to harvesting. Notice that the weather and simulation data provide daily measurements, whereas the S2 time steps are approximately every 5 days. In contrast, S2 imagery

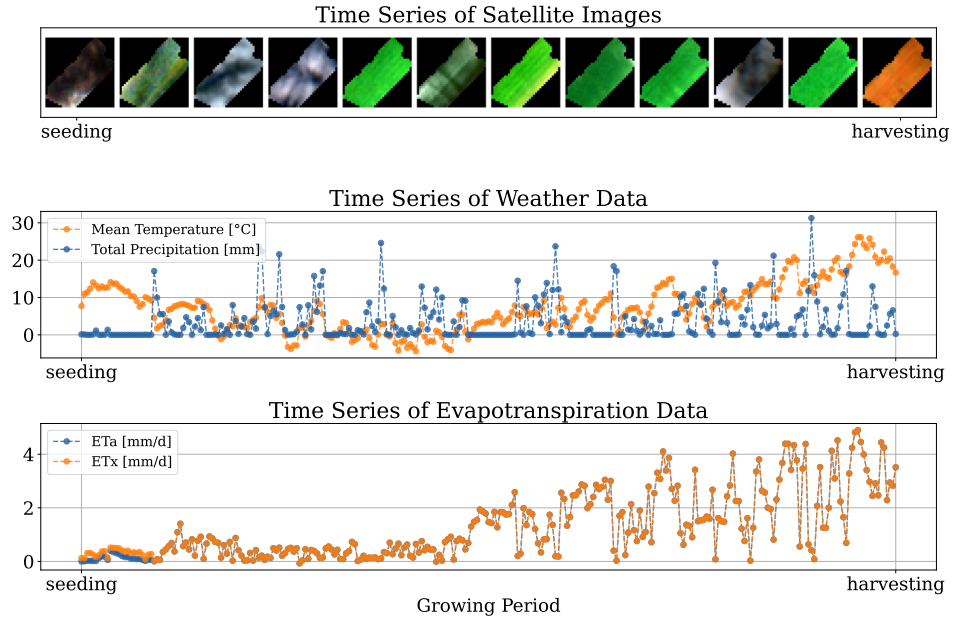


Figure 6.3: Example time series from seeding to harvesting of a randomly selected field. Top: Time series of S2 images in RGB. Center: Time series of weather data. Bottom: Simulated maximum and actual ET. Notably, the maximum and actual ET values are close.

provides spatial information about each field. Importantly, ET exhibits a correlation with temperature. With rising temperatures, the ET also increases. This further results in more biomass accumulation, depicted in the S2 images. Importantly, we highlight that only small differences are noticeable between the maximum and actual ET. This underscores the challenge of accurately solving the yield response to water function (Equation 6.2) at high spatial resolution using only simulation models.

6.3.3 Evaluation

As the baseline experiment, the standard K-fold CV (K=10) is performed, with results presented as the average across folds. Moreover, temporal transferrability is evaluated using a LOYO CV scenario, in which one year is held out during training and used solely for evaluation. For quantitative evaluation, the standard regression metrics are used. Additionally, the predicted ET values are evaluated by agricultural experts.

The PG-LSTM is compared against several models for crop yield prediction that contain no physical components or regularization. This includes a the LSTM and Transformer [98] as described in Section 4.4 and Subsection 5.2.1.

Moreover, we include a simple linear regression model.

Uncertainty estimation

Since we aim to estimate two components (ET_a and K_y), this problem becomes ill-defined. This increases the solution space and may introduce uncertainty into the predictions. Moreover, additional sources of error are present in the data, including measurement errors and noise in the ground-truth yield data and simulation data. To account for the uncertainty, a Deep Ensemble (DE) approach [158] approach is used. In Chapter 8, a detailed overview of uncertainty estimation with DEs is provided. We train 10 separate ensemble members to account for uncertainty in the model.

6.4 Results

6.4.1 Drought Stress

Figure 6.4 illustrates the temporal predictions of the simulated ET_x and predicted ET_a values for the same field as shown in Figure 6.6. The mean prediction for ET_a and K_y over all ensemble members is shown with a buffer of $\pm 2\sigma$ to assess the temporal uncertainty. Note that the simulated ET_a is consistently higher than or equal to the predicted ET_x. A strong consistency between the simulated ET_x and predicted ET_a values is observed that follows the physical conditions that are enforced over the model (see Equation 6.5). This indicates that the model captured important agronomic properties, such that ET_a values must be consistently lower than or equal to ET_x. In this example,ET_a is consistently lower than ET_x, which indicates yield-limiting conditions, such as water scarcity. On the right plot, the estimated K_y values are shown alongside the NDVI. The NDVI is derived from the satellite imagery and serves as an indicator of vegetation density and plant health, and should reflect the biological activity in each sample. More specifically, an increase in ET should be accompanied by higher NDVI values until senescence. Interestingly, a consistent increase in ET over the growing period is observed, which correlates with an increase in NDVI and K_y. This further indicates that the model captured important biological properties and further indicates higher susceptibility to water scarcity at later growth stages. Moreover, it is noted that the uncertainty for the ET_a values is lower compared to the K_y values. This can be explained by the fact that K_y is treated as a free parameter, which can lead to higher uncertainties. In Figure 6.5, simulated ET_x and predicted ET_a values averaged over the dataset are illustrated. A similar pattern to that in the single field example is observed. Additionally, on the right, the

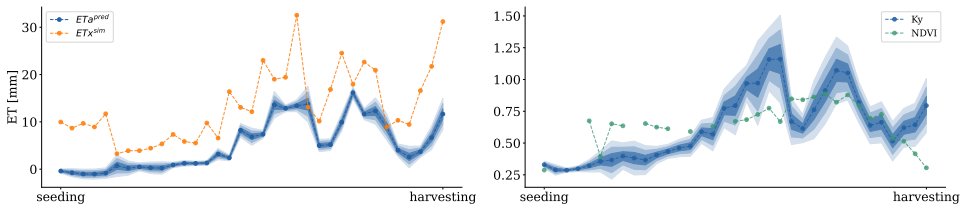


Figure 6.4: Visualization of the simulated and predicted ET for a single field. Left: Temporal simulation of maximum (ET_x^{sim}) and the predicted actual (ET_a^{pred}) crop ET. Right: Visualization of the predicted susceptibility to water scarcity (K_y) and the NDVI. To visualize the uncertainty, the predictions are illustrated with $\pm 2\sigma$.

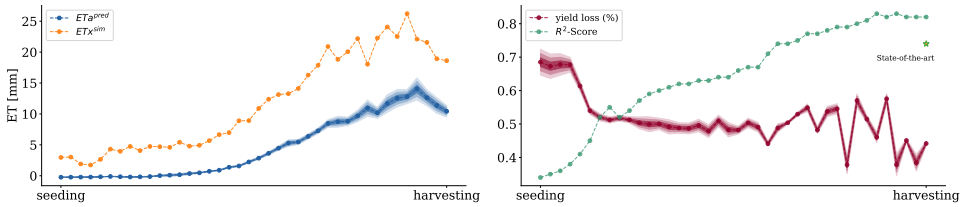


Figure 6.5: Simulated and predicted ET over the entire time period, averaged over the entire dataset. Left: Temporal simulation of maximum (ET_x^{sim}) and the predicted actual (ET_a^{pred}) ET. Right: Derived yield loss over the growing period using the yield response to water function (Equation 6.2). Additionally, the performance is illustrated over the growing period, as shown in the R^2-score over time. To visualize the uncertainty, the predictions are illustrated with $\pm 2\sigma$.

yield loss (in %) at each time point is illustrated, which is derived following Equation 6.2. Importantly, the predicted yield loss negatively correlates with the predicted ET, with a Pearson correlation of -0.45, indicating that the model effectively captured the relationship between a reduction in ET and the corresponding reduction in yield. Finally, the yield prediction performance, expressed as the R^2-score over time, is shown, calculated from the derived yield loss. Importantly, the accuracy increases over time since the estimation of the actual yield loss is more accurate at the end of the growing period.

Furthermore, in Figure 6.6 it is shown that the predicted ET_a values now have a spatial resolution of $10\,\mathrm{m} \times 10\,\mathrm{m}$. Interestingly, the predicted ET_a values correlate significantly with the target yield data at later time steps, further underlining that the model learned the relationship between yield and ET at the pixel level. More specifically, high ET_a values correlate with a higher yield. In contrast, pixels with low ET_a correspond to low yields, indicating crop stress. This underlines the ability to estimate crop yield reduction over time by learning the important features of crop water use. Overall, the model exhibits high in-field variability that closely aligns with the target data.

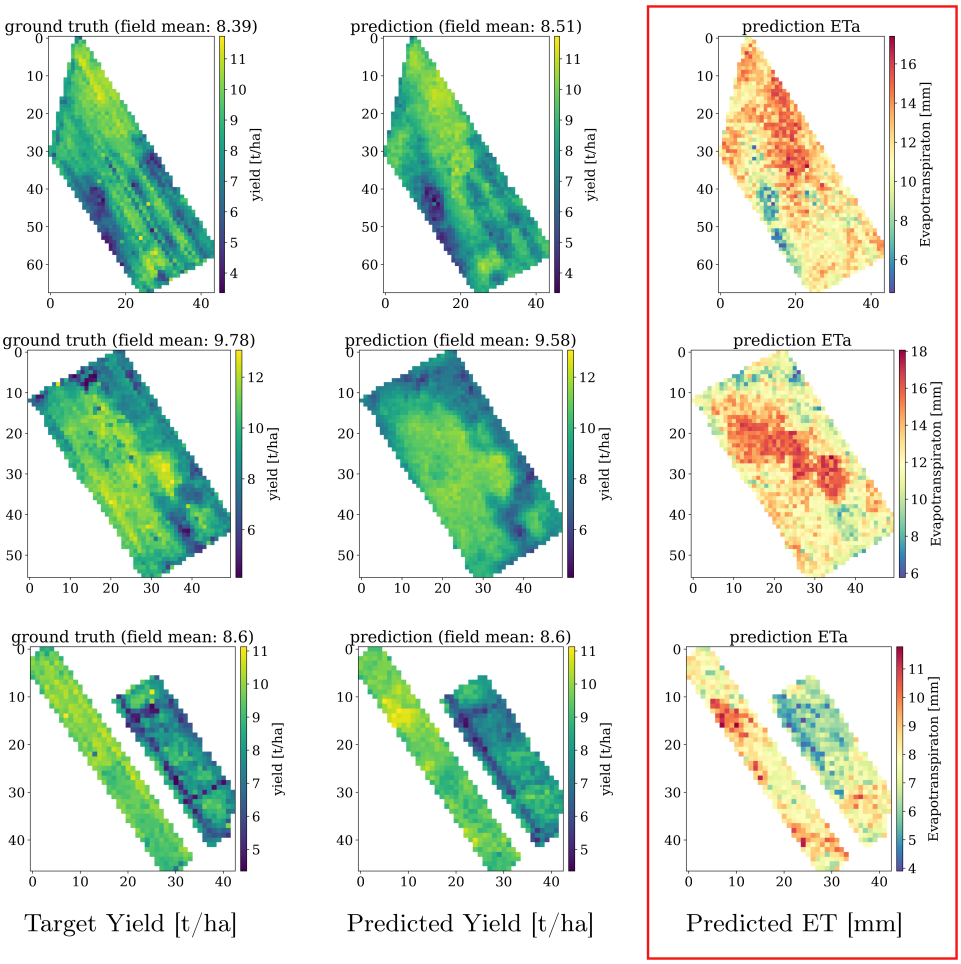


Figure 6.6: Qualitative results for three randomly selected fields. Left to right: Ground truth yield map, predicted yield map, and predicted actual ET.

Table 6.2: Overview of yield prediction performance.

Option	R^2 (↑) -	MAE (↓) t/ha	MAPE (↓) %	RMSE (↓) t/ha	BIAS (↓) t/ha
PG-LSTM^{attn}	**0.82**	**0.59**	**0.11**	**0.86**	-0.01
PG-LSTM	0.81	0.59	**0.11**	0.87	0.07
Transformer	0.73	0.74	0.14	1.05	0.41
LSTM	0.80	0.62	0.12	0.90	0.04
Linear Regression	0.70	0.81	0.17	1.10	**0.00**

6.4.2 Yield Prediction

In Table 6.2, the PG approach is compared against the baseline models. For the PG models, the last time steps are used to calculate the regression metrics. Surprisingly, the PG models outperform all baseline models, including the Transformer, LSTM, and linear regression models. For instance, the PG-LSTM^{attn} model improves 9 pp over the Transformer model in the R^2 score,

Table 6.3: Overview of the model performance without estimating the crop susceptibility to water scarcity (K_y).

Option	R^2 (↑) -	MAE (↓) t/ha	MAPE (↓) %	RMSE (↓) t/ha	BIAS (↓) t/ha
PG-LSTM^{attn}	0.26	1.28	0.21	1.73	-0.18
PG-LSTM	0.74	0.72	0.14	1.02	0.13

Table 6.4: Performance overview for the Leaf-One-Year-Out cross-validation scenario.

Option	R^2 (↑) -					RMSE (↓) t/ha				
	2017	2018	2019	2020	2021	2017	2018	2019	2020	2021
PG-LSTM^{attn}	0.09	-0.38	0.19	**0.10**	**0.29**	1.87	1.93	1.60	2.24	2.92
PG-LSTM	0.14	0.04	**0.41**	-0.71	0.26	1.82	1.48	**1.38**	2.16	**1.89**
Transformer	0.31	**0.30**	0.31	-0.31	-0.61	1.63	1.27	1.48	1.89	2.79
LSTM	0.24	0.29	0.25	-0.68	-0.91	1.71	**1.27**	1.55	**2.14**	3.03
Linear Regression	**0.37**	-0.36	-0.30	-1.63	-0.99	**1.56**	1.76	2.04	2.68	3.10

and 2 pp over the standard LSTM model. Moreover, a reduction in RMSE of 0.19 t/ha is achieved for the PG-LSTM^{attn} over the Transformer model. As expected, the linear regression model performs the worst among all models, yet still achieves an R^2-score of 0.7. Only minor differences are observed between the PG-LSTM^{attn} and the PG-LSTM model. However, the PG-LSTM^{attn} performance is slightly better.

6.4.3 Ablation Studies

The influence of the yield response factor: Table 6.3 evaluates the PG methods that predict only the ET_a values. This experiment evaluates the importance of the free parameter K_y, which describes the yield response to water scarcity. For this, K_y, a constant of 1.05, is defined in Equation 6.2, as provided in [148]. As expected, both models show a decrease in performance. Surprisingly, the PG-LSTM outperforms the PG-LSTM^{attn} by 48 pp in R^2 and by 0.71 t/ha in RMSE. However, PG-LSTM achieves an R^2 of 0.74, thereby outperforming the Transformer and linear regression model, as shown earlier. This suggests that, while K_y is a meaningful parameter to estimate, previously reported values still provide sufficient information to study the relationship between crop stress and yield reduction. Moreover, it shows that the model captured real ET values and is not just fitting the data.

Temporal transferability

Table 6.4 evaluates the temporal transferability of all models using the LOYO scenario. Surprisingly, all models show significantly lower performance when applied to years outside their training set. The maximum R^2 is achieved by PG-LSTM in 2019 with a value of 0.41. In particular, the linear regression

model performs worst when applied to unknown years. In contrast, PG approaches achieve better performance compared to existing baseline models, outperforming them in 2019 and 2021.

6.5 Discussion

In this section, we demonstrated that PG learning enforces physical consistency for crop stress and crop yield forecasting, thereby enhancing the model performance and explainability. This section focused on RQ2 and the subquestions RQ2.1 and RQ2.2. We showed that we can couple physical laws with process-based simulation models and DL models. Specifically, it was demonstrated that the reduction in crop yield can be attributed to the reduction in ET, a proxy for drought stress. This relationship was earlier defined in [147]. Here, the ground truth yield data enhances the estimation of the actual ET, which in turn enables more accurate yield estimation, even at the pixel level with $10\,\mathrm{m} \times 10\,\mathrm{m}$ resolution. For this, low-resolution ET simulations are upsampled using high-resolution S2 data and pixel-level yield data by solving the relation between yield and ET. The results support the hypothesis that learning from data as well as from prior knowledge is beneficial for ML techniques [17]. Additionally, the findings support related studies that report improved performance and robustness by integrating crop stress information into the learning pipeline [154]. However, here, prior knowledge is directly integrated into the learning algorithm via a novel loss function. Consequently, this enforces the NN to approximate the solution of the governing equation between crop stress and crop yield. The results indicate potential for overcoming limitations in both simulation models and data-driven methods. First, the estimated actual ET values are more accurate than the simulated values, with higher spatial resolution and lower computational cost at inference time. Secondly, the resulting yield estimations follow the governing relationship between ET and crop yield even with low uncertainty (Figure 6.5). This thereby increases the transparency of the model. This significantly improves the model interpretability and finally contributes to the trustworthiness of the predictions. Notably, the approach outperforms strong baseline methods such as LSTM and Transformer on several regression metrics (Table 6.2). Additionally, integrating an attention mechanism improved performance. This could be attributed to the limitation of LSTM in capturing long-range dependencies, primarily due to their limited capability of sustaining useful gradients over long sequences [159, 160]. The attention mechanism potentially mitigates this issue

by enabling the model to focus on relevant time steps. Thereby supporting the concentration of important gradient signals. Surprisingly, the attention mechanism fails when predicting only the ET (Table 6.3). This could indicate overfitting to the attention scores. This observation is further supported by the LOYO experiment, where the LSTM demonstrates more robust and consistent performance. Therefore, the inclusion of additional architectural components must be carefully evaluated before being deployed into practice.

Limitations

Although this approach is a step towards more explainable and transparent ML methods, limitations remain that must be considered. First, the dataset sizes are limited, hindering the development and evaluation of more powerful, scalable models. More data is required to fully assess the importance of this work, particularly from water-scarce regions. Limited data can cause dramatic performance degradation, such as demonstrated in the LOYO experiments (Table 6.4). This, moreover, underscores the need for further research and the integration of additional transfer learning techniques [121]. More importantly, we observe that the ground truth data is noisy and associated with artifacts. This could lead to severe negative effects, including severe overfitting to noise, degraded model performance, increased uncertainty, and miscalibration. Unfortunately, in the presence of highly noisy data, popular regularization techniques, such as weight decay [161], dropout [162], or batch-normalization [163], are commonly not sufficient to mitigate the impact of noisy labels [164]. Consequently, it is necessary to adopt methods that mitigate the impact of noisy data, including improved data acquisition practices and learning methods explicitly designed to handle label noise [164].
Finally, modeling ET using process-based models is difficult and requires calibration [155]. Therefore, more effort must be made to calibrate the employed simulation methods to avoid miscalibration and, consequently, model overfitting. Furthermore, ground truth samples of the actual ET are required to estimate the potential of this method.

6.6 Conclusion

Integrating prior knowledge directly into the learning algorithm holds significant potential for crop modeling, offering enhanced adaptability to challenging environmental conditions. This section presented a novel approach to modeling crop productivity under environmental constraints by incorporating agronomic

properties directly into the learning algorithm. We demonstrated promising experimental outcomes. The presented approach supports industry, policymakers, and farmers in achieving more sustainable and resilient agriculture.

7 NETWORK PRIORS & GENERATIVE MODELS

Chapter Highlights:

- Plant growth is ambiguous and depends on diverse environmental conditions. Such conditions can be encoded as a conditional, interpretable latent distribution.
- Generated images of future plant growth offer significant advantages over traditional approaches.
- Conditional DMs can be used for pixel-wise time series regression in EO tasks.

In the previous chapter, we integrated prior knowledge directly into the learning algorithm to enforce the model to approximate a solution that satisfies physical principles. This can be achieved by integrating formalized natural laws. However, sometimes the underlying laws are not known but can be described by data. This implicit knowledge can also be integrated into a learning system by using network priors as an additional form of regularization. In this chapter, we continue investigating the integration of prior knowledge into the learning algorithm for plant growth modeling and time series regression. We focus on the latent space, which serves as a conditional prior during the generation process. This section continuous with RQ1 and the subquestions RQ2.1 and RQ2.2.

7.1 Controlled Image Generation for Plant Growth

In the previous chapter, we discussed that plant growth is highly dependent on limiting environmental conditions, such as climate, nutrient supply, pest pressure, or management practices [165, 166]. Often, limiting conditions are not constant throughout the growing period. For instance, drought stress may occur only during certain time intervals, affecting plant growth in different ways [165, 166]. However, future conditions are commonly unknown, making the forecasting of future appearances highly ambiguous and uncertain. To bridge this uncertainty, plant growth should be modeled by a distribution of

potential appearances.
In this section, we present the controllable GAN (contGAN) [167], a model that maps a single input to a distribution of outputs under controlled conditions. This is defined as one-to-many mapping and based on the GAN framework [168].

> **One-to-many mapping**
> One-to-many mapping describes a relationship where a single input is associated with multiple outputs, while each output can only be linked to a single input.

Here, plant growth is formulated as an image-to-image translation task between two image domains, where each of them represents a different point in time. To introduce ambiguity, a low-dimensional latent distribution is learned that conditions the image generation process and that can be sampled during inference. This latent vector encodes prior knowledge about the target and guides the prediction to generate realistic and diverse outputs.

Image generation offers many advantages

Generated images of future plant growth offer several advantages over previous methods that estimate individual parameters (e.g., crop yield). For instance, they illustrate the complete above-ground coverage, and various parameters can be derived [169, 170, 171]. Parts of the results in this section have been already obtained during my master thesis and have been extended and published during this PhD.

7.1.1 Conditional Image Generation

Adversarial training

The conditional Generative Adversarial Networks (cGANs) [168] method is an established framework for image-to-image translation that follows an adversarial learning philosophy. This task learns the translation between two image domains using a NN, $f_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{Y}$. GANs simultaneously train two NNs: a generative model $\mathcal{G}_\theta$ and a discriminative model $\mathcal{D}_\delta$ that are parameterized by θ and δ. While the generative model aims to generate data indistinguishable from reality, the discriminative model aims to distinguish between real and generated instances. In its basic form, GANs learn a mapping from a random

---

Parts of this section, including figures and tables, have been published already in [167].
This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324 and partly funded by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

noise vector to an output instance. In contrast, a cGAN learns a mapping from an input image and random noise to an output image $\mathcal{G}_\theta : \{X, \epsilon\} \rightarrow Y$. Thus, the input image serves as a condition that guides the image generation process. An important approach was proposed by the Pix2Pix architecture [172], where the Generator tries to fool the discriminator by generating samples that are indistinguishable from the training data.

Conditional learning

Generative modeling is a well-studied field of research for synthetic image generation, also for EO applications. For example, Drees et al. [173] presented a method for plant growth modeling based on cGANs. Regardless, existing methods are limited to predicting a single output rather than a probability distribution. Yet, many problems in EO are ambiguous, and a single input may correspond to a distribution of possible outputs. Individual methods, such as those presented by Zhu et al. [174], address the problem of stochastic image generation by distilling ambiguity into a low-dimensional latent distribution that can be sampled randomly during inference. However, such methods are stochastic, leading to uncontrolled output generation.

To address this limitation, this section presents a framework for controlled image generation by explicitly including prior conditions. We provide more detailed information on the cGAN framework in the Appendix (Subsubsection A.1.1).

7.1.2 One-to-Many Image-to-Image Translation

This approach uses the notation of the *image dataset* as defined in Section 3.1. The objective of this approach is two-fold. First, the model should capture the relationship between an image of a present growth stage and an image of a future growth stage. This is achieved by mapping the present stage X to the future stage Y using a conditional latent representation $\boldsymbol{z}$. Additionally, the latent space should provide stochasticity such that several future appearances could be synthesized. However, latent representations commonly lack interpretability to the user [175]. Consequently, the user cannot evaluate likely potential scenarios. To achieve an interpretable and disentangled latent representation, the objective is to learn the information contained in both image domains and in the latent distribution that describes future appearance. These prior conditions that influence the process of plant growth are defined as Factors of Variation (FOV).

The latent space describes the future appearance

A multimodal mapping function is learned between two image domains $f^\theta(\cdot) : \mathcal{X} \longmapsto \mathcal{Y}$, given paired training data $\{(X \in \mathcal{X}, \boldsymbol{z} \in \mathcal{Z}), Y \in Y\}$. Specifically, $\boldsymbol{z}$ follows a joint probability distribution $p(\boldsymbol{z}|Y, \boldsymbol{\epsilon})$ that serves as a prior condition together with the input image. Here, $p(\epsilon)$ refers to the FOV that describe

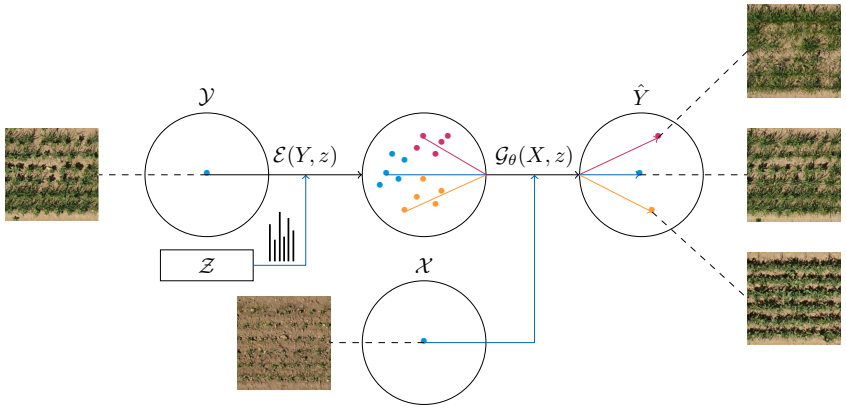


Figure 7.1: Overview of temporal plant growth modeling using close-range remote sensing imagery of two time steps. A target image Y and the domain of Factors of Variation (FOV) $\boldsymbol{\epsilon}$ is used to span a disentangled latent space Z using an Encoder network, $\mathcal{E}(Y, \boldsymbol{\epsilon}) \rightarrow \boldsymbol{z}$. Note that Z follows the distribution of to the FOV. A sample from the latent space, and an input image X, produces deterministic plant growth predictions, $\mathcal{G}_\theta(X, \boldsymbol{z}) \rightarrow Y$ using a cGAN. Sampling from the interpretable latent space, maps one input image to a distribution of possible outputs.

the target image. For example, $\boldsymbol{z}$ contains information about the expression of biomass, the sown mixture ratio, average plant height, and environmental conditions, including soil, water, and nutrient supply. Sampling from $\mathcal{Z}$ during test time results in a distribution of potential outputs.

The model was inspired by the BicycleGAN [174], which generates a random distribution of potential outputs based on a latent noise vector $\boldsymbol{z}$ sampled from a Gaussian distribution. In contrast, the contGAN model does not sample from a random distribution, but instead from a known prior distribution of potential future appearances.

The prior conditional distribution of future appearances (target image Y and FOV $\boldsymbol{\epsilon}$) are encoded into a low-dimensional latent space $\mathcal{Z}$, using a probabilistic multi-layer Variational Autoencoder (VAE) [176], $\mathcal{E}$. The conditional prior distribution of the latent vector $\boldsymbol{z}$ is given by the encoder model:

$$z \sim \mathcal{E}(Y, \boldsymbol{\epsilon}) = \mu + \sigma\boldsymbol{\epsilon}, \tag{7.1}$$

Interpretable latent space

with $\boldsymbol{\epsilon} \sim p(\epsilon)$ given as the FOV describing Y. A Kullback-Leibler (KL) divergence [177] loss between the latent distribution and the ground truth FOV is used to enforce an interpretable latent space that can be sampled during test time:

$$\mathcal{L}_{\mathrm{KL}}(\mathcal{E}) = \mathbb{E}_{\mathcal{Y}}\left[\mathcal{D}_{\mathrm{KL}}(\mathcal{E}(Y, \boldsymbol{\epsilon}) \parallel p(\epsilon))\right] \tag{7.2}$$

Further loss components are introduced to ensure that the latent space aligns with the FOV and is densely populated [167] for sampling. The final objective function is derived in Subsubsection A.1.2.
Finally, together with a randomly sampled latent vector, a generator produces a prediction $\hat{Y}$ which is realistically close to the target image:

$$\mathcal{G}_\theta : \{X, z\} \rightarrow \hat{Y}. \tag{7.3}$$

An overview of the entire workflow is given in Figure 7.1. The figure shows that sampling from the latent distribution produces a distribution of outputs for a single input image. The ambiguity is encoded in the low-dimensional latent vector, which serves as a prior conditional distribution. Importantly, the latent space aligns with the interpretable ground-truth FOV to serve as prior knowledge, guiding the generator to produce realisitc results.

7.1.3 Model Evaluation

The generated images are qualitatively evaluated by human experts, following the standard guidelines described in Section 3.3. For quantitative evaluation, the Ground Cover Fraction (GCF) is used to estimate the fractional green canopy between the generated and the reference image. The GCF estimates canopy development and correlates with above-ground biomass, and is calculated using an automatic color-threshold classification [178]. The GCF evaluates whether the generated images exhibit useful phenotypic traits similar to those in the target image. The L1 loss between the GCF of the reference and target images is quantitatively evaluated. Additionally, the FID_∞, termed FID infinity score [84] is evaluated between generated and reference images, which is an unbiased estimator of the FID. More details about the FID score are given in Section 3.3.
The model is compared with a traditional cGAN, namely the Pix2Pix model [172], which has demonstrated strong capabilities for plant growth modeling [173]. Further, the model is compared with the multimodal BicycleGAN [174], which produces a distribution of outputs.

7.1.4 Data

In this section we use the data from the MixedCropping experiment as described in Subsection 2.2.5. Specifically, image pairs from 4 weeks after plant emergence (domain $\mathcal{X}$) and 8 weeks after emergence (domain $\mathcal{Y}$) are used. The images

Table 7.1: Overview about calculated FID infinity scores ($\mathcal{N}_r$, $\mathcal{N}_g$) (↓) for different experiments.

FID_{Pix2Pix}	$FID_{\text{BicycleGAN}}$	FID_{contGAN}
17.94	25.80	**16.32**

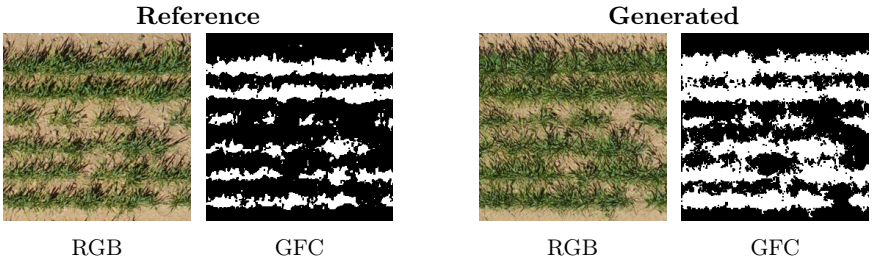


Figure 7.2: Example image for a real (left) and generated sample (right), in RGB (left) and the derived GCF (right) for a SW sample. The generated samples are generated with the contGAN model. White pixels indicate green canopy, while black is defined as bare soil.

define a balanced dataset that contains all three crop types as classes: spring wheat (SW) and faba bean (FB) monocultures, and mixtures (MX) of both. In total, the dataset contains 1057 aligned image pairs of size $[256 \times 256]$ and is randomly split into a training and a separate test set.

7.1.5 Experiments

For the generator model, a U-Net architecture with skip connections is used [179]. For the encoder network, a ResNet architecture with multiple residual blocks [180] are employed. The discriminator model is defined as a PatchGAN [181] implementation with an overlapping patch size of $[70 \times 70]$. Further details are given in Subsubsection A.1.2.

Temporal Plant Growth Modeling

We first evaluate the image generation performance of future time steps against the two baseline models using the FID infinity score. The results are given in Table 7.1. Notably, the contGAN outperforms the other models. The poorest performance is achieved by the BicycleGAN model with an FID score of 28.80. In contrast, the contGAN achieves a FID score of 16.32.
In Figure 7.2, a reference and generated sample are displayed for the SW class. Next to each RGB image, the GCF is displayed, which illustrates the above-ground biomass for the reference and generated sample. Notably, the generated sample and the GCF closely align with the reference image, under-

Table 7.2: L1 error (↓) for GCF generated images.

L1 Error	Pix2Pix	BicycleGAN	contGAN
SW	0.27	0.42	**0.18**
MX	0.25	0.40	**0.23**
FB	0.27	0.39	**0.21**

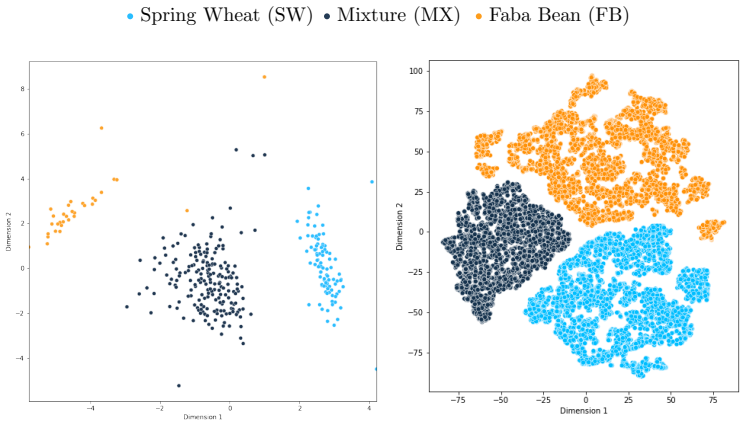


Figure 7.3: Left: Low-dimensional embedding of the FOV that are collecting during the growing period (Subsection 2.2.5) and infused during training. Right: t-SNE embedding of the latent vectors $z = \mathcal{E}(Y, \epsilon)$ with ϵ as the FOV uniquely describing Y. (Source: [167])

scoring the potential of the contGAN model to forecast future appearances of plant growth. In Table 7.2, the L1 error between the GCF of the reference images and the GCF of the generated images is given for the three compared models. The results indicate that the Pix2Pix and the contGAN model are better in estimating the fractional green canopy compared to the BicycleGAN model, with the contGAN model outperforming both.

Controlled Image Generation with Priors

This section demonstrates the impact of including prior conditions on the latent space to model a distribution of potential outputs. This is illustrated in Figure 7.3. The left figure displays a low-dimensional embedding of the ground truth FOV that are inserted as a prior condition during training. Notably, a clear separation between the three classes is visible, namely *SW*, *FB*, and *MX*. The right plot displays samples from the resulting latent space $\mathcal{Z}$ that was sampled during training. Notably, the latent space forms three distinct and separated classes aligning the FOV. In contrast, when only adding random Gaussian noise, the latent space is randomly distributed and not interpretable [167].

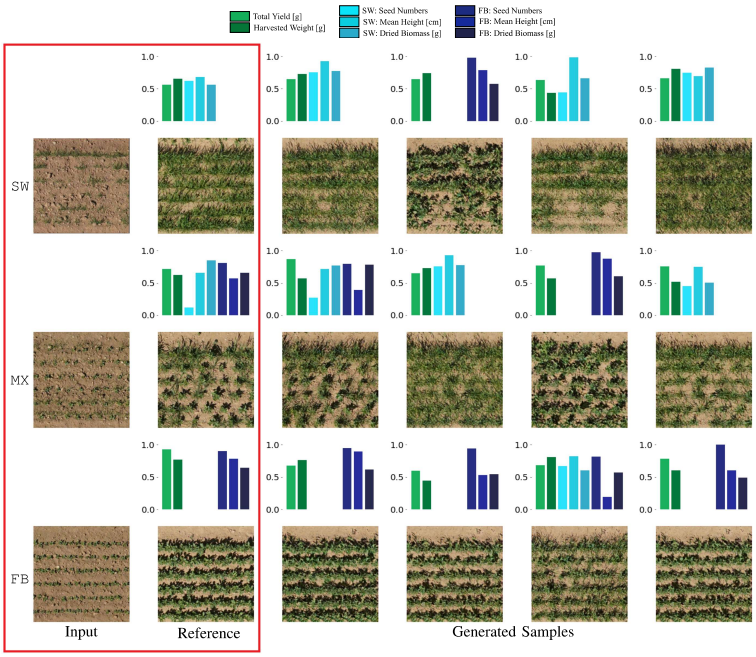


Figure 7.4: Example results of controlled image generation for plant growth modeling. The influence of different FOV on the prediction is illustrated. The first two columns (red box) show the input and the reference images for all the three classes: 1) *SW*, 2) *MX*, and 3) *FB*. On top of each reference image, the normalized FOV are shown as a bar chart that exactly matches the reference image. Next, the generated images using the contGAN model are displayed using different FOV as prior condition. On top of each image, the conditional FOV as a bar chart are displayed that were randomly sampled during test time. Note that feature vectors are interpretable and explainable. Note that the generated samples are controlled by the chosen FOV. (Source: [167])

Figure 7.4 displays the influence of controlled sampling from the interpretable latent space. The model is conditioned on an input image, depicted in the left column, and the FOV, depicted as bar charts above each image. Each bar defines a unique prior condition for the target image (e.g, total yield). The features are further colored according to each class and according to the expression of traits. The image shows that by sampling from the FOV, a distribution of potential outputs can be generated, following the description of the FOV precisely. For instance, reducing the ration of features that are associated with *SW* results in generated images of *FB*. Additionally, the images are highly realistic, with clear morphological details.

7.1.6 Discussion

This section discusses the potential of image generation for plant growth modeling. We showed that images offer several advantages over existing methods and previously presented methods. First, they are easy to interpret, especially for non-experts. Second, different parameters can be derived. This was exemplified by evaluating the GCF between reference and generated images. Moreover, we highlighted that plant growth is inherently ambiguous, and a single input may correspond to a distribution of outputs. To acknowledge this important property, prior knowledge can be encoded into a low-dimensional, interpretable latent space that serves as a condition during image generation. This enables the modeling of a distribution of potential appearances (Figure 7.4). Notably, the proposed method outperforms existing methods for plant growth modeling using image generation [173] while also providing a distribution of outputs. Nevertheless, we see some important limitations in this study. First, the data collection process is very expensive, requiring manual data collection using UAVs. Additionally, the FOV are sampled expensively in the field. Therefore, the dataset is considerably small and only contains a single year and region. This significantly limits the scalability and generalizability of the method. Additionally, the discussed model can only translate between two time points. However, often multiple time steps are required to adequately capture the growing period, as discussed in [182]. Nevertheless, this study contributes towards more explainable ML, thereby supporting the acceptance among researchers and farmers.

7.2 Regression Diffusion for Earth Observation

In the previous section, we studied the impact of latent priors for image generation to model a distribution of outputs for plant growth modeling. For this, a GAN-based model for image-to-image translation was used. However, the model was restricted to mapping a single input time step to a single future time step. Nevertheless, we have seen that many regression tasks in EO require the processing of long time series data to identify patterns and dynamics evolving over time. As demonstrated earlier, crop yield prediction involves analyzing sensor data from seeding to harvesting to uncover patterns that correlate with the measured yield. This requires advanced architectures that can handle long time series inputs.

DMs have achieved outstanding performance in many generative tasks, even outperforming GANs [184]. Inspired by diffusion processes in physical and biological systems, DMs effectively captures spatial and temporal dynamics, with applications ranging from generating photorealistic images [185, 184, 186] to addressing regression problems [187, 188, 189].

Recently, DMs have demonstrated promising performance in EO. For instance, Gao et al. [190] proposes a conditional DM for precipitation nowcasting by explicitly incorporating domain-specific physical constraints. However, most modern generative models do not account for time series inputs and are rarely designed for pixel regression tasks. Furthermore, DMs often generates random samples from the target distribution and lacks strong conditions that are required for image regression tasks. Therefore, the Regression Diffusion (RegDiff) is discussed in this section. RegDiff is a conditional DM tailored for time series regression applications for EO tasks. This method leverages conditional generation for image regression tasks, achieving performance comparable to or surpassing baseline models for crop yield prediction. More importantly, a key advantage of DMs lies in their ability to estimate distributions over predictions that enable robust uncertainty quantification [191, 188, 190]. Leveraging this property, we demonstrate that sampling around the conditional means provides stochasticity that enables uncertainty estimation.

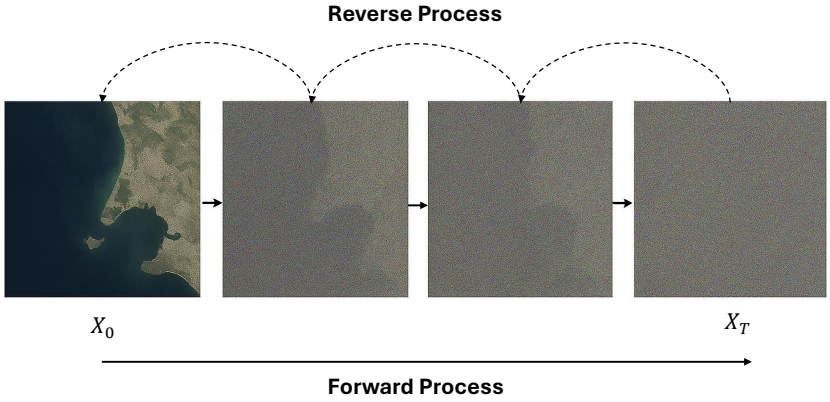


Figure 7.5: Illustration of the forward and reverse process of DMs. The forward process iteratively corrupts the input with noise while the reverse process restores the initial data from noise. (Image partly generated with [41])

7.2.1 Conditional Diffusion for Time Series Regression

In a nutshell, DMs consists of two processes. First, the forward process iteratively corrupts an input image with random noise. Second, a reverse process tries to reconstruct an image from noise that follows the same distribution of the training data. Here, a network $\epsilon_\theta(\cdot)$ estimates the noise component that is subtracted from corrupted image until the image is reconstructed. This method was proposed by Ho et al. [185] and is described in detail in the Appendix (Subsubsection A.1.1). We further provide additional information on the notation. Figure 7.5 schematically shows the forward and reverse processes. Importantly, during inference, the reverse process generates random samples from the target distribution when initialized with noise, resulting in outputs that lack alignment with specific conditions. This stochastic nature makes the standard reverse process unsuitable for regression tasks, which require predictions to be conditioned. Commonly, the literature moved towards text-to-image generation, where the image synthesis is conditioned by a text description [192, 193]. For the yield prediction example, such a model would generate a random yield sample that might not follow the conditions during the growing period. Consequently, such methods are not useful for practical applications such as image regression. Ultimately, stronger conditions are required to ensure that the generated sample explicitly aligns with the input data.

To overcome this drawback, we follow the implementation in [188] and integrate a conditional mean estimator, guiding the reverse diffusion process

Parts of this section originate from a master's thesis that I supervised, extended, and published during this PhD. Parts of this chapter, including figures and tables, have been published already in [183]

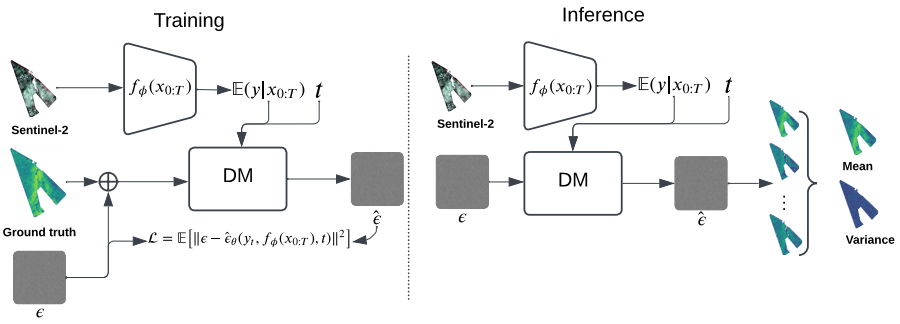


Figure 7.6: Schematic overview of the training and inference process of the RegDiff model using an expert model $f_\phi(X_{0:T})$ and the noise network $\epsilon_\theta(\cdot)$. During inference, multiple samples can be generated from random noise to estimate uncertainty, leveraging the stochastic nature of the sampling process.

to generate accurate estimates of the ground truth. This mean estimator acts as an expert that guides the subsequent estimations based on the ground truth samples. Additionally, this approach is extended to time series data for regression application in EO.

Conditional diffusion

Specifically, the task is to predict the target $Y_0 \in \mathcal{Y}$ from the input data $X \in \mathcal{X}$ by defining data pairs, $\{(X \in \mathcal{X}, Y \in \mathcal{Y}\}$. In the time series image datasets (Section 3.1), the input is given as $X = [X_1, X_2, ..., X_T]$, where T are the number of available time steps in the input sequence. The goal is to find a conditional function that maps the time series input to the target $f_\phi(X_{0:T}) : \mathcal{X} \longmapsto \mathcal{Y}$, by modeling the conditional probabilistic distribution $p(Y_0|X_{0:T})$. This function acts as prior knowledge within the diffusion process [188]. During the reverse diffusion process, a noise network estimates the noise level at each step, conditioned on the prior network's state. At the end of the diffusion process, the generated image should be sufficiently close to the actual target. A schematic overview of the training and inference process is illustrated in Figure 7.6. The figure shows that during training, the DM predicts the noise level of the corrupted ground truth data based on the condition of the prior (expert) network. During inference, a prediction is generated from a random sample by using solely the condition from the prior network as guidance. Interestingly, random initialization generates different results that can be leveraged for uncertainty estimation.

Uncertainty Quantification

As elaborated, DMs are characterized by a stochastic sampling process, which enables uncertainty estimation [188]. To do so, a random sample around the

conditional mean using the expert network is taken, and several predictions of the ground truth are generated via the reverse diffusion process. For uncertainty estimation, a single distribution is defined by taking the mean and standard deviation over all generated samples. The model uncertainty is quantified as the variance around each prediction, as illustrated in Figure 7.6. We will elaborate on this further in Chapter 8. For uncertainty estimation, 10 reverse diffusion processes are defined as proposed in [188].

Data

For the data, we use data from the YieldSAT 2.2.1 dataset, encompassing all three crop types: soybean, corn, and wheat from Argentina. For simplicity, only S2 imagery is used as the model input.

Architecture, Training & Evaluation

The conditional mean network is parameterized by a pre-trained Convolutional LSTM (ConvLSTM) model [194] with a 3D convolution, following the implementation as described by Helber et al. [195]. For this, a 3D convolution is performed on each input of the time series data, increasing the feature size to 64. Following, a ConvLSTM model is applied with a single layer and a kernel size of 9. Finally, a convolution on the output is applied to generate a single value per pixel. This model has the main advantage of integrating convolutional operations directly into the LSTM structure, thereby capturing spatio-temporal dynamics more efficiently than LSTMs.

ConvLSTM

The diffusion network is defined by a lightweight network, consisting of three FCL, each with 128 hidden units. The first FCL takes the conditional mean and the noise-corrupted target. Additionally, the Hadamard product is computed between the FCL output vector and the timestep embedding at each time step. Subsequently, a Softplus function is applied to introduce non-linearity. As the final step, the last FCL predicts an output vector of dimension 1. This represents the predicted forward diffusion noise, $\boldsymbol{\epsilon}$, for $\boldsymbol{Y}$ at each reverse time step.

At the inference level, a sample from the conditional mean is taken $y_T \sim \mathcal{N}(f_\phi(\boldsymbol{X}_{0:T}), \boldsymbol{I})$ and subsequently the predicted noise level is subtracted over 1000 iterations. The final output should be close to the ground truth using the cosine noise schedule. The models are trained on image patches of size 9×9. The Models are trained for each dataset separately using 10-fold CV. We report all metrics as the average over folds. The standard set of regression metrics, as defined earlier in this thesis, is used for evaluation. To evaluate the quality of

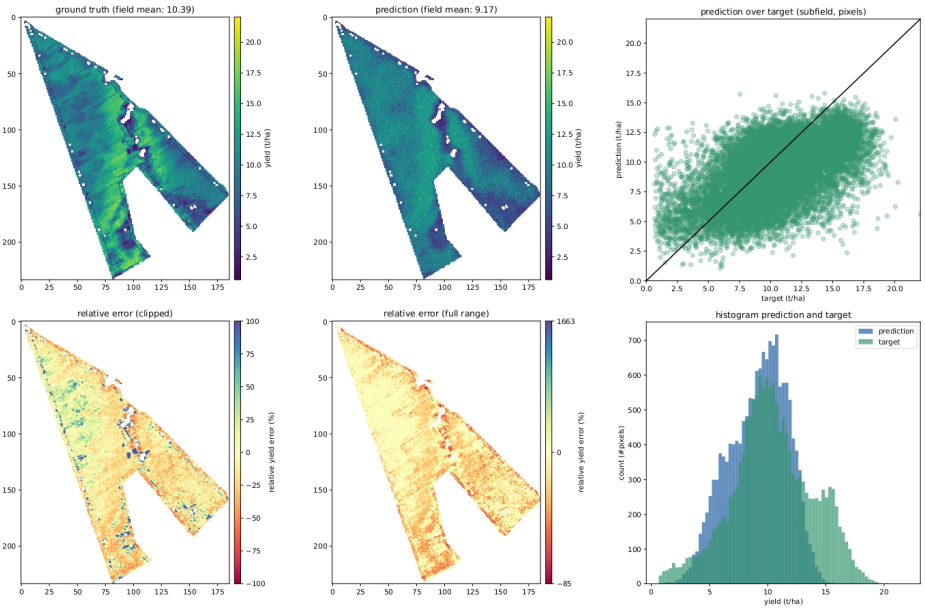


Figure 7.7: Example prediction for a single yield map. Top left: the ground truth yield map, followed by the predicted yield map and prediction-over-target plot. Bottom left: clipped relative error to 100%, absolute error, and histogram of predicted and target values.

the uncertainty estimates, the Quantile Interval Coverage Error (QICE) [188] is used. This score measures how well the learned conditional distribution aligns with the true conditional distribution. Ideally, QICE approximates 0 to indicate an optimal agreement between the predicted and true distributions. The model is compared against a Transformer model following the implementations in [121] and a DM without the conditioning network [185] (DDPM).

7.2.2 Results

Table 7.3: Quantitative results for the *RegDiff* model.

		Field-Level				Subfield-Level			
Dataset	**Model**	MAE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	RRMSE (↓) %	MAE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	RRMSE (↓) %
Corn	Transformer	0.99	0.76	**1.26**	**0.14**	**1.74**	**0.58**	**2.29**	**0.26**
	DDPM	2.60	-0.76	3.11	0.51	4.73	-1.90	5.93	0.92
	regDiff	**0.91**	**0.77**	1.29	0.15	2.16	0.42	2.77	0.33
Soybean	Transformer	0.39	**0.76**	**0.51**	**0.13**	**0.68**	**0.61**	**0.91**	**0.24**
	DDPM	1.04	-0.04	1.17	0.62	1.90	-1.65	2.51	1.06
	regDiff	**0.43**	**0.76**	0.63	0.18	1.13	0.19	1.44	0.42
Wheat	Transformer	0.40	0.92	0.49	0.09	**0.70**	**0.77**	**0.98**	**0.18**
	DDPM	1.77	-0.66	2.16	0.44	3.17	-2.83	3.96	0.80
	regDiff	**0.39**	**0.93**	**0.49**	**0.09**	1.10	0.51	1.42	0.26

The quantitative results are shown in Table 7.3. As expected, the unconditioned DM (DDPM) performs poorly on all regression metrics due to the missing

Table 7.4: Model performance for different numbers of diffusion steps. The results are shown for the *Corn* dataset.

Options	Field-Level				Subfield-Level			
Time steps	MAE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	RRMSE (↓) %	MAE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	RRMSE (↓) %
1000	0.93	0.77	1.32	0.15	2.13	0.43	2.76	0.32
5000	**0.91**	**0.81**	**1.20**	**0.14**	**2.04**	**0.48**	**2.64**	**0.31**

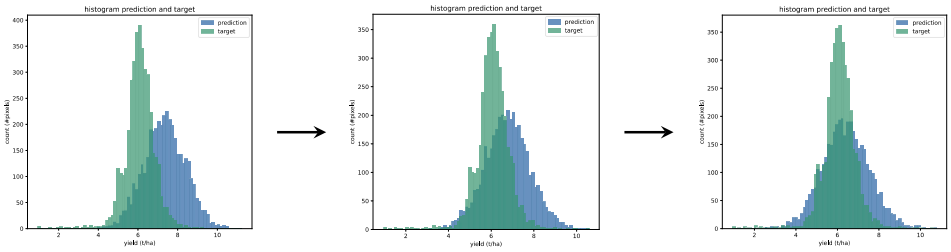


Figure 7.8: Histogram of predicted and target values for different numbers of diffusion steps for 1000, 3000, and 5000 (left to right).

Table 7.5: QICE score for the regDiff model.

Corn	Soybean	Wheat
0.04	0.05	0.05

condition that would guide the model towards faithful predictions. In contrast, the Transformer model and the RegDiff model perform notably well on all regression metrics. Notably, the RegDiff model shows an improvement of 1 pp in R^2 on the corn and wheat dataset at the field level. Surprisingly, on the subfield level, noticeable differences between the Transformer and the DM are evident. We attribute this to the limited number of denoising steps during the reverse process. This can result in remaining noise levels in the prediction and ultimately a mismatch between ground truth and predicted yield pixels. To further evaluate this, the different numbers of diffusion steps are evaluated. The results are shown in Table 7.4. The results are shown for the corn dataset exemplarily. Notably, significant improvements in field- and subfield-level predictions are demonstrated. Increasing the diffusion steps to 5000 results in improvements of 5 pp in R^2 at the subfield and 4 pp in R^2 at the field-level. In Figure 7.8, the alignment of the predicted and target distributions over the diffusion steps is illustrated using a histogram plot. The figure highlights that an increasing number of diffusion steps results in a better alignment between the predicted and target distribution.

In Table 7.5, the QICE metric for all three datasets is evaluated. The best results are achieved for the corn dataset, with a score of 0.049. Overall, the

results indicate that the model provides a strong basis to estimate the true conditional distribution of the target data.

7.2.3 Discussion

In this section, we evaluated the potential of DMs for time series regression tasks. We highlighted that traditional DMs often overlook time series data, especially in image regression tasks. Additionally, we highlighted that DMs requires consistent conditioning with prior knowledge to be applicable for regression tasks in EO. To adress RQ2, we showed that this can be done using conditional expert models. This was demonstrated with RegDiff model. This model is a prior conditioning network that guides the diffusion process and is integrated directly into the learning system. Thus, the processing of time series image data can be used as a condition. The approach demonstrates improved performance on three crop yield prediction datasets in Argentina. It was demonstrated that a conditioned DM surpasses both an unconditional DM and a Transformer model. Furthermore, it was demonstrated that the stochastic nature of the sampling process enables a simple way for uncertainty estimation. Interestingly, this can be applied to any NN, ultimately contributing to more reliable ML in safety-critical applications.

Limitations

We see noticeable limitations in this work that must be evaluated before implementing it. First, a weaker performance on the pixel level is observed. This was attributed to the limited number of diffusion steps. Here, increasing the number of diffusion steps resulted in a significantly better performance. However, increasing the number of diffusion steps also increases computational complexity. Moreover, DMs introduce additional hyperparameters, such as the variance schedule. Therefore, a systematic study of their influence on the model performance must be a key focus. Additionally, the conditioning mechanisms must be further explored, such as alternative architectures and soft conditional settings, as suggested in [187, 196]. This could potentially simplify the training process and reduce the computational demands. Additionally, the quality of uncertainty estimates must be more thoroughly studied, including comparisons with alternative methods.

7.3 Conclusion

Conditional learning provides unique potential for plant growth modeling and crop yield prediction by contributing to more trustworthy and explainable

ML. We showed two different approaches that integrated prior knowledge as a condition during generation process.

PART IV

KNOWING WHAT WE DO NOT KNOW

8 INTRODUCTION TO UNCERTAINTY ESTIMATION

Chapter Highlights:

- Uncertainty estimation can be used to increase trust in ML for EO regression tasks.
- Naturally occurring missing time steps can be leveraged for uncertainty estimation.
- Bayesian inference and prior knowledge can overcome the severe performance collapse under distribution shifts in yield prediction.

This chapter acknowledges the unknown through uncertainty quantification. A key aspect of ML is the recognition that not everything is known, or as Plato said: *"For I was conscious that I knew practically nothing..."*. This highlights the understanding shared by most domain experts: some aspects are inherently unknown. It is, however, fundamental to express your uncertainty.
In this chapter, we provide a general overview of uncertainty estimation, focused on EO regression tasks. The following chapters address the RQ3 using the following subquestions:

- **RQ3.1** How can we quantify the unknown for EO regression tasks?
- **RQ3.2** How can we leverage prior knowledge to quantity uncertainty?
- **RQ3.3** How can we leverage prior knowledge and uncertainty estimation to overcome real-world challenges in EO tasks?

8.1 Uncertainty Estimation in Earth Observation

In this section we will address RQ3.1 and provide a general overview of uncertainty estimation focused to regression tasks. In the previous chapter, the outstanding potential of ML for EO tasks was demonstrated. Nevertheless, traditionally ML methods do not provide confidence about their predictions. This, however, is fundamental for critical applications. Additionally, modern NNs are often badly calibrated, which can reduce the trust in ML-based

predictions [197, 198]. To overcome this, ML research has established the discipline of uncertainty estimation [18]. Although uncertainty estimation methods are often simple to formulate, they are highly difficult to train for large NNs [199], especially for EO datasets [200]. This is primarily due to computational complexities and training instabilities [201]. Additionally, the EO field has very specific requirements that we have discussed earlier and that must be considered. First, the availability of EO data increased dramatically in recent years, resulting in the Era of *Big Geo Data* [202]. While the availability opens new possibilities in assessing the status of planet Earth, the abundance of data, however, requires scalable and efficient solutions. Second, EO data is multimodal and therefore requires flexible architectures and data fusion schemes. Those architectures must be compatible with an uncertainty estimation method. Finally, EO data is noisy, characterized by missing information, and prone to severe data shifts [100]. All those aspects make uncertainty estimation difficult in EO.

Earth Observation requirements

8.2 Sources of Uncertainty in Earth Observation

First, we provide a brief overview of sources of uncertainties that frequently occur in EO tasks. Many factors introduce uncertainties that are frequently overlooked in the current EO literature. The sources of uncertainty are inspired by [18].

1. Variability in the Environment: Real-world systems constantly change. For example, the reflection of a plant can be significantly different after a rain compared to after a drought. Additionally, a crop might yield significantly differently depending on location and management conditions. These differences are often difficult to detect from EO sensors.

2. Sensor Errors: Often, sensors exhibit a degree of inconsistency that can introduce uncertainty. A common problem in EO is the spatial resolution of optical and multispectral satellite imagery. Since resolution is often limited, a single pixel often captures mixed information from multiple surface features (e.g., a road next to a field). Moreover, label sensors are often affected by noise. In Subsection 2.2.1, we have seen that yield data from combine harvesters is often affected by poorly calibrated yield sensors that result in incorrect labeling. Moreover, incorrect positioning systems or delays in the measurement process can cause mismatches between data modalities, such as S2 data and yield measurements.

3. Noisy or Missing Information: EO data is often affected by noisy and

missing information. For instance, optical satellite images are often obscured by clouds and haze, hindering continuous monitoring of the Earth's surface. More specifically, approximately 55 % of the Earth's surface is continuously covered by clouds, with significantly higher coverage during meteorological winter seasons [203]. This can considerably impact the information content in optical satellite data and ultimately introduce uncertainty.

3. Unknown Data: Unknown data is a commonly known problem in ML. The most prominent example is training a classifier on images of cats and dogs. In this training setup, we assume that the space contains only those two categories. However, showing the model an image of a bird will result in complete failure, since this category originates from a completely different space.

4. **Model Structure & Training Pipeline:** errors in the model structure and errors in the raining pipeline can introduce uncertainty. This could, for instance, include the architecture, that can not sufficiently model the underlying process. Additionally, the number of hidden layers, or the number of epochs for training can even cause uncertainty.

8.2.1 Modeling Uncertainties

We already discussed the sources of uncertainty that can arise from the data or the model itself. Similarly, the current ML literature distinguishes at least between two types of uncertainty, namely the data uncertainty, also called *aleatoric uncertainty*, and the model uncertainty, also called *epistemic* uncertainty [204]. Both can be modeled independently. The aleatoric uncertainty captures the randomness in the data collection process. One example of aleatoric uncertainty is the occlusion of optical satellite imagery by clouds or haze. First, cloud occurrence has a random component, and second, occlusion makes it difficult to infer underlying information (e.g., crops or urban areas). Consequently, a model can only guess what lies beneath. Importantly, in this case, the uncertainty cannot be reduced by simply collecting more data of the same kind (i.e., more cloudy images).

Uncertainty can be disentangled into data and model uncertainty

Data uncertainty cannot be reduced with more data

Epistemic uncertainty, in contrast, arises from a lack of knowledge that can in principle be reduced by adding more information. In the classification example of a cloud-corrupted image, we could provide the model with cloud-free observations from before or after the occlusion. With this additional information, the model can learn to reason that if a location was once classified as crop, it is likely to remain the same during the cloudy period as well.

Model uncertainty can be reduced with more data

While this distinction sounds plausible at first glance, it is difficult to clearly

distinguish between the different types of uncertainty in practice. Their interpretation is often context-dependent, and the boundary between aleatoric and epistemic uncertainty should not be seen as a fixed notion [204]. In the example of a cloudy satellite image, what is treated as irreducible randomness in one setting may become reducible in another (e.g., when combining temporal sequences or multisensor data). Furthermore, Valdenegro-Toro and Mori [205] pointed out that disentangled uncertainties do interact in practice.

Predictive uncertainty is the uncertainty that is propagated into a prediction from various sources.

Regardless, in practice, we are primarily interested in the uncertainty that propagates into a prediction. Therefore, it is common in ML to group both types of uncertainty together under the notion of *predictive uncertainty*. Kendall and Gal [206] provide a general framework of modeling both types of uncertainty. This will be later discussed in more detail in Subsection 8.2.5.

Many methods have been proposed that focus on scalable uncertainty estimation [207, 18]. First, different uncertainty estimation methods are compared and evaluated in terms of their ability to handle the complex nature of EO data and the quality of their uncertainty estimates. According to Gawlikowski et al. [18], uncertainty estimation methods can be divided into different categories, including i) *Bayesian methods*, ii) *Ensemble methods*, and iii) *Test time data augmentation*. In the following, we provide an overview of each category.

8.2.2 Bayesian Neural Networks

The Bayesian framework is a popular mechanism for reasoning about uncertainty [18]. A BNN aims to find a probability distribution over the model parameters. More specifically, the goal is to find the posterior distribution over the model parameters θ using Bayes' Theorem, given data and a prior distribution:

$$p(\theta|x, y) = \frac{p(y|x, \theta)p(\theta)}{p(y|x)}, \tag{8.1}$$

Here, $p(\theta)$ is the prior distribution expressing the prior belief about the posterior. Additionally, $p(y|x, \theta)$ is referred to as the likelihood of the data given the model parameters. Finally, $p(y|x)$ is a normalization constant, referred to as the model evidence, and is defined as:

$$p(y|x) = \int p(y|x, \theta)p(\theta)d\theta. \tag{8.2}$$

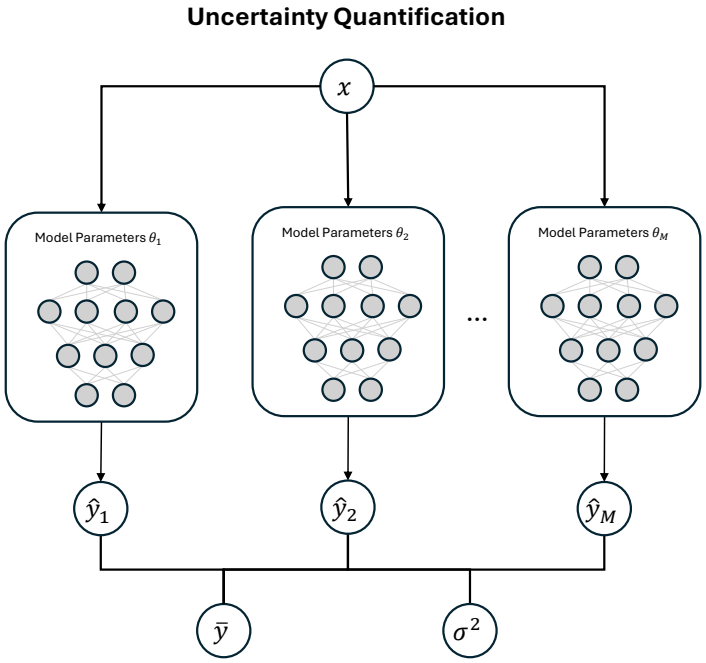


Figure 8.1: Diagram of an uncertainty estimation method, consisting of M models with distinct model parameters θ. This can be either a sample from a Bayesian Neural Network (BNN) or a deterministic ensemble member. Each model generates a prediction for an input sample. The final prediction is the mean across all models. The model uncertainty is given by the variance across the predictions. (Figure derived from [18])

The estimated posterior distribution is later used to make a prediction of an output for a new input using Bayesian Model Averaging (BMA):

$$p(\hat{y}|x,y) = \int \underbrace{p(\hat{y}|x,\theta)}_{\text{Data}} \underbrace{p(\theta|\mathcal{D})}_{\text{Model}} d\theta. \tag{8.3}$$

This is a direct application of the law of total probabilities and enables the computation of uncertainty. In practice, Equation 8.3 is intractable to evaluate for large NNs. Therefore, approximation methods are commonly applied. A common strategy is to employ Monte Carlo (MC) approximations that follow the law of large numbers. The expected value is approximated by evaluating M deterministic models that are parameterized by M samples of the posterior distribution:

$$\mathbb{E}[\hat{y}] \approx \bar{f} = M^{-1} \sum_{i=1}^{M} f_{\theta_i}(x), \tag{8.4}$$

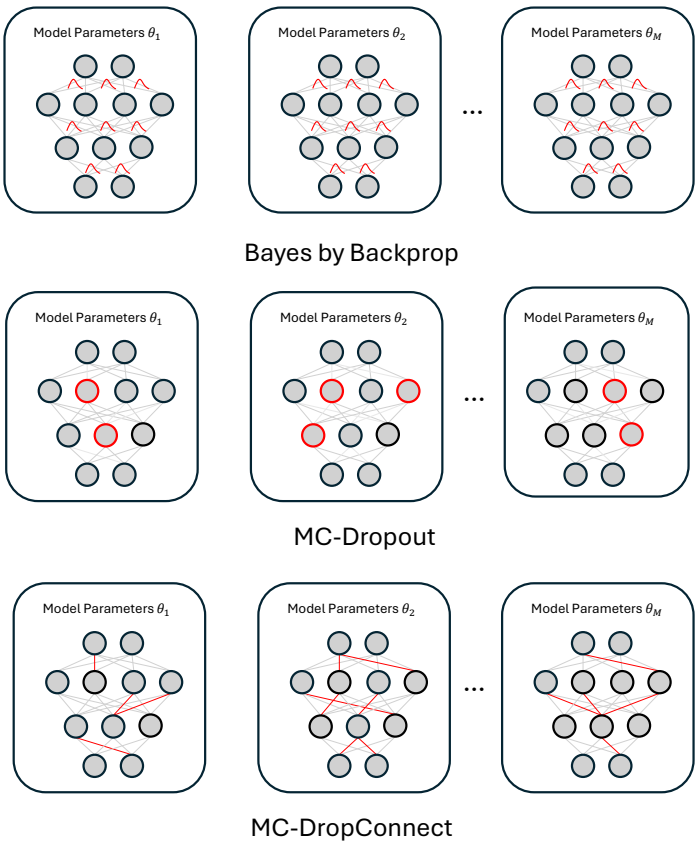


Figure 8.2: Diagram of individual BNNs, including Bayes By Backprop (BBB), MC-Dropout and MC-DropConnect. Note that only the BBB method models a real distribution over model parameters.

giving the MC sample mean. More explicitly, $\bar{f}$ denotes the empirical mean over M samples of the model output $f_\theta(x)$. This approach is also known as ensemble learning and offers a principled way of calculating the model uncertainty by taking the variation around the approximate predictive mean:

$$V(\hat{y}) = M^{-1} \sum_{i=1}^{M} (f_{\theta_i}(x) - \bar{f})^2. \tag{8.5}$$

This formula expresses the uncertainty of the prediction. A schematic overview of calculating uncertainties is given in Figure 8.1. Unfortunately, the probabilities of the posterior distribution are difficult to compute for large NNs [208, 206]. To overcome this drawback, approximate Bayesian Inference methods are utilized in practice. A common approach is approximate Variational Inference. This method introduces a simpler distribution, $q(\theta)$, approximating

the true posterior by minimizing the KL divergence between q and the true posterior:

$$KL(q(\theta) \mid\mid p(\theta|\mathcal{D})). \tag{8.6}$$

Since the posterior is unknown, the KL divergence cannot be minimized directly. However, minimizing the KL is shown to be equal to minimizing the *Evidence Lower Bound (ELBO)*, defined as:

$$\mathcal{L} := -\int q(\theta)\ \log(y|x,\theta)\ d\theta + KL(q(\theta)|p(\theta)). \tag{8.7}$$

Here, the first term is referred to as the likelihood part, enforcing accurate predictions, and the second to as the complexity part, enforcing the output to follow the prior distribution, preventing the model from overfitting. The underlying concept is that the loss function of the NN is defined as the ELBO. An important contribution to the ELBO was made by [209] who introduced the *reparameterization trick* that enables the reduction of variances in stochastic gradients.

Bayes By Backprop (BBB) [210] was introduced to quantify the uncertainty in the model weights by approximating the posterior distribution with a simpler distribution, as described earlier. However, BBB is still computationally expensive [208]. Consequently, many applications rely on last-layer uncertainty approaches, meaning that only the last layers inside the NN are made Bayesian.

MC-Dropout [208] was introduced to approximate Bayesian Inference with low computational costs using a dropout variational distribution. Dropout is a frequently used regularization technique that randomly drops units of the NN by sampling from a Bernoulli random variable with probability p [162]. Gal and Ghahramani [208] showed that dropout approximates a Gaussian process. Randomly sampling from a Bernoulli distribution and dropping out units of the network, approximately integrates over the model parameters and thus approximates a BNN.

ConcreteDropout [211] addresses limitations of MC-Dropout. One limitation of MC-Dropout is that the dropout probability p can be seen as a tunable hyperparameter. Moreover, p is directly linked to the predicted uncertainty. Thus, a search over the parameter space is necessary to achieve good uncertainty estimates. Therefore, MC-Dropout was extended to optimize the parameters of a Bernoulli distribution by using continuous relaxation. This was achieved by treating p as a continuous sigmoid function [211].

MC-DropConnect [212] resembles MC-Dropout but instead of randomly dropping activations, it randomly drops weights to zero with probability p. The

dropout layer is turned on during inference, allowing the generation of samples from the Bayesian posterior distribution. In Figure 8.2, a schematic overview of the presented BNNs is provided.

8.2.3 Ensemble Methods

Ensemble Methods consist of training multiple copies of the same architecture and then combining their predictions, which usually produces a better decision than a single model [207]. Combining the predictions of ensemble members could simply involve taking the average over each individual prediction. The mean prediction over M ensemble members is given as in Equation 8.4. Likewise, the uncertainty can be retrieved by taking the variance around the individual members, as given in Equation 8.5. To introduce variability into the ensemble of learners, random initialization and data shuffling are sufficient [18]. Lakshminarayanan et al. [158] demonstrated that ensembles work well in practice and also have good uncertainty quantification properties. Wilson and Izmailov [213] even pointed out that DEs are characterized by effective Bayesian marginalization properties. Most likely because they tend to sample from different modes and are characterized by high function space diversity [214]. Here, modes refer to the different, often diverse peaks in the loss landscape of the parameter space that a NN can find. In contrast, variational methods and deterministic NNs often fail to explore multiple modes, which may explain why they can fail in practice. In contrast, ensemble methods explore different local optima and thus multiple modes in the weight space, resulting in more stable predictions and uncertainty estimates.

Neural Network Modes

Nevertheless, ensemble methods are characterized by increased computational costs since several copies of the same architecture must be trained. This can easily reduce the applicability of ensemble methods, especially in large and high-dimensional EO datasets.

8.2.4 Test-Time Data Augmentation

Test-time data augmentation is another technique used to quantify uncertainty. It involves creating multiple augmented versions of each test sample to generate different views of the input. For each *view*, a prediction is generated, and the variability across these predictions is then used to estimate the uncertainty. This method is often easy to implement, as only a single model is required and

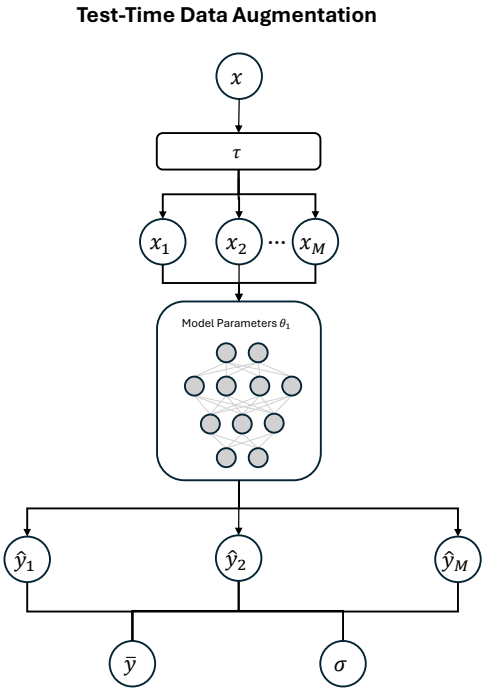


Figure 8.3: Schematic overview of test-time data augmentation for uncertainty estimation. The augmentation function τ creates multiple views of the input, and for each view, a prediction is generated.

no additional modifications are needed. For an augmentation function, the mean prediction across multiple views is given by:

$$\mathbb{E}[\hat{y}] \approx \bar{f} = M^{-1} \sum_{i=1}^{M} f_\theta(\tau(x)), \tag{8.8}$$

where τ is an augmentation function with a stochastic component to create random views for the input. The uncertainty for the prediction is similarly given by calculating the variance. Although the test-time data augmentation method is a simple way to estimate uncertainty, it is important that the augmentation function produces only valid views of the input. Test-time data augmentation is more common in the medical domain, as it already relies on data augmentation techniques [18]. In contrast, in the EO field, this method for uncertainty estimation is less explored. A schematic overview of test-time data augmentation is given in Figure 8.3.

8.2.5 Combining Model & Data Uncertainty

As mentioned earlier, Kendall and Gal [206] provide a principled way of combining aleatoric and epistemic uncertainty in a single model. For this, the data uncertainty is learned directly from the data using loss attenuation by including a second component that estimates the noise inherent in the data. Assume a model that outputs the parameters of a Gaussian distribution $f_\theta(x) = [\hat{y}, \sigma^2]$, reflecting the predictive mean and predictive variance, respectively. Thus, the total uncertainty of M samples of the posterior distribution is given as:

$$V(\hat{y}) \approx \underbrace{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2}_{\text{Data}} + \underbrace{\frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i - \bar{\hat{y}})^2}_{\text{Model}}, \qquad \bar{\hat{y}} = \frac{1}{M}\sum_{i=1}^{M}\hat{y}_i. \tag{8.9}$$

Here, the first term, σ^2, expresses the inherent noise in the data that can be learned as a function of the data. Similarly, the second component expresses the model uncertainty that is given by taking multiple samples of the posterior distribution, as elaborated in Equation 8.5.

Since the noise in the data is a learnable parameter, it must be optimized. This can be done using the Gaussian Negative Log-Likelihood (NLL) [215] loss. For a single sample from the posterior distribution (or ensemble member), the NLL is given as:

$$\mathcal{L}_{\text{NLL}}(x, y) = \frac{1}{2}\left[\log \sigma_i^2 + \frac{(y - \hat{y}_i)^2}{\sigma_i^2}\right] \tag{8.10}$$

The NLL acts as a proper scoring rule and optimizes both the prediction and the variance under the Gaussian assumption by balancing both terms. Interestingly, no uncertainty labels are required to learn the uncertainty. Instead, the uncertainty is learned implicitly. More importantly, this mechanism reduces the impact of erroneous labels, making the model more robust to noisy data [206].

8.2.6 Evaluating the Quality of Uncertainty

We already discussed in Section 3.3 how to evaluate an NN. Nevertheless, evaluating the quality of uncertainty estimates is similarly important. A detailed description of evaluation measures is provided in [216, 18].

In general, evaluating the quality of uncertainty estimates involves evaluating the mean and variance vector for each prediction. This is done to determine

whether the true value lies within the prediction interval with confidence level $1-\alpha$. The prediction interval is given by:

$$\left[\underbrace{y^L = \hat{y} - \tfrac{1}{2}\Phi^{-1}(1-\alpha)\cdot\sigma}_{\text{Lower Bound}} \; ; \; \underbrace{y^U = \hat{y} + \tfrac{1}{2}\Phi^{-1}(1-\alpha)\cdot\sigma}_{\text{Upper Bound}}\right], \tag{8.11}$$

where $\Phi^{-1}(1-\alpha)$ is the quantile function. Therefore, the model does not only predict a single value, but a range within which the true value is expected to lie with confidence level $1-\alpha$. Subsequently, the Mean Prediction Interval Width (MPIW) can be calculated for a confidence level:

$$MPIW = N^{-1}\sum_{i=1}^{N}(y_i^U - y_i^L), \tag{8.12}$$

with y_i^U and y_i^L being the lower and upper bounds es described in Equation 8.11. Since the goal is to learn a model that predicts low uncertainty, the MPIW should be minimized. Nevertheless, the MPIW does not tell anything about whether the prediction interval covered the ground truth value or not. Therefore, we must also report the Prediction Interval Coverage Probability (PICP) that tells us the fraction of true values that are captured by the prediction interval. The PICP is calculated as:

$$PICP = \frac{c}{N}, \tag{8.13}$$

with c being the number of samples that are captured by the predicted interval and N the total number of samples. Consequently, while the MPIW has to be minimized, the PICP has to be maximized.

Calibration is the capability to accurately reflect the true probabilities of observed outcomes.

We can evaluate the PICP for different confidence levels and thus evaluate the calibration of a model. For being calibrated, a model should accurately reflect the true probabilities of observed outcomes. For instance, for an 80% confidence, an event should also occur 80% of the time. Late research revealed that most modern NNs are poorly calibrated and often produce unreliable predictions due to over-confidence or under-confidence [197, 198]. To evaluate model calibration, the accuracy can be shown as a function of confidence, providing us with a calibration curve. A schematic overview of different calibration diagrams is depicted in Figure 8.4. Although calibration diagrams are visually easy to evaluate it often makes sense to express the model calibration in a single metric. For this, the Expected Calibration Error (ECE) [217] is commonly

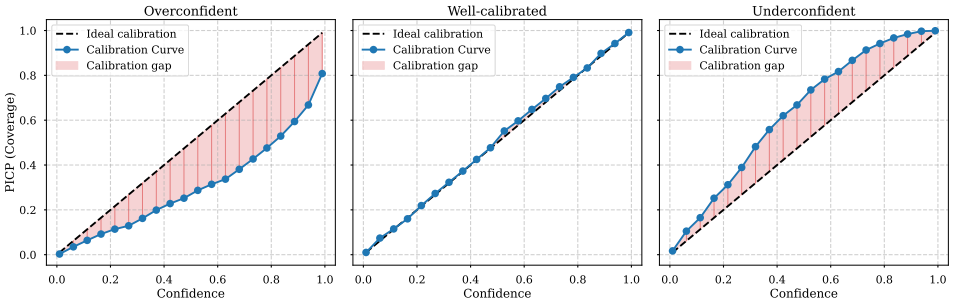


Figure 8.4: Example reliability diagram illustrating an overconfident, well-calibrated, and underconfident model.

used. The ECE is a binning-based metric and expresses the average violation of calibration across bins (confidence), weighted by the number of bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \Big|\mathrm{PICP}(\mathrm{B_m}) - \mathrm{conf}(B_m)\Big|, \tag{8.14}$$

with M being the number of bins (nominal confidence levels), N the total number of samples, and $|B_m|$ the number of samples in bin m. Nevertheless, binning has several limitations that should be considered. First, the choice of the number of bins can significantly affect the ECE. For instance, too few bins can result in an underestimation of the ECE. Additionally, the ECE captures only the magnitude of miscalibration. Since the ECE relies on absolute difference, it ignores whether the model is overconfident or underconfident. Thus, an important property of miscalibration is lost. Finally, the ECE is biased to the sample size, which can lead to unstable estimates.

8.3 Conclusion

In this section we addressed RQ3.1 and presented an introduction to uncertainty estimation for regression applications. We presented different methods and evaluation techniques for uncertainty estimation that can be used in EO regression tasks.

9 UNCERTAINTY QUANTIFICATION WITH MISSING DATA

Chapter Highlights:

- Missing data pose a significant challenge in EO applications.
- Monte Carlo Temporal Dropout (MC-TD) is a novel method for uncertainty estimation that leverages missing time steps.
- MC-TD improves the model performance while increasing the robustness to missing time steps.

In this chapter, we focus on RQ3.1 and RQ3.2 and present a novel method for uncertainty estimation that leverages prior knowledge about missing data in EO.

9.1 Analyzing Temporal Dropout for Regression Tasks

In this section, we highlight the impact of missing time steps. We have already seen that processing time series data is critical for understanding the changes and dynamics of our planet [218]. Moreover, we have discussed that sensors can experience anomalies (e.g., cloud occlusion in optical data) that result in missing information over certain individual time steps [99, 118, 130]. As mentioned, clouds corrupt optical sensors (e.g., S2), as on average, 55% of Earth's land surface is covered by clouds [132]. Inserting such time steps into a model can propagate uncertainty into the predicted outputs, as the model cannot extract meaningful information, as discussed in Subsection 8.2.1. Therefore, addressing the impact of missing data in time series is an urgent challenge. Therefore, various solutions and preprocessing techniques exist to mitigate the impact of missing data and ensure accurate predictions. While various techniques exist that adress missing data [118, 219], most methods overlook the explicit quantification of the introduced uncertainty. However, addressing this gap is crucial for improving the reliability of a model.
In this section, we analyze two methods that simulate missing data across

Parts of this chapter, including figures and tables, have been published already in [44]

time, applied during training and inference. First, during training, it acts as an augmentation method that increases the robustness to missing time steps. Second, during inference, generating multiple MC samples with different *views* serves as a mechanism for uncertainty quantification.
The first technique is based on the dropout variational distribution [208] applied across time, referred to as Monte Carlo Temporal Dropout (MC-TD). Here time steps are randomly dropped using a specific dropout ratio. Since, the optimal dropout ratio is a resource expensive and difficult hyperparameter that depends on each dataset and missing data modality [219], Monte Carlo Concrete Temporal Dropout (MC-cTD) is proposed. Instead of manually searching the best dropout ratio, the *concrete dropout* distribution [211] is used that learns the optimal dropout ratio using standard gradient descent. The concrete dropout distribution approximates a Bernoulli distribution using a continuous relaxation, enabling the dropout probability to be optimized through gradient-based learning [220, 211]. Unlike traditional dropout-based approaches, this model automatically learns an optimal dropout distribution, avoiding costly hyperparameter tuning.

9.1.1 Missing Data in Earth Observation

Irregularly sampled time series are a common phenomenon in signal processing [221]. Numerous studies in DL have focused on developing methods for time series imputation, such as BRITS [222], mTAN [223], and SAITS [224]. In contrast, other methods focused on ignoring the missing data, such as D-GRU [225] and MissFormer [226]. In EO, missing spatial and temporal data negatively affect the performance of predictive models, where more missing data translates to worse predictions [227, 228]. However, some strategies in the EO field mitigate the negative impact of missing data, such as by incorporating features from multiple sensors or using dropout techniques [229, 118]. The Temporal Dropout (TD) technique, which involves randomly dropping time steps, has been used to enhance the model performance [118, 230, 195]. Furthermore, a few studies evaluate missing data as a data augmentation technique to enhance the robustness to missing information. For instance, Weilandt et al. [231] randomly masks the end of a time series to improve early crop classification. On the other hand, some studies have focused on reconstruction tasks that recover the missing time steps. For instance, Chen et al. [232] uses a polynomial fit based on the Savitzky–Golay filter, while others use DL models based on MLPs [233] or convolutions [234].
Regression tasks in EO are commonly underexplored compared to classification

tasks with time series data [235]. One reason might be that DL models are less effective for regression than for classification, as shown by Tan et al. [236]. Another reason might be the limited number of available benchmark datasets [237]. Regardless, the inherent challenge of predicting a continuous rather than a categorical value is often overlooked. Nevertheless, in EO, numerous applications involve pixel-level regression on time series data. For instance, Nguyen et al. [238] use multispectral data and weather time series for pixel-wise crop yield prediction. They use a MLP model for processing the time series data as individual features. However, previous work has shown that RNNs (e.g., LSTMs) achieve better results (see Chapter 4). Furthermore, Maimaitijiang et al. [117] considers the use of drone-based optical time series data for the same task, obtaining images unaffected by cloud occlusion. Another pixel-wise task studied in the literature is cloud removal. This task has been explored with multispectral optical time series data [239], as well as including radar time series [130]. Similarly, in the Earth surface forecast, multispectral optical and weather time series have been used with spatio-temporal models to obtain accurate predictions [240], such as using ConvLSTM models [241].

We have already discussed the importance of uncertainty estimation in EO applications (see Chapter 8). Reliable uncertainty estimates improve decision-making by identifying predictions with high confidence and those that should be treated with caution [206, 204]. It is important to note that the efficacy of uncertainty estimates heavily depends on the chosen method and the specific application [242, 200]. This makes the model selection an important aspect. Although uncertainty quantification methods are readily applicable to time series data [243, 244], their primary application is typically in the hidden layers of a NN. Additionally, most of these studies overlook the uncertainty arising at the input level, which can be crucial in improving model reliability and robustness.

To overcome this limitation, we explore the simulation of missing data in EO time series for regression tasks.

9.1.2 Temporal Dropout for Uncertainty Estimation

We define the problem of learning the uncertainty in the predicted output, propagated through the uncertainty in the input and the model. To estimate model (epistemic) uncertainty, we model the expected output and variance across multiple stochastic forward passes, as described in Chapter 8, by explicitly creating multiple views of the input. This was described in Subsection 8.2.4. Similarly, to estimate the data (aleatoric) uncertainty, which arises from the

data, we assume a normal distribution that is parameterized by two output heads. For this the NLL is minimized, as described in Subsection 8.2.5. This approach captures aleatoric uncertainty in $\sigma^2(x)$. Valdenegro-Toro et al. [245] provides a relationship between input and output uncertainty, showing that input uncertainty is propagated through the model, ultimately resulting in model uncertainty via stochastic MC sampling. We follow this argumentation. Finally, the Predictive Uncertainty (PU) is the combined effect of all uncertainties (see Equation 8.9).

Temporal Dropout

We leverage TD to estimate uncertainty. The dropout technique [162] is commonly used in hidden layers of DL models during training. Nevertheless, applying this technique at the input level over time has shown strong regularization performance in EO applications [74, 118, 230, 219, 195]. This prevents the model from focusing on individual time steps while simultaneously handling missing time steps. Thus, the multivariate time series data $\boldsymbol{x}$ is masked out as

$$\hat{\boldsymbol{x}} = \boldsymbol{x} \odot (1 - \boldsymbol{p}), \tag{9.1}$$

where $\boldsymbol{p} \in [0, 1]^T$ is the binary mask, which is drawn from a Bernoulli distribution, i.e. $p_i \sim \text{Bern}(\alpha), \forall i = 1, ..., T$. Here, α is defined as the dropout ratio. Thus, the input $\boldsymbol{x}$ becomes a random variable $\hat{\boldsymbol{x}}$, providing additional network regularization. In contrast, we estimate the uncertainty arising from missing time steps by enabling TD at the inference level. Gal and Ghahramani [208] demonstrated that dropout at inference approximates Bayesian inference in any NN architecture. Similarly, we extend this formulation along the temporal dimension, named MC-TD. This corresponds to sampling L different dropout masks for prediction, $p_i^{(l)} \sim \text{Bern}(\alpha)$, $\forall l = 1, \ldots, L$, and applying them to the input data as $\hat{\boldsymbol{x}}^{(l)} = \boldsymbol{x} \odot (1 - \boldsymbol{p}^{(l)})$.

Concrete Temporal Dropout

When using the TD technique, an optimal dropout ratio must be identified. This is a computationally expensive process, especially when working with large models, which are commonly used in the EO. Gal et al. [211] proposes a continuous relaxation of the dropout technique, allowing the dropout ratio to be a learnable parameter. The soft dropout mask $\tilde{\boldsymbol{p}}$ is defined as:

$$\tilde{\boldsymbol{p}} = CS(\alpha, \boldsymbol{u}, \tau) = \sigma\left(\left(\left(\log \frac{\alpha}{(1-\alpha)} + \log \frac{\boldsymbol{u}}{(1-\boldsymbol{u})}\right)\frac{1}{\tau}\right), \right. \tag{9.2}$$

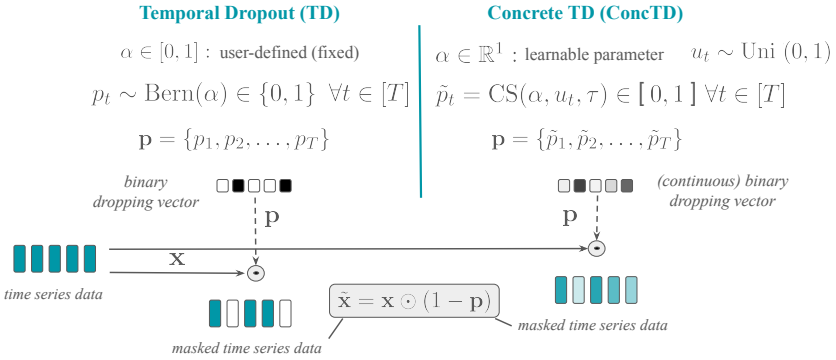


Figure 9.1: Illustration of both TD techniques over time series data for uncertatiny estimation. Left: MC-TD. Right: MC-cTD.

Table 9.1: Dataset description and statistics.

Dataset	Samples	Series length	Features	Avg. target	Std. target
SwissYield [40]	54,098	16–55	15	7.356	2.001
LFMC [48]	2,578	4 (fixed)	61	103.987	39.562
PM2.5 [52]	167,309	120 (fixed)	9	73.673	68.546

with σ as the sigmoid function, $\tau \in \mathbb{R}$ as a temperature value that controls the smoothness of the approximation, $\alpha \in \mathbb{R}$ a learnable parameter in the model, and $\boldsymbol{u}$ a uniform random variable, $u_i \sim \text{Uni}(0,1), \forall i = 1, ..., T$. The latter acts as the auxiliary variable in the sampling reparametrisation [220]. We use this sampling strategy in the dropout technique across time. Furthermore, to get the uncertainty, we use this technique at inference time, referred to as MC-cTD. This corresponds to obtaining L samples of the Uniform distribution, $u_i^{(l)} \sim \text{Uni}(0,1), \forall l = 1, \ldots, L$, and forward over the model with data as $\hat{\boldsymbol{x}}^{(l)} = \boldsymbol{x} \odot (1 - CS(\alpha, \boldsymbol{u}^{(l)}, \tau))$. Additionally, in Figure 9.1 we schematically illustrate MC-TD (left) and MC-cTD (right). Note that MC-TD randomly samples a dropout mask with probability p, whereas MC-cTD provides a continuous binary dropout vector with a learnable dropout ratio. Additionally, in Figure 8.3 we schematically illustrate uncertainty estimation with test-time data augmentation. Here, TD is considered the augmentation function.

Datasets

The described methods are evaluated across three EO regression applications that involve processing temporal sensor data, namely Life Fuel Moisture Content (LFMC) estimation for wildfire risk assessment, air pollution forecasting using Particle Matter 2.5 (PM2.5) predictions, and pixel-wise crop yield prediction. All three applications and datasets are described in Section 2.2. Characteristics each the datasets are presented in Table 9.1.

Wildfire, air pollution, and yield prediction

Table 9.2: Results of the two variants of dropout across time. The MC-TD model uses a dropout ratio of 0.3.

Dataset	Model	R^2(↑)	RMSE (↓)	MAE (↓)	ECE (↓)	PU (↓)
SwissYield	MC-TD	0.76	0.99	0.72	2.14	1.00
	MC-ConcTD	0.79	0.93	0.66	2.23	1.23
LFMC	MC-TD	0.69	21.92	15.67	3.77	1.01
	MC-ConcTD	0.72	20.80	14.76	3.41	8.01
PM2.5	MC-TD	0.71	36.36	24.56	2.92	16.00
	MC-ConcTD	0.77	31.77	21.61	2.15	5.80

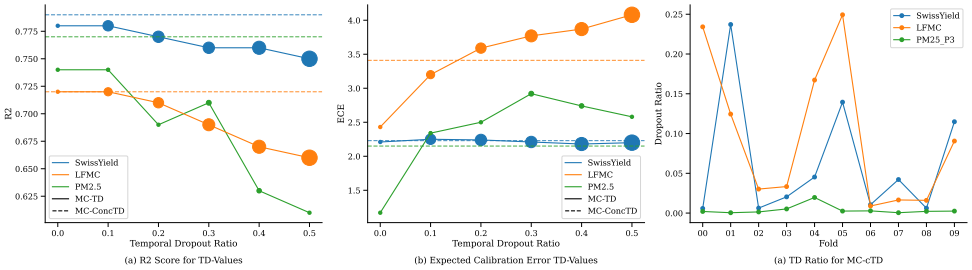


Figure 9.2: Model performance in R^2 and ECE under different TD settings, where marker size indicates the normalized PU.

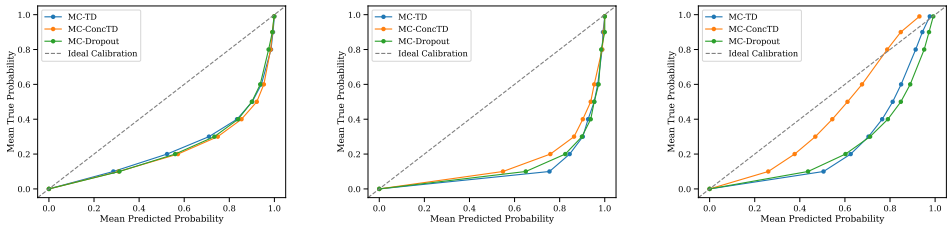
Experimental Setup

Implementation details are provided in Subsubsection A.1.2. We first evaluate MC-TD and MC-cTD against each other, and then compare them to standard uncertainty quantification methods for time series regression tasks, including MC-Dropout [208] and BBB [210]. For a fair comparison, all models use the same architecture. Unless otherwise specified, a TD ratio of 0.3 is used for the MC-TD model to balance regularization and model capacity.

9.1.3 Results

We compare the performance of MC-cTD and MC-TD models across all datasets in Table 9.2. Both approaches demonstrate strong performance across all three datasets. However, MC-cTD consistently outperforms MC-TD on all regression metrics across all datasets. For instance, a maximum improvement of 6 pp in R^2 is observed on the PM2.5 dataset.

In Figure 9.2 (a), we evaluate the performance of various dropout ratios in the MC-TD model. The results indicate that a lower dropout ratio consistently improves performance across all datasets, with a clear decline in model performance as the ratio increases. Additionally, we observe an increase in the PU with increasing TD ratio and decreasing performance. In Figure 9.2 (b), the ECE for every TD ratio is illustrated, together with the constant ECE for

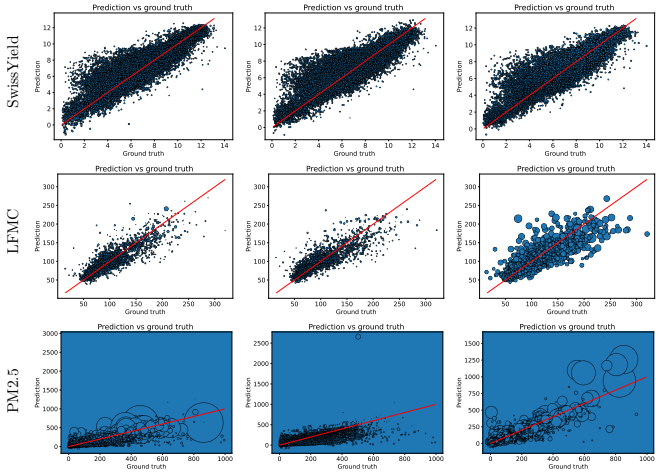


(a) SwissYield (b) LFMC (c) PM2.5

Figure 9.3: Model calibration across confidence levels for all datasets.

the MC-cTD model. Marker sizes indicate the normalized PU. Similarly, we observe an increase in calibration error with increased uncertainty, except for the Yield dataset, and decreased performance. Ultimately, in Figure 9.2 (c), the learned dropout ratios of the MC-cTD model across all folds are illustrated. Notably, the dropout ratios remain below 0.25, yet exhibit high instability across folds and datasets. For instance, the dropout ratio approaches zero for the PM2.5 dataset. The results underline the difficulty of manually tuning the dropout ratio.

Figure 9.3 shows the model calibration. We observe overconfidence in the SwissYield and LFMC datasets. Nevertheless, the MC-cTD model demonstrates relatively better calibration compared to the remaining methods. Conversely, for the PM2.5 dataset, the models exhibit better calibration, especially for the MC-cTD model. Figure 9.4 shows a scatter plot of the predicted and target values for MC-Dropout, the MC-TD, and MC-cTD for all applications. To analyze uncertainty estimates, the predictions are scaled based on the uncerainty. Notably, the PM2.5 dataset exhibits very high uncertainties, while the yield prediction dataset shows low uncertainties. Moreover, we observe that higher uncertainties are associated with lower alignment between predicted and target values, particularly for the MC-cTD model. Finally, we compare the both methods against MC-Dropout and BBB, two common baseline uncertainty quantification methods. The results are summarized in Table 9.3. Notably, the BBB performs poorly on most regression metrics. The MC-cTD model achieves the best R^2 scores on the Yield and PM2.5 datasets, while being equal to the MC-Dropout on the LFMC dataset. Across all evaluation metrics, we find that MC-cTD achieves the best overall results.



(a) MC-Dropout (b) MC-TD (c) MC-ConcTD

Figure 9.4: Prediction over target plots for three regression datasets. The marker size indicates the uncertainty in each prediction that explains the entire coverage in single plots.

Table 9.3: Performance comparison for various uncertainty quantification methods. The best score is highlighted in bold.

Dataset	Model	R^2 (↑)	RMSE (↓)	MAE (↓)	ECE (↓)	PU (↓)
SwissYield	MC-Dropout	0.78	0.94	0.67	2.21	1.02
	MC-TD	0.76	0.99	0.72	2.14	**1.00**
	MC-ConcTD	**0.79**	**0.93**	**0.66**	2.23	1.23
	BBB	0.51	2.02	1.57	**0.73**	23.57
LFMC	MC-Dropout	**0.72**	**20.69**	**14.64**	**2.43**	1.11
	MC-TD	0.69	21.92	15.67	3.77	**1.01**
	MC-ConcTD	**0.72**	20.80	14.76	3.41	8.01
	BBB	0.53	27.00	20.00	4.48	4.01
PM2.5	MC-Dropout	0.74	33.89	22.45	**1.17**	3.99
	MC-TD	0.71	36.36	24.56	2.92	16.00
	MC-ConcTD	**0.77**	**31.77**	**21.61**	2.15	5.80
	BBB	0.30	57.00	39.00	3.57	**2.49**

9.1.4 Discussion

In this chapter, we demonstrated how we can exploit naturally occurring missing time steps for uncertainty estimation and to improve the effectiveness of a model. We showed how we can leverage this prior knowledge to answer RQ3.1 and RQ3.2. We showed that TD at the inference level improves model performance over common uncertainty quantification methods on various EO regression tasks (Table 9.2). However, calibrating the dropout ratio in any DL model is challenging as it requires expensive hyperparameter tuning. This holds for MC-TD. Consequently, MC-TD can only find the optimal

dropout ratio at high computational costs. Interestingly, a higher number of missing time steps consistently results in reduced performance, increased uncertainty, and greater calibration error. We find an optimum value of around 0.1. Notably, the optimal value must balance predictive performance, uncertainty, and calibration error. This underlines the need for effective management of missing instances. In particular, determining the optimal ratio remains challenging, especially when time windows are short (LFMC). Nevertheless, MC-cTD demonstrates flexibility by learning different values based on the validation scenario (fold). This flexibility arises from the model's ability to dynamically adjust the dropout value using the concrete distribution [211]. This results in improved performance and calibration. Nevertheless, we observe high variability across datasets. Moreover, we notice that there may be multiple optimal dropout ratios, as illustrated in Figure 9.2 (c). For instance, for the yield dataset, the learned ratio of the second fold is approximately 0.25, whereas in the sixth fold, it is approximately 0.01. We attribute this variability to the distinct characteristics of each dataset, including varying time series length, number of features, and dataset sizes (Table 9.1). For instance, the LFMC dataset is characterized by only four time steps, which may explain the poor performance of TD-based models on this dataset. In contrast, the PM2.5 dataset has 120 time steps but has only short temporal dependencies compared to the yield dataset. Therefore, dropping the previous time steps can significantly reduce performance. As a result, TD is less effective in this context, potentially explaining the low TD ratios in Figure 9.2 (c). Overall, MC-cTD enhances the robustness of uncertainty estimation across different EO datasets by continuously adapting to unique data characteristics while reducing the computational burden of the hyperparameter search, making uncertainty quantification more accessible for EO applications with time-series data.

Limitations

The proposed methods and experimental setup have limitations that should be considered. We validate the models using EO data without real missing values in the time series, which could potentially differ from scenarios involving actual missing data. This limitation may affect how the models perform in real-world settings with missing values. Additionally, we use only an LSTM encoder to learn the temporal patterns. Moreover, only a few uncertainty quantification methods are used for comparison. However, the primary goal of this research is to demonstrate the effectiveness of TD and MC-cTD in enhancing predictive performance and uncertainty quantification, rather than identifying the optimal architecture for each dataset and use case. Nevertheless, additional models and architectures must be evaluated in the future, including Transformers and other uncertainty quantification methods that have been described in

Chapter 8. Moreover, while TD has shown significant improvements in the related literature [118, 195], we observe only small improvements in this study across individual datasets. This may be attributed to the task and the unique data characteristics, including short time series length (LFMC), short temporal dependencies PM2.5), or noise in remote sensing data. More datasets must be considered to better understand the robustness of the proposed methods. Finally, while we demonstrate improvements in our models regarding the literature, we emphasize that poor calibration remains an ongoing challenge that requires careful evaluation, an issue commonly encountered in DL [246, 198]. Specifically, EO data is often impacted by noisy and uncertain measurements and spatio-temporal distribution shifts, which can lead to poor calibration when the model is applied to unknown environments. Further exploration of model calibration, including post-hoc calibration, will be required before deploying models in practice.

9.1.5 Conclusion

This work underscores the importance of input uncertainty in time series data for EO regression applications. To address this, we introduced two novel uncertainty estimation methods, namely MC-TD and MC-cTD. Both methods account for input uncertainty in time series data by applying TD at the inference level using Monte Carlo sampling. While MC-TD requires manual and expensive tuning of the dropout ratio, MC-cTD learns this automatically as a free parameter. We empirically demonstrate their effectiveness in enhancing predictive performance in various EO datasets. While these models improve the accessibility of uncertainty estimation in EO applications, they also present challenges that require further investigation. Future research will focus on validating these approaches across diverse EO applications and by considering model calibration.

10 BAYESIAN INFERENCE FOR CROP YIELD PREDICTION

Chapter Highlights:

- Uncertainty quantification methods improve performance over purely data-driven baselines in crop yield prediction. DEs and Dropout-based methods are promising approaches for uncertainty estimation. However, selecting the most appropriate is challenging.
- Distribution shifts are a major concern in EO, often leading to complete model failure.
- Incorporating prior knowledge can provide strong inductive biases, resulting in more structured posteriors in function space and more stable predictions on out-of-distribution data.

In this chapter, we present an extensive analysis of uncertainty estimation methods for crop yield prediction. We focus on RQ3.1 and provide a thorough comparison of different uncertainty estimation techniques. Moreover, we address RQ3.3 by evaluating the impact of distribution shifts in real-world scenarios. Finally, we explore how uncertainty estimation, combined with prior knowledge, can help mitigate these challenges.

10.1 Large-Scale Evaluation

Yield estimation has primarily focused on evaluating model accuracy, overlooking uncertainty estimation [73]. However, this offers significant potential when deploying such models into practice. Uncertainty estimation acknowledges the limitations in our understanding of the model and contributes to more trustworthy ML. Therefore, there is a pressing requirement to increase the confidence in ML-based predictions [207, 18]. Although diverse uncertainty quantification methods exist, studies often showcase instability and poor calibration [198]. Exploiting the full potential of uncertainty estimation requires careful comparison of the various existing methods [247].
Only a few studies evaluate uncertainty estimation in crop yield prediction. For

Parts of this chapter, including figures and tables, have been published already in [29].

instance, in [75], a model for corn yield prediction using different EO modalities is proposed, based on a DE method that quantifies the uncertainty of predicted yield values. Similarly, in [108], a model based on MC-Dropout [208] is proposed for regional wheat yield prediction. In [73], a phenology-guided model with uncertainty estimates is proposed, illustrating improved performance in soybean yield prediction. Nevertheless, the systematic evaluation of uncertainty estimation methods remains underexplored.
This section highlights significant improvements in yield prediction enabled by uncertainty estimation. We compare various uncertainty estimation methods against a baseline model. Experiments further demonstrate that incorporating multimodal data enhances prediction accuracy and reduces uncertainty.

10.1.1 Methodology

Data

For all experiments, a subset of the soybean dataset collected in Argentina between 2017 and 2022 is used. This dataset is selected for its high data quality in Argentina. This provides a better understanding of the model behavior. Additionally, training multiple models on the entire dataset is computationally expensive. We integrate S2 imagery as the baseline input and use the monthly time series sampling method, as described in Chapter 4. Additionally, we evaluate the integration of ADMs by using the IF method, as described in Chapter 4.

Architecture, Training & Evaluation

We assess five approaches: Variational Inference using BBB [210], MC-Dropout [208], Concrete Dropout [211], DEs [158], and MC-DropConnect [212]. The model architecture and training are described in Subsubsection A.1.2. At inference, 20 stochastic forward passes are realized for each test sample. Except for the DE, 5 separate models are trained, as recommended in [158]. The output variance is used to calculate the predictive uncertainty, as described in Equation 8.9, and is reported as a standard deviation (t/ha). We report both components, epistemic and aleatoric uncertainty, combined, as we find both being correlated. Models are trained at the pixel level, using stratified, grouped K-fold CV, where pixels are grouped by field and stratified by region. We compare the model performance of different uncertainty estimation methods against a deterministic baseline (LSTM). We use the same architecture as described earlier in Section 4.4.

Table 10.1: Overview of model performance for 10-fold cross-validation for Sentinel-2 and additional modalities. The best score is highlighted in bold.

Evaluation	Sentinel-2				Sentinel-2 + ADM			
	MAE (↓) t/ha	RMSE (↓) t/ha	MAPE (↓) %	R^2 (↑) -	MAE (↓) t/ha	RMSE (↓) t/ha	MAPE (↓) %	R^2 (↑) -
Deep Ensemble	0.39	0.52	0.11	0.75	0.37	0.48	**0.10**	0.79
BBB	0.65	0.81	0.20	0.41	0.61	0.78	0.19	0.44
MC-DropConnect	**0.37**	**0.51**	**0.10**	**0.77**	**0.36**	**0.46**	**0.10**	**0.81**
Concrete Dropout	0.40	0.52	0.11	0.75	0.38	0.49	0.11	0.78
MC-Dropout	**0.37**	**0.51**	**0.10**	0.76	**0.36**	0.47	**0.10**	0.8
Baseline LSTM	0.40	**0.51**	0.11	0.74	0.40	0.52	0.11	0.76

10.1.2 Results

The results are divided into two sections. The first section assesses prediction accuracy, and the second assesses the quality of uncertainty quantification and calibration. Further, we assess the impact of reducing the number of available time steps on the model performance and uncertainty.

Table 10.1 presents quantitative metrics at the field level, where pixel-wise predictions are averaged per field and compared with the ground truth. The table is divided into two cases: i) the models are trained with S2 as input only, ii) the models are trained with S2 and ADMs (all available, see Section 4.4). The best scores are highlighted. We first notice that most probabilistic models outperform the baseline LSTM on all regression metrics. For instance, the MC-DropConnect model achieves an R^2 of 0.77 using only S2 data as input. This marks an increase of 3 pp compared to the baseline model. Only the BBB method performs notably poorly with an R^2 of only 0.41. We notice that including ADMss enhances model performance across all models. For instance, in terms of R^2, Dropconnect shows a 4 pp improvement from 0.77 to 0.81. Additionally, all models, except BBB, either perform similarly or surpass the baseline model. Particularly noteworthy is the significant 5 pp improvement in R^2 achieved by MC-Dropconnect compared to the baseline when ADMss are included. Likewise, the DE method improves by 3 pp in R^2.

In Table 10.2, the uncertainty estimations are quantitatively evaluated. We observe an improvement across models when integrating ADMs, similar to the improvements in predictive accuracy. The models consistently deliver sharper PIs (MPIW). The PICP shows marginal reductions for the BBB and Concrete Dropout model. The PU reduces for MC-Dropconnect and DE, remains constant for MC-Dropout, and increases for BBB. Note that the Concrete Dropout method produces unrealistically high uncertainties (indicated with †). We found that this was the case for very few fields with outlying characteristics compared to the rest of the dataset. In Figure 10.1, we present an example field, showing the ground-truth yield map, the predicted yield map, the error

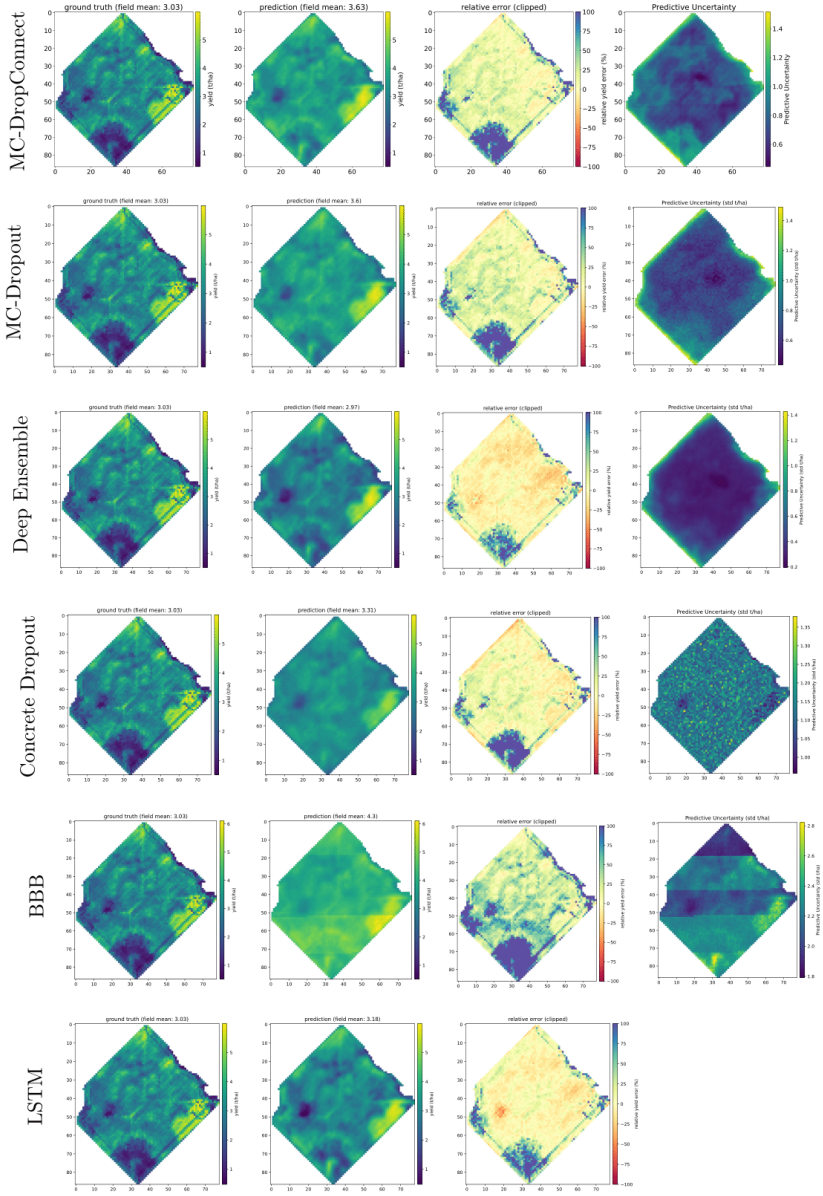


Figure 10.1: Example ground truth yield map and predictions for a field, harvested in 2021. The ground truth yield map is illustrated on the left. Next to it are the pixel-wise predictions, followed by the relative error and the predictive uncertainty (t/ha) maps.

Table 10.2: Uncertainty estimation overview for S2 and ADMs. † indicates a value higher than 50 t/ha

Evaluation	Sentinel-2					Sentinel-2 + ADM				
	MPIW (↓) t/ha	PICP (↑) %	PU (↓) t/ha	NLL (↓) -	ECE (↓) -	MPIW (↓) t/ha	PICP (↑) %	PU (↓) t/ha	NLL (↓) -	ECE (↓) -
Deep Ensemble	**2.63**	0.83	0.93	1.56	0.28	2.60	0.84	0.92	1.64	0.37
BBB	4.44	1.00	1.62	1.35	0.58	4.19	**0.92**	1.76	1.34	0.79
MC-DropConnect	2.67	0.86	**0.92**	1.11	0.17	**2.57**	0.87	**0.83**	1.05	**0.09**
Concrete Dropout	†	**0.91**	†	1.03	**0.15**	†	0.89	19.02	1.08	0.11
MC-Dropout	2.95	0.89	0.94	**0.98**	0.26	2.92	0.90	0.94	**0.79**	0.79

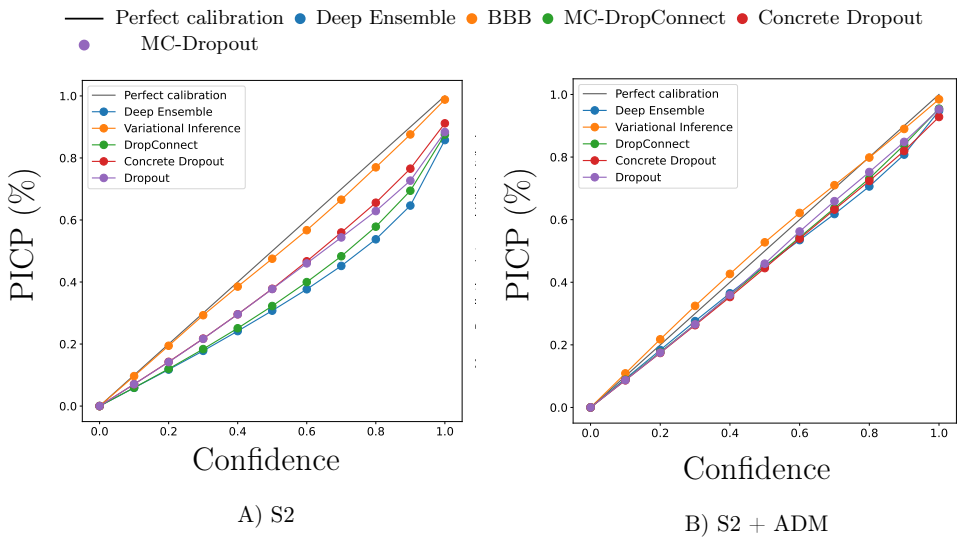


Figure 10.2: Calibration evaluation overview. Left: reliability plot (PICP) at different confidence levels for S2. Right: reliability plot (PICP) at different confidence levels for S2 and ADMs.

map, and the predictive uncertainty (t/ha) for each pixel. The predictions are generated for each model trained on S2 imagery. All models accurately predict the field mean and capture infield variability. Examining the error map reveals higher errors at boundaries and in low-yield areas. Notably, inaccuracies at the boundaries are common, often due to noisy labels in these regions. Notably, the BBB and Concrete Dropout methods demonstrate lower infield variability and higher pixel-wise errors.

We emphasize that the MC-Dropout, MC-DropConnect, DE, and BBB models reveal a clear correlation with both the ground truth data and the error map. Notably, higher uncertainty is observed at the border level, underscoring the greater difficulty of modeling these regions. Additionally, areas with lower yields tend to exhibit higher uncertainties. Similarly, the uncertainty values demonstrate a strong correlation with the model error, underscoring the utility of the uncertainty estimates as indicators of predictive confidence. Nevertheless, the Concrete Dropout model demonstrates a poor correlation between model error and uncertainty estimates. Additionally, the BBB method shows artifacts in the prediction that we refer to as deviating sampling distributions during inference. Note that the deterministic LSTM provides no estimations of uncertainty.

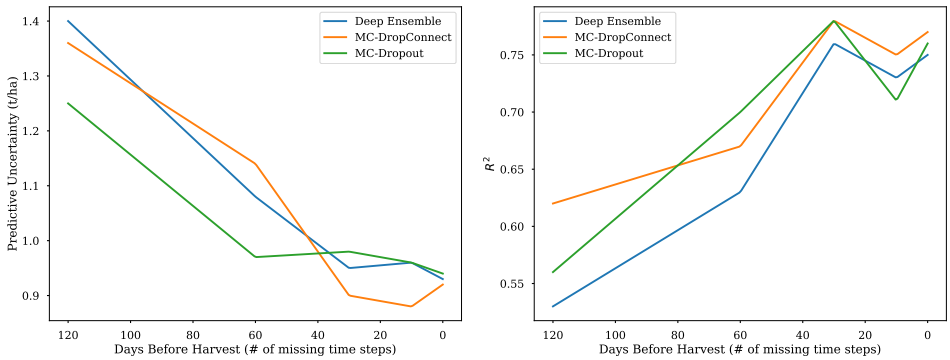


Figure 10.3: Forecasting error and uncertainty at various time steps before harvest. Left: RMSE at each time frame. Right: Predictive uncertainty at each time frame.

Calibration

Figure 10.2 shows the model calibration for a specific region in Argentina across both evaluation scenarios. The left plot shows the PICP at various confidence levels using only S2 as input. A perfectly calibrated model would follow the diagonal line. The BBB model demonstrates nearly perfect calibration across all confidence levels, whereas MC-DropConnect, DE, and MC-Dropout appear overconfident. The right plot presents the same experiments for models trained with both S2 and ADMs as inputs. We observe a substantial improvement when integrating multimodal data: all models that were previously overconfident now exhibit significantly improved calibration.

Error & Uncertainty

To analyze the link between model error and uncertainty, we conduct forecasting experiments by training models with a truncated time series. More specifically, models are assessed by progressively reducing the number of available time steps up to 120 days before the harvest. Specifically, making the task more challenging by masking subsequent time steps and reducing the available training data. Figure 10.3 shows the R^2 for each scenario, on the right plot. We focus the analysis on the MC-Dropout, MC-DropConnect, and DE models, since these methods have consistently performed well in earlier experiments. For all models, the error increases as the number of time steps decreases. We further illustrate predictive uncertainty for each forecasting scenario in the left plot. We stress that the uncertainty left plot correlates well with the error curves, confirming our previous observation. More specifically, the predictive uncertainty reduces as more input time steps are available. The

correlation between uncertainty and accuracy is confirmed through an analysis of uncertainty and different error metrics. The MC-Dropconnect model achieves the highest correlation, with an estimated correlatoin of 0.98 between R^2 and uncertainty. Nevertheless, the remaining models exhibit similarly good scores.

10.1.3 Conclusion

In conclusion, this section highlights the importance of uncertainty quantification in yield prediction. Nevertheless, selecting an uncertainty estimation method remains a challenge. We compared various uncertainty estimation methods and found that, while BBB offers a strong theoretical foundation, it performs poorly and incurs higher computational costs in practice. Similarly, while Concrete Dropout offers advantages over its predecessor MC-Dropout, its uncertainty estimations are unstable and unreliable. In contrast, to answer RQ3.1, including MC-DropConnect, MC-Dropout, and DE collectively exhibit enhanced yield prediction accuracy even compared to strong deterministic baselines.

10.2 Deep Ensembles & Distribution Shifts

In the previous section, we evaluated different methods for estimating uncertainty. Here, we focus on distribution shifts, a common phenomenon in EO. We address RQ3.3 and discuss potential mitigation strategies that leverage both uncertainty estimation and prior knowledge.

> **Distribution Shifts**
> *Distribution shifts refer to a change in the properties of the testing data compared to the training data.*

ML models are highly susceptible to distribution shifts and exhibit overconfidence and degraded performance when exposed to such data [248, 249]. One reason for the lack of generalizability to distribution shifts is the scarcity of dedicated training samples [16]. For instance, crop yields are often affected by distribution shifts between years and regions, rooted in different management practices, environmental conditions, and yearly fluctuating climate conditions. Climate variability accounts for approximately 30 % of the global yield variability [250]. Consequently, generalization to unknown years and regions that exhibit a different data distribution causes a severe degradation in model

performance. For example, Perich et al. [40], noted that a model was not able to accurately predict crop yields for unseen years. Only Mena et al. [125] showed that transferring to unknown years can be improved over the standard IF by adaptive fusion of multimodal EO data.

In this section, we investigate the model performance under distribution shifts. We show that the cause for poor performance lies in the distance weight space of the NN that hinders the generalization to unknown data. To bridge this gap, we show that DEs can improve the poor generalization capacity of modern NNs. We argue that this lies in the higher diversity in weight space, which explores multiple *modes* and thus reduces the generalization gap. A *mode* refers to the different local optima in the solution landscape, a NN can converge to during training [251]. Each mode produces a function that defines a unique mapping between the input and the output. Additionally, we show that the inclusion of prior knowledge further improves the generalization capacity by adding additional inductive bias that anchors the posterior predictive distribution in closer locations in weight space.

Deep Ensembles and modes in weight space

10.2.1 Distribution Shifts in Training Data

In Subsection 2.2.1, we have already shown that the YieldSAT dataset is characterized by significantly different data distributions between crops and countries. Likewise, we observe differences between years and regions for a single country and crop. To assess the severity of distribution shifts across regions and years, the input data and target yield data are first analyzed for these shifts. In Figure 10.4, we visualize a low-dimensional embedding using t-SNE of the surface reflectance of the S2 input. This includes the time series of S2 imagery and ADMs. The data are colored by year (left) and region (right). We highlight that in both cases, a clear separation between individual years and regions is visible. More specifically, the data form clusters within each group. Moreover, subclusters are visible. This indicates unique characteristics within a year-region combination. Below is the kernel density estimation plot of the target yield distribution. The data is grouped by year (left) and region (right). Note that each group's distribution is unique. This is evidenced by the Kruskal-Wallis test, which compares the distributions of yield data between regions and years. The results are depicted in Table 10.3. For both tests, a significantly high p-value is reported ($p < 0.0001$). This means that at least one year and one region have a significantly different yield distribution compared to the rest of the groups. For all experiments, the same dataset is used as described in Section 10.1. To compare each pair of yield

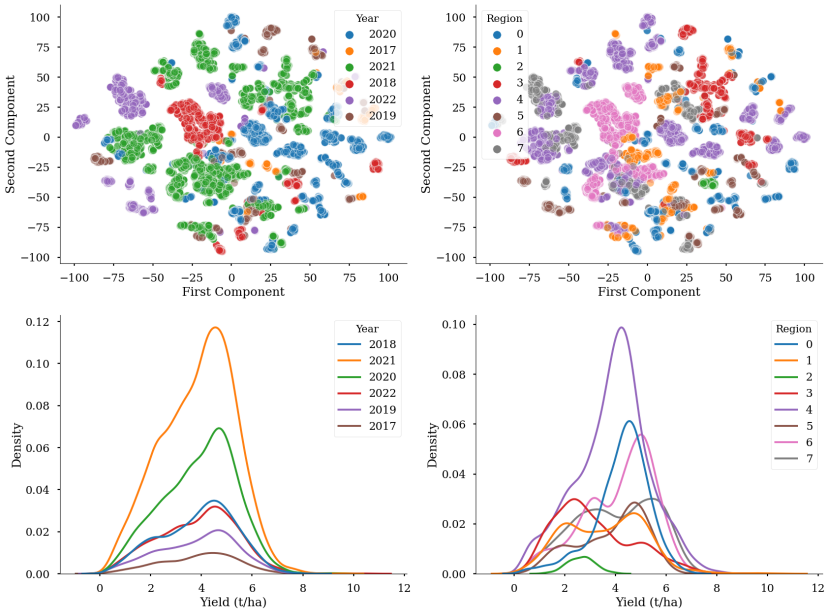


Figure 10.4: Visualization of the training data distribution grouped by years and regions. Top left: t-SNE plot of the input data (S2 and ADMs) grouped by **years**. Top right: Top left: t-SNE plot of the input data (S2 and ADMs) grouped by **region**. Bottom left: Kernel density estimation plot for the target yield data grouped by **years**. Bottom right: Kernel density estimation plot for the target yield data grouped by **regions**.

Table 10.3: Kruskal–Wallis H-test between yield distributions grouped by years and by regions.

Evaluation	Year	Region
p-value	p<0.0001 ****	p<0.0001 ****

distributions, a post hoc comparison of the distributions using Dunn's test is performed. The pairwise comparison between the years is depicted in Table 10.4. We highlight that most pairwise comparisons indicate significantly different distributions. Only between 2018 and 2019, no significance is found. The pairwise comparison between the regions is depicted in Table 10.5. Similar to the years, the data distribution between regions is mostly significantly different. However, individual regions do not show significantly different distributions. For instance, region 5 shows no significant difference between regions 4, 6, and 7. In conclusion, the results undermine the assumption that the distribution across years and regions is mostly significantly different, which increases the difficulty of generalization.

Table 10.4: Pairwise post-hoc comparisons of the yield distributions between individual **years**. Each cell displays the statistical significance level of the difference between two years based on the Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Year	2017	2018	2019	2020	2021	2022
2017	-	****	***	****	****	****
2018	****	-	ns	***	****	****
2019	***	ns	-	**	****	****
2020	****	***	**	-	****	****
2021	****	****	****	****	-	****
2022	****	****	****	****	****	-

Table 10.5: Pairwise post-hoc comparisons of the yield distributions between individual **regions**. Each cell displays the statistical significance level of the difference between two regions based on the Dunn's test using the Holm–Bonferroni correction. ns = no significance ($p \geq 0.05$), $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$, $**** = p < 0.0001$.

Region	0	1	2	3	4	5	6	7
0	-	****	****	****	****	****	**	****
1	****	-	****	****	****	****	****	****
2	****	****	-	***	****	****	****	****
3	****	****	***	-	****	****	****	****
4	****	****	****	****	-	ns	****	ns
5	****	****	****	****	ns	-	ns	ns
6	**	****	****	****	****	ns	-	***
7	****	****	****	****	ns	ns	***	-

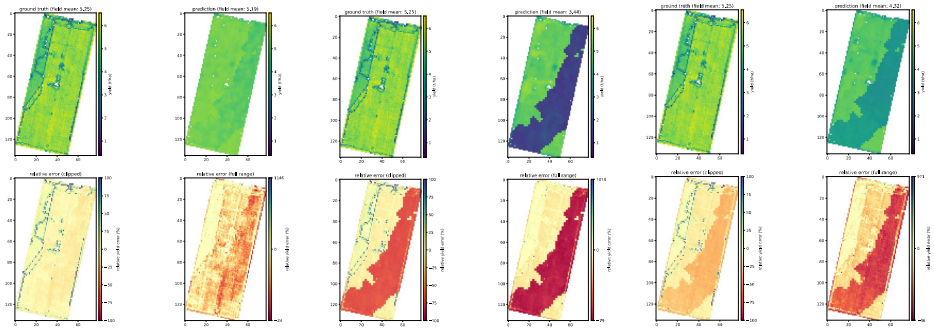
10.2.2 Learning Under Distribution Shift

For all experiments, the same LSTM architecture and training are used as described in Section 10.1. However, to analyze the impact of shifted data, only the DE model is compared against the baseline LSTM (see Section 4.4). Further, we assess the impact of different network architectures, including the integration of local-neighborhood (3D-LSTM) information, as described in Section 4.5. We restrict the analysis to DEs, since they are becoming the gold standard for uncertainty quantification under distribution shifts [213]. Moreover, they represent uncertainty via multiple independent parameter sets that simplify the analysis posterior distribution. In contrast, the other BNN approaches represent uncertainty via stochastic sampling or distributions over weights. This makes direct weight space comparisons ill-posed. For more information see see Chapter 8. To assess the performance under distribution shift, each model is evaluated on unseen years and regions, using the LOYO and LORO CV. For this, a single year or region is held out during training and used only for evaluation. These scenarios assess the generalization capacity in

Deep Ensemble

Table 10.6: Overview for crop yield prediction at the field level using temporal (Leave-One-Year-Out) and spatial splitting (Leave-One-Region-Out). All models, except the baseline LSTM, are defined as a DE with 5 members.

Evaluation	Leave-One-Year-Out		Leave-One-Region-Out	
	RMSE (↓) t/ha	R^2 (↑) -	RMSE (↓) t/ha	R^2 (↑) -
LSTM (DE)	0.70	0.55	0.65	0.62
3D-LSTM (DE)	**0.23**	**0.63**	**0.19**	**0.73**
LSTM (Baseline)	0.72	0.53	0.74	0.50



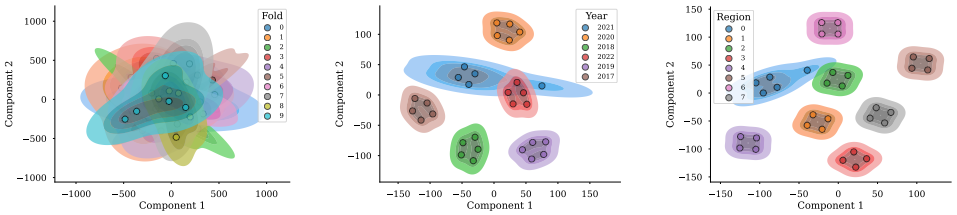
(a) Standard 10-Fold CV (b) LOYO CV (c) LORO CV

Figure 10.5: Example predicted field for different experimental setups to assess the impact of distribution shifts. Left: standard CV, center: LOYO CV, right: LORO CV. For each experiment, the ground truth yield map, predicted yield map are shown. Below the relative error (clipped and full range).

real-life settings. We define a region as the data points that belong to a single farmer.

Performance under Distribution Shift

The quantitative results are given in Table 10.6 and confirm that the model performance significantly degrades under distributional shift, especially for the baseline LSTM. For instance, for the LOYO experiment, an overall reduction in R^2 of 21 pp is reported, compared to the standard CV experiments, where an R^2 of 0.74 was reported (see Table 10.1). Similarly, for the LORO experiment, a significant reduction of 24 pp in R^2 is shown. Nevertheless, the DE methods exhibit improved performance compared to the baseline model in both settings, as shown in Table 10.6. In the LOYO setting, the DE improves by 2 pp R^2 over the baseline model. Likewise, the DE model exhibits a 12 pp increase over the baseline in the LORO experiment. This suggests that the uncertainty quantification methods still exhibit significantly higher generalization capabili-



(a) 10-Fold CV (b) LOYO (c) LORO

Figure 10.6: t-SNE of the network parameters for the DE model for different CV scenarios. Left: standard 10-CV, center: LOYO, right: LORO. Note that LOYO and LORO are training scenarios under data shift. The weights are colored by fold.

ties on shifted data. Interestingly, adding spatial context with the 3D-LSTM further widens the gap between the baseline and the DE approach. Especially in the LORO scenario, an improvement of 23 pp is achieved, resulting in an overall R^2 of 0.73. This score is almost equal to the R^2 of the baseline model in the standard CV scenario (R^2 of 0.74). Additionally, in the LOYO setting, an improvement of 10 pp in R^2 is shown compared to the baseline model. Similarly, both models show a significant reduction in RMSE.
Figure 10.5 shows a qualitative example of a predicted field under distribution shift. Compared to the standard CV, the model exhibits regions of complete performance collapse under distribution shift, as evidenced by high pixel-wise error.

10.2.3 Weight Space Diversity and Distribution Shift

To investigate the reason for the degrading performance and the difference between probabilistic and deterministic models, we analyze the model parameters in Figure 10.6. The plot shows a low-dimensional t-SNE embedding of the trained DE model weights. The weights are displayed for the standard 10-fold CV (left), LOYO (center), and LORO (right) scenario. Additionally, the weights are colored by folds. For example, in the LOYO scenario, a model is colored by the year in the validation set. The plot reveals interesting insights. While the weight distribution of the ensemble members overlaps entirely in the standard 10-fold CV, a clear separation is observed for the LOYO and LORO scenario. Each fold forms a distinct cluster in weight space with essentially no overlapping.
We conclude two main things from this. First, we argue that this may explain the poor performance under distribution shift since the model is less capable

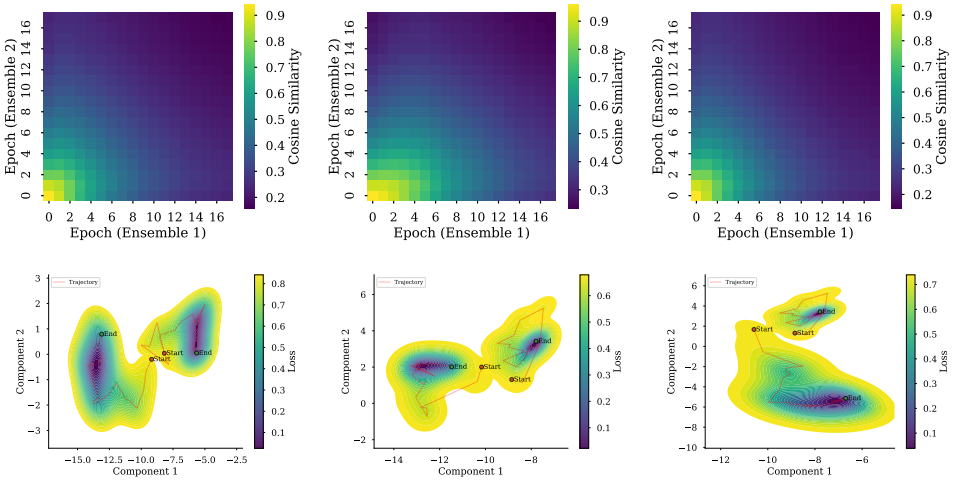


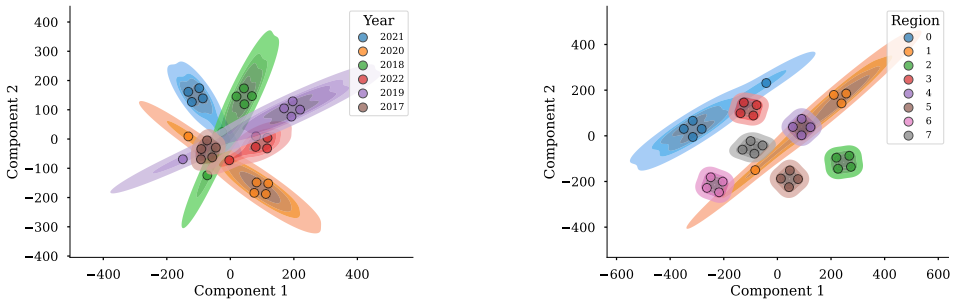
Figure 10.7: Visualization of the diversity of the weight space during model training. Top: Cosine similarity between pairs of ensemble members during training. Bottom: t-SNE plot of the weight space during training, together with the loss. The trajectory in the weight space is highlighted in red from start to end.

of generalizing to unknown data distributions. Second, DEs explores multiple modes in weight space, which explains its better performance compared to the deterministic baseline, which explores only a single mode.

To investigate the diversity in weight space, the model weights during training are visualized in Figure 10.7. The top row shows the cosine similarity between two ensemble members over the training epochs. Each comparison shows that the similarity in weight space is high at the start of the training and decreases throughout the training. At the end of the training, the weight space is clearly separated between the two ensemble members. The lower plot shows the trajectories in weight space for the two ensemble members, along with the validation loss. The plot illustrates that the ensemble members are initialized randomly in nearby regions. Additionally, the plot illustrates that each ensemble member explores unique modes in weight space, characterized by low loss values. This indicates that each ensemble member explores several optimal solutions.

10.2.4 Improving Generalization with Prior Knowledge

Although we have seen that DEs offer a significant advantage over the deterministic baseline model by exploring different modes, we still observe a decrease in performance under distribution shift. To mitigate this, we included



(a) LOYO

(b) LOFO

Figure 10.8: t-SNE of the network parameters for the DE model trained with injected prior knowledge (TD, Spatial) under distribution shift. Left: LOYO, right: LORO.

spatial context using the 3D-LSTM model. The resulting weight spaces for both scenarios are depicted in Figure 10.8. The figure shows that adding prior knowledge reduces the distance and anchors solutions in nearby regions. Consequently, the generalization performance under distribution shift increases significantly, as evidenced in Table 10.6. Still, the elongated clusters indicate remaining epistemic uncertainty, but it is structured, aligned along dimensions that matter under the inductive bias.

10.3 Discussion

In this section, the difficulty of crop yield prediction under distribution shift was analyzed to adress RQ3.3. We simulated a real-world scenario of transferring across unknown years and regions and observed severe performance collapse. We argue that the decrease in performance is mainly due to significantly different data distributions across years and regions. To mitigate this, DEs were analyzed which showed significantly improved performance compared to the deterministic baseline. DEs are known for being more robust to distribution shifts than single-mode methods [252]. Following Fort et al. [214], we argue that the improved performance lies in the diversity of the weight space, which produces unique functions of the NN. DEs explore multiple modes that are characterized by low-loss landscapes. In contrast, deterministic baselines explore only a single mode, which reduce their generalizability. Nevertheless, we argue that this is not sufficient to bridge the generalization gap to shifted data. We still observe severe performance decrease when applying a NN to unknown years or regions. Consequently, this is a critical research question

before deploying NNs into practice. Helber et al. [121] argued that additional transfer learning methods are required.
In contrast, we argue that selecting an appropriate prior can mitigate this issue. The distance between solutions in function space, as obtained by independently trained ensemble members, reflects the degree to which a model's prior constrains the hypothesis space. In the absence of a strong prior, distribution shifts induce large divergences between ensemble members, indicating unstable and inconsistent generalization to shifted data. In contrast, well-chosen priors can constrain the posterior to a more compact, structured region in weight space. As illustrated, this can reduce inter-model distance and lead to more stable solutions and improved generalization under shifted conditions. However, as Izmailov et al. [253] pointed out, that we commonly think of a prior as the distribution over parameters, as elaborated in Chapter 8 for BNNs. In contrast, the authors argue that what is important is the prior over functions induced by a vague prior over parameters, coupled with the architecture of the NN, which incorporates sufficient prior knowledge to converge to a good solution. This alone can provide a strong inductive bias. Or differently, the support of a NN should be large while the inductive bias must be specific for a given problem [213]. We showed that by changing the architecture to account for spatial dependencies (see Section 4.5), we can anchor the function space across a common solution. Consequently, we observe a significant increase in the performance under data shift by simply changing the functional form of the NN.

Limitations

The presented study has limitations that must be considered. First, we only investigated this phenomenon on a single dataset. This is because training DEs is expensive, especially in an already expensive CV training setting. In the future, more datasets will be evaluated. Additionally, cheaper DE implementations must be developed to reduce the computational complexity. Finally, we have only investigated single prior components that improve the inductive bias of the model. In the future, additional functional forms of the model architecture must be investigated that support this work.

10.4 Conclusion

Distribution shifts remain a challenge in crop yield prediction. To answer RQ3.3. we showed that DEs offers a potential solution. Coupled with a strong prior, the generalization gap can be reduced.

PART V

CONCLUSION

11 | SUMMARY

In this thesis, the integration of prior knowledge into Machine Learning (ML) was studied, defined as Domain-Informed Learning (DIL). We investigated different types, representations, and integration practices of prior knowledge into the ML pipeline. We proposed novel approaches that address ongoing challenges in ML research for EO tasks. Specifically, we studied (1) the integration of prior knowledge to enrich the data space (PART II), (2) the integration of prior knowledge to enforce knowledge conformity (PART III), and (3) the integration of prior knowledge into the model hypothesis for uncertainty estimation (PART IV). This thesis focused on agricultural applications, including crop yield prediction, plant phenotyping, and crop stress modeling. Additionally, individual methods were studied for further EO applications, including air pollution forecasting, and the analysis of wildfire risk estimation was conducted.

11.1 Key Contributions

Each research question was addressed by various experiments and discussed in light of the current literature. We quantified the success of each contribution and research question, using the evaluation criteria: (1) superior performance, (2) increased explainability, (3) increased trustworthiness, and (4) novel per-

Table 11.1: Summary of the thesis contributions with regard to the evaluation criteria. ✓ indicates criteria fulfilled, (✓) indicates partial fulfillment, ✗ indicates no fulfilment.

Part	Chapter	Short Description	Superior Performance	More Explainable	More Trustworthy	Novel Perspectives
PART I	Section 2.2	*YieldSAT* Dataset	✗	✗	✗	✓
PART II	Chapter 4	Multimodal Crop Yield Prediction	✓	(✓)	✗	(✓)
	Chapter 5	Time Series Analysis	(✓)	✓	✗	(✓)
PART III	Chapter 6	Physics-Guided Drought Stress Estimation	✓	✓	✓	✓
	Chapter 7	Controlled Crop Growth Modeling	✓	✓	(✓)	✓
	Chapter 7	Time Series Regression with Diffusion	(✓)	✗	(✓)	✗
PART IV	Chapter 9	Uncertainty Estimation with MC-Temporal Dropout	✓	✗	✓	✗
	Chapter 10	Bayesian Inference & Distribution Shifts	✓	✓	✓	✓

spectives. The evaluation criteria are partially oriented by the evaluation catalogue proposed by von Rueden et al. [254] for informed learning algorithms. In Table 11.1, we summarize the contributions of each chapter with respect to the evaluation criteria. In summary, every chapter provides considerable contributions and novelties to the field of domain-informed ML for EO applications. This thesis makes notable contributions to large-scale crop yield prediction at the field and subfield levels. In Section 2.2, a novel large-scale satellite benchmark dataset for crop yield prediction was introduced, referred to as *YieldSAT*. This chapter provides *novel perspectives* for future research in crop yield prediction. In the past, yield prediction was mainly driven by process-based crop models or localized ML approaches focused on individual crop types, countries, and years. Moreover, most models were refined to regional yield prediction by relying on reported statistics. In contrast, this thesis presents a significant contribution by bringing yield prediction to the subfield level while maintaining a large scale. *YieldSAT* marks a significant step in the current literature and is expected to contribute to fruitful future research.

Novel Dataset

PART II studied the enrichment of the data space for crop yield prediction to answer RQ1. In Chapter 4, we developed baseline models for large-scale crop yield prediction with ML. We extensively studied various input modalities and data fusion methods and presented two fusion methods: (1) a simple input fusion (Section 4.4) and (2) an advanced attention-based feature fusion method (Section 4.5). Further, we showed that Copernicus Sentinel-2 (S2) satellite data alone is highly important for yield prediction and requires only minimal preprocessing. Nevertheless, we showed the complementary importance of auxiliary data sources. Moreover, it was demonstrated that including spatial context significantly improves the model performance. Surprisingly, we experimentally demonstrated that highly enriched features (e.g., Vegetation Index (VI) or simulation results) do not significantly improve model performance. Instead, enriching the data space with such data can even result in decreased performance and unnecessarily high implementation efforts.

Extensive analysis of input modalities and models

This chapter mainly contributed to *superior performance* in crop yield prediction. Nevertheless, we also partially improved the *explainability* by including an attention mechanism. Finally, this chapter provided *novel perspectives* for crop yield modeling with ML in general.

In Chapter 5, we explored the importance of time series representations for crop yield prediction. A novel method was introduced that aligns the time series representation with the physiological processes of the crop's growing cycle. We showed that this method *improved performance* in terms of computational

Novel time series sampling methods

efficiency while maintaining high accuracy compared to baseline approaches. Moreover, *improved explainability* was demonstrated by leveraging the intrinsic explainability of the self-attention mechanism of the Transformer model [98].

In PART III, we studied the integration of prior knowledge as a condition to answer RQ2. We emphasized that integrating prior knowledge into the data space can significantly enhance the model performance. However, models often remain a 'black-box' and may fail to capture important physical processes (e.g., natural laws). We highlighted that constraining ML models with physical principles is essential to *increase trust* in ML methods. This part further emphasized that most biological processes are subject to uncertainty and that their outputs commonly differ due to random variation. Although the methods presented in this part did not explicitly model different types of uncertainty, we still introduced variability through different methods that partially contributed to increased *trustworthiness*.

Output variability

In Chapter 6, we presented a novel method for crop stress forecasting by coupling ML and simulation models to satisfy important physical principles of plant growth. Moreover, a novel, Physics-Guided (PG) loss term was introduced. We demonstrated that crop yields can be modeled as a function of temporal drought stress and showcased significant improvements in *performance* and *explainability*. This chapter further contributes to *increased trustworthiness* through physical consistency and the inclusion of uncertainty. Moreover, physics-guided learning provides *novel perspectives* for the field of EO.

Physics-Guided Learning

In Chapter 7, we showed the potential of Generative Adversarial Networks (GANs) for crop phenotyping and plant growth modeling using image-to-image translation. We showcased that the prediction of the future plant appearance can be controlled using a conditional, low-dimensional latent vector that encodes physical properties of the future. This method *improved performance* over baseline models and showed *increased explainability and trustworthiness* through an explainable latent space. Further, we provided *novel perspectives* by framing crop growth modeling as an image-generation task. Section 7.2 presented a conditional Diffusion Model (DM) for time series regression using a prior conditional network. This model showed the potential of including conditional prior for DMs by demonstrating *superior performance* over baseline models.

Controlled one-to-many mapping

Conditional Diffusion Models

PART IV quantified the unknown to answer RQ3. We explicitly modeled different types of uncertainty using Bayesian inference for crop yield prediction, air pollution monitoring, and wildfire risk assessment. Uncertainty estimation contributes to *increased trustworthiness*, since it tells the user which predictions to trust and which not. In Chapter 9, we presented a novel uncertainty estimation method that leverages naturally occurring missing time steps, defined as *Monte-Carlo Temporal Dropout (MC-TD)*. We demonstrated *improved performance* and greater robustness to missing time steps than existing uncertainty estimation methods. In Chapter 10, we performed an extensive analysis of Bayesian inference for crop yield prediction. We assessed performance, model calibration, and uncertainty quality, and demonstrated *improved performance* compared to deterministic baseline models. Finally, we investigated the severity of distribution shifts and showcased a significant collapse in model performance. We demonstrated that Deep Ensemble (DE) methods, when coupled with prior knowledge, can mitigate the negative impact of distribution shifts. This was due to: 1) the exploration of multiple modes in weight space, which reduces the risk of overconfident predictions, and 2) the prior knowledge reduces the generalization gap. This section contributed to *improved performance and explainability*. This chapter further provides *novel perspectives* for *Domain-Informed Learning* by explicitly including Bayesian principles.

Missing data for uncertainty estimation

Uncertainty estimation & distribution shifts in yield prediction Deep Ensembles

11.2 Future Perspectives

We see several future directions for *Domain-Informed Learning* for EO applications.

11.2.1 Hybrid Models with Bayesian Inference

We have demonstrated that prior knowledge is represented and injected in various ways, making the assessment and evaluation of prior knowledge difficult. We lack a principled understanding of how prior knowledge interacts within a model and influences outcomes. More research is required that systematically quantifies the impact of injecting prior knowledge into the ML pipeline. Existing approaches, such as in [254], must be extended by considering Bayesian inference. From a Bayesian perspective, prior knowledge is the belief about the hypothesis, which is updated after observing data. We believe that this provides substance for a more general framework of learning with prior knowledge while further accounting for the model uncertainty. Izmailov et al. [253]

already argued that what matters is the prior over functions that is induced by combining the prior over the parameters with the functional form of the model. Likewise, the injection of prior knowledge certainly has a major effect on the performance that may be estimated in the induced function space.

11.2.2 Hybrid Foundation Models

Most ML approaches are tailored to a single downstream application. However, recently, Foundation Models (FMs) emerged as a highly promising research direction that addresses several EO applications within a single model. FMs can process various data sources and can be fine-tuned to many downstream applications within a single model [255, 256]. Nevertheless, FMs are still in their early stages and require significant research. Specifically, the integration of uncertainty quantification and physical consistency is commonly overlooked in the current research. However, this is a highly desirable aspect [22] and fundamental to increasing explainability and trust and ultimately to more acceptance of ML among users.

11.2.3 Innovation Transfer with Open Science and Open Data

Although huge amounts of EO data are collected every second, openly accessible, labeled, and challenging datasets are still lacking, which hinders research and ultimately the transition of ML into industry innovation. In this thesis, we discussed the example of crop yield prediction. However, many other EO applications suffer from limited data, especially in regression tasks [237]. There is still a barrier to sharing data due to high acquisition costs, inconsistent data quality, skepticism about sharing data, and data privacy concerns. Overcoming this data shortage is crucial to ensuring the development of ML algorithms and the transfer of innovation. We believe that a more open mentality towards open science and open data will largely contribute to a fruitful future development.

11.3 Final Remarks

Domain-Informed Learning is highly promising to address social and economic challenges. We hope this work contributes, in its own small way, to scientific development.

PART VI

APPENDIX

A FURTHER INFORMATION

A.1 Models

A.1.1 Common Neural Network Architectures

Most modern DL architectures consist of the basic building blocks described in Chapter 3 and offer different inductive biases. Therefore, the building blocks are combined into architectures that serve a common task. For example, we can combine the spatial inductive bias from convolutional layers with the temporal inductive bias from recurrent layers. The resulting *convolutional LSTM* model processes spatio-temporal patterns [194]. We briefly discuss the most important architectures that are necessary to know, to follow the content of this work.

Long-Short Term Memory

Recurrent layers in RNNs [257] assume temporal locality, which can hinder the propagation of information over long sequences. As a result, signals from early time steps may not effectively influence later computations, leading to the *vanishing gradient problem* [258], particularly in deep or long networks. Consequently, the model struggles to extract information from distant entities within the sequence, leading to degraded performance. To address this limitation, Hochreiter and Schmidhuber [259] introduced the LSTM architecture, which enables information to be stored and accessed over longer temporal sequences. The LSTM uses a memory cell that maintains relevant information across time steps, and a set of gates that regulate information flow: (1) the input gate, (2) the forget gate, and (3) the output gate. The forget gate explicitly controls which information is discarded, thereby allowing the network to retain only the most relevant temporal dependencies.

vanishing gradient problem

Transformer

The Transformer model is another architecture designed to process sequential information. It was introduced by Vaswani et al. [98] and is based entirely on the attention mechanism [260], applied over time. Unlike RNNs, the

Transformer models dependencies without regard to the distance between elements in the input or output sequences. The attention mechanism enables the network to capture long-range dependencies by attending to all previously observed or generated positions in the sequence. Interestingly, the Transformer architecture relies solely on fully connected (MLP) layers, without employing convolutional or recurrent operations. Each attention layer computes three representations for every input token: the *Query (Q)*, the *Key (K)*, and the *Value (V)*. Conceptually, the query can be seen as a search request, the keys as index entries, and the values as the retrieved content. The attention operation computes similarity scores between queries and keys to determine how strongly each value should contribute to the output representation. Typically, multiple attention mechanisms are combined in the multi-head attention module, which allows the model to attend to information from different representational subspaces. A single attention head may not capture all aspects of complex dependencies. By contrast, multi-head attention employs several independent heads, each learning its own set of queries, keys, and values, thereby enhancing the model's capacity to represent diverse relationships within the sequence.

Conditional Generative Adversarial Network

cGANs [168] are an established method for image-to-image translation following an adversarial learning philosophy. The Pix2Pix architecture [172] is formalized as:

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(\mathcal{G}_\theta, \mathcal{D}_\delta) = \mathbb{E}_{\mathcal{X},\mathcal{Y}}[\log \mathcal{D}_\delta(X, Y)] \\ + \mathbb{E}_{\mathcal{X},\mathcal{Z}}[\log(1 - \mathcal{D}_\delta(X, \mathcal{G}_\theta(X, z)))],\end{aligned} \tag{A.1}$$

Further, an $L1$ loss is included to force the generator not only to fool the discriminator but also to produce samples which are close to the target image:

$$\mathcal{L}_{L1}(\mathcal{G}_\theta) = \mathbb{E}_{\mathcal{X},\mathcal{Y},\mathcal{Z}}[||Y - \mathcal{G}_\theta(X, \epsilon)||_1]. \tag{A.2}$$

This leads to the final objective described by:

$$\mathcal{G}_\theta^*, \mathcal{D}_\delta^* = \arg\min_{\mathcal{G}_\theta} \max_{\mathcal{D}_\delta} \mathcal{L}_{\text{cGAN}}(\mathcal{G}_\theta, \mathcal{D}_\delta) + \lambda_I \mathcal{L}_{L1}(\mathcal{G}_\theta), \tag{A.3}$$

where λ_I is a hyperparameter used to control the weighting of the $L1$ loss.

Diffusion Models (DM)

The training of DMs involves two key processes: a forward process q and a reverse process p, both of which are modeled as Markov chains. In the forward

process, also called the noising process, an initial data distribution $X_0 \sim q(\mathcal{X}_0)$ is progressively corrupted with noise over K time steps. In this context, the time step refers to the steps that are required to corrupt a single image with noise, denoted as K. Similarly, X_0 refers only to the image instance that is not corrupted by noise. By the final step, the data is transformed into a standard Gaussian distribution $X_K \sim \mathcal{N}(0, \boldsymbol{I})$. At each time step k, the process depends solely on the previous step and is given by:

$$q(X_k|X_{k-1}) := \mathcal{N}(X_k; \sqrt{1-\beta_k}X_{k-1}, \beta_k \boldsymbol{I}). \tag{A.4}$$

Here, β_k is the variance schedule, defining the amount of added noise at step k. Doing this over K steps is given by:

$$X_k \sim q(X_k|X_0) = \mathcal{N}(X_k; \sqrt{\bar{\alpha}_k}X_0, (1-\bar{\alpha}_k)\boldsymbol{I}), \tag{A.5}$$

with $\alpha_k := 1 - \beta_k$ and $\bar{\alpha_k} := \prod_k^K \alpha_k$. Following [185], using a reparameterization trick allows writing X_k directly:

$$X_k = \sqrt{\bar{\alpha}_k}X_0 + \sqrt{1-\bar{\alpha}_k}\epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \boldsymbol{I}). \tag{A.6}$$

Similar to the forward process, the reverse process is also defined as a Markov chain, aiming to iteratively restore data from noise by learning model parameters θ [185]. The reverse process is represented by:

$$p_\theta(X_{k-1}|X_k) = \mathcal{N}(X_{k-1}; \mu_\theta(X_k, k), \Sigma_\theta(X_k, k)), \tag{A.7}$$

where $\mu_\theta(X_k, k)$ and $\Sigma_\theta(X_k, k))$ are parameterized by the model. Following, we can predict X_{k-1} from X_k by:

$$X_{k-1} = \frac{1}{\sqrt{\alpha_k}}(X_k \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(X_k, k)) + \sigma_k z, \text{ with } z \sim \mathcal{N}(0, \boldsymbol{I}). \tag{A.8}$$

Note that $\epsilon_\theta(X_k, k)$ estimates the noise component that will be subtracted from X_k to reach the next step in the iterative denoising process [185].

A.1.2 Model Configurations & Training

All DL models are trained using the ADAM optimizer.

LightGBM

For the LightGBM [122], the preprocessed data is vectorized by concatenating all timesteps into a single vector. The LightGBM model is trained with the regression objective, *gbdt* boosting, a learning rate of 0.1, and an early-stopping round of 10. For the LSTM, the input data retains its sequential structure: each sample is a sequence of concatenated feature vectors, one for each time step.

LSTM

The LSTM processes sequential data using 2 stacked layers with 128 hidden units, followed by 2 FCLs with 128 hidden neurons and 1 output neuron, respectively. ReLU activation and batch normalization are applied before the final prediction layer. The training utilizes a fixed learning rate of 0.001, a batch size of 1024, and 50 epochs. Early stopping halts training if validation performance does not improve for 8 epochs.

Transformer Encoder

The Transformer encoder architecture [98] consists of 2 layers. Additionally, a hidden size of 128, 8 attention heads, and dropout with a dropout rate of 0.2 are incorporated. The output is passed through a linear layer with an output of 1, reflecting the predicted yield. We train the model for a maximum of 50 epochs with a batch size of 256, a learning rate of 0.0003, and a reduce-on-plateau learning rate scheduler. For regularization, early stopping is applied after 10 consecutive epochs, with no improvement on the validation set.

3D-LSTM

A conv3d block is used that applies 3D convolution across multiple input planes to capture the locality within the field. Next, a kernel size of $(1, 5, 5)$ is used to process the spatial dimension and expand the number of channels to 64, incorporating batch normalization and LeakyReLU activation. Further, the output is passed through a LSTM model as described in Subsubsection A.1.2 with 64 hidden units instead of 128.
Model training is run for a maximum of 50 epochs with a learning rate of 0.006 and a batch size of 2048. The training incorporates a reduce-on-plateau learning rate scheduler. Additionally, early stopping is applied after 10 consecutive epochs with no improvement on the validation set to avoid overfitting. Moreover, during training, data augmentation is applied to increase the generalization

performance. This includes a random rotation of the input window of 90 degrees, and TD with a 0.2 probability is employed for temporal features.

Multi-Modal Attention Fusion (MMAF)

A padding value of -1 is used to harmonize the time series length of the temporal modalities. Each block utilizes batch normalization and ReLU activation. For S2 the 3D-LSTM block is used. Static features are processed with a conv2d block with a $(5, 5)$ kernel size, followed by a FCL. Each modality-specific encoder outputs a feature representation of $N \times 64$. Sropout is applied to the attention weights with a probability of 0.2. The final fused representation is fed into a linear layer to predict the final yield value.
The training is similar to the 3D-LSTM as described in Subsubsection A.1.2.

contGAN

For the contGAN model [167], a cVAE-GAN and a cLR-GAN are combined to jointly learn the encoding from latent code to the output and back to the latent code ($Y \rightarrow z \rightarrow \hat{Y}$ and $z \rightarrow \hat{Y} \rightarrow \hat{z}$). Explicitly encouraging the connection between output and latent code to be invertible is thought to prevent many-to-one mapping, known as mode collapse [174]. An $L1$ loss term enforces the encoder to produce outputs which are close to the test time distribution in an $L1$ sense. We then search for the best generator/encoder pair that minimizes the following loss term:

$$\begin{aligned} \mathcal{G}^*, \mathcal{E}^*, \mathcal{D}^* = \arg\min_{\mathcal{G},\mathcal{E}} \max_{\mathcal{D}} & \mathcal{L}_{\text{GAN}}^{\text{VAE}}(\mathcal{G}, \mathcal{D}, \mathcal{E}) + \lambda_I \mathcal{L}_1^{\text{VAE}}(\mathcal{G}, \mathcal{E}) \\ & + \mathcal{L}_{\text{GAN}}^{\text{cLR}}(\mathcal{G}, \mathcal{D}) + \lambda_I \mathcal{L}_1^{\text{cLR}}(\mathcal{G}, \mathcal{E}) \\ & + \lambda_{\text{latent}}\, \mathcal{L}_1^{\text{latent}}\,(\mathcal{G}, \mathcal{E}) + \lambda_{\text{KL}}, \mathcal{L}_{\text{KL}}(\mathcal{E}) \end{aligned} \tag{A.9}$$

where the parameters λ control the weighting of all terms. The full loss term of the cVAE-GAN is given as:

$$\mathcal{G}^*, \mathcal{D}^*, \mathcal{E}^* = \arg\min_{\mathcal{G},\mathcal{E}} \max_{\mathcal{D}} \mathcal{L}_{\text{cGAN}}^{\text{VAE}} \quad + \lambda_I\, \mathcal{L}_{L1}^{\text{VAE}}(\mathcal{G}) + \lambda_{\text{KL}}\, \mathcal{L}_{\text{KL}}(\mathcal{E}). \tag{A.10}$$

To increase the impact of the latent space, a conditional Latent Regressor GAN (cLR-GAN) was proposed by [174]. The cLR-GAN forces the latent distribution to follow the test-time distribution, which can be randomly sampled. This is realized by using an L1 loss in the latent space between the FOV and the encoding of the synthesized future appearance:

$$\mathcal{L}_{L1}^{\text{latent}}(\mathcal{G}) = \mathbb{E}_{\mathcal{X},\mathcal{Z}}[||\boldsymbol{z} - \mathcal{E}(\mathcal{G}(\boldsymbol{X}, \boldsymbol{z}))||_1]. \tag{A.11}$$

The hyperparameters λ are set to $\lambda_{\mathrm{I}} = 0.5$, $\lambda_{\mathrm{latent}} = 10$, $\lambda_{\mathrm{KL}} = 0.01$. We train on 400 epochs with a learning rate of 2e−4.

RegDiff

The training is executed for a maximum of 100 epochs, with early stopping applied after 10 consecutive epochs without improvement on the validation set. A learning rate of 0.006 is used with a reduce-on-plateau learning rate scheduler, and a batch size of 2048. The models are trained on image patches of size 9×9.

Physics-Guided LSTM

Training is conducted for a maximum of 100 epochs. The learning rate is set to 0.001, and the batch size is 512. A reduce-on-plateau learning rate scheduler is employed during training. For regularization, early stopping is applied if no improvement is observed on the validation set for 10 consecutive epochs.

MC-TD

An IF strategy is used that concatenates all features along the time steps, as described in Chapter 4. For the architecture, we use a 2-layer LSTM to extract temporal information. Then, an additional layer is used to map the last hidden state to a hidden dimension. Finally, two prediction heads are employed to estimate the mean and variance, respectively. All layers consist of 128 units, including 20% of dropout. In addition, we use batch normalization layers after the LSTM model. We train the models for 100 epochs with an early stopping criterion based on a patience of five. The optimization is carried out with a batch size of 128 over the MSE function. For MC sampling, 20 samples are used as in [208].

UQ-LSTM

For the uncertainty LSTM, the same backbone architecture and training details as for the standard LSTM model (Subsubsection A.1.2). However, at the end, two distinct FCLs are used with 1 output unit, corresponding to the predicted yield and output variance, respectively. For MC-Dropout and MC-DropConnect, the optimal dropout probability p is determined to be 0.3.

A.2 Growth Stages & Vegetation Indices

The growth stage data was provided by Xarvio[1] within the project agreement. For this, proprietary models estimate daily cultivar-specific crop growth stages. These proprietary models are created per country and crop, using local field trial observations and weather data to reflect regional growth conditions. For each field, the start and end point of each growth stage was given. Only the 10 major growth stages were used. For soybean, a translation to the BBCH system was done. An overview of the major growth stages for each crop type is given in Table A.1.

Table A.1: Major growth stages for each crop type (Source: [261])

#	**Rapeseed** (*Brassica napus L. ssp. napus*)	**Wheat** (*Triticum aestivum*)	**Soybean** (*Glycine max L.*)
0	Germination	Germination	Germination
1	Leaf development	Leaf development	Unifoliolate
2	-	Tillering	1st-25th Trifoliolate
3	Stem elongation	Stem elongation	-
4	-	Booting	-
5	Heading	Heading	-
6	Flowering	Flowering	Flowering
7	Fruit development	Fruit development	Fruit development
8	Ripening	Ripening	Ripening
9	Senescence	Senescence	Senescence

Table A.2 shows commonly used VIs for crop yield prediction.

[1] www.xarvio.com [accessed: April 17, 2026]

Table A.2: Overview of commonly used VIs in yield prediction. R = Red, G = Green, B = Blue, N = NIR, RE = Red Edge. (Source: [87])

Vegetation Index	**Formula**
Chlorophyll Index Green (CIG)	$(N/G) - 1.0$
Chlorophyll Index Red Edge (CIRE)	$(N/RE1) - 1$
Green Normalized Difference Vegetation Index (GNDVI)	$(N - G)/(N + G)$
Normalized Difference Vegetation Index (NDVI)	$(N - R)/(N + R)$
Normalized Difference Vegetation Index (NDYI)	$(G - B)/(G + B)$
Ratio Vegetation Index (RVI)	$RE2/R$
Wide Dynamic Range Vegetation Index (WDRVI)	$(0.1 * N - R)/(0.1 * N + R)$
Normalized Green Red Difference Index (NGRDI)	$(G - R)/(G + R)$
Modified Chlorophyll Absorption Ratio Index / Optimized Soil-Adjusted Vegetation Index (MCARI/OSAVI)	$\frac{(((RE2-RE1)-0.2*(RE2-G))*(RE2/RE1))}{(1.16*(RE2-RE1)/(RE2+RE1+0.16))}$
Transformed Chlorophyll Absorption Ratio Index / Optimized Soil-Adjusted Vegetation Index (TCARI/OSAVI)	$\frac{(3*(RE2-RE1)-0.2*(RE2-G))*(RE2/RE1)}{(1.16*(RE2-RE1)/(RE2+RE1+0.16))}$

BIBLIOGRAPHY

[1] Michael Lawrence, Thomas Homer-Dixon, Scott Janzwood, Johan Rockström, Ortwin Renn, and Jonathan F Donges. Global polycrisis: the causal mechanisms of crisis entanglement. *Global Sustainability*, 7:e6, 2024.

[2] FSIN and Global Network Against Food Crises. Grfc 2025. Food Security Information Network and Global Network Against Food Crises, 2025.

[3] Kibrom A Abay, Clemens Breisinger, Joseph Glauber, Sikandra Kurdi, David Laborde, and Khalid Siddig. The russia-ukraine war: Implications for global and regional food security and potential policy responses. *Global Food Security*, 36:100675, 2023.

[4] FAO. Fao food price index, 2025. URL https://www.fao.org/worldfoodsituation/foodpricesindex/en/. Accessed: April 17, 2026.

[5] Matteo Coronese, Francesco Lamperti, Klaus Keller, Francesca Chiaromonte, and Andrea Roventini. Evidence for sharp increase in the economic damages of extreme natural disasters. *Proceedings of the National Academy of Sciences*, 116(43):21450–21455, 2019.

[6] Damien Delforge, Valentin Wathelet, Regina Below, Cinzia Lanfredi Sofia, Margo Tonnelier, Joris AF van Loenhout, and Niko Speybroeck. Em-dat: the emergency events database. *International Journal of Disaster Risk Reduction*, page 105509, 2025. (Accessed: 12-10-2025).

[7] FAO. The impact of disasters on agriculture and food security 2023: Avoiding and reducing losses through investment in resilience, 2023. Accessed: April 17, 2026.

[8] United Nations. World population prospects 2024: Summary of results (un desa/pop/2024/tr/no. 9.), 2024.

[9] United Nations, Department of Economic and Social Affairs, Population Division. World urbanization prospects 2018: Highlights, 2019.

[10] Miriam E Marlier, Amir S Jina, Patrick L Kinney, and Ruth S DeFries. Extreme air pollution in global megacities. *Current Climate Change Reports*, 2(1):15–27, 2016.

[11] Eurostat. Farms and farmland in the european union — statistics, 2022. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Farms_and_farmland_in_the_European_Union_-_statistics. Accessed: April 17, 2026.

[12] Eurostat. Farmers and the agricultural labour force - statistics, 2022. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Farmers_and_the_agricultural_labour_force_-_statistics. Accessed: April 17, 2026.

[13] United Nations General Assembly. Transforming our world: The 2030 agenda for sustainable development, 2015. URL https://www.refworld.org/legal/resolution/unga/2015/en/111816. Accessed: April 17, 2026.

[14] Serco Italia. Sentinel data access annual report 2023, 2024. Accessed: April 17, 2026.

[15] David J Lary, Gebreab K Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, et al. Machine learning applications for earth observation. *Earth observation open science and innovation*, 165, 2018.

[16] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

[17] Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

[18] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty

in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589, 2023.

[19] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[20] Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. Knowledge-guided machine learning: Current trends and future prospects. *arXiv preprint arXiv:2403.15989*, 2024.

[21] Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

[22] Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth and climate foundation models. *arXiv preprint arXiv:2405.04285*, 2024.

[23] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[24] Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, 8(2):127–150, 1979.

[25] Bo-Cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266, 1996.

[26] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[27] Tom G Farr and Mike Kobrick. Shuttle Radar Topography Mission produces a wealth of data. *Eos, Transactions American Geophysical Union*, 81(48):583–585, 2000.

[28] Laura Poggio, Luis M De Sousa, Niels H Batjes, Gerard 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.

[29] Miro Miranda, Deepak Pathak, Patrick Helber, Benjamin Bischke, Hiba Najjar, Cristhian Sanchez, Francisco Mena, Valdenegro Toro, Marlon Nuske, Marcela Charfualan, and Dengel Andreas. Yieldsat: A multi-modal benchmark dataset for high-resolution crop yield prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. (Accepted for publication).

[30] Cristhian Sanchez, Deepak Pathak, Miro Miranda, Marcela Charfuelan, Patrick Helber, Marlon Nuske, Benjamin Bischke, Peter Habelitz, Nafisur Rahman, Francisco Mena, et al. Influence of data cleaning techniques on sub-field yield predictions. In *IGARSS- IEEE International Geoscience and Remote Sensing Symposium*, pages 4852–4855. IEEE, 2023.

[31] FAO. *The State of Food Security and Nutrition in the World 2025*. FAO, 2025.

[32] Naveen Kumar Arora. Impact of climate change on agriculture production and its sustainable solutions. *Environmental sustainability*, 2(2):95–96, 2019.

[33] Amy Molotoks, Pete Smith, and Terence P Dawson. Impacts of land use, population, and climate change on global food security. *Food and Energy Security*, 10(1):e261, 2021.

[34] Gurdeep Singh Malhi, Manpreet Kaur, and Prashant Kaushik. Impact of climate change on agriculture and its mitigation strategies: A review. *Sustainability*, 13(3):1318, 2021.

[35] A. Devot, L. Royer, B. Arvis, D. Deryng, E. Caron Giauffret, L. Giraud, V. Ayral, and J. Rouillard. Research for agri committee – the impact of extreme climate events on agriculture production in the eu, 2023.

[36] Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nian-Feng Tzeng. An open and large-scale dataset for multi-modal climate change-aware crop yield predictions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5375–5386, 2024.

[37] Erhu He, Yiqun Xie, Licheng Liu, Weiye Chen, Zhenong Jin, and Xiaowei Jia. Physics guided neural networks for time-aware fairness: an applica-

tion in crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14223–14231, 2023.

[38] Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in europe. *Field Crops Research*, 276:108377, 2022.

[39] Domenico Benfenati, Domenico Amalfitano, Cristiano Russo, Cristian Tommasino, and Antonio Maria Rinaldi. Data centric artificial intelligence for agrifood domain: A systematic mapping study. *Computers and Electronics in Agriculture*, 239:110847, 2025.

[40] Gregor Perich, Mehmet Ozgur Turkoglu, Lukas Valentin Graf, Jan Dirk Wegner, Helge Aasen, Achim Walter, and Frank Liebisch. Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. *Field Crops Research*, 292:108824, 2023.

[41] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[42] Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, and Manan Shah. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial intelligence in agriculture*, 4:58–73, 2020.

[43] Corentin Leroux, Hazaël Jones, Anthony Clenet, Benoit Dreux, Maxime Becu, and Bruno Tisseyre. A general method to filter out defective spatial observations from yield mapping datasets. *Precision Agriculture*, 19:789–808, 2018.

[44] Miro Miranda, Francisco Mena, and Andreas Dengel. An analysis of temporal dropout in earth observation time series for regression tasks. In *Advances in Intelligent Data Analysis XXIII: 23rd International Symposium on Intelligent Data Analysis, IDA 2025, Konstanz, Germany, May 7–9, 2025, Proceedings*, volume 15669, page 389. Springer Nature, 2025.

[45] Richard Barnes. *RichDEM: Terrain Analysis Software*, 2016. URL http://github.com/r-barnes/richdem. Accessed: April 17, 2026.

[46] Martin Kopeckỳ, Martin Macek, and Jan Wild. Topographic wetness index calculation guidelines based on measured soil moisture and plant species composition. *Science of the Total Environment*, 757:143785, 2021.

[47] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[48] K. Rao, A. P. Williams, J. F. Flefil, and A. G. Konings. SAR-enhanced mapping of live fuel moisture content. *Remote Sensing of Environment*, 245:111797, 2020.

[49] Dean Gesch, Michael Oimoen, Susan Greenlee, Charles Nelson, Michael Steuck, and Dean Tyler. The national elevation dataset. *Photogrammetric engineering and remote sensing*, 68(1):5–32, 2002.

[50] Marc Simard, Naiara Pinto, Joshua B Fisher, and Alessandro Baccini. Mapping forest canopy height globally with spaceborne lidar. *Journal of Geophysical Research: Biogeosciences*, 116(G4), 2011.

[51] Shishi Liu, Yaxing Wei, Wilfred M Post, Robert B Cook, Kevin Schaefer, and Michele M Thornton. The unified north american soil map and its implication on the soil organic carbon stock in north america. *Biogeosciences*, 10(5):2915–2930, 2013.

[52] Song Chen. PM2.5 Data of Five Chinese Cities. UCI Machine Learning Repo., 2017.

[53] Madhuri R Paul, Dereje T Demie, Sabine J Seidel, and Thomas F Doering. Effects of spring wheat/faba bean mixtures on early crop development. *Plant and Soil*, 506(1):311–326, 2025.

[54] Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.

[55] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[56] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[57] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2): 187–212, 2022.

[58] David H Wolpert, William G Macready, et al. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

[59] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[60] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[61] Frank Rosenblatt et al. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, volume 55. Spartan books Washington, DC, 1962.

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

[63] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2): 179–211, 1990.

[64] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[65] Guanyuan Shuai and Bruno Basso. Subfield maize yield prediction improves when in-season crop water deficit is included in remote sensing imagery-based models. *Remote Sensing of Environment*, 272:112938, 2022.

[66] Malte von Bloh, David Lobell, and Senthold Asseng. Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture*, 227:109606, 2024.

[67] Felipe A. Lopes, Vasit Sagan, and Flavio Esposito. PlantPlotGAN: A Physics-Informed Generative Adversarial Network for Plant Disease Prediction . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7051–7060. IEEE, 2024. DOI: 10.1109/WACV57701.2024.00691.

[68] George Worrall, Jasmeet Judge, Kenneth Boote, and Anand Rangarajan. In-season crop phenology using remote sensing and model-guided machine learning. *Agronomy Journal*, 115(3):1214–1236, 2023.

[69] Qi Yang, Licheng Liu, Junxiong Zhou, Rahul Ghosh, Bin Peng, Kaiyu Guan, Jinyun Tang, Wang Zhou, Vipin Kumar, and Zhenong Jin. A flexible and efficient knowledge-guided machine learning data assimilation (kgml-da) framework for agroecosystem prediction in the us midwest. *Remote Sensing of Environment*, 299:113880, 2023.

[70] David Shulman, Assaf Israeli, Yael Botnaro, Ori Margalit, Oved Tamir, Shaul Naschitz, Dan Gamrasni, Ofer M Shir, and Itai Dattner. Physics-guided inverse regression for crop quality assessment. *Journal of Agricultural, Biological and Environmental Statistics*, pages 1–24, 2024.

[71] Issoufou Liman Harou, Cory Whitney, James Kung'u, and Eike Luedeling. Crop modelling in data-poor environments – a knowledge-informed probabilistic approach to appreciate risks and uncertainties in flood-based farming systems. *Agricultural Systems*, 187:103014, 2021. DOI: https://doi.org/10.1016/j.agsy.2020.103014.

[72] Mengjia Qiao, Xiaohui He, Xijie Cheng, Panle Li, Qianbo Zhao, Chenlu Zhao, and Zhihui Tian. Kstage: A knowledge-guided spatial-temporal attention graph learning network for crop yield prediction. *Information Sciences*, 619:19–37, 2023.

[73] Chishan Zhang and Chunyuan Diao. A phenology-guided bayesian-cnn (pb-cnn) framework for soybean yield estimation and uncertainty analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:50–73, 2023.

[74] Nando Metzger, Mehmet Ozgur Turkoglu, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Crop classification under varying cloud cover with neural ordinary differential equations. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

[75] Yuchi Ma, Zhou Zhang, Yanghui Kang, and Mutlu Özdoğan. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. *Remote Sensing of Environment*, 259:112408, 2021.

[76] Erhu He, Yiqun Xie, Licheng Liu, Zhenong Jin, Dajun Zhang, and Xiaowei Jia. Knowledge guided machine learning for extracting, preserving, and adapting physics-aware features. In *Proceedings of the 2024 SIAM*

International Conference on Data Mining (SDM), pages 715–723. SIAM, 2024.

[77] Mohammad Hossain Dehghan-Shoar, Gabor Kereszturi, Reddy R. Pullanagari, Alvaro A. Orsi, Ian J. Yule, and James Hanly. A physically informed multi-scale deep neural network for estimating foliar nitrogen concentration in vegetation. *International Journal of Applied Earth Observation and Geoinformation*, 130:103917, 2024. DOI: https://doi.org/10.1016/j.jag.2024.103917.

[78] Nikhil Vemuri. Developing a hybrid data-driven and informed model for prediction and mitigation of agricultural nitrous oxide flux hotspots. *Frontiers in Environmental Science*, 12:1353049, 2024.

[79] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[80] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[82] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

[83] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[84] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[85] Miro Miranda, Deepak Pathak, Marlon Nuske, and Andreas Dengel. Multi-modal fusion methods with local neighborhood information for

crop yield prediction at field and subfield levels. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4307–4311. IEEE, 2024.

[86] Deepak Pathak, Miro Miranda, Francisco Mena, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Hiba Najjar, Jayanth Siddamsetty, Diego Arenas, Michaela Vollmer, Marcela Charfuelan, Marlon Nuske, and Andreas Dengel. Predicting Crop Yield with Machine Learning: An Extensive Analysis of Input Modalities and Models on a Field and Sub-Field Level. In *IGARSS- IEEE International Geoscience and Remote Sensing Symposium*, pages 2767–2770, 2023. DOI: 10.1109/IGARSS52108.2023.10282318.

[87] Hiba Najjar, Miro Miranda, Marlon Nuske, Ribana Roscher, and Andreas Dengel. Explainability of sub-field level crop yield prediction using remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[88] Alexander Münzberg and Miro Miranda Lorenz. Vorhersage von landwirtschaftlichen erträgen und wachstum. In *Hybride KI mit Machine Learning und Knowledge Graphs: Innovative Lösungen aus der Praxis*, pages 153–167. Springer Fachmedien Wiesbaden Wiesbaden, 2025.

[89] Jiajia Li, Mingle Xu, Lirong Xiang, Dong Chen, Weichao Zhuang, Xunyuan Yin, and Zhaojian Li. Foundation models in smart agriculture: Basics, opportunities, and challenges. *Computers and Electronics in Agriculture*, 222:109032, 2024.

[90] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multimodal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[91] Lucien Wald. Some terms of reference in data fusion. *IEEE Transactions on geoscience and remote sensing*, 37(3):1190–1193, 2002.

[92] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. Common practices and taxonomy in deep multiview fusion for remote sensing applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4797–4818, 2024.

[93] Ozgul Calicioglu, Alessandro Flammini, Stefania Bracco, Lorenzo Bellù, and Ralph Sims. The future challenges of food and agriculture: An

integrated analysis of trends and solutions. *Sustainability*, 11(1):222, 2019.

[94] Ribana Roscher, Lukas Roth, Cyrill Stachniss, and Achim Walter. Data-centric digital agriculture: A perspective. *arXiv preprint arXiv:2312.03437*, 2023.

[95] Kaiyu Guan, Jin Wu, John S Kimball, Martha C Anderson, Steve Frolking, Bo Li, Christopher R Hain, and David B Lobell. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote sensing of environment*, 199:333–349, 2017.

[96] Priyanga Muruganantham, Santoso Wibowo, Srimannarayana Grandhi, Nahidul Hoque Samrat, and Nahina Islam. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9):1990, 2022.

[97] Marie Weiss, Frédéric Jacob, and Grgory Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, 236:111402, 2020.

[98] 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*, 30, 2017.

[99] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, 2015.

[100] Francisco Mena, Diego Arenas, Marcela Charfuelan, Marlon Nuske, and Andreas Dengel. Impact assessment of missing data in model predictions for earth observation applications. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 967–971, 2024. DOI: 10.1109/IGARSS53475.2024.10640375.

[101] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020.

[102] Merryn L Hunt, George Alan Blackburn, Luis Carrasco, John W Redhead, and Clare S Rowland. High resolution wheat yield mapping using sentinel-2. *Remote Sensing of Environment*, 233:111410, 2019.

[103] Ahmed Kayad, Marco Sozzi, Simone Gatto, Francesco Marinello, and Francesco Pirotti. Monitoring within-field variability of corn yield using sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23):2873, 2019.

[104] Amit Kumar Srivastava, Nima Safaei, Saeed Khaki, Gina Lopez, Wenzhi Zeng, Frank Ewert, Thomas Gaiser, and Jaber Rahimi. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific reports*, 12(1):3215, 2022.

[105] Juan Cao, Zhao Zhang, Yuchuan Luo, Liangliang Zhang, Jing Zhang, Ziyue Li, and Fulu Tao. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. *European Journal of Agronomy*, 123:126204, 2021.

[106] Xanthoula Eirini Pantazi, Dimitrios Moshou, Thomas Alexandridis, Rebecca L Whetton, and Abdul Mounem Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and electronics in agriculture*, 121:57–65, 2016.

[107] Keyhan Gavahi, Peyman Abbaszadeh, and Hamid Moradkhani. Deepyield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, 184:115511, 2021.

[108] Xinlei Wang, Jianxi Huang, Quanlong Feng, and Dongqin Yin. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sensing*, 12(11):1744, 2020.

[109] Seongchan Kim, Seungkyun Hong, Minsu Joh, and Sa-kwang Song. DeepRain: ConvLSTM Network for Precipitation Prediction using Multichannel Radar Data. In *Proceedings of the 7th International Workshop on Climate Informatics*, pages 89–92, November 2017. DOI: 10.5065/D6222SH7.

[110] Dhivya Elavarasan and PM Durai Raj Vincent. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. *Journal of Ambient Intelligence and Humanized Computing*, 12(11):10009–10022, 2021.

[111] Seyed Mahdi Mirhoseini Nejad, Dariush Abbasi-Moghadam, Alireza Sharifi, Nizom Farmonov, Khilola Amankulova, and Mucsi László. Multispectral crop yield prediction using 3d-convolutional neural networks and attention convolutional lstm approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:254–266, 2022.

[112] Luning Bi, Owen Wally, Guiping Hu, Albert U Tenuta, Yuba R Kandel, and Daren S Mueller. A transformer-based approach for early prediction of soybean yield using time-series images. *Frontiers in Plant Science*, 14: 1173036, 2023.

[113] Claudia Vallentin, Katharina Harfenmeister, Sibylle Itzerott, Birgit Kleinschmit, Christopher Conrad, and Daniel Spengler. Suitability of satellite remote sensing data for yield estimation in northeast germany. *Precision Agriculture*, 23(1):52–82, 2022.

[114] D Moravec, J Komárek, J Kumhálová, M Kroulík, J Prošek, P Klápště, et al. Digital elevation models as predictors of yield: comparison of an UAV and other elevation data sources. *Agronomy Research*, 15(1): 249–255, 2017.

[115] Raí A Schwalbert, Telmo Amado, Geomar Corassa, Luan Pierre Pott, PV Vara Prasad, and Ignacio A Ciampitti. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern brazil. *Agricultural and Forest Meteorology*, 284:107886, 2020.

[116] Luwei Feng, Yumiao Wang, Zhou Zhang, and Qingyun Du. Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sensing of Environment*, 262:112514, 2021.

[117] Maitiniyazi Maimaitijiang, Vasit Sagan, Paheding Sidike, Sean Hartling, Flavio Esposito, and Felix B Fritschi. Soybean yield prediction from uav using multimodal data fusion and deep learning. *Remote sensing of environment*, 237:111599, 2020.

[118] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multimodal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187: 294–305, 2022.

[119] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS journal of photogrammetry and remote sensing*, 169:421–435, 2020.

[120] Juncheng Ma, Binhui Liu, Lin Ji, Zhicheng Zhu, Yongfeng Wu, and Weihua Jiao. Field-scale yield prediction of winter wheat under different irrigation regimes based on dynamic fusion of multimodal uav imagery. *International Journal of Applied Earth Observation and Geoinformation*, 118:103292, 2023.

[121] Patrick Helber, Benjamin Bischke, Peter Habelitz, Cristhian Sanchez, Deepak Pathak, Miro Miranda, Hiba Najjar, Francisco Mena, Jayanth Siddamsetty, Diego Arenas, et al. Crop yield prediction: An operational approach to crop yield modeling on field and subfield level with machine learning models. In *IGARSS - IEEE International Geoscience and Remote Sensing Symposium*, pages 2763–2766. IEEE, 2023.

[122] 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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 3149–3157, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[123] Richard G Allen, Luis S Pereira, Dirk Raes, Martin Smith, et al. Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. *Fao, Rome*, 300(9):D05109, 1998.

[124] Kelly R Thorp. pyfao56: Fao-56 evapotranspiration in python. *SoftwareX*, 19:101208, 2022.

[125] Francisco Mena, Deepak Pathak, Hiba Najjar, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Miro Miranda, Jayanth Siddamsetty, Marlon Nuske, Marcela Charfuelan, Diego Arenas, Michaela Vollmer, and Andreas Dengel. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. *Remote Sensing of Environment*, 318:114547, 2025. DOI: https://doi.org/10.1016/j.rse.2024.114547.

[126] Hiba Najjar, Patrick Helber, Benjamin Bischke, Peter Habelitz, Cristhian Sanchez, Francisco Mena, Miro Miranda, Deepak Pathak, Jayanth Siddamsetty, Diego Arenas, et al. Feature attribution methods for multivariate time-series explainability in remote sensing. In *IGARSS 2023-2023*

IEEE International Geoscience and Remote Sensing Symposium, pages 5014–5017. IEEE, 2023.

[127] Miro Miranda, Francisco Mena, Marcela Charfuelan, and Andreas Dengel. Informed learning for efficient crop yield prediction. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2025)*. IEEE, 2025.

[128] Cristhian Sanchez, Francisco Mena, Marcela Charfuelan, Marlon Nuske, and Andreas Dengel. Assessment of sentinel-2 spatial and temporal coverage based on the scene classification layer. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4099–4103. IEEE, 2024.

[129] Hiba Najjar, Deepak Pathak, Marlon Nuske, and Andreas Dengel. Intrinsic explainability of multimodal learning for crop yield simulation. *Computers and Electronics in Agriculture*, 239:111003, 2025.

[130] Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2085–2095, 2023.

[131] Uwe Meier, Hermann Bleiholder, Liselotte Buhr, Carmen Feller, Helmut Hack, Martin Heß, Peter D Lancashire, Uta Schnock, Reinhold Stauß, Theo Van Den Boom, et al. The bbch system to coding the phenological growth stages of plants–history and publications. *Journal für Kulturpflanzen*, 61(2):41–52, 2009.

[132] Michael D King, Steven Platnick, W Paul Menzel, Steven A Ackerman, and Paul A Hubanks. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, 2013.

[133] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[134] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[135] Edmundo Acevedo, Paola Silva, and Herman Silva. Growth and wheat physiology, development. *Laboratory of Soil-Plant-Water Relations. Faculty of Agronomy and Forestry Sciences. University of Chile. Casilla*, 1004, 2006.

[136] PV Biscoe and JN Gallagher. A physiological analysis of cereal yield. i. production of dry matter. *Agricultural Progress*, 53:34–50, 1978.

[137] Miro Benjamin Miranda, Mauricio Charfuelan, Matias Valdenegro-Toro, and Andreas Dengel. Informed learning for estimating drought stress at fine-scale resolution enables accurate yield prediction. In *ECAI 2025 – 27th European Conference on Artificial Intelligence*, pages 5384–5391. IOS Press, 2025.

[138] Miro Miranda, Marcela Charfuelan, and Andreas Dengel. Exploring physics-informed neural networks for crop yield loss forecasting. In *NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024.

[139] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

[140] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[141] Tony Fischer, Derek Byerlee, and Greg Edmeades. Crop yields and food security: will yield increases continue to feed the world? In *Capturing Opportunities and Overcoming Obstacles in Australian Agronomy*, 2012.

[142] Yinhong Kang, Shahbaz Khan, and Xiaoyi Ma. Climate change impacts on crop yield, crop water productivity and food security–a review. *Progress in natural Science*, 19(12):1665–1674, 2009.

[143] Richard G Allen, Masahiro Tasumi, and Ricardo Trezza. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (metric)—model. *Journal of irrigation and drainage engineering*, 133(4):380–394, 2007.

[144] Mojtaba Naghdyzadegan Jahromi, Shahrokh Zand-Parsa, Fatemeh Razzaghi, Sajad Jamshidi, Shohreh Didari, Ali Doosthosseini, and Hamid Reza Pourghasemi. Developing machine learning models for wheat yield predic-

tion using ground-based data, satellite-based actual evapotranspiration and vegetation indices. *European Journal of Agronomy*, 146:126820, 2023.

[145] Richard G Allen, Luis S Pereira, Terry A Howell, and Marvin E Jensen. Evapotranspiration information reporting: I. factors governing measurement accuracy. *Agricultural Water Management*, 98(6):899–920, 2011.

[146] Azeem Khan, Claudio O Stöckle, Roger L Nelson, Troy Peters, Jennifer C Adam, Brian Lamb, Jinshu Chi, and Sarah Waldo. Estimating biomass and yield using metric evapotranspiration and simple growth algorithms. *Agronomy journal*, 111(2):536–544, 2019.

[147] J Doorenbos and AH Kassam. Yield response to water. *Irrigation and drainage paper*, 33:257, 1979.

[148] Pasquale Steduto, Theodore C Hsiao, Elias Fereres, Dirk Raes, et al. *Crop yield response to water*, volume 1028. fao Rome, Italy, 2012.

[149] Luis S Pereira, Richard G Allen, Martin Smith, and Dirk Raes. Crop evapotranspiration estimation with fao56: Past and future. *Agricultural water management*, 147:4–20, 2015.

[150] Guoyong Leng and Jim W Hall. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental research letters: ERL [Web site]*, 15(4):044027, 2020.

[151] Francisco Mena, Deepak Pathak, Hiba Najjar, Cristhian Sanchez, Patrick Helber, Benjamin Bischke, Peter Habelitz, Miro Miranda, Jayanth Siddamsetty, Marlon Nuske, et al. Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction. *Remote Sensing of Environment*, 318:114547, 2025.

[152] Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

[153] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, 12(1):1040, 2022.

[154] Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V Archontoulis. Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt. *Scientific reports*, 11(1):1606, 2021.

[155] LS Pereira, Paula Paredes, DJ Hunsaker, R López-Urrea, and Z Mohammadi Shad. Standard single and basal crop coefficients for field crops. updates and advances to the fao56 crop water requirements method. *Agricultural Water Management*, 243:106466, 2021.

[156] Jiabing Cai, Yu Liu, Tingwu Lei, and Luis Santos Pereira. Estimating reference evapotranspiration with the fao penman–monteith equation using daily weather forecast messages. *Agricultural and Forest Meteorology*, 145(1-2):22–35, 2007.

[157] Gijs Simons, Reinier Koster, and Peter Droogers. Hihydrosoil v2. 0-high resolution soil maps of global hydraulic properties. *Future Works.[online] Available from https://www. futurewater. eu/projects/hihydrosoil*, 2020.

[158] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[159] Jingyu Zhao, Feiqing Huang, Jia Lv, Yanjie Duan, Zhen Qin, Guodong Li, and Guangjian Tian. Do rnn and lstm have long memory? In *International Conference on Machine Learning*, pages 11365–11375. PMLR, 2020.

[160] Thirupathi Kandadi and G Shankarlingam. Drawbacks of lstm algorithm: A case study. *Available at SSRN 5080605*, 2025.

[161] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

[162] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

[163] Francis Bach. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc 32nd Int Conf Mach Learn*, volume 37, page 448, 2015.

[164] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11): 8135–8153, 2022.

[165] John S Boyer. Plant productivity and environment. *Science*, 218(4571): 443–448, 1982.

[166] Wu Gang, Wei Zhen-Kuan, Wang Yong-Xiang, Chu Li-Ye, and Shao Hong-Bo. The mutual responses of higher plants to environment: physiological and microbiological aspects. *Colloids and Surfaces B: Biointerfaces*, 59 (2):113–119, 2007.

[167] Miro Miranda, Lukas Drees, and Ribana Roscher. Controlled Multimodal Image Generation for Plant Growth Modeling. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 5118–5124. IEEE, 2022.

[168] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[169] Laura Zabawa, Anna Kicherer, Lasse Klingbeil, Reinhard Töpfer, Heiner Kuhlmann, and Ribana Roscher. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164:73–83, 2020.

[170] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147: 70–90, 2018.

[171] Lukas Drees, Dereje T Demie, Madhuri R Paul, Johannes Leonhardt, Sabine J Seidel, Thomas F Döring, and Ribana Roscher. Data-driven crop growth simulation on time-varying generated images using multi-conditional generative adversarial networks. *Plant Methods*, 20(1):93, 2024.

[172] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[173] Lukas Drees, Laura Verena Junker-Frohn, Jana Kierdorf, and Ribana Roscher. Temporal prediction and evaluation of brassica growth in the field using conditional generative adversarial networks. *Computers and Electronics in Agriculture*, 190, 2021.

[174] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017.

[175] Patrick Esser, Robin Rombach, and Bjorn Ommer. A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2020.

[176] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[177] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[178] Andres Patrignani and Tyson E Ochsner. Canopeo: A powerful new tool for measuring fractional green canopy cover. *Agronomy Journal*, 107(6): 2312–2320, 2015.

[179] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[180] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[181] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016.

[182] Lukas Drees, Immanuel Weber, Marc Rußwurm, and Ribana Roscher. Time dependent image generation of plants from incomplete sequences with cnn-transformer. In *DAGM German Conference on Pattern Recognition*, pages 495–510. Springer, 2022.

[183] Miro Miranda, Akshay Dinesh, David N. Lesmes-Leon, Fernando Mena, Mauricio Charfuelan, and Andreas Dengel. regdiff: Regression diffusion for earth observation. In *Proceedings of the IGARSS 2025 – IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2025.

[184] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794, 2021.

[185] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851, 2020.

[186] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[187] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021.

[188] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

[189] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[190] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[191] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[192] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10157–10166, 2023.

[193] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36:41693–41706, 2023.

[194] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[195] Patrick Helber, Benjamin Bischke, Carolin Packbier, Peter Habelitz, and Florian Seefeldt. An operational approach to large-scale crop yield prediction with spatio-temporal machine learning models. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4299–4302. IEEE, 2024.

[196] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[197] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[198] Matias Valdenegro-Toro. I find your lack of uncertainty in computer vision disturbing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1263–1272, 2021.

[199] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[200] Codrut-Andrei Diaconu and Nina Maria Gottschling. Uncertainty-aware learning with label noise for glacier mass balance modeling. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

[201] Gianni Franchi, Olivier Laurent, Maxence Leguéry, Andrei Bursuc, Andrea Pilzer, and Angela Yao. Make me a bnn: A simple strategy for estimating bayesian uncertainty from pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12194–12204, 2024.

[202] Jae-Gil Lee and Minseo Kang. Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2):74–81, 2015.

[203] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, 2013. DOI: 10.1109/TGRS.2012.2227333.

[204] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

[205] Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516, 2022. DOI: 10.1109/CVPRW56347.2022.00157.

[206] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in neural information processing systems*, 30, 2017.

[207] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[208] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[209] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[210] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

[211] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

[212] Aryan Mobiny, Pengyu Yuan, Supratik K Moulik, Naveen Garg, Carol C Wu, and Hien Van Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific reports*, 11(1):5458, 2021.

[213] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

[214] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. arxiv 2019. *arXiv preprint arXiv:1912.02757*, 2019.

[215] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. DOI: 10.1109/ICNN.1994.374138.

[216] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR, 2018.

[217] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. DOI: 10.1609/aaai.v29i1.9602.

[218] Lynn Miller, Charlotte Pelletier, and Geoffrey I Webb. Deep learning for satellite image time-series analysis: A review. *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[219] Francisco Mena, Diego Arenas, and Andreas Dengel. Increasing the robustness of model predictions to missing sensors in earth observation. *arXiv preprint arXiv:2407.15512*, 2024.

[220] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[221] Sophocles J Orfanidis. *Introduction to signal processing*. Prentice-Hall, Inc., 1995.

[222] Wei Cao, Dong Wang, Jian Li, Hao Zhou, et al. Brits: Bidirectional recurrent imputation for time series. *Advances in Neural Information Processing Systems*, 31, 2018.

[223] Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

[224] Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[225] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[226] Stefan Becker, Ronny Hug, Wolfgang Huebner, Michael Arens, and Brendan Tran Morris. Missformer:(in-) attention-based handling of missing observations for trajectory filtering and prediction. In *International Symposium on Visual Computing*, pages 521–533, 2021.

[227] Jordi Inglada, Arthur Vincent, Marcela Arias, and Claire Marais-Sicre. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8(5):362, 2016.

[228] Felipe Ferrari, Matheus Pinheiro Ferreira, Cláudio Aparecido Almeida, and Raul Queiroz Feitosa. Fusing Sentinel-1 and Sentinel-2 images for deforestation detection in the Brazilian Amazon under diverse cloud conditions. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

[229] Stella Ofori-Ampofo, Charlotte Pelletier, and Stefan Lang. Crop type mapping from optical and radar time series using attention-based deep learning. *Remote Sensing*, 13(22):4668, 2021.

[230] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah R Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023.

[231] Frank Weilandt, Robert Behling, Romulo Goncalves, Arash Madadi, Lorenz Richter, Tiago Sanona, Daniel Spengler, and Jona Welsch. Early crop classification via multi-modal satellite data fusion and temporal attention. *Remote Sensing*, 15(3):799, 2023.

[232] Jin Chen, Per Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. A simple method for reconstructing a high-quality

NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment*, 91(3-4):332–344, 2004.

[233] Monidipa Das and Soumya K Ghosh. A deep-learning-based forecasting ensemble to predict missing data for remote sensing analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5228–5236, 2017.

[234] Giuseppe Scarpa, Massimiliano Gargiulo, Antonio Mazza, and Raffaele Gaetano. A CNN-based fusion method for feature extraction from Sentinel data. *Remote Sensing*, 10(2):236, 2018.

[235] Navid Mohammadi Foumani, Lynn Miller, Chang Wei Tan, Geoffrey I Webb, Germain Forestier, and Mahsa Salehi. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9):1–45, 2024.

[236] Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I Webb. Time series extrinsic regression: Predicting numeric values from time series data. *Data Mining and Knowledge Discovery*, 35(3):1032–1060, 2021.

[237] Xizhe Xue and Xiao Xiang Zhu. Regression in earth observation: Are vlms up to the challenge? *Geoscience and Remote Sensing Magazine*, 2025.

[238] Long H Nguyen, Jiazhen Zhu, Zhe Lin, Hanxiang Du, Zhou Yang, Wenxuan Guo, and Fang Jin. Spatial-temporal multi-task learning for within-field cotton yield prediction. In *Advances in Knowledge Discovery and Data Mining*, pages 343–354, 2019.

[239] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1796–1805, 2020.

[240] Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1132–1142, 2021.

[241] Codruț-Andrei Diaconu, Sudipan Saha, Stephan Günnemann, et al. Understanding the role of weather data for earth surface forecasting using

a ConvLSTM-based model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1362–1371, 2022.

[242] Levente Foldesi and Matias Valdenegro-Toro. Comparison of Uncertainty Quantification with Deep Learning in Time Series Regression. In *Workshop on Robustness in Sequence Modeling at NeurIPS 2022*, 2017.

[243] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

[244] Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian Recurrent Neural Networks. In *Women in Machine Learning Workshop (WiML) at NeurIPS 2017*, 2017.

[245] Matias Valdenegro-Toro, Ivo Pascal de Jong, and Marco Zullich. Unified uncertainties: Combining input, data and model uncertainty into a single formulation. *arXiv e-prints*, pages arXiv–2406, 2024.

[246] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[247] Mohamed Farag, Jana Kierdorf, and Ribana Roscher. Inductive conformal prediction for harvest-readiness classification of cauliflower plants: A comparative study of uncertainty quantification methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–659, 2023.

[248] Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 914–927. Curran Associates, Inc., 2021.

[249] Burak Ekim, Girmaw Abebe Tadesse, Caleb Robinson, Gilles Hacheme, Michael Schmitt, Rahul Dodhia, and Juan M. Lavista Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2265–2274, June 2025.

[250] Deepak K Ray, James S Gerber, Graham K MacDonald, and Paul C West. Climate variation explains a third of global crop yield variability. *Nature communications*, 6(1):5989, 2015.

[251] Remus Pop and Patric Fulop. Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles. *arXiv preprint arXiv:1811.03897*, 2018.

[252] Florian Seligmann, Philipp Becker, Michael Volpp, and Gerhard Neumann. Beyond deep ensembles: A large-scale evaluation of bayesian deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 36:29372–29405, 2023.

[253] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

[254] Laura von Rueden, Jochen Garcke, and Christian Bauckhage. How does knowledge injection help in informed machine learning? In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[255] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[256] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[257] James M. Keller, Derong Liu, and David B. Fogel. *Recurrent Neural Networks*, pages 77–100. John Wiley & Sons, 2016.

[258] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[259] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[260] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[261] Uwe Meier. *Growth stages of mono-and dicotyledonous plants*. Federal Biological Research Centre for Agriculture and Forestry, 2001.

Curriculum Vitæ

Miro Benjamin Miranda Lorenz

Education

RPTU University of Kaiserslautern-Landau since 2022
PhD in Computer Science
German Research Center for Artificial Intelligence

University of Bonn 2019 – 2021
M. Sc. in Life Science Informatics
Bonn-Aachen International Center for Information Technology

University of Giessen 2016 – 2019
B. Sc. in Agricultural Sciences

Experience

German Research Center for Artificial Intelligence since 2022
Smart Data & Knowledge Services Department
Position: Researcher

University of Groningen 2025
Bernoulli Institute
Position: Visiting Researcher

University of Bonn 2021 – 2022
Remote Sensing Group
Position: Researcher