Phenology-Aligned Representation and Normalization for Cross-Country Crop Yield Prediction

Parna Asadi

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Master of Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Abstract

Accurate crop yield forecasting across heterogeneous geographic regions is significantly hindered by domain shift, driven by regional variations in crop calendars, season lengths, and climatic conditions. Traditional calendar-based time series indexing fails to account for these differences, misaligning physiological growth stages and degrading the cross-country generalization of machine learning models. To overcome this limitation, this thesis introduces PhenoNorm, a phenology-aligned preprocessing pipeline designed to mitigate temporal misalignment. PhenoNorm re-indexes multi-source environmental features, including weather, vegetation indices, and soil moisture, along a unified physiological coordinate system using normalized cumulative Growing Degree Days (nGDD). This transformation enables observations from diverse agro-ecological regions across different countries to be compared at equivalent developmental stages. This study systematically evaluates the interaction of this biological alignment with various feature normalization strategies (global versus country-wise) across multiple model architectures, including Random Forest, XGBoost, LSTM, PatchTST, and a pretrained time series foundation model, Moirai. Empirical results demonstrate that global normalization outperforms country-wise normalization by preserving critical scale differences and informative spatial variations required for cross-domain transfer. Sequence-based models outperformed tabular baselines, with the transformer-based PatchTST achieving the highest overall predictive performance by effectively capturing stage-dependent temporal interactions and early-season vegetation signals. Furthermore, while the Moirai foundation model failed to generalize under domain shift in a zero-shot setting, it demonstrated strong adaptation efficiency and became highly competitive after task-specific fine-tuning. Overall, this research establishes that temporal representation and normalization can strongly affect cross-country generalization.

Acknowledgements

This thesis marks the end of a research journey, but it also continues a curiosity I have held for a long time. Long before I began studying machine learning, I was already drawn to questions of agriculture. It is deeply meaningful to return to this field in a new form, bringing together engineering, data science, and agriculture.

I would like to express my deepest gratitude to Professor Kiringa for introducing me to the initial idea behind this work and for opening the door to this area of research. I am equally grateful to Professor Kantere for her thoughtful guidance, steady encouragement, and invaluable support throughout the process of shaping the problem and developing a solution. Their consistent presence and guidance gave this work both direction and strength. This study would not have been possible without their mentorship.

I am also profoundly thankful to my parents, whose love and support carried me through this journey across distance and difficulty. I owe special thanks to my mother for her constant encouragement, warmth, and unwavering belief in me, especially during the times when we were far apart. Her love was a source of comfort and strength in every stage of this work.

Finally, I would like to thank my father, an excellent professor and a lifelong source of wisdom, whose deep knowledge of agriculture helped me see the value of this research. His insight not only inspired many of my ideas, but also guided me toward asking better questions and seeking better answers.

To all of them, I offer my sincere gratitude.

Human beings are members of a whole,
In creation of one essence and soul.
If one member is afflicted with pain,
Other members uneasy will remain.
- Saadi

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Accurate crop yield prediction is a fundamental component of food security planning, agricultural management, and climate-resilient decision-making, particularly for staple crops such as maize, which is cultivated across diverse agro-ecological[1] systems. In recent years, machine learning (ML) and deep learning (DL) approaches have substantially improved the ability to estimate yields from heterogeneous data sources, including weather observations, soil properties, and remote sensing products. Unlike traditional statistical or process-based models, these approaches capture agricultural dynamics at much finer spatial and temporal scales [53, 15].

Potential users of regional yield forecasts include agricultural extension services, government agencies, insurers, commodity and supply-chain organizations, and food-security planners. At these spatial scales, yield estimates can support the early identification of production risk, allocation of field-inspection and advisory resources, crop-insurance assessment, harvest, storage, and transportation planning, and the anticipation of regional production shortfalls. The output should be interpreted alongside agronomic expertise, local management information, and uncertainty estimates before it is used to support consequential decisions [2].

Tree-based ensembles and gradient boosting methods have demonstrated strong results when combined with carefully engineered seasonal features [38], and sequence-based deep learning architectures, including recurrent and transformer-based time series models, have shown strong capability in modeling temporal crop-development patterns and long-range seasonal dependencies from sequential observations [45, 44].

Despite these advances, robust prediction across countries remains a major challenge. Agricultural time series differ substantially across regions in terms of crop calendars, season length, climate regimes, feature distributions, and yield variability, leading to pronounced domain shift between environments [43, 70]. In such settings, models trained on one group of countries often fail to generalize reliably to unseen regions.

To address temporal dependencies, sequence models such as recurrent neural networks (RNNs) have been increasingly adopted, showing strong potential for learning complex

[1] Agro-ecological: Refers to the interaction between agricultural systems, environmental conditions, climate, soil, and crop-management practices within a specific region.

temporal patterns from multi-source agricultural data [24, 47]. However, their effectiveness often depends on how seasonal signals are represented prior to model training. In many existing pipelines, preprocessing steps such as temporal alignment, in-season extraction, and normalization are treated as secondary implementation details.

This thesis explores the impact of temporal representation and data preprocessing on cross-country predictive modeling. Traditional calendar-based indexing often fails to capture regional variations in crop development, while standard normalization techniques may obscure valuable cross-country differences. To overcome these challenges, this study proposes a standardized representation that maintains a consistent physiological reference frame across diverse geographic regions.

1.1 Problem Definition

This section formalizes the multi-country crop yield prediction task and outlines the key challenges.

1.1.1 Notation and Setup

Consider a multi-country dataset comprising C countries indexed by $c \in \{1, 2, \ldots, C\}$. For each country c, we observe N_c region–year samples, where each sample corresponds to a specific administrative region in a given growing season. The dataset is partitioned into disjoint sets of training countries $\mathcal{C}_{\text{train}}$ and held-out test countries $\mathcal{C}_{\text{test}}$, such that $\mathcal{C}_{\text{train}} \cap \mathcal{C}_{\text{test}} = \emptyset$. This setup reflects a cross-country generalization scenario in which models must transfer to previously unseen regions.

For a given region–year sample i in country c, the input consists of a multivariate time series of environmental and remote-sensing variables observed over the growing season:

$$\mathbf{X}_{c,i} = \{\mathbf{x}_{c,i,1}, \mathbf{x}_{c,i,2}, \ldots, \mathbf{x}_{c,i,T_c}\}, \tag{1.1}$$

where $\mathbf{x}_{c,i,t} \in \mathbb{R}^d$ denotes a d-dimensional feature vector at time step t, and T_c is the season length in country c. The feature vector may include meteorological variables (e.g., temperature), vegetation indices (e.g., NDVI), soil-moisture indicators, and static soil or topographic attributes commonly used in yield prediction studies [30, 70].

The target variable for sample i in country c is the end-of-season yield ($y_{c,i} \in \mathbb{R}_+$), typically measured in **tonnes per hectare** (t/ha).

Learning Objective The objective is to learn a predictive function $f_\theta : \mathbb{R}^{T \times d} \to \mathbb{R}_+$, parameterized by θ, that maps seasonal time series observations to yield estimates:

$$\hat{y}_{c,i} = f_\theta(\mathbf{X}_{c,i}). \tag{1.2}$$

The model parameters are learned by minimizing an empirical risk over samples from the training countries:

$$\min_{\theta} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{i=1}^{N_c} \mathcal{L}\left(y_{c,i}, f_\theta(\mathbf{X}_{c,i})\right), \tag{1.3}$$

where $\mathcal{L}$ denotes a regression loss function, such as root mean squared error (RMSE). Model performance is evaluated on held-out countries to assess generalization under distributional shift [49].

1.1.2 Key Challenges

The multi-country yield prediction setting introduces several challenges:

Challenge 1: Temporal misalignment across countries. Crop calendars and season lengths vary across regions. When time series are indexed by calendar time (e.g., day of year), observations at the same index t may correspond to different physiological stages. For example, a given day may represent early vegetative growth in one region and flowering in another. Formally, for countries c and c':

$$\text{Stage}(t, c) \neq \text{Stage}(t, c') \quad \text{for } t \in [1, T_c] \cap [1, T_{c'}]. \tag{1.4}$$

This inconsistency complicates the learning of patterns consistent with agronomic expectations [74].

Challenge 2: Domain shift in feature and target distributions. Agricultural systems differ across countries in climate, soil properties, management practices, and crop varieties. As a result, the joint distribution of inputs and outputs varies across regions:

$$P_c(\mathbf{X}, y) \neq P_{c'}(\mathbf{X}, y), \quad \text{for } c \neq c'. \tag{1.5}$$

Such domain shift is known to degrade model performance when transferring to unseen environments [49, 65]. In this setting, country-level domain shift is multidimensional. It may arise from differences in climatic regimes, planting dates and phenological progression, soil characteristics, cultivar choice and management practices, the spatial and temporal coverage or quality of input data, and the distribution and reliability of reported yields.

Challenge 3: Sensitivity to preprocessing and normalization. Feature scaling and normalization significantly affect model behavior, optimization stability, and cross-domain transferability. In heterogeneous time series settings, different normalization strategies can lead to substantially different performance outcomes [35, 52].

A common choice is country-wise normalization, where features are standardized using statistics computed per country, versus global normalization, which uses dataset-wide

statistics. The former removes local scale differences, while the latter preserves inter-country magnitude relationships by enforcing a shared scale. Despite their widespread use, the impact of these strategies on cross-country generalization in crop yield prediction remains insufficiently studied.

1.2 Research Questions

To address the challenges outlined above, this thesis investigates the following research questions:

1. **RQ1: Phenology alignment and cross-country generalization.** How does aligning seasonal observations using nGDD, rather than calendar time, affect predictive performance and cross-country generalization?
2. **RQ2: Normalization under domain shift.** How do country-wise and global normalization strategies compare in terms of generalization to unseen countries within a phenology-aligned pipeline? Which strategy better supports transfer across regions with differing climate regimes and yield distributions?
3. **RQ3: Foundation models versus task-specific models.** How does a pretrained time series foundation model compare to task-specific baselines under cross-country domain shift? In particular, how does their performance vary between zero-shot and fine-tuned settings, and how sensitive are they to the amount of fine-tuning data?

To answer these questions, we propose *PhenoNorm*, a preprocessing pipeline for multi-country crop yield prediction. The pipeline integrates quality-controlled preprocessing, crop season mapping, in-season extraction, temporal alignment via nGDD, and consistent feature normalization to construct a unified representation across countries. By establishing this biologically aligned feature space, the pipeline ensures that observations remain physiologically consistent regardless of the specific machine learning algorithm applied.

The proposed solution enables a more controlled experimental setting in which representation and normalization choices can be evaluated independently of model architecture. Models are trained on a set of countries and evaluated on disjoint held-out countries, using a combination of standard regression metrics and hydrology-inspired skill scores such as Nash–Sutcliffe efficiency [42] and Kling–Gupta efficiency [14].

1.3 Research Design Overview

This study was organized into four interlinked phases, each addressing a specific aspect of multi-country crop yield prediction. The overall objective was to develop a phenology-aligned and normalization-consistent preprocessing pipeline, examine its interaction with

different model families, and assess the applicability of time series foundation models in this setting.

The first phase focused on data characterization and quality control. We started from a multi-country crop yield panel constructed from sub-national statistics, crop calendars, meteorological data, vegetation indices, soil-moisture products, and static soil layers. The dataset was examined to identify valid regions and years, detect outliers and implausible values, and assess temporal and cross-country variability. The outcome of this phase was a curated dataset of region–year samples used in subsequent analysis.

The second phase developed a representation and preprocessing pipeline designed for cross-country comparability. Seasonal observations were mapped to crop seasons, restricted to in-season periods, and aligned using GDD to index developmental progress. From this representation, two complementary input forms were constructed: Stage-aggregated tabular features for tree-based models and fixed-length multivariate sequences for sequence-based models. In addition, multiple normalization schemes were defined, including country-wise and global normalization for tabular inputs, and country-wise, global, and bin-wise normalization for sequence inputs [35].

The third phase evaluated the behavior of the proposed preprocessing pipeline across different model families, including Random Forest, XGBoost, LSTM, and PatchTST, under a unified experimental setup.

The fourth phase extended the methodology to a pretrained time series foundation model. Pretrained Moirai was considered alongside task-specific models, and was studied under both zero-shot and fine-tuned settings to assess its behavior [69].

Along with the proposed preprocessing pipeline (shown in Figure 1.1), a consistent evaluation protocol was applied, based on a country-level train–test split, validation within the training countries, and a set of evaluation metrics including RMSE, MAE, and R^2, together with hydrological skill scores [42, 14].

1.4 Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 surveys the relevant literature on crop yield prediction, including traditional and machine learning-based approaches, sequence modeling for agricultural time series, and challenges related to cross-country generalization.

Chapter 3 introduces the proposed methodology. It presents the dataset and preprocessing workflow, describes the phenology-aligned representation and normalization strategies, and details the predictive models and evaluation protocol used to study cross-country crop yield prediction.

Chapter 4 presents the empirical results and comparative analysis. It evaluates the impact of the proposed representation on cross-country comparability, compares model families under a common experimental protocol, and analyzes the behavior of pretrained foundation models in both zero-shot and fine-tuned settings.

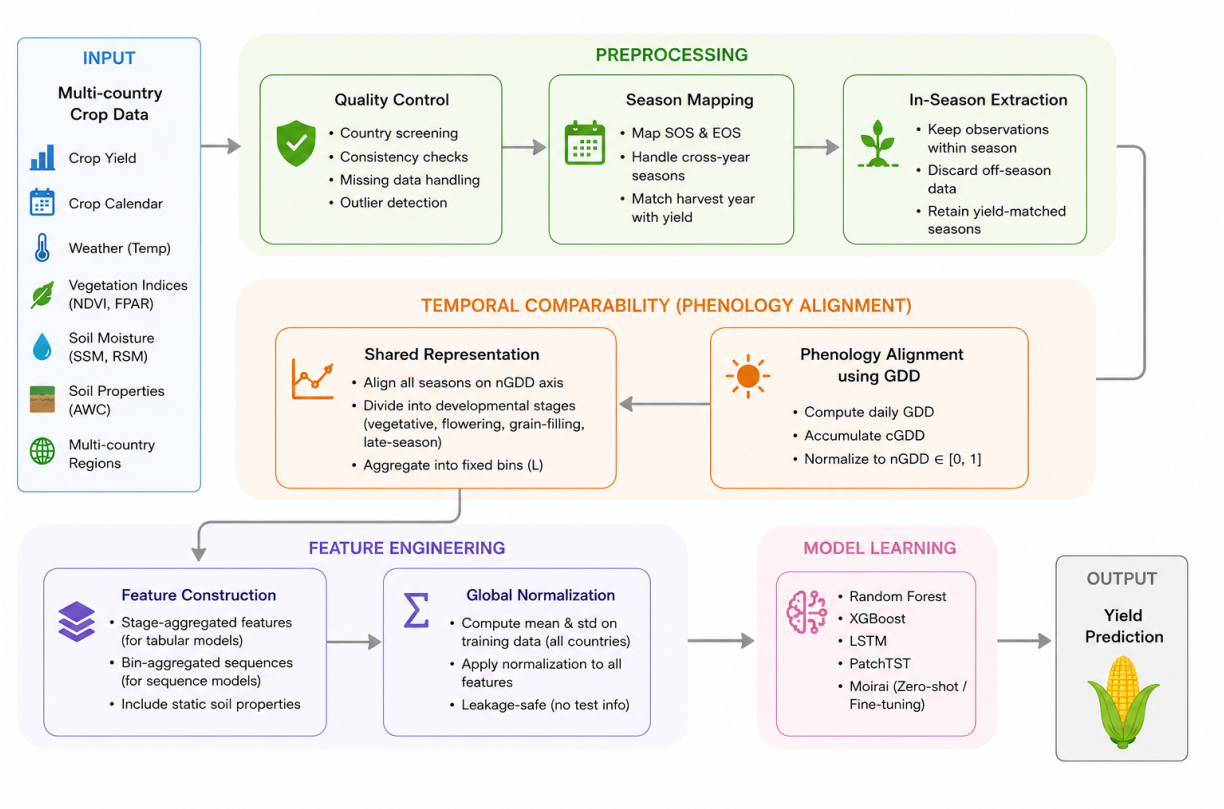


Figure 1.1: Overview of the proposed PhenoNorm

Chapter 5 shows the main conclusions of the thesis and highlights promising directions for future research.

Chapter 2

Literature Review

First, we review crop yield prediction as an agricultural forecasting problem, highlighting its role in precision agriculture. Second, we examine the use of machine learning and deep learning models for agricultural prediction, with a focus on multimodal time series modeling using weather, soil, and satellite-derived variables. Third, we discuss representation and preprocessing strategies for cross-region prediction, including temporal alignment, feature normalization, and domain adaptation techniques.

2.1 Crop Yield Prediction in Agricultural AI

2.1.1 Yield Prediction in Precision Agriculture

Precision agriculture aims to optimize crop management through site-specific decisions, where inputs such as fertilizer, irrigation, and seeding rates are adapted to local environmental and crop conditions [56]. Within this paradigm, crop yield prediction serves as a central task that links sensing infrastructure to decision-making processes. At the field scale, spatially explicit yield forecasts enable variable-rate management by identifying intra-field variability and guiding targeted interventions [2]. At larger scales, yield predictions support scenario analysis, allowing practitioners to evaluate the impact of alternative management strategies on productivity and resource use [48].

Modern precision agriculture systems integrate heterogeneous data sources, including satellite imagery, proximal sensors, weather observations, and farm management records, into unified analytical pipelines [2]. Within such systems, yield prediction operates as a downstream task that aggregates multi-source information into actionable estimates of production. These estimates are subsequently used in related tasks such as anomaly detection, stress monitoring, and decision support, making yield prediction one of the key components of data-driven agricultural workflows.

Beyond field-level applications, yield prediction also plays a critical role in regional and global decision-making. Forecasts inform food security planning, supply-chain logistics, and policy interventions designed to mitigate the impact of climate variability and market

fluctuations [28]. At these scales, accurate prediction is particularly challenging due to spatial heterogeneity and evolving environmental conditions, which motivate the use of data-driven approaches.

Importantly, precision agriculture does not imply direct control over all factors that affect crop yield. Environmental variables such as temperature, precipitation, humidity, and extreme-weather events cannot generally be controlled at the field scale; however, their observed and forecasted effects can provide early indicators of crop stress and production risk [45, 37]. Yield forecasts can therefore support data-driven decisions about controllable management actions, such as prioritizing field scouting, adjusting irrigation where infrastructure is available, targeting fertilizer or pest-management interventions, and planning harvest operations [56, 28]. At regional and national scales, the same forecasts can support the allocation of advisory resources, storage and logistics planning, insurance risk assessment, and food-security preparedness [28, 45].

2.1.2 Traditional Yield Prediction Approaches

Traditional yield prediction methods can be broadly categorized into process-based crop models and empirical statistical approaches. Process-based models simulate plant growth by explicitly tracking biophysical and physiological mechanisms, such as phenological development, photosynthesis, carbon assimilation, and water and nutrient uptake [45]. These frameworks integrate daily weather observations, specific soil profiles, and local management practices to mathematically estimate biomass accumulation and final yield. Due to their mechanistic nature, they offer high interpretability and have been widely utilized in agronomic research and climate impact assessments [9].

Despite their rigorous representation of complex biological interactions, process-based models are inherently constrained by their data requirements. They demand extensive, high-resolution input parameters and precise local calibration, which are often unavailable, highly uncertain, or computationally prohibitive to acquire at large spatial scales. This limitation severely restricts their applicability in operational, regional, or cross-country forecasting, particularly in data-sparse environments [45].

Empirical approaches, in contrast, bypass explicit biological simulation by learning direct statistical relationships between observed environmental predictors and final yield outcomes. Early predictive models predominantly relied on linear or polynomial regressions applied to aggregated seasonal variables, such as total precipitation and average temperature [39]. To capture the complex realities of agricultural systems, these classical statistical frameworks commonly incorporated nonlinear components, such as quadratic terms for thermal and water-related indicators, alongside general time trends [71].

By employing these nonlinear functional forms, researchers could effectively model the parabolic response of crops to environmental drivers without needing a formal mathematical equation. For instance, such statistical designs inherently reflect the biological reality that increases in heat and moisture are highly beneficial for crop development up to an optimal physiological threshold. Once that optimum is exceeded, further temperature increases or excess rainfall rapidly become detrimental to the final yield. Consequently,

these statistical approaches have been instrumental in identifying critical environmental thresholds and quantifying the risks of climate extremes [54].

However, traditional empirical models remain heavily reliant on aggregated seasonal features, which fundamentally limits their capacity to capture fine-grained intra-seasonal dynamics and temporal variability. Because they compress dynamic weather events into broad seasonal summaries, they often fail to account for the precise timing of environmental stressors, such as a brief but severe heatwave occurring specifically during the highly sensitive flowering stage. This inability to natively process complex, temporally ordered growth patterns has strongly motivated the transition toward more flexible, data-driven machine learning and sequence-modeling approaches.

2.1.3 Key Predictive Features and Their Definition

Modern yield prediction models rely on a combination of temperature-based, vegetation-based, and moisture-related indicators to characterize crop growth conditions over time [28]. In this study, five of the primary dynamic variables are used: average temperature (Tavg), normalized difference vegetation index (NDVI), fraction of absorbed photosynthetically active radiation (FPAR), surface soil moisture (SSM), and root-zone soil moisture (RSM) [36]. Alongside these dynamic features, static soil properties such as available water capacity (AWC) are incorporated to provide a baseline for local soil characteristics.

Thermal conditions are represented using average temperature (*Tavg*). Tavg corresponds to the daily mean air temperature, and is computed as:

$$T_{\text{avg}} = \frac{T_{\max} + T_{\min}}{2}, \tag{2.1}$$

where $T_{\max}$ and $T_{\min}$ denote the daily maximum and minimum air temperatures, respectively.

Vegetation dynamics are represented using NDVI and FPAR, both derived from remote sensing products. NDVI is defined as:

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{red}}}{\rho_{\text{NIR}} + \rho_{\text{red}}}, \tag{2.2}$$

where ρ_{NIR} and ρ_{red} denote reflectance in the near-infrared and red spectral bands, respectively. Healthy vegetation reflects strongly in the near-infrared region and absorbs strongly in the red region due to chlorophyll activity, meaning higher NDVI values generally indicate denser and more active vegetation [20].

FPAR measures the fraction of incoming photosynthetically active radiation absorbed by the vegetation canopy. Unlike NDVI, which primarily reflects canopy greenness, FPAR more directly captures radiation interception and photosynthetic activity. Together, NDVI and FPAR provide complementary information about crop growth, canopy structure, and biomass accumulation.

Mathematically, FPAR is defined as:

$$\mathrm{FPAR} = \frac{\mathrm{APAR}}{\mathrm{PAR}}, \tag{2.3}$$

where APAR denotes the absorbed photosynthetically active radiation and PAR represents the incoming photosynthetically active radiation.

Water availability is represented using a combination of dynamic moisture indicators (SSM and RSM) [11] and time-invariant soil properties (AWC) [3]. SSM describes moisture conditions in the upper soil layer and reflects short-term responses to rainfall and evaporation, while RSM captures deeper subsurface moisture accessible to plant roots over longer periods. These dynamic variables are particularly important in rainfed agricultural systems, where water stress strongly influences crop productivity. To complement these temporal observations, AWC is included as a static measure that defines the inherent capacity of the soil to store plant-available water. By incorporating AWC as a time-invariant predictor, models can better contextualize the dynamic moisture variables, as identical rainfall or surface moisture levels will have vastly different impacts on crop stress depending on the soil's underlying water retention capacity.

Empirical studies show that fusing weather, vegetation indices, and soil properties typically yields better performance than using any single source alone [28, 36]. For example, combinations of MODIS-based NDVI metrics with seasonal temperature and precipitation summaries have been used to forecast regional yields in the Canadian Prairies and other large production regions with good accuracy several weeks before harvest [36]. Similar multi-source feature sets have been adopted in field-scale experiments, where multi-spectral and structural features from UAVs are combined with meteorological and soil information to map within-field yield variability [2, 73, 31]. In most of these works, the design emphasis is on identifying informative data sources and summary statistics, while the temporal structure of growth is handled through seasonal aggregates or heuristic time series descriptors [28, 68].

2.2 Machine Learning for Agricultural Prediction

In machine learning settings, crop yield prediction is typically formulated as an end-of-season regression problem, where yields are estimated from environmental and physiological indicators observed during the growing season [38]. Common predictors include meteorological variables, vegetation indices, and soil-related attributes, which together capture the temporal evolution of crop growth and environmental stressors [36, 20]. While these predictors have demonstrated strong empirical performance, their effectiveness depends critically on how temporal information is represented and aggregated, an aspect that remains relatively underexplored in multi-region settings.

2.2.1 Machine Learning in Yield Prediction

ML has become a central component of agricultural AI, providing flexible models capable of integrating heterogeneous data sources while capturing complex, nonlinear relationships

between environmental conditions, management practices, and crop performance [62, 45]. In contrast to purely process-based crop models, ML approaches can learn directly from historical yield observations and environmental covariates, making them particularly useful in settings where large volumes of observational data are available but detailed information on cultivar characteristics and management practices remain incomplete or uncertain [38, 36]. Within this broader trend, crop yield prediction has emerged as one of the most active application areas, alongside crop-type mapping, phenology detection, and disease and stress monitoring [44, 56].

A wide range of ML architectures have been explored for crop yield prediction, ranging from classical ensemble methods to modern deep learning and sequence modeling approaches [28, 44]. Classical models such as linear regressions have gradually been complemented by tree-based ensembles, kernel methods, and neural networks (NN), reflecting the increasing complexity and volume of available agricultural data [45, 40]. Studies consistently report that ensemble methods (e.g., Random Forests and Gradient Boosting) and deep-learning architectures (e.g., convolutional and recurrent networks) tend to outperform simple baselines when sufficient training data and engineered features are available [62].

Among these, tree-based ensemble models remain widely used due to their robustness, interpretability, and strong performance on heterogeneous agricultural datasets containing mixed environmental and remote-sensing features.

ML models are deployed at multiple spatial and temporal scales. At field and farm scale, they are used to exploit dense sensor networks, high-resolution drone imagery, and machinery telematics[1] as inputs to yield and biomass prediction models [2, 73]. At regional and continental scales, they combine gridded weather data, satellite-derived vegetation indices, and soil maps to generate subnational yield forecasts that support market and policy decisions [38, 36]. Across these scales, the common objective is to convert heterogeneous sensing data into predictive representations that capture crop-development dynamics and environmental variability for use in supervised learning models [44, 56].

These advances have led to the adoption of a broad range of ML architectures for crop yield prediction:

Random Forest

Random Forest (RF) is an ensemble learning algorithm that constructs a large collection of decision trees during training and combines their predictions through averaging [5]. Each tree is trained using a bootstrap sample of the training data, while random subsets of features are considered at each split. This combination of bagging and random feature selection reduces overfitting and improves generalization performance.

RF models are particularly well-suited for agricultural prediction tasks because they can capture nonlinear relationships and interactions among heterogeneous environmental variables without requiring strong assumptions about data distributions. In crop yield

[1]Machinery telematics: A system integrating GPS, sensors, and cellular/satellite communication to remotely monitor heavy equipment's location, health, and performance in real time.

prediction, RF has been widely applied to integrate meteorological variables, vegetation indices, and soil information, often demonstrating strong robustness across diverse geographic regions [38, 45].

XGBoost

Extreme Gradient Boosting (XGBoost) is a scalable implementation of gradient-boosted decision trees [10]. Unlike bagging-based methods such as Random Forest, boosting methods construct trees sequentially, where each new tree attempts to correct the prediction errors of the previous ensemble.

XGBoost incorporates several algorithmic improvements, including regularization, sparse-aware learning, parallelized tree construction, and optimized handling of missing values [10]. These properties allow the method to achieve high predictive accuracy while maintaining computational efficiency on large datasets.

In agricultural applications, XGBoost has been used for crop yield forecasting, drought analysis, and environmental modeling tasks involving heterogeneous climate and remote-sensing data [40, 62]. Its ability to model complex nonlinear relationships and interactions between environmental drivers makes it particularly suitable for multi-source agricultural datasets. In many comparative studies, XGBoost achieves performance competitive with or superior to traditional ensemble methods and shallow neural networks.

2.2.2 Sequence Models and Growth-Pattern Learning

The increasing availability of dense time series data has motivated the use of sequence models for agricultural prediction, particularly for tasks that depend on within-season dynamics [45, 44]. RNNs and LSTM networks have been widely applied to crop yield forecasting, where they process temporally ordered sequences of weather and remote-sensing features to learn dependencies that cannot be captured by static seasonal summaries alone [2, 29]. In these models, each time step corresponds to a day or other fixed interval, and the network learns how patterns of temperature, moisture, and vegetation indices over the season relate to final yield [45, 36].

More recently, transformer-based architectures and hybrid CNN–RNN models have been explored to better capture long-range dependencies and multi-scale temporal patterns in crop growth [28, 46]. Reviews of deep-learning methods for yield prediction report that LSTM, CNN, and their combinations (e.g., CNN–LSTM) are among the most commonly used architectures, often paired with NDVI time series, radar backscatter, or fused optical–radar sequences [62]. These models can, in principle, learn rich growth-pattern representations from raw or lightly processed time series, but in practice they are usually trained on calendar-aligned sequences defined by fixed time steps (for example, day-of-year or image acquisition dates) [46, 45]. As a result, differences in crop calendars and season length across regions are often handled implicitly or ignored, and the learned patterns may reflect a mixture of phenological signals and dataset-specific timing artifacts [28, 44].

Long Short-Term Memory Networks

LSTM networks are a specialized form of RNN introduced to address the difficulty of learning long-range temporal dependencies [24]. LSTMs use memory cells and gating mechanisms to regulate the flow of information through time, allowing relevant signals from earlier stages of a sequence to influence later predictions.

In agricultural yield prediction, LSTM models are commonly applied to sequential weather observations, vegetation indices, and soil-moisture measurements collected throughout the growing season [2, 29]. Their sequential processing structure enables the model to learn temporal relationships between crop-development stages and environmental conditions, including delayed effects of drought, heat stress, or vegetation growth patterns.

Because LSTM models operate directly on ordered temporal inputs, they are sensitive to how seasonal sequences are aligned and represented. Differences in planting dates, harvest timing, or season duration across regions may therefore affect the temporal consistency of the learned representations. This characteristic makes LSTM-based approaches particularly relevant for evaluating phenology-aligned representations.

2.2.3 Time Series Modeling and Transferability Limitations

Time series forecasting has been extensively studied in a wide range of domains, including finance, energy systems, transportation, healthcare, and industrial monitoring [67]. In particular, transformer variants have achieved strong performance in long-term forecasting tasks by capturing long-range dependencies and complex temporal interactions [59, 75].

Despite these advances, most time series models developed for these domains are tailored to data regimes that differ significantly from agricultural yield prediction. For example, financial and energy forecasting often rely on high-frequency, densely sampled signals with relatively consistent temporal resolution, while transportation and sensor-based applications benefit from continuous or near-continuous monitoring [59]. In contrast, crop yield prediction is characterized by low-frequency observations and highly irregular temporal dynamics driven by environmental, biological, and management processes.

Moreover, many state-of-the-art time series models assume relatively stationary or weakly non-stationary processes, where historical patterns are assumed to generalize to future observations [67]. However, agricultural systems exhibit strong non-stationarity due to climate variability, changing agronomic practices, and inter-annual variability in weather extremes. These characteristics limit the direct applicability of models trained on generic benchmarks such as electricity demand or traffic flow datasets.

Some studies on transformer-based time series forecasting further highlight that model effectiveness is highly domain-dependent, and that simple architectures can outperform complex deep models when inductive biases do not align with the underlying data-generating process [72]. This observation reinforces the need for careful adaptation when transferring time series methodologies across domains.

As a result, although general-purpose time series forecasting models provide a strong methodological foundation, their direct application to crop yield prediction is limited. Agricultural forecasting requires models that can explicitly handle sparse temporal sampling, strong seasonal effects, and multi-source exogenous drivers, which are not fully addressed in standard time series benchmarks or architectures developed for other application domains.

Patch Time Series Transformer

In the broader time series forecasting literature, patch-based transformer designs have also been used to better capture dynamic correlations across subsequences and to improve robustness in long-horizon settings [12].

Patch Time Series Transformer (PatchTST) is a transformer-based architecture for multivariate time series forecasting [47]. Instead of processing individual time steps directly, PatchTST divides the input sequence into smaller temporal patches, which are then embedded and processed using self-attention mechanisms. This patching strategy reduces computational complexity while enabling the model to capture both local and long-range temporal dependencies.

Given an input sequence, PatchTST divides it into overlapping or non-overlapping temporal patches, which are then processed by Transformer encoder layers. Unlike recurrent models, transformers process all sequence positions simultaneously through attention operations, allowing them to model complex temporal interactions more efficiently over long sequences. PatchTST additionally employs channel-independent processing, where each variable is treated as a separate univariate sequence before information is combined across channels [47].

In agricultural applications, transformer-based models have increasingly been explored for crop yield prediction and environmental time series analysis because of their ability to capture multi-scale seasonal patterns from remote sensing and climate data [46, 28].

In parallel, recent work has explored large-scale pretrained time series foundation models that aim to learn generalizable temporal representations across domains. Models such as TimeGPT [22], TimesFM [16], and Moirai [69] are trained on diverse collections of time series data and can be applied in zero-shot or fine-tuned settings for downstream forecasting tasks. These models leverage transformer architectures and large-scale pretraining to capture generic temporal patterns, reducing the need for task-specific feature engineering. While early results in domains such as finance, energy, and demand forecasting are promising, their application to agricultural forecasting remains relatively underexplored.

Moirai

Moirai is a transformer-based time series foundation model designed for zero-shot and transfer-learning forecasting across diverse domains [69]. Unlike task-specific forecasting models trained on a single dataset, Moirai is pretrained on large collections of heterogeneous time series data and then adapted to downstream prediction tasks with minimal additional training.

The architecture follows the transformer paradigm, where temporal dependencies are modeled through self-attention operations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.4}$$

where Q, K, and V denote query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

Given an input sequence $X = (x_1, x_2, \ldots, x_T)$, the model learns a contextual representation $H = (h_1, h_2, \ldots, h_T)$, where each hidden representation h_t captures information from multiple temporal positions through attention-based interactions.

A distinguishing characteristic of Moirai is its use of large-scale pretraining across multiple time series domains, enabling the model to learn generalized temporal representations that can transfer to unseen datasets and forecasting tasks [69]. In agricultural prediction settings, this capability is relevant because crop yield datasets often exhibit limited labeled samples, strong regional heterogeneity, and varying environmental conditions.

2.2.4 Transfer Learning and Generalization Gaps

Training machine learning models that generalize across countries or ecological regions remains an underexplored but increasingly important direction in agricultural forecasting. Unlike standard supervised learning settings, cross-country yield prediction is characterized by strong domain shift. In addition, labeled data availability is highly uneven across regions, making it difficult to train robust global models without introducing bias toward data-rich countries.

Improving generalization across regions, seasons, and management systems has led to growing interest in transfer learning and domain adaptation for yield prediction and related agricultural tasks [51, 25]. Transfer-learning approaches typically pre-train deep networks on data-rich crops or regions and then fine-tune them on target domains with limited labeled data [29, 7]. These strategies have been applied to winter wheat and other cereals using multi-source remote sensing and climate data, where multi-task and 3D CNN–LSTM architectures are used to transfer information between crops and sites [19, 7]. Domain-adaptation methods, including adversarial and discrepancy-based techniques, explicitly aim to align feature distributions between source and target regions so that a model trained in one environment can maintain performance in another [51, 37, 26].

At the same time, work on domain adaptation and transfer learning in agriculture has largely focused on bridging differences in image appearance, sensor characteristics, or climate regimes. Studies on crop-type mapping and weed recognition, for example, show that adversarial domain adaptation can reduce performance drops when models are applied to new fields with different backgrounds or illumination conditions [27, 26]. For yield prediction specifically, recent contributions highlight that deep models trained with standard random splits may have limited transferability when evaluated under spatial or temporal hold-out designs, and that validation strategies have a strong impact on the perceived robustness of learned models [58, 37, 43].

While these methods provide powerful mechanisms for improving generalization, they reach their full potential when applied to representations specifically designed for cross-region consistency. By addressing temporal and scale misalignments at the source, these strategies can more effectively isolate true phenological patterns from region-specific artifacts. This highlights a significant opportunity to enhance cross-country generalization through a principled approach to representation design, where precise temporal alignment serves as the foundation for more robust model adaptation.

2.3 Representations for Cross-Country Yield Prediction

Beyond model architecture, the way features are scaled and temporally aligned plays a critical role in cross-country generalization [35, 45].

2.3.1 Phenology-Based and Growth-Stage Representations

A long tradition in crop science represents crop development using thermal time and phenological stages [54]. Growing Degree Days (GDD) [41] are widely used to index crop progress, and agronomic guidelines routinely map cumulative GDD to key stages such as emergence, vegetative growth, flowering, grain filling, and maturity [34, 39].

GDD quantifies accumulated heat above a base threshold required for crop development:

$$\mathrm{GDD}_t = \max(T_{\mathrm{avg},t} - T_{\mathrm{base}},\, 0)\,. \tag{2.5}$$

Here, $T_{\mathrm{avg},t}$ denotes the average temperature at time t, and T_{base} is the minimum temperature required for crop growth. The max operator ensures that temperatures below the threshold contribute zero.

Remote-sensing studies often adopt phenology-aligned representations by aligning vegetation index trajectories to estimated growth stages or by extracting stage-specific metrics, such as peak greenness and senescence timing, that correspond to transitions aligned with known crop-development patterns [73, 76].

In this study, we select maize as our primary focus crop to evaluate the proposed pipeline. For maize specifically, cumulative GDD thresholds are commonly used to approximate vegetative development stages. Agronomic guidelines report that maize typically requires approximately 100–120 GDDs for emergence after planting, followed by roughly 82–85 GDDs per additional leaf collar through the last growth stage [34]. Consequently, cumulative thermal time can provide a practical approximation of crop developmental progression throughout the season. However, the exact GDD requirements may vary across cultivars and environmental conditions, and stress factors such as drought can alter growth rates, making GDD-based staging an approximate rather than exact representation of phenology.

In the yield-prediction literature, these ideas are typically incorporated through engineered features. Many empirical models include seasonal GDD, heat-stress indices, and their interactions with precipitation or soil moisture to capture temperature thresholds and combined environmental stresses [54, 53]. Other studies aggregate vegetation indices and moisture variables over predefined growth stages (for example, vegetative, reproductive, and grain-filling phases), improving the interpretability of predictors in relation to plant physiology [20, 76]. However, these approaches are usually defined within individual regions or datasets and do not explicitly enforce a shared notion of developmental progress across countries with different crop calendars and season lengths [28, 58].

The key objective is to construct a temporal axis that reflects crop developmental progress, so that both tabular and sequence-based models operate on inputs corresponding to comparable growth stages across countries [34, 76].

2.3.2 Normalization and Scaling in Heterogeneous Time Series

Normalization and feature scaling fundamentally shape how ML models interpret variability across heterogeneous datasets. In time series learning, preprocessing choices directly influence optimization stability, feature geometry, and the statistical relationships learned by predictive models [35, 52]. Benchmarking studies in forecasting and sequence modeling have shown that normalization strategies, including global z-score normalization, min–max scaling, and per-series standardization, can substantially alter predictive performance and model robustness, particularly when data distributions vary across domains or temporal regimes [35, 52].

In remote sensing and agricultural modeling, normalization is commonly applied to account for differences in sensor characteristics, environmental conditions, and regional variability [36]. Many yield-prediction studies standardize features independently within regions, fields, or years to stabilize training and reduce local variance [30]. While this approach improves numerical conditioning and prevents variables with larger magnitudes from dominating optimization, it may also suppress meaningful large-scale differences associated with climate regimes, soil properties, or management practices [38]. Consequently, region-wise normalization can inadvertently remove part of the signal required for cross-country generalization.

Conversely, global normalization preserves inter-regional contrasts by computing scaling statistics across all training samples [45, 11]. Such representations retain large-scale climatic and environmental gradients, which may benefit transferability across regions. However, global statistics may also become biased toward dominant countries or highly represented production systems, leading to feature spaces that underrepresent smaller or climatically distinct regions [58]. Similar concerns have been discussed in domain adaptation literature, where distributional imbalance between source and target domains can negatively affect generalization performance [26].

From a modeling perspective, normalization can be interpreted as an implicit form of domain adaptation. The chosen scaling strategy defines the feature space in which samples from different regions are compared and therefore determines which patterns are treated

as similar or distinct. Recent studies in representation learning emphasize that preprocessing decisions influence both optimization dynamics and the geometry of learned latent spaces [35, 52]. In cross-country agricultural prediction tasks, these effects become especially important because climatic distributions, management practices, and phenological trajectories differ substantially across regions.

These observations suggest that normalization should be designed jointly with the temporal representation rather than treated as an isolated preprocessing step. Applying consistent scaling to phenology-aligned representations may help disentangle transferable agronomic signals from region-specific scale effects, thereby reducing the tendency of models to learn country-dependent offsets instead of generalizable growth patterns [58]. This perspective motivates the representation pipeline explored in this study.

2.3.3 Shared Representations for Cross-Domain Generalization

Beyond conventional feature scaling, temporal alignment also influences how heterogeneous agricultural time series are compared across regions. Crop observations exhibit strong seasonality, differing planting schedules, and variable growth durations across countries and production systems [43, 58]. Consequently, aligning sequences purely by calendar date may associate observations from substantially different developmental stages, introducing inconsistencies into the learned representation. To mitigate these effects, several remote-sensing studies incorporate phenology-aware preprocessing strategies, including temporal smoothing, stage-based aggregation, and alignment using estimated crop-development indicators [73, 76]. These approaches improve the physiological interpretability of temporal signals.

Related challenges have also been studied in domain adaptation and transfer learning. Existing approaches commonly focus on the model level by learning domain-invariant representations or aligning feature distributions between source and target domains through adversarial or discrepancy-based objectives [51, 26]. In agricultural applications, such methods have been explored for tasks including yield prediction, crop mapping, and crop–weed discrimination, demonstrating that reducing distributional mismatch can improve transferability across environments [27, 7].

Despite these advances, temporal representations in many existing pipelines remain organized around acquisition dates or fixed sampling intervals, while normalization is typically applied within individual datasets or regions. Under these representations, differences in season length and crop progression are not explicitly encoded, requiring models to infer such relationships implicitly. Studies evaluating spatial and temporal transferability report that models performing well under random train–test splits often degrade substantially under cross-region or cross-year evaluation settings [58, 43], indicating that representation-level inconsistencies may remain an important source of domain shift.

Table 2.1 summarizes the key limitations of existing agricultural forecasting representations and their implications for cross-country generalization.

From this perspective, representation design directly influences cross-domain generalization. Pipelines that enforce both temporal and statistical consistency across regions aim

Table 2.1: Comparison of existing approaches and their limitations under cross-country shift

Approach	**Strength**	**Limitation**	**Cross-Country Issue**
Calendar-based alignment	Simple and widely used	Ignores phenological variation	Temporal stage mismatch across regions
Region-wise normalization	Stabilizes local training	Removes inter-region variability	Weak shared representation across countries
Seasonal aggregation	Reduces noise	Loses temporal dynamics	Weak stage-specific representation
Deep sequence models	Captures temporal dependencies	Sensitive to representation quality	Transferability degradation under domain shift
Foundation models	Strong large-scale temporal priors	Limited agricultural adaptation	Domain mismatch in sparse seasonal data

to construct an input space in which samples from different countries are directly comparable. Such representations can complement sequence-modeling approaches by reducing spurious variation before learning takes place.

Chapter 3

PhenoNorm Workflow and Methods

This chapter begins with a description of the multi-country dataset, including data requirements, preprocessing steps, and quality-control procedures. It then introduces the phenology-aligned representation based on growing degree days and presents the feature scaling strategies considered in this work, including global, country-wise, and bin-wise normalization. It then describes the predictive models, including Random Forest, XGBoost, LSTM, PatchTST, and the Moirai foundation model. Finally, it presents the common training and evaluation protocol used to ensure a fair comparison across models.

3.1 Dataset Description

3.1.1 Data Requirements and Applicability

PhenoNorm is designed to be broadly applicable, provided that certain key data characteristics are satisfied. In particular, this pipeline requires temporally resolved datasets that can be aligned with crop development stages.

At a minimum, the input data must include:

- Time series observations at a daily or near-daily resolution, enabling the construction of temporally consistent sequences.
- Calendar-based information, such as sowing and harvesting dates or crop calendar records, to define the crop season and restrict analysis to in-season periods.
- Temperature data, including either daily minimum and maximum temperatures or daily average temperature, to compute GDD and support phenology-based alignment.
- Environmental or remotely sensed variables (e.g., vegetation indices, soil moisture, or climate indicators) that capture crop growth dynamics over time.

Given these components, the pipeline can be applied to any dataset where crop development can be represented as a function of accumulated thermal time. This enables consistent alignment of crop growth stages across regions with differing climatic conditions and planting schedules.

Importantly, the pipeline is not restricted to a specific data source or geographic region. As long as the required temporal resolution, temperature variables, and seasonal context are available, this method can be extended to other crops, countries, or integrated data products with minimal modification.

Extending the pipeline to a new crop would require crop-specific parameterization. In particular, the crop calendar, growing-degree-day base temperature, expected thermal requirements, and phenological-stage boundaries should be selected according to the target crop and, where possible, local cultivar and management conditions. The feature set can also be expanded when suitable data are available. Relevant additions include higher-resolution static soil properties, irrigation occurrence or volume, fertilizer and other management indicators, crop-specific phenology observations, drought and heat-stress indices, and indicators of extreme events such as heat waves, flooding, or damaging precipitation. These variables should be aligned to the target crop season, subjected to the same quality-control procedures, and scaled using statistics estimated from the training partition only.

However, a limitation of the proposed pipeline is its dependence on reliable crop calendar and temperature information. In regions where sowing dates, harvest dates, or daily temperature observations are unavailable or highly uncertain, this method may become less accurate. Similarly, crops whose development is weakly correlated with thermal accumulation may require alternative alignment strategies.

3.1.2 Benchmark Dataset for Evaluation

To evaluate the pipeline, a representative multi-country crop yield benchmark dataset is used. This study uses CY-Bench [50], which provides sub-national crop yield data combined with environmental and soil-related predictors, and serves as a standardized reference for comparison with existing approaches. The dataset integrates multiple data sources, including crop yield statistics, crop calendar information, meteorological variables, remote sensing products, and static soil properties. In this study, the maize subset of the benchmark is considered.

For each region–year sample, the target variable is end-of-season yield, typically reported in tonnes per hectare. The associated predictors include:

- Meteorological variables such as near-surface air temperature and related climate indicators, obtained from the AgERA5 dataset [4];
- Remotely sensed vegetation indices, including NDVI derived from MOD09CMG [64] and FPAR products provided through the European Commission Joint Research Centre (JRC) [17];

- Soil moisture indicators, including SSM and RSM, obtained from the GLDAS hydrological dataset [55];
- Static soil properties such as available water capacity (AWC), derived from the WISE Soil Database [3];
- Crop calendar and crop mask information obtained from the WorldCereal initiative [21, 63].

These data sources provide complementary information about crop growth conditions, capturing temperature dynamics, vegetation development, and water availability across diverse agro-ecological regions, while also enabling comparison against established baselines defined within the benchmark.

3.1.3 Quality Control

To ensure reliability of the dataset, a quality-control step is applied across two levels:

Country-level screening

Country directories are first scanned for the presence of all required data, including meteorological records, crop calendar, and yield observations. Countries missing any of these are excluded prior to any modeling. Among the remaining countries, yield series are evaluated against three quantitative criteria: a mean yield greater than 0.05 t/ha, a yield standard deviation greater than 0.01 t/ha, and a zero-valued observation fraction below 90%. Countries failing any of these thresholds are classified as invalid and excluded from the modeling dataset. This step removes yield series that are too sparse, near-constant, or dominated by zero entries, which would provide no discriminative signal for regression.

Feature-level validation

Missingness and value ranges are examined across all dynamic features, including average temperature, NDVI, FPAR, SSM, and RSM. Missing values within retained seasons are carried through the pipeline and represented via a binary observation mask, which is used directly by the sequence models to distinguish observed from imputed positions. Static soil features, such as available water capacity (AWC), are imputed with the country-level median where missing, as their absence would otherwise exclude entire administrative regions from modeling. Outliers are identified using conservative statistical thresholds but are not removed automatically; they are treated as part of the natural variability inherent in a multi-country agricultural dataset unless values are clearly implausible (e.g., negative NDVI below -0.3 or temperature values outside the physiological range of the crop). Overall missingness across the retained dataset remains low, and no feature exhibits excessive sparsity after season filtering.

3.1.4 Season Mapping

For each administrative region, crop season boundaries are defined using crop calendar information containing the start-of-season (SOS) and end-of-season (EOS) day-of-year values. Daily meteorological observations are matched to the corresponding growing season by determining whether each observation date falls within the valid seasonal interval.

Both standard and cross-year growing seasons are supported. For seasons where the growing cycle extends across calendar years, early-year observations are assigned to the preceding season start year to preserve continuity of crop development. The harvest year associated with each season is then determined using the seasonal structure and matched against available yield records.

In-Season Extraction

Only region–season pairs with corresponding yield observations are retained. Dynamic predictors outside the identified crop season interval are discarded prior to feature extraction and temporal aggregation. This process ensures that retained observations align with crop-development periods while reducing the influence of off-season environmental variability.

3.2 Temporal Representation Construction

This section describes how raw environmental time series are transformed into a phenology-aligned representation suitable for cross-country comparison and model input construction. The overall pipeline is summarized in Figure 3.1.

3.2.1 Cumulative Growing Degree Days

To account for differences in crop development rates across environments, thermal time is used instead of calendar time. GDD is computed as:

$$\mathrm{GDD}_t = \max\Big(\frac{T_{\max,t} + T_{\min,t}}{2} - T_{\text{base}},\, 0\Big), \tag{3.1}$$

where $T_{\max,t}$ and $T_{\min,t}$ are daily maximum and minimum temperatures, and T_{base} is the base temperature for crop growth, and t denotes the current day within the growing season.

Daily GDD values are accumulated over time within each region and season to produce cumulative GDD trajectories:

$$\mathrm{cGDD}_t = \sum_{\tau=1}^{t} \mathrm{GDD}_\tau. \tag{3.2}$$

The resulting cumulative trajectories provide a representation of thermal progression throughout each growing season.

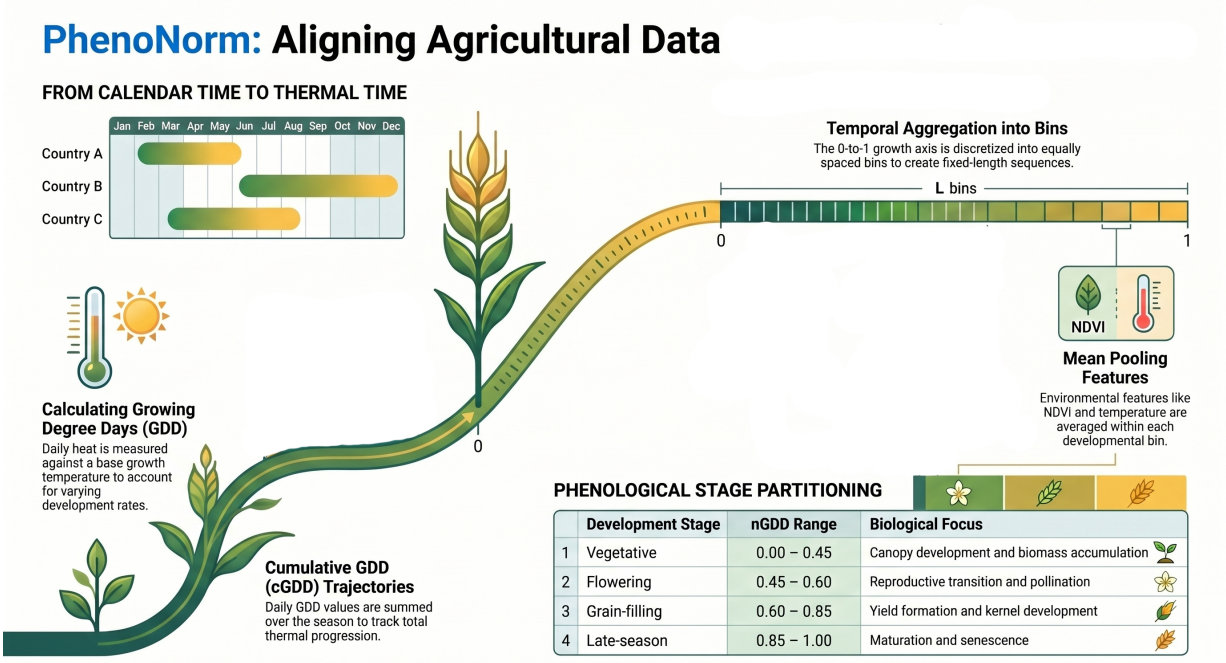


Figure 3.1: Overview of the phenology-aligned temporal representation construction process.

3.2.2 Phenology-Aligned Normalization

For each crop season, cumulative GDD is then normalized to obtain a relative developmental index:

$$\mathrm{nGDD}_t = \frac{\mathrm{cGDD}_t}{\mathrm{cGDD}_T}, \tag{3.3}$$

where T is the final day of the crop season. This produces a normalized scale from 0 to 1 representing crop development progress.

Using nGDD, all temporal observations are re-indexed along a common developmental axis. This transformation aligns crop seasons across countries by expressing time in terms of relative growth progression.

3.2.3 Phenological Stage Partitioning

The resulting representation allows observations from different regions to be compared at approximately equivalent phenological stages. In the proposed pipeline, the normalized cumulative GDD trajectory is divided into four broad developmental stages:

- **Vegetative stage** ($0.00 \leq n_{\mathrm{GDD}} < 0.45$), representing early canopy development and biomass accumulation;

- **Flowering stage** ($0.45 \leq n_{\mathrm{GDD}} < 0.60$), corresponding to reproductive transition and pollination-related growth processes;
- **Grain-filling stage** ($0.60 \leq n_{\mathrm{GDD}} < 0.85$), during which assimilates are allocated toward kernel development and yield formation;
- **Late-season stage** ($0.85 \leq n_{\mathrm{GDD}} \leq 1.00$), representing maturation and senescence near harvest.

The stage boundaries are intended as approximate phenological partitions rather than exact physiological transition points. They were selected based on commonly reported maize developmental progression patterns in agronomic literature [34], where vegetative growth occupies the largest portion of cumulative thermal accumulation, followed by reproductive transition, grain filling, and maturation stages. Because phenological progression varies across cultivars, climatic conditions, and management practices, the proposed boundaries should be interpreted as heuristic approximations designed to provide consistent cross-country temporal alignment rather than precise biological stage delineation.

3.2.4 Temporal Aggregation into Bins

To construct fixed-length representations suitable for sequence models, the nGDD axis is discretized into L equally spaced bins over the interval $[0, 1]$. Multiple values of L are evaluated experimentally to study the effect of temporal resolution on predictive performance.

Dynamic variables are then grouped according to their associated nGDD bin. Within each bin, feature values are aggregated using mean pooling:

$$\bar{x}_{b,f} = \frac{1}{N_b} \sum_{o \in b} x_{o,f}, \tag{3.4}$$

where $x_{o,f}$ denotes the value of feature f for observation o, and N_b is the number of observations assigned to bin b.

The resulting representation forms a fixed-length multivariate sequence of shape $L \times F$, where F denotes the number of dynamic features. This binning strategy preserves the relative progression of crop development while enabling direct comparison between seasons with different calendar lengths or planting schedules.

After aggregation, yield observations are joined using the corresponding (region_id, harvest_year) pair.

3.3 Feature Scaling Strategies

This section defines the normalization strategies considered in this study. Because feature distributions vary across countries and across the seasonal axis, different scaling approaches may affect model behavior and cross-country generalization.

All normalization statistics are computed using training data only and are applied unchanged to validation and test sets to ensure a leakage-safe pipeline.

3.3.1 Global Normalization

In global normalization, feature-wise means and standard deviations are computed by pooling all training samples across countries. Each feature is then standardized using these global statistics. This approach preserves relative differences between countries while placing all features on a common scale.

For feature dimension j, global normalization is defined as:

$$\tilde{x}_{c,i,t,j} = \frac{x_{c,i,t,j} - \mu_j^{\text{global}}}{\sigma_j^{\text{global}}}, \tag{3.5}$$

where $x_{c,i,t,j}$ denotes the original value of feature j for country c, sample i, and time step t, while $\tilde{x}_{c,i,t,j}$ denotes the normalized value. The quantities μ_j^{global} and σ_j^{global} represent the global mean and standard deviation of feature j, respectively, computed across all countries, samples, and time steps in the training set.

The global mean for feature j is given by:

$$\mu_j^{\text{global}} = \frac{1}{\sum_{c \in \mathcal{C}_{\text{train}}} N_c T_c} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{i=1}^{N_c} \sum_{t=1}^{T_c} x_{c,i,t,j}, \tag{3.6}$$

where $\mathcal{C}_{\text{train}}$ denotes the set of training countries, N_c is the number of samples associated with country c, and T_c is the number of temporal observations for each sample in country c.

The corresponding global standard deviation is:

$$\sigma_j^{\text{global}} = \sqrt{\frac{1}{\sum_{c \in \mathcal{C}_{\text{train}}} N_c T_c} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{i=1}^{N_c} \sum_{t=1}^{T_c} \left(x_{c,i,t,j} - \mu_j^{\text{global}} \right)^2}. \tag{3.7}$$

3.3.2 Country-wise Normalization

In country-wise normalization, feature-wise means and standard deviations are computed separately for each training country using only the training partition. Samples belonging to countries observed during training are normalized using the statistics corresponding to their country. For unseen test countries, where country-specific statistics are unavailable, the global training statistics are applied instead. This design preserves a leakage-safe evaluation protocol while still allowing country-adaptive scaling for countries represented in the training data.

For feature dimension j in country c, country-wise normalization is defined as:

$$\tilde{x}_{c,i,t,j} = \frac{x_{c,i,t,j} - \mu_{c,j}}{\sigma_{c,j}}, \tag{3.8}$$

where $x_{c,i,t,j}$ denotes the original feature value for country c, sample i, time step t, and feature j, while $\tilde{x}_{c,i,t,j}$ denotes the normalized value. The quantities $\mu_{c,j}$ and $\sigma_{c,j}$ represent the country-specific mean and standard deviation of feature j, computed only from training samples belonging to country c. If a country is not present in the training partition, normalization falls back to the global training statistics.

The mean for country c and feature j is given by:

$$\mu_{c,j} = \frac{1}{N_c T_c} \sum_{i=1}^{N_c} \sum_{t=1}^{T_c} x_{c,i,t,j}, \tag{3.9}$$

where N_c denotes the number of samples associated with country c, and T_c denotes the number of temporal observations per sample.

The corresponding standard deviation is:

$$\sigma_{c,j} = \sqrt{\frac{1}{N_c T_c} \sum_{i=1}^{N_c} \sum_{t=1}^{T_c} (x_{c,i,t,j} - \mu_{c,j})^2}. \tag{3.10}$$

3.3.3 Bin-wise Normalization

For sequence inputs, bin-wise normalization is defined along the phenology-aligned axis. For each feature and each temporal bin, normalization statistics are computed across all training samples. Each value is then standardized relative to the corresponding bin-specific statistics. This approach accounts for the fact that feature distributions may vary across different stages of crop development. Bin-wise normalization is applied after temporal alignment and binning, and operates directly on the fixed-length sequence representation.

Let $b \in \{1, \ldots, L\}$ denote the bin index of the phenology-aligned sequence representation. For feature j at bin b, bin-wise normalization is defined as:

$$\tilde{x}_{c,i,b,j} = \frac{x_{c,i,b,j} - \mu_{b,j}}{\sigma_{b,j}}, \tag{3.11}$$

where $x_{c,i,b,j}$ denotes the value of feature j for country c, sample i, and phenological bin b, while $\tilde{x}_{c,i,b,j}$ denotes the normalized value. The quantities $\mu_{b,j}$ and $\sigma_{b,j}$ represent the bin-specific mean and standard deviation of feature j, computed across all training samples for bin b. The parameter L denotes the total number of phenology-aligned bins in the fixed-length sequence representation.

The bin-specific mean is given by:

$$\mu_{b,j} = \frac{1}{\sum_{c \in \mathcal{C}_{\text{train}}} N_c} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{i=1}^{N_c} x_{c,i,b,j}, \tag{3.12}$$

where $\mathcal{C}_{\text{train}}$ denotes the set of training countries and N_c is the number of samples associated with country c.

The bin-specific standard deviation is:

$$\sigma_{b,j} = \sqrt{\frac{1}{\sum_{c \in \mathcal{C}_{\text{train}}} N_c} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{i=1}^{N_c} (x_{c,i,b,j} - \mu_{b,j})^2}. \tag{3.13}$$

3.4 Task-Specific Tabular and Sequence Models

This section describes the predictive models used in this study. All models operate on inputs derived from the proposed PhenoNorm representation, using either tabular or sequence-based formats depending on the model family.

The primary objective of the modeling setup is to evaluate the proposed pipeline rather than to develop state-of-the-art architectures. Consequently, we restrict attention to standard, widely used models (Random Forest, XGBoost, LSTM, and a transformer) with modest hyperparameter tuning, and we deliberately avoid complex or hybrid architectures as well as extensive task-specific optimization or fine-tuning. This design choice helps ensure that observed performance differences can be attributed primarily to the input representation and normalization strategy, instead of the unique details of particular model setups.

The model output is a scalar estimate of end-of-season maize yield, expressed in tonnes per hectare, for a given sub-national administrative region and harvest year. The prediction is therefore intended as a regional decision-support indicator.

3.4.1 Random Forest

Random Forest [5] is used as one of the tree-based baselines for tabular yield prediction. The model operates on stage-aggregated features constructed from the phenology-aligned representation, including summaries of temperature, vegetation indices, soil moisture, cumulative GDD, and static soil properties.

For a given region—year sample with feature vector $x_{c,i}$ in country c, the Random Forest predictor is defined as:

$$\hat{y}_{c,i} = \frac{1}{K} \sum_{k=1}^{K} T_k(x_{c,i}), \tag{3.14}$$

where K is the number of trees in the ensemble and $T_k(\cdot)$ denotes the prediction of the k-th regression tree. This aggregation reduces variance relative to individual trees and has been shown to provide strong performance for crop yield prediction across diverse environments [38].

Hyperparameter tuning

Hyperparameters are selected using a grid search over:

- Number of trees (`n_estimators`) between 150 and 300: The chosen values allow the ensemble to reach a regime where additional trees yield diminishing returns in accuracy while keeping computational costs manageable, consistent with typical Random Forest tuning practice.
- Maximum tree depth (`max_depth`) between 12 and 20: This range provides sufficient capacity to model nonlinear interactions among climate and soil predictors without fully grown trees that tend to overfit noisy agricultural data.
- Minimum samples required for a split (`min_samples_split`) between 5 and 10: The chosen range prevents very small leaves, which helps regularize the model and improve generalization in settings with heterogeneous field conditions.
- Minimum samples required at a leaf node (`min_samples_leaf`) between 2 and 4: The same rationale applies as for `min_samples_split`.
- Number of features considered at each split (`max_features`) set to 0.25, 0.3, or `sqrt`: These values explore both the commonly recommended `sqrt` heuristic and smaller feature fractions, promoting tree diversity.

3.4.2 XGBoost

XGBoost [10] is used as a gradient-boosted tree baseline for tabular crop yield prediction. The model operates on the same stage-aggregated features derived from the phenology-aligned representation as Random Forest. The model is trained in a pooled multi-country setting using both globally-normalized and country-wise normalized tabular representations. Gradient boosting is implemented using the `XGBRegressor` framework with histogram-based tree construction, and the objective function is set to squared error regression.

For a given region-–year sample with feature vector $x_{c,i}$ in country c, the XGBoost predictor is defined as:

$$\hat{y}_{c,i} = \sum_{m=1}^{M} \eta \, f_m(x_{c,i}), \tag{3.15}$$

where M is the number of boosting rounds, η is the learning rate, and each $f_m \in \mathcal{F}$ is a regression tree from the space of CART (Classification and Regression Trees) models [6]. The model is trained with the squared-error regression objective (`reg:squarederror`), which minimizes the mean squared difference between predicted and observed yields. The learning rate η scales the contribution of each tree, enabling a trade-off between convergence speed and robustness to overfitting.

Hyperparameter tuning

Hyperparameters are optimized using randomized search over the validation split. The search space includes:

- Learning rate (`learning_rate`) between 0.02 and 0.20: This range balances convergence speed with generalization, with smaller values providing more conservative updates and reducing the risk of overfitting.
- Maximum tree depth (`max_depth`) between 4 and 10: The chosen range controls model complexity, allowing trees to capture nonlinear interactions while limiting very deep trees that tend to overfit noisy signals.
- Minimum child weight (`min_child_weight`) between 1 and 7: Larger values enforce a minimum amount of training signal in each leaf, acting as a regularizer that prevents overly specific splits on small sample subsets.
- Row subsampling ratio (`subsample`) between 0.6 and 1.0: Subsampling rows introduces stochasticity between trees, which can improve robustness and reduce overfitting, while ensuring that each model still sees a substantial portion of the training data.
- Feature subsampling ratio (`colsample_bytree`) between 0.6 and 1.0: Subsampling features at the tree level decorrelates trees and encourages diverse split structures, which is beneficial when predictors such as climate and soil variables are correlated.
- L_1 Regularization (`reg_alpha`) between 0 and 1: This coefficient controls sparsity in leaf weights, allowing the model to shrink less informative leaves toward zero and thereby improve interpretability and generalization.
- L_2 Regularization (`reg_lambda`) between 0.5 and 3.5: This parameter penalizes large leaf weights, smoothing the fitted function and reducing sensitivity to noise in training targets.

A total of 20 random hyperparameter configurations are evaluated. Each configuration is trained with up to 5000 boosting rounds using early stopping based on validation RMSE, which halts training once additional trees no longer improve performance. The final model is retrained on the complete training set using the best validation configuration and the optimal number of boosting iterations.

3.4.3 LSTM

LSTM-based [24] sequence models have been successfully applied to crop yield prediction using combined meteorological and remote sensing inputs [61]. In this study, the model operates on fixed-length multivariate sequences constructed by discretizing the phenology-aligned axis into L bins. Each input has shape $L \times F$, where L is the number of bins

and F is the number of features. The model outputs a scalar prediction of end-of-season yield. The architecture consists of one or more stacked LSTM layers followed by a fully connected regression head. Dropout is applied to reduce overfitting. The model is trained using mean squared error loss and the Adam optimizer, with early stopping based on validation performance.

Given an input sequence $\{x_t\}_{t=1}^{M}$, the LSTM encoder produces a sequence of hidden states $\{h_t\}_{t=1}^{M}$ by recurrently updating the hidden and cell states:

$$(h_t, c_t) = \text{LSTM}(x_t, h_{t-1}, c_{t-1}; \theta), \tag{3.16}$$

where h_t and c_t denote the hidden and cell states at time t, and θ denotes the LSTM parameters. The final hidden state is then passed through a regression head to obtain the yield prediction:

$$\hat{y}_{c,i} = W h_M + b, \tag{3.17}$$

where W and b are the weights and bias of the fully connected output layer.

Hyperparameter tuning

Hyperparameters are sampled randomly from:

- Hidden size (`hidden_size`) chosen from $\{128, 192, 256, 320, 384\}$: This range provides sufficient capacity to capture nonlinear temporal dependencies in the environmental sequences, while avoiding excessively large hidden states that would increase computational cost and overfitting risk.
- Number of layers (`num_layers`) between 1 and 3: This explores shallow to moderately deep LSTM stacks, allowing the model to learn hierarchical temporal features without incurring the optimization difficulties often observed in very deep recurrent networks.
- Dropout rate (`dropout`) between 0.0 and 0.2: Low to moderate dropout is applied between layers to regularize the network and reduce overfitting, while preserving most of the representational capacity of the recurrent layers.
- Regression head size (`head_dim`) chosen from $\{64, 128, 256\}$: The hidden dimension of the fully connected regression head controls the degree of nonlinear transformation applied to the final state before prediction, balancing expressiveness and parameter efficiency.
- Learning rate (`learning_rate`) between 6×10^{-5} and 10^{-3}: This interval centers around the commonly used Adam default of 10^{-3}, allowing for both more conservative and more aggressive updates depending on validation performance.
- Weight decay (`weight_decay`) chosen from $\{0.0, 10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$: These values provide a range of L_2 regularization strengths, enabling the model to control weight magnitudes and improve generalization in the presence of noisy yield observations.

Hyperparameter optimization is performed using random search over 12 trials. Each trial is trained for up to 60 epochs using validation RMSE for model selection, with early stopping applied when validation performance fails to improve. The final model is selected based on the best validation RMSE.

3.4.4 PatchTST

PatchTST [47] is used as a transformer-based baseline for sequence-based yield prediction. The model operates on the same phenology-aligned sequences as the LSTM model, where each input is represented as a multivariate time series of length L. The input sequence is divided into overlapping patches along the temporal axis, and each patch is treated as a token that is processed by a transformer encoder with self-attention:

Let $X_{c,i} = (x_1, \ldots, x_L) \in \mathbb{R}^{L \times F}$ denote the phenology-aligned sequence for region—year sample i in country c. The sequence is first partitioned into P patches:

$$z_p = \text{Patch}(X_{c,i}; p), \quad p = 1, \ldots, P, \tag{3.18}$$

where each $z_p \in \mathbb{R}^{P_{\text{len}} \times F}$ corresponds to a local temporal window of length P_{len}. The resulting sequence of patch embeddings $Z = (z_1, \ldots, z_P)$ is then processed by a transformer encoder to produce contextualized patch representations $H = \text{Transformer}(Z; \theta)$. A regression head maps a pooled representation of these patches to a scalar yield prediction:

$$\hat{y}_{c,i} = W \, \text{Pool}(H) + b, \tag{3.19}$$

where $\text{Pool}(\cdot)$ denotes averaging over patch tokens, and W, b are the parameters of the output layer. The model is trained using mean squared error loss with the Adam optimizer and early stopping based on validation RMSE.

Hyperparameter tuning

Hyperparameters are sampled from:

- Model dimension (`d_model`) chosen from $\{192, 256, 320\}$: This range controls the width of the transformer layers, providing sufficient capacity to model complex temporal dependencies without incurring prohibitive computational cost.
- Number of layers (`num_layers`) between 4 and 8: This allows the encoder depth to vary from relatively shallow to moderately deep, enabling the model to stack multiple self-attention blocks while avoiding over-parameterized architectures.
- Dropout rate (`dropout`) between 0.05 and 0.15: Low to moderate dropout is applied to attention and feed-forward sublayers to regularize the transformer without severely limiting its expressive power.

- Learning rate (`learning_rate`) chosen from $\{1, 2, 3\} \times 10^{-4}$: These values are typical for Adam-based training of transformer models and balance stable convergence with training speed.
- Weight decay (`weight_decay`) chosen from $\{0.0, 10^{-4}, 5 \times 10^{-4}\}$: This range of L_2 regularization strengths controls the magnitude of model weights and helps mitigate overfitting in settings with limited samples per country.
- Number of attention heads (`nhead`) set to 4 or 8: These settings allow the multi-head attention mechanism to learn multiple temporal interaction patterns in parallel, while keeping the per-head dimension compatible with the chosen `d_model`.

Hyperparameter optimization is performed using random search over 8 trials[1]. Each trial is trained for up to 10 epochs using validation RMSE for model selection, with early stopping applied when validation performance fails to improve. Model selection is based on validation RMSE under the same training protocol as the other models.

PatchTST is selected for this study because its architectural design is well matched to the structure of the PhenoNorm representation. The patching mechanism reduces effective sequence length while preserving localized temporal patterns, which allows the transformer encoder to capture dependencies across the full crop cycle without the sequential bottleneck of recurrent models [47]. This makes it particularly suitable for yield prediction, where yield is influenced by interactions among environmental conditions distributed across multiple developmental stages. Accordingly, PatchTST is included as the main attention-based sequence model to evaluate whether biologically aligned temporal representations can be exploited more effectively by transformer architectures than by tabular or recurrent baselines.

3.5 Foundation Model: Zero-Shot and Fine-Tuning

The final methodological phase extends the evaluation from task-specific models to a pretrained time series foundation model. The objective of this phase is to examine whether a pretrained forecasting model can generalize to cross-country yield prediction, either directly in a zero-shot setting or after supervised adaptation on limited task-specific data [26].

More specifically, this phase addresses two questions: First, whether a pretrained foundation model can provide competitive predictions without task-specific training; and second, how much labeled data are required for fine-tuning to surpass the best task-specific sequence model. These questions are studied under the same representation, country-level split, and evaluation protocol, so that comparisons with LSTM and PatchTST remain consistent.

[1]The number of trials is intentionally limited to prioritize fair representation comparison under consistent computational budgets rather than exhaustive architecture optimization.

3.5.1 Foundation Model Selection and Task Formulation

The general methodological idea is to assess foundation models as pretrained sequence learners that may transfer useful forecasting priors across domains [26]. In principle, such models can offer two advantages in the present setting: Improved robustness under cross-country heterogeneity and improved data efficiency when task-specific labeled data are limited.

In practice, however, foundation models differ substantially in their expected input structure, forecasting assumptions, support for multivariate covariates, and compatibility with fixed-length phenology-aligned agricultural sequences. This study focuses on **Moirai**, since it is the model that best matches the current data representation and forecasting setup [69]. In particular, Moirai is well-suited to multivariate time series forecasting, can be adapted to the phenology-aligned sequence representation used in this work, and supports zero-shot as well as fine-tuned evaluation.

Moirai is originally formulated as a sequence-to-sequence forecaster that takes a context window of past observations and predicts a future horizon of the same target series [69]. For end-of-season yield prediction, this forecasting setup is adapted so that the model consumes multivariate environmental sequences and produces a scalar yield prediction. The experiments consider both a zero-shot formulation and a supervised fine-tuning formulation.

To interface the dataset with Moirai, each sample is recast as a daily multivariate forecasting instance identified by country, administrative unit, and harvest year. A synthetic start timestamp is assigned at the beginning of the harvest year, and the observed sequence is represented using the fixed-length environmental covariates available over the season. The five dynamic predictors are passed as real-valued covariates, and their binary observation masks are appended as additional dynamic channels so that missing values are made explicit to the model rather than left implicit in zero-filled inputs. Since Moirai expects a temporal target series, the zero-shot formulation uses a valid univariate sequence channel to satisfy the forecasting interface, whereas the scalar yield variable is preserved as the supervision label used to assess downstream predictive utility.

For the experiment, the pretrained `Salesforce/moirai-1.0-R-small` checkpoint was first used without updating its parameters. Each region-year sample was represented as a fixed-length sequence of 50 phenology-aligned bins containing five dynamic variables. A binary observation mask was concatenated to the five dynamic variables, yielding 10 dynamic input channels. Missing or non-finite feature values were replaced by zero after the mask had been retained, allowing the model to distinguish an imputed zero from an observed value. The target supplied to Moirai was the 50-bin tavg sequence, while the end-of-season yield label was retained separately and used only for post-hoc evaluation of the forecast-derived scalar prediction. The model used a context length of 50 bins, prediction length of 5, automatic patch-size selection, 100 probabilistic forecast samples, and inference batch size 16. The pretrained model was evaluated separately on the validation and held-out test partitions using the same country-level split metadata as the task-specific models.

3.5.2 Zero-Shot Evaluation Under Domain Shift

The first set of experiments evaluates Moirai in a zero-shot regime. Here, zero-shot means that the model is not trained or fine-tuned on the crop dataset or on any of the countries used in the task-specific training pipeline; instead, it is applied directly with its pretrained parameters to the phenology-aligned sequence inputs.

The zero-shot evaluation is conducted under the same country-level split and metric protocol used for the task-specific models. Input sequences are constructed from the PhenoNorm. Predictions are then generated for the held-out test countries and compared against the strongest task-specific sequence baseline in order to assess whether pretrained forecasting priors transfer effectively to this specialized agricultural prediction problem.

3.5.3 Fine-Tuning and Data-Efficiency Experiments

To investigate data efficiency, Moirai is fine-tuned on progressively larger subsets of the training data. Fine-tuning regimes are defined by the fraction of available labeled training samples:

$$\mathcal{D}_\alpha \in \{0.5\%,\ 1\%,\ 2\%,\ 3\%,\ 5\%,\ 7\%,\ 10\%,\ 15\%,\ 20\%,\ 30\%,\ 50\%,\ 75\%,\ 100\%\}, \tag{3.20}$$

where α denotes the proportion of training region–year samples used for supervised adaptation.

For each $\mathcal{D}_\alpha$, the model is initialized from pretrained weights, fine-tuned on the selected subset, and evaluated on the same held-out test countries. Performance is compared both to the zero-shot version of Moirai and to the best non-foundation model, allowing the study to determine the smallest training fraction for which the foundation model becomes competitive or superior.

Fine-tuning used the predefined country-level training partition, while the validation partition was used for model selection and early stopping; the held-out test countries were not used during adaptation. The objective was to assess whether adapting the pretrained temporal representations to the agricultural yield-prediction setting improved performance relative to the zero-shot model and task-specific baselines.

3.5.4 Computing Environment

All experiments were executed using an NVIDIA H100 NVL GPU with approximately 95.8 GB of GPU memory. The computational environment used NVIDIA driver version 570.195.03 and CUDA version 12.8. Core software packages included NumPy 2.2.6, pandas 2.3.3, Matplotlib 3.10.8, and Seaborn 0.13.2. The GPU-accelerated implementation was used for the deep-learning and foundation-model experiments, while memory-mapped data storage was used to enable efficient access to the sequence tensors.

3.6 Evaluation and Interpretation

This section extends the core evaluation protocol by introducing additional validation strategies and model interpretation methods.

3.6.1 Train–Test and Validation Protocols

All experiments use a country-level train–test split to evaluate cross-country generalization. Countries are first grouped according to yield statistics and normalized yield profiles, and held-out test countries are selected through stratified sampling across these groups. This ensures that low, medium, and high yield regimes are represented in both training and test partitions while maintaining distributional differences between the two sets. Consequently, all models are evaluated on countries that are not observed during training.

Hyperparameter optimization is performed using a fixed validation split derived from the training partition. Randomized hyperparameter search is combined with early stopping based on validation RMSE. The final model configuration is then retrained using the full training data before evaluation on the held-out test countries.

The additional leave-one-country-out (LOCO) and k-fold cross-validation analyses are conducted specifically for the Random Forest baseline to provide a more detailed assessment of geographic generalization and model robustness. Prior work in geospatial and environmental machine learning has shown that conventional random validation strategies can produce overly optimistic estimates when spatial or regional dependencies exist between training and validation samples [15, 30]. Consequently, spatially or geographically aware validation protocols such as leave-location-out or grouped cross-validation are recommended for evaluating predictive performance under distribution shift.

Random Forest is selected for these additional analyses because it serves as the primary non-deep-learning baseline and provides relatively stable and computationally efficient repeated retraining across multiple validation folds. In contrast, other models require substantially higher computational cost due to iterative optimization, early stopping, and hyperparameter search, making exhaustive LOCO evaluation computationally prohibitive at the scale of the dataset.

All preprocessing steps, including feature construction, representation generation, and normalization statistics, are computed exclusively from the training partition and applied unchanged to validation and test data, preventing information leakage across partitions.

3.6.2 Performance Metrics

Model performance is evaluated using a combination of standard regression metrics and hydrological efficiency metrics in order to assess not only average prediction error, but also the ability of the models to reproduce the variability, structure, and magnitude of observed yield dynamics across heterogeneous environments. This mixed evaluation strategy

follows recommendations from both machine-learning and environmental modeling literature, where relying on a single metric is often insufficient for robust model assessment [33].

Standard Regression Metrics

Three standard regression metrics are used to evaluate predictive accuracy: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

- **Root Mean Squared Error (RMSE):** RMSE measures the average magnitude of prediction error while assigning greater weight to large deviations due to the squared-error formulation. Because crop yield prediction can contain occasional extreme failures under severe environmental conditions, RMSE is useful for evaluating overall robustness to large prediction errors. RMSE is computed as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}. \tag{3.21}$$

- **Mean Absolute Error (MAE):** MAE measures the average absolute prediction error and provides a more directly interpretable estimate of mean deviation in yield units. Unlike RMSE, it is less sensitive to outliers and therefore complements RMSE by providing a more stable measure of average predictive accuracy:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|. \tag{3.22}$$

- **Coefficient of determination (R^2):** The coefficient of determination quantifies the proportion of variance in observed yields explained by the model relative to a baseline predictor given by the mean of the observations. Values closer to 1 indicate stronger explanatory performance, whereas values near or below 0 indicate weak predictive skill:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{3.23}$$

In all equations above, N denotes the total number of evaluated samples, y_i denotes the observed yield for sample i, $\hat{y}_i$ denotes the corresponding predicted yield, and $\bar{y}$ denotes the mean observed yield across the evaluation set.

These metrics are computed at both pooled and per-country levels. RMSE and MAE primarily assess absolute predictive accuracy, while R^2 evaluates the ability of the model to reproduce observed yield variability. Together, they provide complementary perspectives on model behavior and are widely used in crop yield prediction and regression-based forecasting studies [28, 38, 40].

Nash–Sutcliffe Efficiency (NSE)

Because crop yield forecasting shares important characteristics with environmental and hydrological prediction problems, the Nash–Sutcliffe Efficiency (NSE) is additionally used as a domain-relevant skill metric. NSE is widely adopted in hydrology to evaluate predictive skill relative to a naive baseline defined by the mean of the observations [42]. It is defined as:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}. \tag{3.24}$$

Here, Y_i and $\hat{Y}_i$ denote observed and predicted yields, respectively, and $\bar{Y}$ is the mean observed yield. NSE values close to 1 indicate strong predictive performance, values near 0 indicate performance comparable to predicting the mean, and values below 0 indicate performance worse than the mean-reference baseline.

In addition to NSE, some hydrological studies also report the coefficient of determination using the squared Pearson correlation formulation rather than the described R^2 in Equation 3.23 [1, 33]. The squared Pearson correlation coefficient is defined as:

$$R^2_{\mathrm{Pearson}} = \left(\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{N}(\hat{Y}_i - \bar{\hat{Y}})^2}} \right)^2. \tag{3.25}$$

Although the residual-based R^2 used in this study is mathematically very similar to NSE under ordinary least-squares evaluation settings, it differs conceptually from the squared Pearson correlation formulation shown above. The residual-based R^2 and NSE both evaluate predictive skill relative to the mean-observation baseline and directly penalize prediction error magnitude, whereas the squared Pearson correlation primarily measures the strength of linear association between observed and predicted values and does not explicitly account for prediction bias or variance mismatch [33].

For this reason, NSE and residual-based R^2 often produce nearly identical values in regression-based prediction tasks, while the squared Pearson correlation coefficient may behave differently under systematic overestimation, underestimation, or scale mismatch. Retaining NSE alongside R^2 therefore improves interpretability across both machine-learning and hydrological research communities.

Kling–Gupta Efficiency (KGE)

To complement RMSE, R^2, and NSE, the Kling–Gupta Efficiency (KGE) is also reported. KGE was introduced to address limitations of NSE by explicitly decomposing predictive performance into correlation, variability agreement, and bias components [14]. KGE is computed as:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \tag{3.26}$$

where r is the Pearson correlation coefficient between predictions and observations, $\alpha = \sigma_{\hat{y}}/\sigma_y$ is the variability ratio, and $\beta = \mu_{\hat{y}}/\mu_y$ is the bias ratio.

Unlike RMSE, which measures only average error magnitude, KGE provides a more diagnostic assessment of predictive behavior by simultaneously evaluating temporal covariation, variability reproduction, and systematic bias. This is particularly valuable in the present multi-country setting, where two models may achieve similar RMSE values while differing substantially in their ability to reproduce cross-country variability patterns or avoid systematic over- and under-estimation [14].

Taken together, the selected metrics provide a balanced evaluation strategy that captures both statistical prediction accuracy and environmental-modeling skill characteristics.

3.6.3 Evaluating Calendar versus Phenological Alignments

To isolate the effect of temporal representation design, we conduct a controlled comparison between two alternative seasonal alignment strategies within the same sequence-modeling pipeline. The first representation uses the nGDD axis, while the second replaces the biological progression axis with a normalized calendar-based seasonal progression. In the calendar-aligned representation, observations are indexed according to their relative position within the crop season, computed from the SOS and EOS boundaries provided by the crop calendar data.

To ensure a fair comparison, all other components of the pipeline are kept unchanged across the two representations. This includes the feature set, temporal aggregation procedure, normalization strategy, sequence length, train/validation/test partitions, hyperparameter optimization budget, model architecture, optimization procedure, and evaluation metrics. The only modified component is the temporal axis used for temporal binning and sequence construction.

For the calendar-aligned representation, normalized seasonal progression is defined as:

$$p_{\mathrm{calendar}} = \frac{\mathrm{DOY} - \mathrm{SOS}}{\text{season length}}, \tag{3.27}$$

where DOY denotes the observation day-of-year and season length is computed using the corresponding SOS and EOS boundaries, including support for cross-year growing seasons. Similar to PhenoNorm, observations are grouped into fixed temporal bins and aggregated into multivariate sequences suitable for sequence modeling.

This experimental design enables a direct evaluation of whether biologically informed temporal alignment improves cross-country generalization beyond what can be achieved using normalized calendar progression alone.

3.6.4 Held-Out-Year Benchmark Evaluation

Beyond assessing spatial transferability to unseen countries, we conduct a secondary set of experiments designed specifically to evaluate PhenoNorm against the established official baseline. To ensure a fair and isolated comparison, all primary methodological choices (including feature alignment, model architectures, optimization strategies, and performance metrics) are strictly maintained. The sole adjustment is replacing the cross-country spatial split with a cross-year temporal split, perfectly aligning with the benchmark's evaluation framework. In this configuration, models are trained on earlier growing seasons within a specific country, validated on subsequent seasons, and evaluated on later, held-out years.

For these temporal generalization tests, the sequence-based models process the exact same phenology-aligned inputs (dynamic features, missingness masks, and static attributes) as in the primary spatial experiments. Hyperparameter tuning procedures are uniformly applied, optimizing for validation RMSE before final evaluation on the test years. By matching the benchmark's exact temporal forecasting constraints, this protocol successfully isolates the impact of our proposed biological alignment and modeling strategy.

3.6.5 Interpretation and Agronomic Evaluation

To interpret the learned models and assess agronomic plausibility, the thesis employs interpretation methods applied to the proposed representation.

Random Forest and XGBoost feature importance

For Random Forest and XGBoost models, feature importance is extracted directly from the trained models using their built-in importance measures [5, 10]. In both methods, importance is derived from how frequently and effectively a feature is used to reduce prediction error across the ensemble of decision trees. We use the standard implementation of mean decrease in impurity (MDI), as provided by the scikit-learn interface. Feature importance values are obtained from the trained model through the `feature_importances_` attribute.

In this thesis, we rank features based on these importance scores and analyze the top contributing variables, focusing on environmental and temporal features that most influence yield prediction performance across the growing season.

Integrated Gradients for sequence models

For deep sequence models, feature importance is analyzed using Integrated Gradients (IG), a gradient-based attribution method that estimates how strongly each input contributes to the final prediction [60]. In this thesis, IG is implemented using `IntegratedGradients` from the `Captum` library, which computes attributions for differentiable model inputs.

The method works by comparing the model output for an input sequence against a baseline reference and accumulating gradients along a straight-line path between them. This approach provides a stable estimate of feature contributions and helps mitigate issues such as gradient saturation in deep networks.

In practice, IG attributions are computed over sequence inputs and then aggregated across phenological stages and along the nGDD axis to identify which variables and crop-development periods contribute most strongly to model predictions. The resulting importance patterns are compared qualitatively with known crop-physiology relationships.

Together, the evaluation and interpretation protocol provides a consistent basis for comparing models, diagnosing the effects of representation and normalization choices, and grounding the learned patterns in agronomic context.

Chapter 4

Experimental Results and Discussion

This chapter presents the experimental evaluation of the proposed pipeline for multi-country crop yield prediction. The results are structured to follow the methodological pipeline introduced earlier, moving from data analysis and representation validation to model evaluation and comparative insights.

4.1 Dataset Characteristics and Validation

4.1.1 Dataset Quality and Coverage

After filtering, the pipeline retained 35 out of 42 countries and ensured that all included regions contribute meaningful variation for modeling [1]. The final dataset contained 60,806 region–year samples and 23 harvest years (2001–2023). The retained countries covered a broad range of agro-climatic conditions, including both high-yield intensive agricultural systems and lower-yield rainfed environments. Figure 4.1 shows that usable samples are constrained by the overlap between yield records and GDD lookup availability.

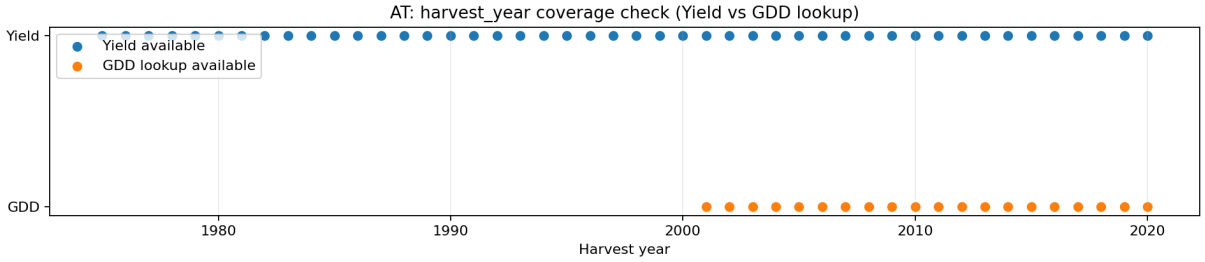


Figure 4.1: Example coverage check for AT, showing overlap between years with yield records and years with GDD lookup availability.

[1] For country codes, see Table 6.1 in Appendix 6.1.

The filtering procedure removed countries with insufficient sample support or degenerate yield distributions. In total, seven countries were excluded because they either contained fewer than 30 samples or exhibited near-constant zero yields, making them unsuitable for supervised modeling. The remaining countries showed substantial variability in yield distributions, supporting meaningful cross-country evaluation.

The final yield distribution ranged from 0 to 24.09 t/ha, with a global mean of 7.11 t/ha and standard deviation of 3.64 t/ha. Zero-valued yields represented approximately 1% of the dataset, indicating that the filtering procedure successfully removed countries dominated by invalid or non-informative targets while preserving realistic variability across agricultural systems.

Feature completeness analysis showed low overall sparsity. Across all extracted predictors, the total missing-value rate remained below 4%, and no feature exhibited more than 20% missing observations. Most static soil variables were available for more than 99% of samples. Only a very small number of samples contained extensive missingness, indicating that the phenology-aligned extraction procedure produced stable feature representations across countries and seasons.

The resulting feature space consisted primarily of a stage-aggregated representation derived from normalized phenological progression. In total, 40 temporal features were generated from NDVI, FPAR, temperature, surface soil moisture, and root-zone soil moisture variables using stage-level aggregation. Two additional static soil variables, available water capacity and bulk density, were included as time-invariant predictors.

Correlation analysis (Table 4.1) further confirmed the relevance of the extracted features. Vegetation-related indicators, particularly FPAR and NDVI during flowering and

Table 4.1: Correlation analysis results for top 15 absolute correlations with yield.

Feature	Correlation
fpar_mean_flowering	0.585
fpar_at_p060	0.579
fpar_at_p045	0.569
Tavg_at_p100	-0.541
Tavg_mean_late	-0.540
fpar_mean_grainfill	0.517
ndvi_mean_flowering	0.503
ndvi_at_p045	0.496
ndvi_at_p060	0.491
ndvi_mean_grainfill	0.450
Tavg_mean_veg	-0.431
fpar_at_p085	0.376
Tavg_at_p085	-0.352
Tavg_mean_grainfill	-0.332
ndvi_at_p085	0.303

grain-filling stages, showed the strongest positive association with yield. Temperature-related variables exhibited moderate negative correlations, reflecting the adverse effect of excessive thermal conditions in some production regions. Soil-moisture variables showed weaker but still informative relationships with yield, particularly under heterogeneous climatic conditions.

4.1.2 Yield Distribution and Cross-Country Variability

Before evaluating predictive models, the dataset was examined to characterize the degree of heterogeneity across countries. Figure 4.2 presents the distribution of maize yields for all countries included in the study.

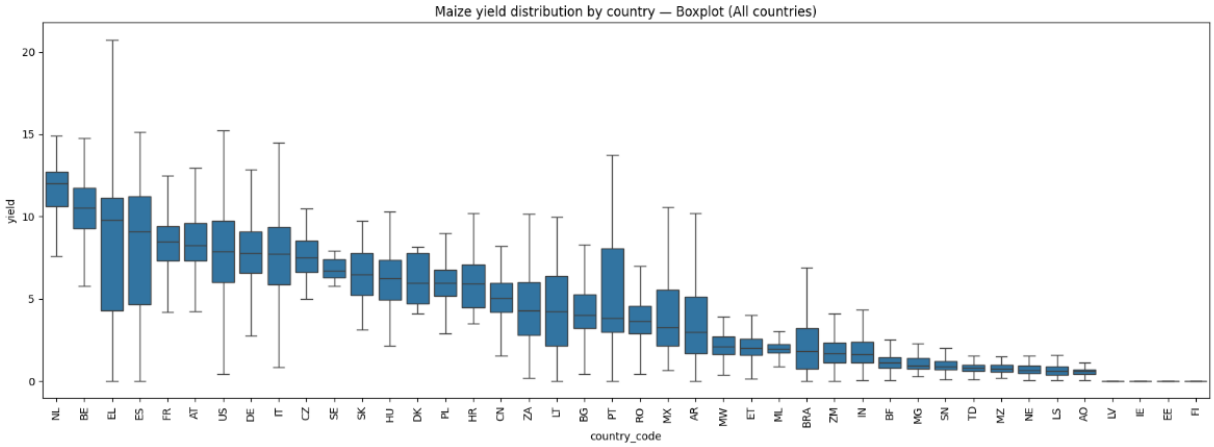


Figure 4.2: Distribution of maize yield across all countries in the dataset. Each boxplot summarizes the yield distribution for a single country.

Substantial differences can be observed both in median yield and variability across countries. Several European countries exhibit consistently high yields, while others show lower productivity and substantially wider distributions. Some countries also contain strong within-country variability, with large interquartile ranges and extreme outliers, indicating heterogeneous environmental and management conditions even within the same national region.

To better understand the heterogeneity of the dataset, exploratory analyses were performed on yield distributions and country-level groupings.

Figure 4.3 shows the distribution of mean yields across predefined yield groups and unsupervised clusters. The tertile-based grouping separates the data into low, medium, and high-yield regions with clearly distinct distributions. The clustering analysis further reveals that countries naturally organize into groups with substantially different productivity ranges, indicating strong cross-country variability in agricultural conditions and yield behavior.

The train–test distribution comparison shows that the held-out countries cover a broad portion of the overall yield range, supporting the validity of the cross-country evaluation

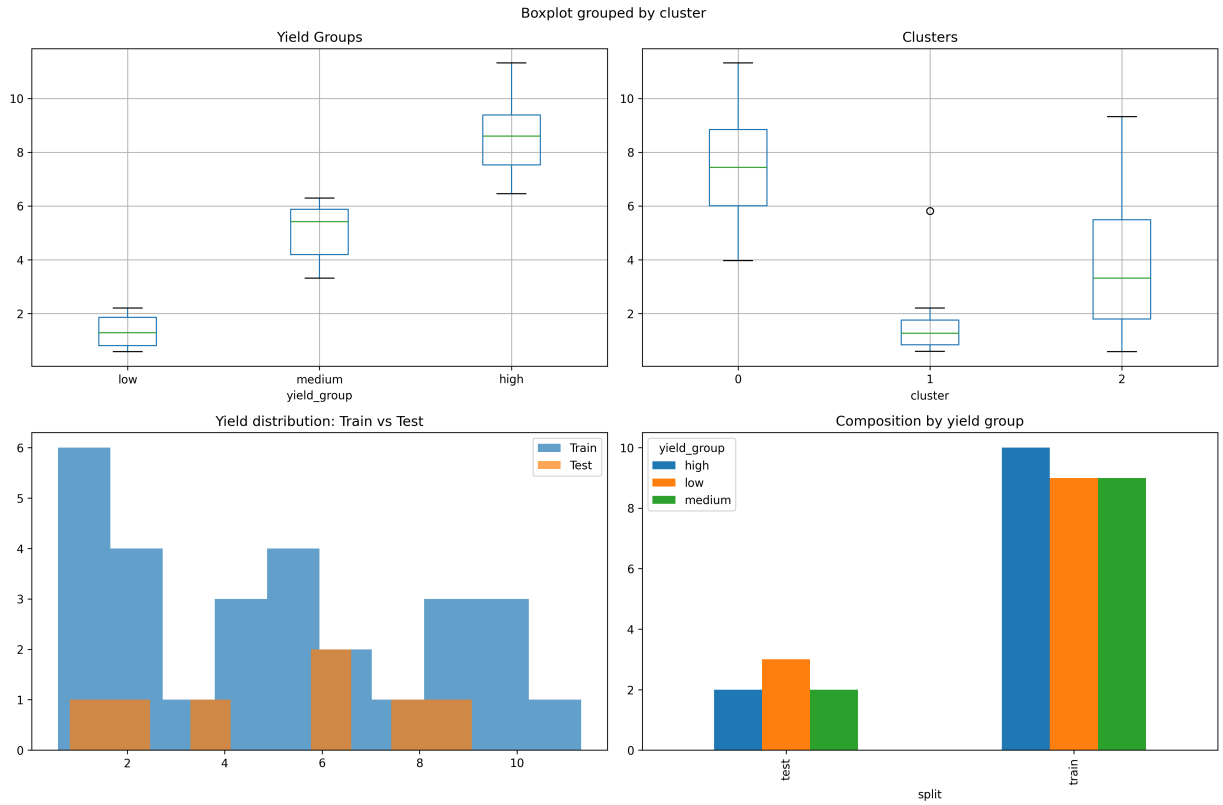


Figure 4.3: Exploratory analysis of yield-group distributions, cluster structure, and train–test composition across the dataset.

setting. In addition, the split composition analysis confirms that all yield groups are represented in both training and testing partitions.

4.1.3 Impact of Temporal Alignment and Normalization

Figure 4.4 shows that calendar-day indexing produces substantial variation in seasonal progression across countries, reflecting differences in planting dates and season length. In contrast, nGDD alignment reduces this temporal mismatch by placing observations on a physiological time scale, making trajectories more comparable while still preserving meaningful cross-country differences.

Figure 4.5 shows how temporal alignment and normalization influence the structure of multi-country seasonal trajectories, using FPAR as a representative example. The same analysis was performed for all temporal variables, and FPAR is shown here because it clearly illustrates the effect. When trajectories are aligned by calendar days, substantial dispersion remains across countries due to differences in planting dates, season duration, and developmental timing. In contrast, cumulative GDD alignment produces more coherent trajectories by expressing time in a physiological coordinate system, thereby improving cross-country comparability.

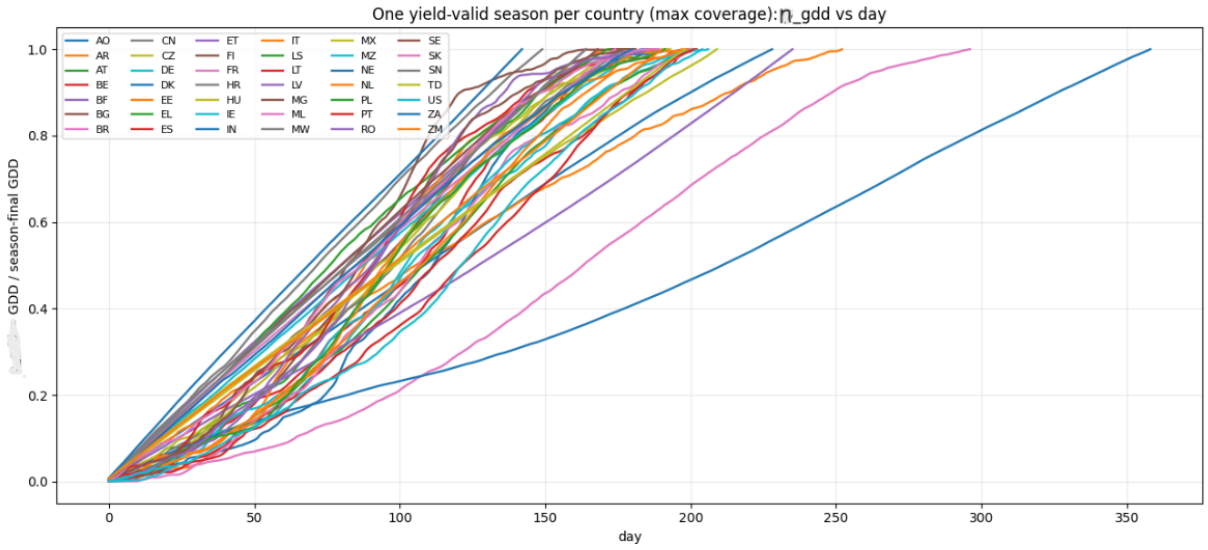


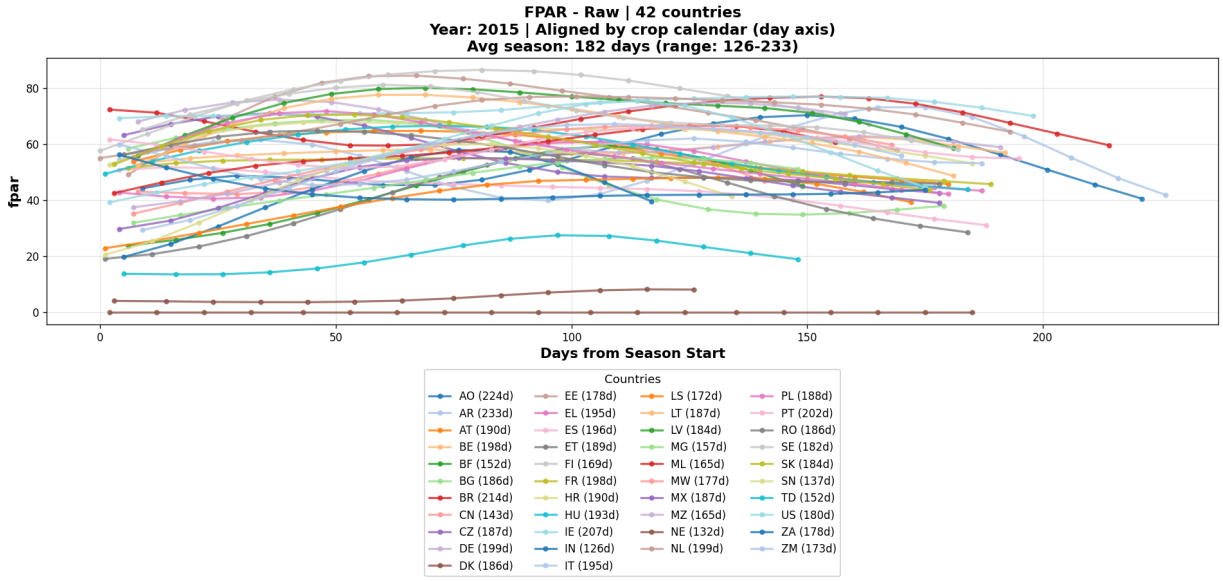
Figure 4.4: One yield-valid season per country, comparing calendar-day indexing with nGDD.

Normalization further affects the shape and comparability of the aligned curves. Global normalization preserves meaningful between-country amplitude differences while maintaining a common scale, whereas country-wise normalization rescales each country independently and can suppress informative magnitude variation. This qualitative pattern is consistent with the quantitative modeling results, which showed that global normalization generally supported better predictive performance.

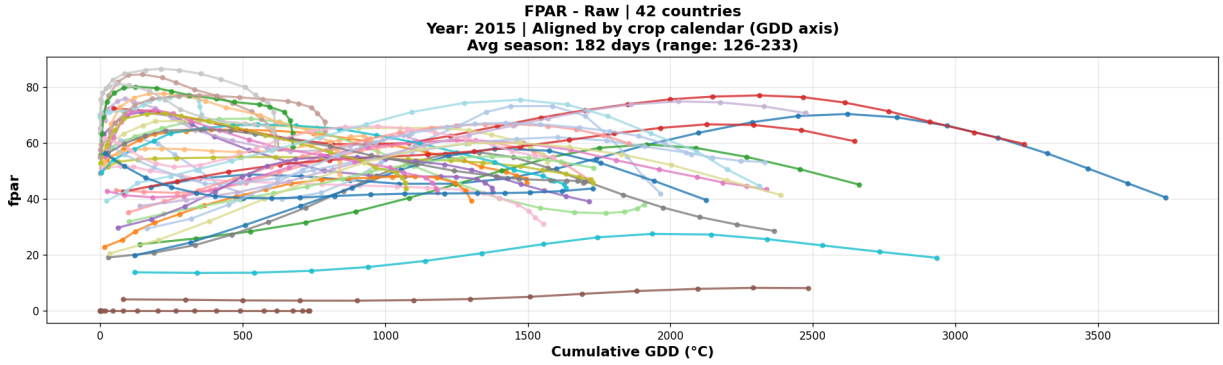
Global normalization preserves differences in the overall magnitude and range of FPAR values between countries on a common scale. In contrast, country-wise normalization emphasizes the relative evolution of FPAR within each country, making differences in the timing, rate of canopy development, peak behavior, and senescence patterns easier to compare. The country-normalized trajectories are bounded between 0 and 1, with values near 0 and 1 representing the lower and upper ends of each country's observed seasonal FPAR range, respectively.

These visualizations were generated independently of the final quality-filtering procedure and directly from the raw database. Consequently, all 42 available countries are shown in Figure 4.5, including countries that were not retained in the final modeling set. The purpose of this analysis is therefore illustrative: It provides qualitative evidence that both temporal alignment and normalization substantially influence the input structure before predictive modeling.

Figure 4.6 illustrates the effect of the proposed temporal representation. Panel (a) displays the original FPAR sequences indexed by days from season start, where substantial differences in trajectory timing and duration remain across countries. Panel (b) expresses the same observations on the nGDD axis, so that each horizontal position corresponds to comparable relative thermal development. Panel (c) then summarizes the aligned trajectories into four phenological stages, preserving the distribution of cross-country variation.

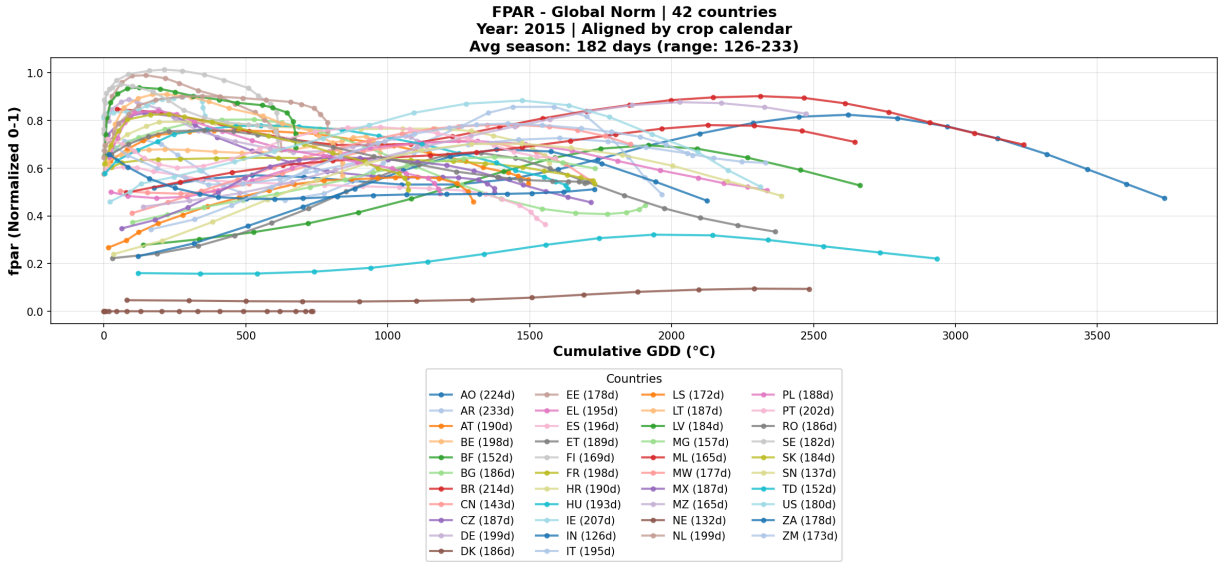


(a) FPAR, day axis, before normalization

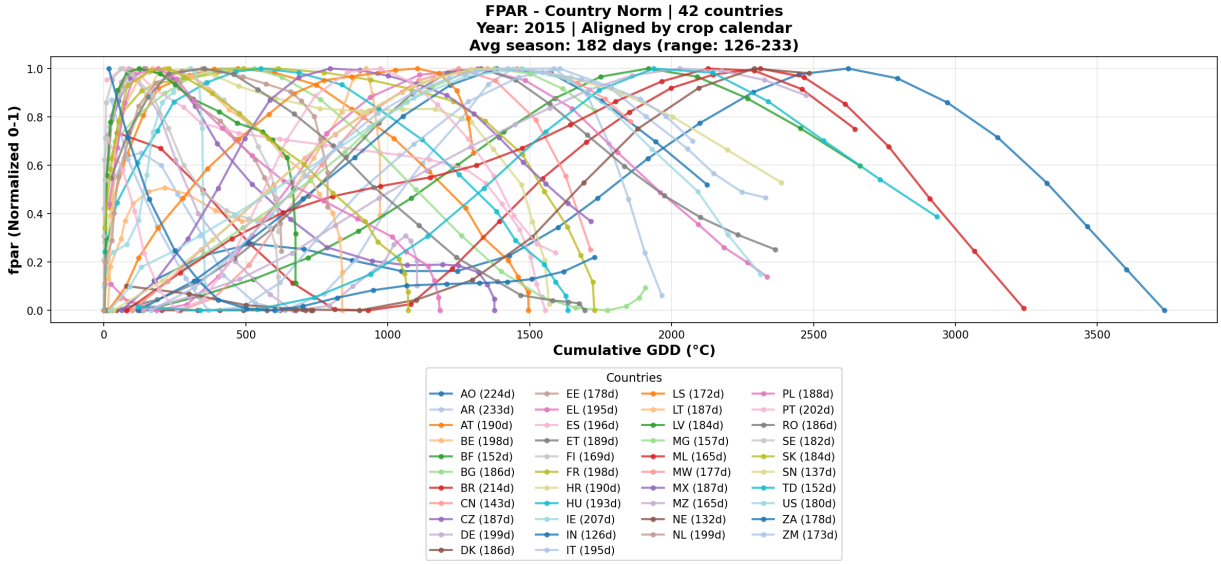


(b) FPAR, GDD axis, before normalization

Figure 4.5: Illustrative comparison of cross-country FPAR trajectories under different alignment and normalization choices.



(c) FPAR, cGDD axis, global normalization



(d) FPAR, cGDD axis, country normalization

Figure 4.5: Illustrative comparison of cross-country FPAR trajectories under different alignment and normalization choices (continued).

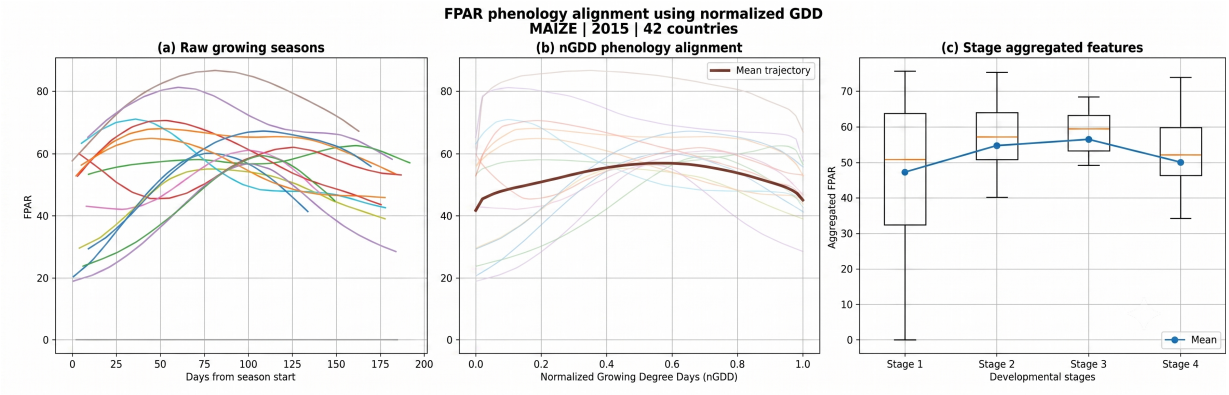


Figure 4.6: FPAR trajectories across 42 countries in 2015: (a) raw season-time indexing, (b) nGDD alignment, and (c) aggregation into four phenological stages.

4.2 Model Evaluation

4.2.1 Random Forest Results

Random Forest was used to evaluate the impact of normalization strategies on tabular representations. Hyperparameter tuning produced similar configurations for the global and raw settings, both selecting 200 trees with a maximum depth of 15 and identical split constraints. In contrast, the country-wise setting favored a deeper and more flexible model (maximum depth 20 with smaller split thresholds), suggesting an attempt to fit more localized patterns. Despite this increased complexity, performance remained substantially worse, indicating that the loss of a shared feature scale cannot be compensated by model capacity.

Although Random Forest does not inherently require feature normalization, we applied the same normalization variants used for the other models to maintain a consistent experimental setting and to isolate the effect of representation choice across the full comparison. In this case, normalization was not introduced to improve Random Forest optimization, but to test whether preserving or removing cross-country differences in feature scale altered the information available to the model. As shown in Table 4.2, global normalization achieved the best performance, while country-wise normalization performed poorly, likely due to the removal of informative cross-country differences.

Table 4.2: RF test performance under different normalization strategies.

Normalization	RMSE	MAE	R^2
Global	2.353	1.931	0.488
Raw	2.356	1.933	0.487
Country-wise	3.453	2.987	-0.102

Feature-importance analysis of the globally normalized model, shown in Table 4.3, provides additional insight into the learned predictive structure. The prominence of these

Table 4.3: Top-ranked Random Forest features under global normalization.

Feature	Importance
Tavg_veg_std	0.111
Tavg_late_std	0.080
Tavg_grainfill_sum	0.067
Tavg_grainfill_std	0.043
GDD_flowering_max	0.039
Tavg_veg_mean	0.033
FPAR_late_std	0.033
GDD_full_final	0.030
GDD_veg_max	0.028
GDD_late_max	0.028

features indicates that the model relied strongly on seasonal temperature dynamics and accumulated thermal conditions, especially during the vegetative, flowering, grain-filling, and late-season stages.

Vegetation and water-related features also contributed, though less strongly than the thermal variables. Among the most important non-thermal predictors were FPAR_late_std, FPAR_veg_std, NDVI_flowering_mean, RSM_grainfill_sum, and awc. This ranking shows that thermal conditions define much of the seasonal growth environment, while canopy activity and water availability provide complementary information about crop status and yield potential.

4.2.2 XGBoost Results

XGBoost was evaluated as a stronger tree-based baseline for the tabular representation to assess whether gradient boosting could better capture cross-country patterns under different normalization strategies. Hyperparameter tuning produced broadly similar configurations for the global and country-wise settings, with both favoring relatively deep trees (max_depth=10), low learning rates (around 0.03), and subsampling of both observations and features. The country-wise model used slightly stronger regularization, reflected in higher values of min_child_weight and reg_alpha, while reg_lambda remained similar across settings. The global model also achieved a lower validation RMSE (≈ 1.07) than the country-wise model (≈ 1.20), although both selected the maximum tested number of estimators (n_estimators=5000). Overall, these results indicate that XGBoost benefited from a flexible yet regularized boosting configuration, and that the global setting was more effective at leveraging shared structure across countries.

While XGBoost is generally insensitive to feature scaling when used with tree boosters, normalization was retained here for consistency across models and to assess the effect of representation choice on performance. As shown in Table 4.4, global normalization produced substantially better generalization than country-wise normalization. Under global normalization, XGBoost outperformed the Random Forest baseline on all reported test metrics,

Table 4.4: XGBoost test performance under different tabular normalization strategies.

Normalization	RMSE	MAE	R^2
Global	2.203	1.735	0.552
Country-wise	3.925	3.345	-0.423

indicating that boosting was better able to exploit the shared cross-country feature space. By contrast, the country-wise setting led to a large degradation in performance, reinforcing the conclusion that independently normalizing each country removes information that is important for transfer to unseen countries.

Feature-importance analysis of the globally normalized model shows a strong concentration on temperature and thermal-time-related predictors. Several additional high-ranking features reflected thermal accumulation and late-season temperature dynamics, indicating that the boosted model relied primarily on seasonal temperature variability and heat accumulation to explain yield differences across countries.

Vegetation and moisture variables contributed complementary but smaller signals. Table 4.5 shows that thermal conditions appear to dominate the broad cross-country signal, while remotely sensed vegetation status and soil-moisture-related variables provide supporting information about crop condition during critical developmental stages.

Table 4.5: Top-ranked XGBoost features under global normalization.

Feature	Importance
`Tavg_veg_std`	0.280
`Tavg_late_std`	0.085
`Tavg_veg_sum`	0.076
`GDD_late_max`	0.061
`FPAR_flowering_max`	0.049
`Tavg_grainfill_sum`	0.020
`GDD_full_final`	0.020
`Tavg_late_mean`	0.018
`RSM_veg_sum`	0.014
`Tavg_late_sum`	0.013

4.2.3 LSTM Results

The sequential inputs for the LSTM were constructed by aggregating daily observations into fixed phenology-aligned bins based on nGDD. Specifically, each sample was represented as a multivariate sequence obtained by averaging dynamic variables within discrete `n_gdd` intervals, which provided a compact and comparable description of crop development across environments. This bin-wise aggregation reduced irregular temporal sampling while

preserving the seasonal progression of each series. Although the preprocessing pipeline was defined generally for different sequence lengths, the LSTM normalization comparison reported here used $L = 100$ bins, while additional experiments on the best-performing global configuration were later conducted to assess the sensitivity of model performance to bin size.

Hyperparameter tuning selected a hidden size of 384, 2 LSTM layers, a dropout rate of 0.1, a regression head size of 64, a learning rate of approximately 3.64×10^{-4}, and a weight decay of 10^{-4}. These results suggest that the best LSTM configuration balanced relatively high model capacity with light regularization.

The globally normalized model achieved the best performance using a moderately deep architecture (2 layers, hidden size 384) with light regularization. The country-wise configuration selected a similar architecture but without dropout and with a larger prediction head, suggesting a tendency to overfit local patterns. In contrast, bin-wise normalization favored a smaller and shallower model (1 layer, hidden size 320) with higher dropout and learning rate, indicating a less stable input representation. These trends further support that global normalization provides a more consistent and learnable feature space for sequence models.

Among the evaluated normalization strategies (Table 4.6), global normalization again achieved the best performance, followed by bin-wise and country-wise normalization.

Table 4.6: LSTM test performance under different normalization strategies.

Normalization	RMSE	MAE	R^2
Global	2.337	1.781	0.566
Bin-wise global	2.404	1.861	0.541
Country-wise	2.483	1.900	0.510

Additional evaluation of the globally normalized LSTM provides a more detailed view of model behavior beyond the aggregate test metrics. As shown in Figure 4.7, the model achieved moderate predictive performance, capturing the broad yield trend but showing reduced calibration at higher yield values. The scatter pattern indicates that the model learns the central structure of the data reasonably well, but underestimates samples in the upper tail of the yield distribution.

4.2.4 PatchTST Results

Based on the normalization experiments with the tree-based baselines and LSTM, global normalization was retained as the most reliable representation for cross-country learning. Rather than repeating the full normalization comparison for PatchTST, we therefore focused on refining the model under this selected preprocessing setting and on evaluating design choices that were more specific to sequence modeling, particularly model configuration and temporal resolution.

Hyperparameter tuning selected the largest tested model dimension ($d_{\text{model}} = 320$), 6 Transformer layers, a dropout rate of 0.15, a learning rate of 10^{-4}, weight decay of 10^{-4}, and 8 attention heads. The final PatchTST model was then trained for up to 20 epochs using automatic mixed precision (AMP). These results indicate that PatchTST performed best with a relatively deep and high-capacity architecture combined with moderate regularization and conservative optimization.

PatchTST achieved the strongest overall performance among all evaluated models. The best configuration was obtained with a sequence length of $L = 50$, this result substantially outperformed the LSTM and tabular baselines, indicating that PatchTST was better able to exploit structured temporal dependencies across phenological development. The improvement also suggests that a patch-based Transformer representation was especially well-suited to capturing stage-dependent interactions in the aggregated seasonal trajectories.

Figure 4.8 shows that PatchTST achieves substantially better agreement between predicted and observed yield than the LSTM baseline (Figure 4.7). The improvement is reflected quantitatively by the higher R^2 and lower RMSE, indicating that the model captures both the central tendency and a larger share of the yield variability. Some underprediction remains at the upper end of the yield range, suggesting that extreme high-yield

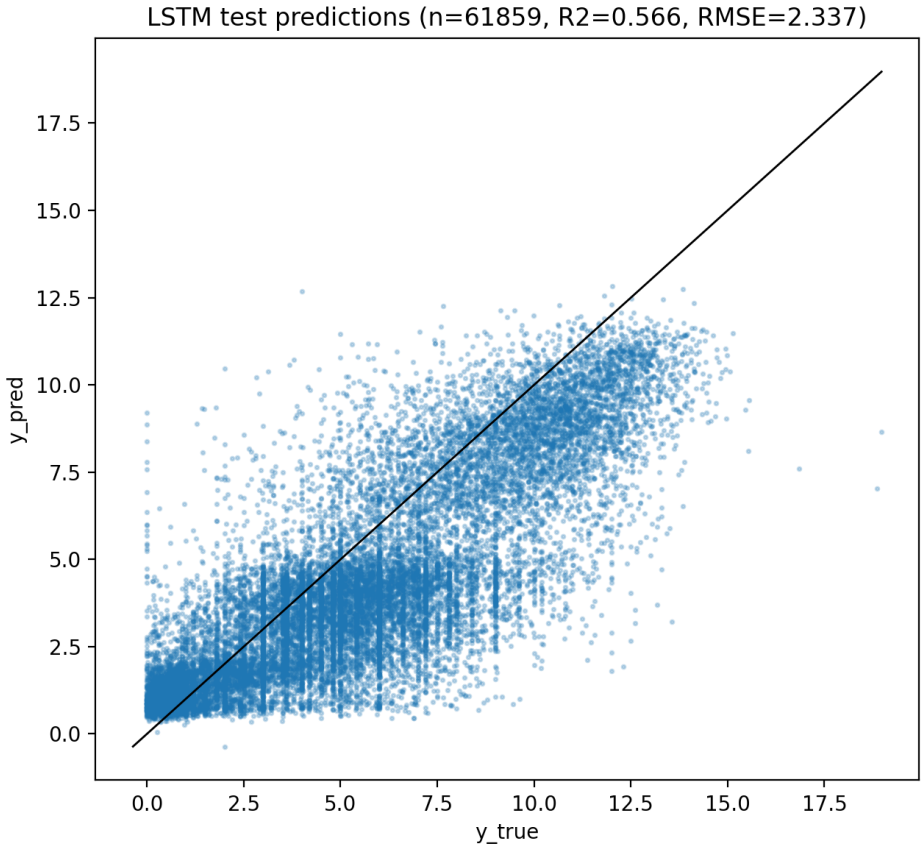


Figure 4.7: LSTM test predictions under global normalization.

cases are still harder to model accurately.

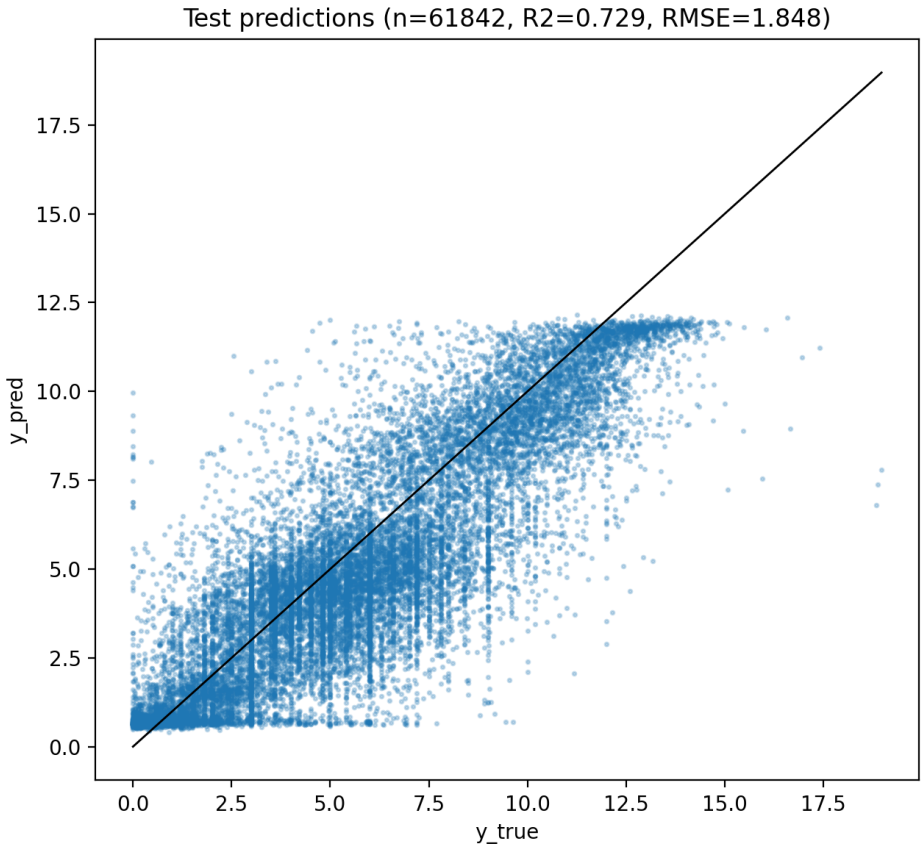


Figure 4.8: PatchTST test predictions under global normalization.

Integrated Gradients analysis provided additional insight into the predictive structure learned by the best PatchTST model. As shown in Figure 4.9, attribution was concentrated primarily in the vegetation-related variables, with NDVI and FPAR consistently receiving the highest importance across much of the phenology-aligned seasonal axis. The heatmap further indicates that attribution was strongest during the early developmental period near sowing and vegetative growth, before gradually decreasing toward later stages of the season. Temperature also contributed substantially, particularly at the beginning and end of the seasonal trajectory, whereas soil-moisture variables (SSM and RSM) remained comparatively less influential throughout most bins.

The stage-wise aggregation in Figure 4.10 confirms these trends. During the vegetative stage, FPAR and NDVI exhibited the highest aggregated attribution values, followed by temperature, while soil-moisture-related variables contributed substantially less. Similar patterns remained during flowering and grain-filling, although the relative importance of vegetation variables gradually decreased. In the late-season stage, NDVI remained highly influential, while the importance of temperature increased relative to earlier stages and became comparable to FPAR. Across all stages, SSM and RSM consistently received the lowest attribution values.

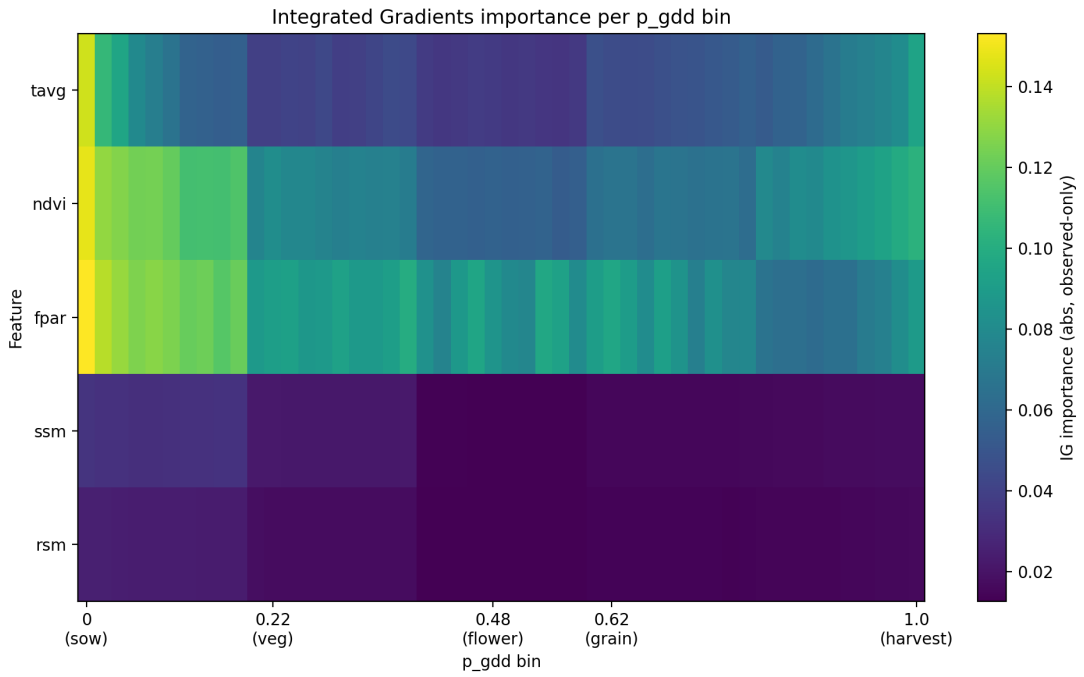


Figure 4.9: Integrated Gradients feature importance across phenological bins (nGDD).

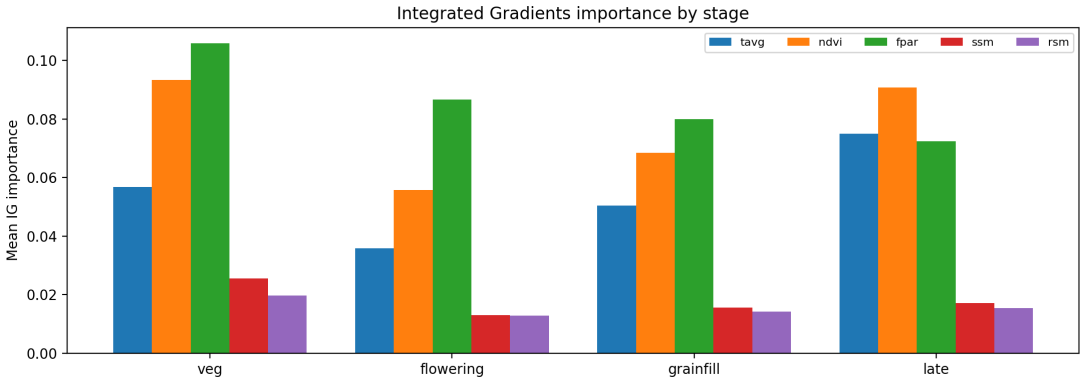


Figure 4.10: Stage-wise aggregated Integrated Gradients importance across four phenological stages.

These attribution patterns are agronomically plausible and help explain the strong predictive performance of the model. Vegetation indices provide a direct proxy for canopy condition, biomass accumulation, and seasonal crop development, which likely enabled PatchTST to identify informative temporal signatures associated with final yield formation. The concentration of attribution during the early and mid-season portions of the phenology-aligned trajectory suggests that the model relied most heavily on canopy-development dynamics established before maturity, while late-season predictions incorporated a more balanced combination of vegetation and temperature-related signals.

Sensitivity to Sequence Length

A sensitivity analysis was conducted over sequence lengths $L \in \{24, 50, 75, 100, 150\}$ in order to determine how temporal resolution affected PatchTST performance. The results (Table 4.7) show that model accuracy was sensitive to this design choice, confirming that sequence length was one of the key hyperparameters for the Transformer-based formulation.

Table 4.7: Sensitivity of PatchTST performance to the number of phenology-aligned bins.

Bins L	Val RMSE	Test RMSE	Test MAE	Test R^2
24	1.849	2.188	1.651	0.620
50	**1.598**	**1.848**	**1.358**	**0.729**
75	1.617	1.965	1.437	0.693
100	1.670	2.029	1.482	0.673
150	1.673	1.930	1.431	0.704

The best performance was achieved at $L = 50$, with the lowest validation and test error as well as the highest test R^2. Shorter sequences such as $L = 24$ likely lost useful temporal detail by over-compressing seasonal development, whereas longer sequences such as $L = 100$ and $L = 150$ introduced a finer representation that did not translate into better generalization. This suggests that excessively coarse binning removes informative phenological variation, while excessively fine binning may increase noise or sparsity without adding enough predictive structure.

4.2.5 Zero-Shot and Fine-Tuned Moirai Performance

To examine whether a pretrained time series foundation model could improve yield prediction, Moirai was evaluated in both zero-shot and fine-tuned settings. For Moirai inference, each sample produced a probabilistic forecast consisting of multiple sampled trajectories over the prediction horizon. To adapt the data for Moirai, each administrative unit–year sample was represented with fixed length $L = 50$, and the dynamic inputs consisted of the environmental variables and their corresponding observation masks. To obtain a single scalar prediction, the mean of the final forecast step across all generated samples was used. This value was then compared with the observed yield to compute RMSE, MAE, and R^2.

Table 4.8: Zero-shot Moirai performance before task-specific fine-tuning.

Split	RMSE	MAE	R^2
Validation	7.353	5.932	-3.375
Test	9.595	6.372	-6.314

The zero-shot results were poor on both validation and test data (Table 4.8), indicating that the pretrained model did not transfer effectively to the multi-country maize yield prediction problem without adaptation.

Table 4.9: Fine-tuning performance of Moirai under different fractions of the training set.

Training fraction	RMSE	MAE	R^2
0.005	2.792	2.245	0.381
0.010	2.622	2.094	0.454
0.020	2.540	1.993	0.487
0.030	2.507	2.006	0.501
0.050	2.395	1.865	0.544
0.070	2.301	1.763	0.579
0.100	2.322	1.783	0.572
0.150	2.210	1.698	0.612
0.200	2.193	1.659	0.618
0.300	2.208	1.697	0.613
0.500	2.163	1.651	0.628
0.750	**2.144**	**1.619**	**0.635**
1.000	2.211	1.677	0.612

Moirai was therefore fine-tuned using different fractions of the available training set in order to assess adaptation efficiency. Fine-tuning led to large improvements at all data fractions, converting the model from unusable zero-shot performance to competitive supervised performance. With only 0.5% of the training data, Moirai already surpassed its zero-shot performance substantially, and improvement continued steadily as more data were introduced.

The strongest results were obtained at intermediate-to-large data fractions, shown in Table 4.9. Increasing the training fraction did not lead to strictly monotonic improvement, and the full-data setting did not produce the best test result. In particular, performance at 100% of the training data was weaker than at 75%, 50%, and 20%, suggesting that the model may have experienced mild overfitting or optimization variability at the largest training fraction, preventing further gains in test accuracy.

Although pretraining provides Moirai with broad temporal representations, it does not guarantee superior performance relative to models designed and optimized specifically for this yield-prediction task. The pretrained model was developed from heterogeneous time-series domains and forecasting objectives, whereas the present application involves phenology-aligned agricultural sequences, sparse and partially missing remote-sensing observations, static cross-country domain shift, and a scalar end-of-season yield target. This mismatch between the pretraining distribution and the target task can limit the direct transferability of pretrained representations, particularly in the zero-shot setting.

Fine-tuning can reduce this mismatch by adapting the model parameters to the feature distributions and temporal patterns of the agricultural dataset. However, adaptation remains constrained by the amount and diversity of labeled training data, the selected optimization configuration, and the extent to which the fine-tuning objective matches the end-of-season yield-regression task. In contrast, task-specific models such as the LSTM and PatchTST are trained directly to minimize yield-prediction error using the phenology-

aligned representation. Their architectures and parameters can therefore specialize fully to the available agricultural inputs and target, which may enable them to outperform a pretrained foundation model when the task-specific dataset and preprocessing pipeline are well matched to the prediction problem. Consequently, the value of Moirai in this study lies not only in its absolute performance, but also in its ability to adapt from a pretrained initialization under limited labeled-data conditions.

These results show that Moirai has strong adaptation potential after supervised fine-tuning, but it did not surpass the best PatchTST model in this study.

4.3 Comparative Analysis and Interpretation

The results highlight three key findings: First, temporal modeling is essential; sequence-based models outperform tabular approaches under the current protocols. Second, representation matters; PhenoNorm inputs can improve learning compared to calendar-base indices, and third, pretraining helps, but only with adaptation; Moirai fails in zero-shot but becomes competitive after fine-tuning with 75% of training data.

4.3.1 Comparing Chronological and Biological Time Scales

To evaluate whether biologically informed temporal alignment contributes meaningfully to cross-country generalization, we compared PhenoNorm PatchTST against a calendar-aligned PatchTST baseline. The only modified component is the temporal axis used to construct the input sequences, and Table 4.10 summarizes the results obtained using the two temporal representations.

Table 4.10: Comparison between calendar-aligned and nGDD under the same sequence-modeling pipeline.

Representation	$R^2 \uparrow$	RMSE ↓	MAE ↓
Calendar-aligned	0.69	1.97	1.45
nGDD	**0.73**	**1.84**	**1.36**

The phenology-aligned representation outperforms the calendar-aligned baseline across all evaluation metrics. Because the underlying model architecture and training procedure remain unchanged, the observed improvement can be attributed primarily to the temporal representation itself.

These results suggest that aligning environmental observations according to crop developmental progression provides a more transferable representation across heterogeneous agricultural regions than normalized chronological progression alone. While normalized calendar time accounts for differences in season length, it does not explicitly capture variations in crop development rates caused by temperature and climatic differences between

regions. In contrast, the nGDD representation aligns observations according to accumulated thermal progression, enabling sequences from different countries to correspond more closely to comparable phenological stages.

The improvement further supports the central hypothesis of this thesis, that representation-level temporal alignment can reduce cross-country temporal inconsistency and improve the ability of sequence models to learn transferable growth-pattern relationships from heterogeneous agricultural datasets.

4.3.2 Comparison with CY-Bench Baselines

To further evaluate the effectiveness of PhenoNorm, the country-level performance of PatchTST and LSTM was compared against the official CY-Bench benchmark baselines provided by the AgML initiative [50]. The CY-Bench benchmark standardizes crop yield forecasting across countries using a common preprocessing framework that includes spatial harmonization, crop-calendar-based seasonal extraction, standardized predictor sources, and country-level evaluation splits. Predictor variables include weather, soil moisture, evapotranspiration, vegetation indices, and soil properties aggregated over administrative units.

It is important to note that the CY-Bench LSTM baseline is implemented as a residual forecasting model rather than a direct yield predictor. Specifically, CY-Bench first estimates a trend component using its `TrendModel`, then trains the LSTM on detrended targets, and finally reconstructs the yield prediction by adding the predicted residual back to the estimated trend at inference time. The benchmark comparison is therefore against a stronger hybrid baseline that combines explicit temporal trend estimation with neural sequence modeling.

Table 4.11 shows that the PhenoNorm models improved performance in several countries relative to the CY-Bench LSTM baseline, particularly when using PatchTST. Large gains were observed in BR, DE, RO, PT, and ZA, where the proposed sequence representations achieved clearly higher R^2 values. Results for BF and MZ are excluded from this interpretation, as critically small sample sizes combined with severe domain shift render model comparisons statistically unreliable for these countries.

Global normalization achieved the strongest performance among the normalization strategies considered. This finding indicates that preserving differences in feature magnitude across countries remains useful even when the test year belongs to countries represented during training. The result is therefore consistent with the cross-country experiments, where global normalization also performed more robustly by retaining inter-country variation relevant to yield prediction.

The results also reveal considerable heterogeneity across countries. While the proposed approach performed strongly in many medium and high-sample regions, performance remained limited in several low-sample or highly variable countries such as Mali, Lesotho, and Burkina Faso. In some cases, the original CY-Bench LSTM baseline remained superior, particularly for countries where the benchmark representation may already align well with local seasonal structure.

Table 4.11: Country-level R^2 comparison between the CY-Bench residual LSTM baseline and the PhenoNorm models

Country	Samples	CY-Bench LSTM	PhenoNorm LSTM	PhenoNorm PatchTST
BR	40000	Not Provided	0.177	0.499
US	13104	0.490	-0.084	0.475
AR	2565	0.270	0.053	-0.275
IN	1518	0.320	0.054	0.083
DE	923	-0.730	-0.648	0.404
IT	567	0.290	0.047	0.043
FR	562	-0.290	-0.536	-0.423
ES	283	0.340	0.578	0.223
CN	240	0.280	-0.104	-1.841
RO	239	-2.160	-1.183	-0.141
EL	214	-0.050	-0.331	0.041
ZM	213	0.230	-0.074	-0.204
ET	185	-0.990	-1.594	-3.141
HU	120	-2.500	-0.984	-1.478
NE	120	-1.230	-0.203	-0.307
PL	97	-1.620	-0.934	-0.260
CZ	84	-7.600	-0.462	-0.285
ML	72	-0.600	-10.889	-15.356
ZA	70	-0.130	-0.056	0.361
MX	69	-0.380	-0.286	-0.045
BE	58	-101.800	-0.172	0.111
AO	54	-2.190	-0.403	-3.988
LT	54	-3.600	-1.539	-5.140
NL	54	-67.700	-0.451	-0.249
AT	54	-9.680	-2.675	-1.991
LS	53	-1.470	-20.793	-9.787
TD	51	-0.040	0.199	-0.939
SN	39	-0.020	0.065	-0.270
BG	36	-3.290	-2.875	-1.584
PT	40	-1.980	0.348	0.498
SK	32	-16.680	-0.695	-0.610
HR	12	-151.690	-0.181	0.573

4.3.3 Physiological Interpretation of Learned Features

The feature-importance patterns identified by both Random Forest and PatchTST are consistent with well-established drivers of maize yield formation. In the Random Forest model, the strongest predictors were dominated by stage-specific temperature statistics and thermal-time variables, including mean and variability of temperature during vegetative, flowering, grain-filling, and late-season periods, as well as multiple GDD indicators. This is

agronomically plausible because maize yield is highly sensitive to thermal conditions across the season [57], with the reproductive and grain-filling phases being especially vulnerable to temperature stress. High temperatures around flowering can reduce pollen viability, silk receptivity, and seed set, while elevated temperatures during grain filling can accelerate development, shorten effective filling duration, and reduce kernel weight [66].

The prominence of vegetative-stage thermal variables is also biologically meaningful [18]. Early and mid-season temperature conditions regulate canopy expansion, leaf-area development, and the rate at which maize progresses toward reproductive stages, thereby influencing the plant's capacity to intercept radiation and accumulate biomass later in the season. Likewise, the importance of late-season temperature variability suggests that the model is capturing stress patterns that affect grain filling efficiency and final dry matter partitioning to kernels [66].

The PatchTST attribution profile complements this interpretation but emphasizes a different aspect of the yield-generation process. In the early and vegetative stages, NDVI and FPAR received the highest integrated-gradient importance, exceeding the contributions of temperature and soil-moisture channels. This suggests that the sequence model relied strongly on early canopy establishment and radiation-interception signals as leading indicators of final yield [13]. Such a pattern is agronomically reasonable because early vigor and rapid canopy development influence cumulative light interception, photosynthetic activity, and biomass formation, which in turn set an upper bound on reproductive potential [18].

This interpretation is reinforced by the NSE, KGE, and R^2_{Pearson} values reported in Table 4.12. The superior performance of PatchTST across all three metrics indicates that its advantage is not limited to lower average error; rather, it also reproduces the structure of observed yield variation more faithfully, with better balance between co-variation, amplitude, and bias. The higher R^2_{Pearson} achieved by PatchTST further suggests that the transformer-based representation captures temporal yield dynamics more consistently across heterogeneous countries and production environments.

Fine-tuned Moirai also performs strongly, showing that pretrained sequence representations can become competitive after adaptation, although they remain less well calibrated than the best task-specific transformer. The comparison between XGBoost and LSTM provides an additional cross-disciplinary insight: Despite relatively similar error-based skill, the higher KGE of LSTM suggests that sequential models may better preserve the dynamic structure and temporal variability of yield fluctuations, not merely their average level. Taken together, these metrics support the broader conclusion that successful cross-country maize yield prediction depends not only on reducing prediction error, but also on capturing the timing, magnitude, and variability of physiologically meaningful seasonal signals.

To conclude, the results demonstrate that combining PhenoNorm representation with transformer-based models can provide the most effective solution for multi-country yield prediction.

Table 4.12: Overall best performance of each model under its optimal configuration.

Model	RMSE	MAE	R^2	R^2_{Pearson}	NSE	KGE
Random Forest	2.353	1.931	0.488	0.519	0.488	0.472
XGBoost	2.203	1.735	0.552	0.553	0.552	0.615
LSTM	2.337	1.781	0.566	0.672	0.566	0.673
PatchTST ($L = 50$)	**1.848**	**1.358**	**0.729**	**0.760**	**0.729**	**0.820**
Moirai (fine-tuned)	2.144	1.619	0.635	0.7224	0.635	0.696

Chapter 5

Conclusion and Future Work

5.1 Key Findings

5.1.1 Phenology alignment and cross-country generalization

The results provide strong evidence that the proposed PhenoNorm pipeline improves the consistency and transferability of seasonal trajectories across countries. A controlled comparison between calendar-aligned and phenology-aligned temporal representations showed that biologically informed alignment based on nGDD outperformed normalized calendar progression. These findings suggest that aligning environmental observations according to developmental progression produces more coherent cross-country representations than chronological alignment alone.

5.1.2 Normalization under domain shift

The experiments demonstrate that normalization strategy has a substantial impact on cross-country generalization, and that no single approach is universally optimal across all model families. Global normalization generally produced the most stable and transferable behavior under unseen-country evaluation, while more adaptive normalization strategies exhibited varying effectiveness depending on the model architecture and representation format. These results highlight normalization as an important representation-level design choice in heterogeneous agricultural time series learning.

5.1.3 Foundation models versus task-specific models

The comparison between foundation models and task-specific models highlights both the limitations and potential of transfer learning in agricultural forecasting. In zero-shot settings, Moirai failed to generalize effectively to crop yield prediction under severe cross-country domain shift. However, fine-tuning substantially improved performance, making

the model competitive with strong supervised baselines. The results suggest that pretrained time series representations do not automatically transfer to specialized agricultural forecasting tasks, but can become highly effective once adapted using sufficient task-specific supervision.

5.2 Conclusion

This thesis investigated how to improve multi-country crop yield prediction under domain shift by designing a preprocessing pipeline aligned with crop developmental progression. The central objective was to determine whether phenology-aware temporal representations and normalization strategies could improve cross-country comparability and support robust prediction across heterogeneous agricultural environments. To address this problem, the study introduced PhenoNorm, which combines crop season mapping, in-season extraction, nGDD alignment, temporal aggregation, leakage-safe normalization, and model-specific sequence construction.

The empirical results demonstrate that temporal representation design plays a critical role in cross-country agricultural forecasting performance. Visual and statistical analyses revealed substantial heterogeneity in planting dates, season duration, and environmental trajectories across countries, confirming that chronological alignment alone is insufficient for constructing comparable seasonal representations.

The comparative modeling experiments also established a clear hierarchy in predictive performance across model families. Random Forest and XGBoost provided useful and interpretable baselines, but remained limited relative to sequence-based approaches. LSTM substantially improved performance, highlighting the importance of modeling within-season temporal dependencies directly. Among the supervised approaches, PatchTST achieved the strongest overall performance, producing the best balance of RMSE, MAE, and R^2 under the evaluated cross-country protocol.

The foundation model experiments provided an additional perspective on transfer learning for agricultural forecasting. In zero-shot settings, Moirai exhibited limited transferability to crop yield prediction, indicating that generic pretrained time series representations do not automatically generalize to specialized agricultural tasks under strong domain shift. However, after fine-tuning, Moirai improved substantially and became competitive with the strongest supervised baselines. Although its best performance did not surpass PatchTST, the results suggest that foundation models can become highly effective for agricultural forecasting when adapted using sufficient task-specific supervision.

Overall, this thesis demonstrates that preprocessing and representation design are fundamental components of robust agricultural machine learning systems. By aligning environmental observations according to crop developmental progression, PhenoNorm improves temporal consistency across heterogeneous agricultural regions and provides a practical foundation for cross-country crop yield prediction.

5.3 Future Work

Several directions remain open for future investigation:

A first direction involves systematic ablation studies to isolate the contribution of individual pipeline components. While the current experiments demonstrate that phenology-aligned temporal indexing improves performance relative to calendar-aligned progression, additional studies could further quantify the independent contributions of temporal aggregation, stage partitioning, and normalization strategy selection. Such analyses would provide deeper insight into which representation components contribute most strongly to cross-country generalization.

The phenology-alignment strategy itself also presents opportunities for refinement. In the current formulation, crop development is represented primarily through cumulative GDD. While thermal accumulation is widely used in agronomy, crop development may also depend on photoperiod sensitivity, soil-water availability, management practices, and stress-related environmental conditions. Future work could therefore investigate adaptive or learned alignment strategies that incorporate additional physiological signals [32]. It could explicitly select, weight, or adapt training countries according to agro-climatic similarity with a target country, using climate, soil, phenological, and production-system indicators.

A third direction is the evaluation of more advanced time series architectures within the phenology-aligned pipeline. Recent developments in the broader time series literature, including decomposition-based Transformers, MLP-mixer architectures, and time series foundation models, have demonstrated strong performance on generic forecasting benchmarks. Evaluating these models under the PhenoNorm preprocessing setting would help determine whether their benefits extend to low-frequency and highly heterogeneous agricultural datasets. Future work should also incorporate explainability methods, calibrated uncertainty estimates, and operational monitoring pipelines, with agronomic oversight to ensure that model outputs are interpreted appropriately before supporting real-world decisions.

Beyond task-specific models, a particularly promising direction involves the development of agricultural time series foundation models through self-supervised or transfer-learning approaches. Current foundation models are typically pretrained on generic domains such as energy demand, finance, or traffic forecasting, and their applicability to agricultural forecasting remains insufficiently explored [69, 23]. Large-scale pretraining on heterogeneous agricultural time series data could enable models to learn generalized crop-development representations that transfer across crops, countries, and climate conditions. Techniques such as masked temporal reconstruction, contrastive seasonal learning, and cross-domain adaptation may further improve robustness in low-data settings [8].

The experiments in this thesis use complete in-season sequences spanning the full crop season. Accordingly, the reported results should be interpreted primarily as end-of-season yield estimation performance. A natural extension is to evaluate progressively truncated sequences: for a selected phenological cutoff, only observations available up to that point would be supplied to the model while the target would remain final end-of-season yield.

Repeating this procedure across successive nGDD thresholds would quantify the trade-off between forecast lead time and predictive accuracy, and would identify the earliest stage at which the model provides sufficiently reliable estimates for operational decision support.

Finally, future evaluations should consider robustness under stronger distribution shifts and operational forecasting settings. Agricultural systems are increasingly affected by climate variability and extreme weather events, which may alter historical relationships between environmental conditions and yield outcomes. Evaluating PhenoNorm under climate-shift scenarios and extreme-weather holdout conditions would provide a more rigorous assessment of generalization capability. Similarly, extending the pipeline toward in-season and uncertainty-aware forecasting may improve its applicability for early-warning systems, food-security monitoring, and agricultural decision support.

Overall, PhenoNorm establishes a foundation for phenology-aware agricultural time series learning, while also highlighting the need for future research at the intersection of temporal representation learning, domain generalization, and agricultural forecasting.

Chapter 6

Appendix

6.1 Country Code Mapping

Table 6.1 lists the country symbols used throughout the thesis and their corresponding country names.

Table 6.1: Country codes used in the thesis.

Code	Country	Code	Country
AO	Angola	ML	Mali
AR	Argentina	MX	Mexico
AT	Austria	MZ	Mozambique
BE	Belgium	NE	Niger
BF	Burkina Faso	NL	Netherlands
BG	Bulgaria	PL	Poland
BR	Brazil	PT	Portugal
CN	China	RO	Romania
CZ	Czech Republic	SK	Slovakia
DE	Germany	SN	Senegal
EL	Greece	TD	Chad
ES	Spain	US	United States
ET	Ethiopia	ZA	South Africa
FR	France	ZM	Zambia
HR	Croatia		
HU	Hungary		
IN	India		
IT	Italy		
LS	Lesotho		
LT	Lithuania		

References

[1] D. L. J. Alexander, A. Tropsha, and David A. Winkler. Beware of r^2: Simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of Chemical Information and Modeling*, 55(8):1316–1322, 2015.

[2] Claudia Aviles Toledo, Melba M Crawford, and Mitchell R Tuinstra. Integrating multimodal remote sensing, deep learning, and attention mechanisms for yield prediction in plant breeding experiments. *Frontiers in plant science*, 15:1408047, 2024.

[3] N. H. Batjes. Harmonized soil property values for broad-scale modelling (wise30sec). *Geoderma*, 269:61–68, 2016.

[4] Hendrik Boogaard, Allard de Wit, Jan te Roller, and Kees van Diepen. Agera5: A climate forcing dataset for agricultural applications. *European Commission Joint Research Centre Data Report*, 2022.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[7] E Bueechi, F Reuß, M Pikl, L Homolova, V Lukas, M Trnka, and W Dorigo. Making optimal use of limited field-scale data for crop yield forecasting using transfer learning and sentinel-1 and 2 data. *Smart Agricultural Technology*, page 101567, 2025.

[8] Angelo Casolaro, Vincenzo Capone, Gennaro Iannuzzo, and Francesco Camastra. Deep learning for time series forecasting: Advances and open problems. *Information*, 14(11):598, 2023.

[9] AJ Challinor, TR Wheeler, PQ Craufurd, JM Slingo, and DIF Grimes. Design and optimisation of a large-area process-based model for annual crops. *Agricultural and forest meteorology*, 124(1-2):99–120, 2004.

[10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[11] Minghan Cheng, Xiyun Jiao, Lei Shi, Josep Penuelas, Lalit Kumar, Chenwei Nie, Tianao Wu, Kaihua Liu, Wenbin Wu, and Xiuliang Jin. High-resolution crop yield

and water productivity dataset generated using random forest and remote sensing. *Scientific data*, 9(1):641, 2022.

[12] Yunyao Cheng, Chenjuan Guo, Bin Yang, Haomin Yu, Kai Zhao, and Christian S Jensen. A memory guided transformer for time series forecasting. *Proceedings of the VLDB Endowment*, 18(2):239–252, 2024.

[13] Di Chu, David B. Lobell, Yang Zhang, Jianxi Wang, Weihong Wang, Xiuqing Wang, Tao Dong, Jian Liu, Xiaowei Zhan, et al. Remote estimation of the fraction of absorbed photosynthetically active radiation in maize canopies. *Journal of Plant Ecology*, 8(4):429–439, 2015.

[14] Guillaume Cinkus, Naomi Mazzilli, Hervé Jourde, Andreas Wunsch, Tanja Liesch, Nataša Ravbar, Zhao Chen, and Nico Goldscheider. When best is the enemy of good–critical evaluation of performance criteria in hydrological models. *Hydrology and Earth System Sciences*, 27(13):2397–2411, 2023.

[15] Andrew Crane-Droesch. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13(11):114003, 2018.

[16] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[17] European Commission Joint Research Centre. Fraction of absorbed photosynthetically active radiation (fpar) products. JRC Data Products, 2024.

[18] Muhammad Farooq, Muhammad Hussain, Sami Ul-Allah, and Kadambot H. M. Siddique. Impact of water deficit stress in maize: Phenology and yield components. *Scientific Reports*, 10:2944, 2020.

[19] Mahdiyeh Fathi, Reza Shah-Hosseini, Hossein Arefi, and Armin Moghimi. A multi–task 3d–resnet–bilstm transfer learning approach for winter wheat yield prediction using multi–source remote sensing data: Evaluating the impact of source crop selection (corn and soybean) in transfer learning. *Remote Sensing Applications: Society and Environment*, page 101766, 2025.

[20] Dayun Feng, Hongye Yang, Kexin Gao, Xiuliang Jin, Zhenhai Li, Chenwei Nie, Guoqiang Zhang, Liang Fang, Linli Zhou, Huirong Guo, et al. Time-series ndvi and greenness spectral indices in mid-to-late growth stages enhance maize yield estimation. *Field Crops Research*, 333:110069, 2025.

[21] Benoit Franch, Kristof Van Tricht, Inbal Becker-Reshef, et al. Worldcereal: A dynamic open global crop map. *Earth System Science Data*, 2022.

[22] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023.

[23] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[25] Tongxi Hu, Xuesong Zhang, Gil Bohrer, Yanlan Liu, Yuyu Zhou, Jay Martin, Yang Li, and Kaiguang Zhao. Crop yield prediction via explainable ai and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, 336:109458, 2023.

[26] Xing Hu, Siyuan Chen, and Dawei Zhang. Domain adaptation in agricultural image analysis: a comprehensive review from shallow models to deep learning. *arXiv e-prints*, pages arXiv–2506, 2025.

[27] Talha Ilyas, Jonghoon Lee, Okjae Won, Yongchae Jeong, and Hyongsuk Kim. Overcoming field variability: unsupervised domain adaptation for enhanced crop-weed recognition in diverse farmlands. *Frontiers in Plant Science*, 14:1234616, 2023.

[28] Md Abu Jabed and Masrah Azrifah Azmi Murad. Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*, 10(24), 2024.

[29] Abhasha Joshi, Biswajeet Pradhan, Subrata Chakraborty, Renuganth Varatharajoo, Shilpa Gite, and Abdullah Alamri. Deep-transfer-learning strategies for crop yield prediction using climate records and satellite image time-series data. *Remote Sensing*, 16(24):4804, 2024.

[30] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019.

[31] Patrick Killeen, Iluju Kiringa, and Tet Yeap. Corn grain yield prediction using UAV-based high spatiotemporal resolution multispectral imagery. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1054–1062, 2022.

[32] Alexander Kocian, Daniele Massa, Samantha Cannazzaro, Luca Incrocci, Sara Di Lonardo, Paolo Milazzo, and Stefano Chessa. Dynamic bayesian network for crop growth prediction in greenhouses. *Computers and electronics in agriculture*, 169:105167, 2020.

[33] P. Krause, D. P. Boyle, and F. Bäse. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5:89–97, 2005.

[34] C. Lee. Corn growth stages and growing degree days: A quick reference guide. Technical Report AGR-202, Cooperative Extension Service, University of Kentucky College of Agriculture, Lexington, KY, 2011.

[35] Felipe Tomazelli Lima and Vinicius MA Souza. A large comparison of normalization methods on time series. *Big Data Research*, 34:100407, 2023.

[36] Jiangui Liu, Ted Huffman, Budong Qian, Jiali Shang, Qingmou Li, Taifeng Dong, Andrew Davidson, and Qi Jing. Crop yield estimation in the canadian prairies using terra/modis-derived crop metrics. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2685–2697, 2020.

[37] Ziao Liu, Liping Di, Ruixin Yang, Liying Guo, Chen Zhang, Hui Li, and Bosen Shao. In-season crop yield prediction: State of the art and future research direction. *International Journal of Applied Earth Observation and Geoinformation*, 146:105129, 2026.

[38] David B. Lobell, George Azzari, Marshall Burke, Sidney Gourlay, Zhengxia Jin, Talip Kilic, and Siobhan Murray. Random forests for global and regional crop yield predictions. *PLOS ONE*, 11(6):e0156571, 2016.

[39] Wei Lu, Wiktor Adamowicz, Scott R Jeffrey, Greg G Goss, and Monireh Faramarzi. Crop yield response to climate variables on dryland versus irrigated lands. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 66(2):283–303, 2018.

[40] Pavithra Mahesh and Rajkumar Soundrapandiyan. Yield prediction for crops by gradient-based algorithms. *Plos one*, 19(8):e0291928, 2024.

[41] Gregory S McMaster and William W Wilhelm. Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology*, 87(4):291–300, 1997.

[42] Lieke A Melsen, Arnald Puy, Paul JJF Torfs, and Andrea Saltelli. The rise of the nash-sutcliffe efficiency in hydrology. *Hydrological Sciences Journal*, 70(8):1248–1259, 2025.

[43] Aparna S Menon, J Aravinth, S Rajendran, and P Kiran. Deep learning-based farm-level crop yield prediction using multi-temporal satellite data for complex engineering application. *Smart Agricultural Technology*, page 101562, 2025.

[44] RNV Mohan, Pravallika Sree Rayanoothala, and R Praneetha Sree. Next-gen agriculture: integrating ai and xai for precision crop yield predictions. *Frontiers in Plant Science*, 15:1451607, 2025.

[45] Alejandro Morales and Francisco J Villalobos. Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, 14:1128388, 2023.

[46] 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.

[47] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv*, abs/2211.14730, 2022.

[48] Milad Nouri and Shadman Veysi. Mitigating crop modeling uncertainties through machine learning in drylands. *Scientific Reports*, 15(1):42663, 2025.

[49] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[50] Dilli Paudel, Michiel Kallenberg, Stella Ofori-Ampofo, Hilmy Baja, Ron van Bree, Aike Potze, Pratishtha Poudel, Abdelrahman Saleh, Weston Anderson, Malte von Bloh, et al. Cy-bench: A comprehensive benchmark dataset for sub-national crop yield forecasting. *Earth System Science Data Discussions*, 2025:1–28, 2025.

[51] Rhorom Priyatikanto, Yang Lu, Jadu Dash, and Justin Sheffield. Improving generalisability and transferability of machine-learning-based maize yield prediction model through domain adaptation. *Agricultural and Forest Meteorology*, 341:109652, 2023.

[52] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*, 2024.

[53] AJ Rigden, N vD Mueller, NM Holbrook, Natesh Pillai, and Peter Huybers. Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting us maize yields. *Nature Food*, 1(2):127–133, 2020.

[54] Susan M Robertson, Scott R Jeffrey, James R Unterschultz, and Peter C Boxall. Estimating yield response to temperature and identifying critical temperatures for annual crops in the canadian prairie region. *Canadian Journal of Plant Science*, 93(6):1237–1247, 2013.

[55] Matthew Rodell, Paul R. Houser, Uwe Jambor, et al. The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381–394, 2004.

[56] Sarmistha Saha, Olga D Kucher, Aleksandra O Utkina, and Nazih Y Rebouh. Precision agriculture for improving crop yield predictions: a literature review. *Frontiers in Agronomy*, 7:1566201, 2025.

[57] Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to u.s. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37):15594–15598, 2009.

[58] Stefan Stiller, Kathrin Grahmann, Gohar Ghazaryan, and Masahiro Ryo. Improving spatial transferability of deep learning models for small-field crop yield prediction. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 12:100064, 04 2024.

[59] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58(80), 2025.

[60] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[61] Huiren Tian, Pengxin Wang, Kevin Tansey, Jingqi Zhang, Shuyu Zhang, and Hongmei Li. An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china. *Agricultural and Forest Meteorology*, 310:108629, 2021.

[62] 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.

[63] Kristof Van Tricht, Inbal Becker-Reshef, Benoit Franch, et al. Worldcereal: Global cropland and crop-type mapping system. *Scientific Data*, 10, 2023.

[64] Eric Vermote. Mod09cmg modis surface reflectance product. NASA MODIS Land Surface Reflectance Product, 2015.

[65] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[66] Muhammad Aamer Waqas, Xuan Wang, Saif Ali Zafar, Muhammad Aslam Noor, Hira Anee Hussain, Muhammad Ali Nawaz, and Muhammad Farooq. Thermal stresses in maize: Effects and management strategies. *Plants*, 10(2):293, 2021.

[67] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[68] Michael Wilton. Crop yield estimation using ndvi: a comparison of various ndvi metrics. 2021.

[69] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR, 2024.

[70] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[71] Seong D Yun and Benjamin M Gramig. Spatial panel models of crop yield response to weather: Econometric specification strategies and prediction performance. *Journal of Agricultural and Applied Economics*, 54(1):53–71, 2022.

[72] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[73] Jianyong Zhang, Yanling Zhao, Zhenqi Hu, and Wu Xiao. Unmanned aerial system-based wheat biomass estimation using multispectral, structural and meteorological data. *Agriculture*, 13(8):1621, 2023.

[74] Xiaoyang Zhang, Mark A Friedl, Crystal B Schaaf, Alan H Strahler, John CF Hodges, Feng Gao, Bradley C Reed, and Alfredo Huete. Monitoring vegetation phenology using modis. *Remote sensing of environment*, 84(3):471–475, 2003.

[75] Jingyuan Zhao, Fulin Chu, Lili Xie, Yunhong Che, Yuyan Wu, and Andrew F Burke. A survey of transformer networks for time series forecasting. *Computer Science Review*, 60:100883, 2026.

[76] Yan Zhao, Ruizhu Jiang, Jason Brider, Scott Chapman, and Andries Potgieter. Characterizing wheat and barley growth and phenology using multi-spectral remote sensing for site-specific precision agriculture. *in silico Plants*, 7(2):diaf013, 2025.