HYBRID PHENOLOGICAL MODELLING FOR ENHANCING PLANT PHENOLOGY PREDICTION

OTOBONG NSE

July, 2025

SUPERVISORS:

dr. M. Khodadadzadeh
dr, M. Farnaghi

HYBRID PHENOLOGICAL MODELLING FOR ENHANCING PLANT PHENOLOGY PREDICTION

OTOBONG NSE

Enschede, The Netherlands, July, 2025

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
dr. M. Khodadadzadeh
dr. M. Farnaghi

THESIS ASSESSMENT BOARD:
prof. dr. R. Zurita Milla
dr. C. Paris

ABSTRACT

Accurate phenological forecasting is vital for ecological monitoring, climate adaptation, and agricultural decision-making. Common modelling approaches, which include Process-Based Models (PBMs) and Machine Learning (ML), offer contrasting strengths: PBMs provide biological interpretability but struggle with ecological complexity and transferability, while ML models are data-driven and adaptable but often lack physiological transparency. This study investigated two hybrid phenology modelling strategies for predicting budburst dates under climate variability. The first strategy applied an ML-derived parameterization of the UNIFORC Process-Based Model, dynamically estimating site- and year-specific parameters. The second combined multiple fixed-parameter UNIFORC variants and machine learning regressors using a linear ensemble metamodel. Both approaches were evaluated over a 15-year period (2001–2015) using a year-forward chaining strategy and windowed climate data. The ML-derived parameterization was found to produce ecologically meaningful parameter dynamics, such as plausible interannual shifts in the thermal forcing threshold (f_crit ranging from ~41 to 48) and forcing onset (t0 ranging from DOY 25 to 29), while maintaining relatively consistent predictive accuracy (mean RMSE = 8.70, MAE = 6.99). In contrast, the ensemble model achieved comparable overall performance (mean RMSE = 8.17, MAE = 6.93), but exhibited a pronounced Regression-to-the-mean effect, evidenced by a lower slope (0.08) and R^2 (0.02) in observed vs. predicted plots. This study concluded that while both models offer predictive value, the ML-derived parameterization provides superior interpretability and biological transparency, making it more suitable for applications requiring ecological insight and climate adaptability.

Keywords: ML-derived Parameterization, Hybrid Modelling, UNIFORC, Year-Forward-Chaining, Parallel Ensemble

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Jesus, my Saviour and Redeemer, for guiding me through this study period. My very deep gratitude goes to my supervisors, Dr. M. Khodadadzadeh and Dr. M. Farnaghi, for their invaluable guidance, feedback, and support throughout this research. Thank you very much.

I am grateful to the ITC and the ITC Excellent Scholarship, University of Twente, for providing me with the scholarship and for providing the academic environment and technical resources essential to the successful completion of this MGEO programme. Special thanks go to the custodians of the PEP725 and ERA5 datasets, without which this research would not have been possible.

Finally, heartfelt thanks to my family and friends, especially my wife and son, for their patience, encouragement, and belief in me during this academic journey. I also wish to thank the Pastors and Members of the Redeemed Christian Church of God, Amazing Grace Sanctuary, Enschede, for giving me another home away from home. God bless you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI – Artificial Intelligence
CDD – Cold Degree Day
CROPGRO – Crop Growth Simulation Model (part of DSSAT)
DOY – Day of Year
DRF – Distributed Random Forest
DSSAT – Decision Support System for Agrotechnology Transfer
DWD – Deutscher Wetterdienst (German Weather Service)
ECMWF – European Centre for Medium-Range Weather Forecasts
ERA-5 – ECMWF Reanalysis 5th Generation
GDD – Growing Degree Day
GEE – Google Earth Engine
GLAM-Parti – General Large Area Model for annual crops (Partitioning variant)
IoT – Internet of Things
ITPM – Improved Temperature and Photoperiod Multiplicative model
L-BFGS-B – Limited-memory Broyden–Fletcher–Goldfarb–Shanno with Box constraints
LightGBM – Light Gradient Boosting Machine
LIME – Local Interpretable Model-agnostic Explanations
MAE – Mean Absolute Error
ML – Machine Learning
NDVI – Normalised Difference Vegetation Index
ORYZA – Rice growth and development model developed by IRRI
PBM – Process-Based Models
PEP725 – Pan European Phenology Network
RCP – Representative Concentration Pathways
RMSE – Root Mean Square Error
RUE – Radiation Use Efficiency
RF – Random Forest
SSP – Shared Socioeconomic Pathways
SALSA – Search, Appraisal, Synthesis, Analysis (literature review method)
SHAP – SHapley Additive exPlanations
SOYDEV – Soybean Development Simulation Model
SVR – Support Vector Regression
UAVs – Unmanned Aerial Vehicles
UNIFORC – Unified Forcing Model for phenology
WoS – Web of Science
XAI – Explainable Artificial Intelligence
XGB – XGBoost (Extreme Gradient Boosting)

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

Plant phenology, the timing of recurring biological events such as budburst, flowering, and leaf senescence, serves as a critical indicator of ecosystem function in response to climate variability (Richardson et al., 2012; Menzel et al., 2020). Shifts in phenological events are among the most sensitive and observable biological responses to climate change, and such shifts have profound implications for carbon cycling, species interactions, and agricultural productivity (Inouye, 2022). Over the past century, warming temperatures have altered the onset of phenophases across temperate and boreal biomes, with flowering and leaf-out now occurring 2–5 days earlier per decade in many regions (Menzel et al., 2020, Zeng et al., 2024). However, these responses are neither uniform nor linear. Species-specific sensitivities, interacting environmental drivers, and legacy effects complicate predictions, particularly as climate regimes grow increasingly non-stationary (Hänninen et al., 2019; Ponti & Sonolo, 2023; Zohner et al., 2023).

The ecological consequences of phenological change are far-reaching. Shifts in growing season length directly influence net primary productivity and carbon sequestration, with earlier spring onset enhancing CO_2 uptake but potentially exacerbating late-season drought stress (Piao et al., 2019; Keenan et al., 2014). Equally critical are the cascading trophic effects; phenological mismatches between plants and pollinators, or between budburst and herbivore emergence, can destabilize long-evolved ecological networks (Forrest & Miller-Rushing, 2010; Stuble et al., 2021). In agricultural systems, where phenology governs yield potential and frost risk, inaccurate predictions carry significant economic costs (Leolini et al., 2020; Devkota et al., 2023). Yet despite its centrality to climate adaptation, phenological forecasting remains fraught with uncertainty, in part because traditional modelling approaches struggle to reconcile mechanistic biological principles with the complexity of real-world environmental interactions (Chuine & Régnière, 2017; Dietze et al., 2018).

Process-based models (PBMs) have long dominated phenological research, encoding physiological knowledge into mathematical frameworks that simulate events like budburst through accumulated temperature (growing degree days, GDD) or chilling-forcing interactions (UniChill, PhenoFlex) (Chuine et al., 2000; Luedeling et al., 2021). For instance, simple models like Growing Degree Day (GDD) estimate budburst by summing daily temperatures above a base threshold (McMaster, 1997), while more advanced models, such as UNIFORC, incorporate nonlinear forcing functions and have been applied to temperate trees and grapevines (Chuine et al., 1999; Leolini et al., 2020).

PBMs are valued for their biological interpretability; their parameters represent physiological processes that can often be empirically estimated (Asse et al., 2020). This makes them well-suited for explanatory modelling and forecasting in data-sparse regions, provided the physiological assumptions remain valid (Chuine & Régnière, 2017). To better simulate spring phenology in cold climates, chilling-based models like Utah (E. A. Richardson et al., 1974) and UniChill (Chuine et al., 2000) were developed, with more recent models like PhenoFlex (Luedeling et al., 2021) unifying chilling and forcing phases for added flexibility.

Despite these strengths, PBMs face several limitations. Parameter estimation often relies on inverse modelling, which may produce physiologically implausible values or mask unmodelled processes (Akhavizadegan et al., 2021). The resulting equifinality can degrade accuracy when models are applied to novel conditions (Beven & Freer, 2001). Many PBMs also rely on oversimplified assumptions, e.g., GDD models may overlook soil moisture, complex photoperiod cues, or legacy effects (Piao et al., 2019; Way & Montgomery, 2015; Ogle et al., 2014). Even advanced models like PhenoFlex can mispredict under non-stationary climates if calibrated on outdated conditions, as shown in spruce forests where models overestimated budburst under recent warming (Gao et al., 2023).

These limitations underscore the need for more flexible, robust, and dynamically parameterised phenology models that remain valid under shifting climate regimes (Hänninen et al., 2019; Asse et al., 2020; Pei et al., 2024).

Machine Learning (ML) has transformed phenological modelling by identifying patterns in data rather than relying on predefined biological rules (Dai, et al., 2019; Garnot et al., 2025). Using predictors such as temperature, photoperiod, and precipitation, ML models like Random Forests and Support Vector Regression (SVR) have achieved budburst prediction errors as low as 7 days, significantly outperforming PBMs, which typically yield errors of 13–26 days (Li et al., 2023; Gao et al., 2023). ML can also detect non-obvious drivers, e.g., the influence of rainfall on cherry flowering, often missed by PBMs (Masago & Lian, 2022; Currier & Sala, 2022).

ML captures complex, nonlinear interactions and adapts to local conditions, especially in data-rich contexts (Tufail et al., 2023). However, its predictive strength comes at the cost of interpretability. Models are often seen as "black boxes," with outputs that may conflict with biological expectations, such as rainfall outweighing temperature as a key predictor, without clear mechanistic justification (Deng et al., 2024; Kim et al., 2025). This opacity is being addressed by explainable AI (XAI) techniques such as SHAP and LIME, though their adoption remains uneven (e.g. Lisboa et al., 2023; Patidar et al., 2024).

ML models also require large, diverse datasets for generalisation; performance degrades with sparse or unrepresentative training data (Maestrini et al., 2022; Mohammed et al., 2025). They struggle in novel climate regimes, often underestimating phenophases under warming scenarios. For instance, a PhenoFormer model trained on baseline climates consistently underestimated spring and autumn timings during warmer years, though it still reduced nRMSE by up to 8% compared to PBMs (Garnot et al., 2025).

Hybrid modelling frameworks have emerged as a promising synthesis, combining PBM's interpretability with ML's flexibility. Approaches range from using ML to parameterize PBM equations (Akhavizadegan et al., 2021) to embedding process-based constraints within neural network architectures (Willard et al., 2022). In viticulture, a hybrid model coupling grapevine phenology submodels with ML-derived climate projections improved harvest date forecasts by 4–7 days compared to PBMs alone (Leolini et al., 2023). However, significant challenges persist, particularly in quantifying and reducing uncertainty. Few models propagate input data errors or parameter uncertainties through to predictions (Dietze et al., 2018), while others neglect spatial autocorrelation in phenological observations, risking overconfident extrapolations (Pei et al., 2024).

As climate change accelerates, the need for biologically grounded models that are adaptable to novel conditions has never been more pressing, a gap this thesis seeks to explore. This research also attempts to advance the field of hybrid phenology modelling by critically exploring two distinct hybrid strategies: a machine learning-derived parameterization approach that dynamically estimates parameters of a plant phenology PBM (UNIFORC), and a parallel ensemble framework that integrates multiple fixed-parameter sets and machine learning regressors via a linear regression meta-learner. By comparing these strategies using a structured climate dataset and a year-forward chaining evaluation protocol, this study examines their predictive accuracy and capacity for biological interpretability, adaptability to climate variability, and robustness under future-like conditions. It contributes to the expanding body of research on hybrid phenological models that reconcile physiological insight with empirical performance.

1.2 Problem Statement

Accurate and interpretable prediction of plant phenology is essential for understanding ecological responses to accelerating climate change. While process-based models (PBMs) offer biological interpretability through mechanistic rules (Chuine et al., 2016; Hänninen et al., 2019), they are limited by rigid functional forms and the difficulty of parameter calibration, which restricts their adaptability across species, sites, and future climate scenarios (Körner & Basler, 2010; Beven & Freer, 2001; Vitasse et al., 2021). Conversely, ML models provide flexible, high-performing alternatives in data-rich settings but often sacrifice causal interpretability and struggle under extrapolation to novel climates or sparse datasets (Reichstein et al., 2019; Patidar et al., 2024; Dronova & Taddeo, 2022).

Hybrid models have emerged to combine PBM interpretability with ML adaptability (Maestrini et al., 2022; Ou et al., 2025). However, many existing hybrids either bypass or oversimplify PBM parameterization, neglecting the opportunity for ML to inform biologically meaningful parameters such as forcing thresholds or chilling sensitivity (Piña-Rey et al., 2023). Additionally, few studies rigorously test hybrid models under changing climatic baselines or systematically compare parameter learning, regression, and ensemble strategies (Garnot et al., 2025; Zhu et al., 2024).

To address these gaps, this research develops a hybrid phenological modelling framework that dynamically estimates PBM parameters using ML, integrates direct DOY prediction, and fuses outputs through interpretable ensemble models. The framework applies a window-based data approach and a year-forward chaining strategy to simulate real-world forecasting under shifting climate conditions. By incorporating iterative learning and multi-model comparisons, this work aims to deliver phenological predictions that are not only accurate but also biologically interpretable, generalizable, and resilient across diverse ecological and climatic contexts (Inouye, 2022; Zohner et al., 2023).

1.3 Research Aim

This research aims to develop and evaluate a hybrid phenological modelling framework that integrates process-based models with machine learning to predict budburst phenophases using a year-forward chaining approach, thereby simulating predictive performance under evolving climate conditions while ensuring accuracy, interpretability, and generalizability.

1.4 Research Objectives

The objectives of this research study are as follows:

1. To review and categorize existing hybrid phenological models according to their coupling strategies, data requirements, spatial and temporal scales, and target phenophases, to identify methodological trends, knowledge gaps, and opportunities for innovation.

2. To develop and evaluate two hybrid modelling approaches that integrate process-based and machine learning components for more accurate, interpretable, and generalizable phenological prediction

3. To evaluate the hybrid model's performance under evolving climate conditions.

1.5 Research Questions

To achieve the stated objectives, this research will attempt to answer the following questions:

1. What are the prevailing strategies in hybrid phenological modelling, and what methodological limitations and knowledge gaps remain?

2. How do different hybrid strategies, such as parallel ensemble modelling and machine learning-driven parameter estimation with feedback, compare their ability to integrate PBM and ML components for accurate and interpretable budburst prediction?

3. How effectively can hybrid models simulate future phenological responses under evolving climate conditions when evaluated using a year-forward chaining strategy and windowed climate data?

1.6 Report Structure

This report is organised in chapters with Chapter 1 providing background on phenology and its relevance under climate change, identifying the problems inherent in phenology modelling approaches, and outlining the research objectives and aim. Chapter 2 examines existing work on hybrid phenology models, structured using the SALSA framework (Scope, Approach, Learning, Scale, Applicability). Key gaps in the literature are identified to situate the current research. Chapter 3 describes the study area, dataset and data sources. Furthermore, we describe the hybrid modelling approaches, including quantitative and qualitative evaluation metrics. In Chapter 4, we present the results from the modelling and evaluation experiments. These results are discussed in Chapter 5, along with limitations of the research. Also, the chapter contains a summary of the research's findings and recommendations for future work.

2. LITERATURE REVIEW

2.1 Review of related works

2.1.1 Search and Appraisal

As shown in Figure 1, Comprehensive searches were conducted on Scopus, Web of Science (WoS), and Google Scholar using keywords such as “phenology modelling”, “process-based phenology modelling”, “machine learning phenology modelling”, and “hybrid phenology modelling”. Over 3,000 documents were initially retrieved (e.g., 1,788 from Scopus, 1,302 from WoS, and 1071 from Google Scholar).

Studies were screened for relevance based on title, abstract, and full-text availability. Deep learning, yield-focused, and remote sensing-only, as well as stand-alone Machine Learning, Process-based and comparative studies were excluded to retain focus on interpretable, biologically grounded hybrid models. Following duplicate removal and eligibility filtering, eight studies were selected for detailed synthesis.

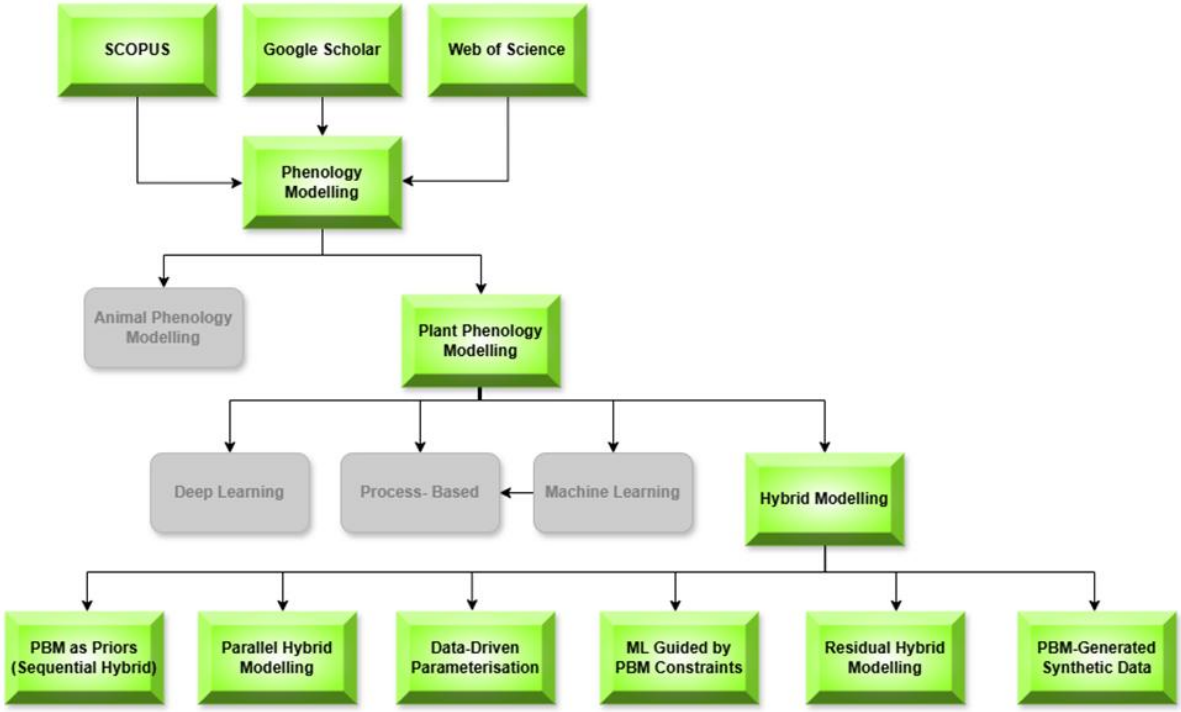


Figure 1 Structured Literature Review Procedure

2.1.2: Research Scopes

Reviewed studies spanned a range of objectives, primarily focusing on agricultural phenology for crops like rice (Zhang et al., 2024; Kawakita et al., 2023; Yu et al., 2025), soybean (McCormick et al., 2021), wheat (Droutsas et al., 2022), and sugarcane (Da Veiga Grubert, 2023), as well as tree phenology (Garnot et al., 2025). The phenological stages modelled include heading, flowering, maturity, leaf emergence, and yield. The common and overarching goal was to enhance prediction accuracy by leveraging PBMs’ mechanistic understanding and ML’s data-driven flexibility, addressing limitations such as nonlinearity in climate-phenology interactions or data scarcity. However, the scope varied in specificity: agricultural studies target actionable outcomes like planting schedules. The agricultural focus is highly relevant given global food security challenges. The lack of focus on underrepresented crops (e.g., millets or legumes beyond soybean) and non-crop plants highlights a gap in addressing diverse agricultural systems, particularly in vulnerable regions.

2.2 Hybrid Modelling Approaches

Six hybrid modelling approaches varying in the coupling strategies of the PBMs and ML components were reviewed, each with distinct mechanisms and theoretical underpinnings (See Table 1). The following sections describe each approach with concrete examples, followed by an analytical evaluation of its efficacy and limitations.

2.2.1 Sequential Hybrid

Sequential hybrid models work by using PBM outputs as inputs or priors for ML models, leveraging the mechanistic foundation of PBMs while using ML to refine predictions. For example, Zhang et al. (2024) used a crop phenology model to predict rice heading and maturity dates, which XGBoost refined using climatic (temperature, rainfall, solar radiation) and varietal data. The PBM captured physiological processes (e.g., thermal time accumulation), while the XGBoost modelled nonlinear interactions, achieving RMSEs of 4.65 and 5.72 days for heading and maturity, respectively. Similarly, Yu et al. (2025) employed the ORYZA PBM developed by the International Rice Research Institute (IRRI) to simulate rice phenological stages, with a Distributed Random Forest (DRF) refining predictions by learning climate-phenology interactions from temperature, solar radiation, and soil moisture data.

Sequential hybrids exploit PBMs' ability to model well-understood physiological processes (e.g., degree-day accumulation) while using ML to capture complex, data-driven patterns (e.g., climate or genetic variability effects). This is particularly effective in data-rich environments, as in Zhang et al. (2024), where 337 locations provide ample training data for XGBoost. However, the approach assumes PBM outputs are sufficiently accurate to serve as reliable inputs. ML may propagate or amplify errors if PBMs are miscalibrated (e.g., due to poor parameterization for local conditions). Additionally, sequential hybrids require extensive data preprocessing to align PBM and ML inputs, which can be computationally intensive.

2.2.2 Parallel Hybrid Modelling

The parallel hybrid strategy uses independent PBM and ML predictions, combined via ensemble methods to produce a final output. McCormick et al. (2021) used soybean PBMs (CROPGRO and SOYDEV) and an Ensemble Super Learner to generate independent predictions of flowering and maturity, weighted and aggregated into a final ensemble. Parallel hybrids mitigate individual model weaknesses by leveraging ensemble diversity. For instance, PBMs may excel in stable climates, while ML models adapt to anomalies, as seen in Arcomano et al. (2022). Ensemble weighting allows dynamic adjustment based on model performance, enhancing robustness. The stacking approach requires careful calibration of weighting schemes, which can be subjective or data-dependent. McCormick et al. (2021) do not specify how weights are determined, raising questions about reproducibility. Additionally, parallel hybrids increase computational complexity, as both models run independently.

2.2.3 ML-derived parameterisation

In this hybrid strategy, ML models are used to estimate or refine key parameters of a PBM, which are then used to drive PBM simulations. This approach leverages ML's ability to capture complex relationships from data to address one of the significant limitations of PBMs, manual or site-specific parameter calibration.

Droutsas et al. (2022) demonstrated this method using Random Forest and XGBoost to predict key physiological parameters, specifically Radiation Use Efficiency (RUE) and the rate of change of harvest index (dHI/dt), for the GLAM-Parti model. Although the primary output of interest was yield, GLAM-Parti is a land surface model with a phenology module, allowing it to simulate various stages of plant development. By accurately predicting RUE and dHI/dt, the ML models significantly improved the PBM's simulation of wheat biomass and yield under variable climatic conditions, capturing 98% of the observed variance. This illustrates how ML can enhance the PBM's performance by overcoming limitations in parameter estimation while preserving its physiological basis.

Similarly, Kawakita et al. (2023) applied an ML-derived parameterisation approach using genomic data to inform phenological model parameters. They focused on 11 heading date-related genes in rice, using haplotype combinations as predictors for calibrating parameters across four phenology models: the Beta model, simplified Beta model, ORYZA2000, and SIMIRIW. Various ML regressors, including support vector machines, random forests, and ridge regression- were used to predict model parameters from genotype information. These parameter-informed PBMs were then run to simulate heading dates, and their outputs were either used individually or combined into a dual-integrated ensemble. The resulting hybrid models achieved improved RMSEs of 5.9 days under known environments and 6.5 days under unseen test conditions, demonstrating strong generalisation. This work highlights the strength of using ML to account for genotype-dependent variability while preserving the mechanistic interpretability of PBMs.

2.2.4 ML Guided by PBM Constraints

In this hybrid approach, the PBM acts as a biological filter or constraint on the ML model, ensuring predictions remain physiologically plausible. Instead of allowing the ML model to learn freely from data, the PBM imposes rules based on known plant biology, for example, preventing the prediction of flowering before the vegetative stage. Worrall et al. (2022) applied this strategy using the Decision Support System for Agrotechnology Transfer (DSSAT) crop model to guide a neural network that predicted corn phenological stages from remote sensing data. DSSAT provided physiological limits and stage sequences that filter or restrict ML outputs during training or prediction. This helped maintain biological realism, especially in real-time in-season monitoring.

However, this approach depends heavily on the accuracy and flexibility of the PBM. If the PBM is poorly calibrated or oversimplified, it may constrain the ML model too much, limiting its ability to detect novel patterns or improve predictions. In Worrall et al.'s case, the study did not report detailed performance metrics, making it difficult to assess how much the PBM constraint improved results. Overall, PBM-guided ML offers greater interpretability and biological coherence, but care must be taken not to over-restrict the ML model's learning potential

2.2.5 Residual Learning

Using Residual Learning, ML refines PBM outputs by learning residuals between predictions and observations. For example, Garnot et al. (2025) used PhenoFormer, a transformer-based model, to refine PBM outputs for tree phenology (leaf emergence, colouration) in Switzerland, achieving a 13% R^2 improvement and 1.1-day RMSE reduction. Residual learning allows ML to focus on correcting PBM errors, leveraging transformers' ability to model temporal dependencies in phenological data. This is particularly effective for tree phenology, where long-term observations (70,000 in Garnot et al., 2025) enable robust learning. The downside of this approach is total reliance on training data. Also, Transformers are computationally intensive and require large datasets, limiting applicability in data-scarce regions. The approach also assumes PBM priors are reasonably accurate, which may not hold for poorly parameterised models.

2.2.6 PBM-Generated Synthetic Data

In this hybrid approach, process-based models (PBMs) are employed to generate synthetic datasets that serve as pretraining inputs for machine learning (ML) models, which are subsequently fine-tuned using observational data. Maestrini et al. (2022) applied this method by simulating crop growth and phenological development under diverse environmental conditions to create a rich training corpus. This PBM-generated data enabled the ML model to learn generalised relationships, which were later refined using real-world measurements to improve yield prediction. Although the study's primary objective was yield estimation, phenology played a critical role, as the timing of developmental stages strongly influences yield outcomes. The synthetic data encapsulated these dynamics, allowing the ML model to internalise biologically meaningful patterns even in data-scarce contexts.

This strategy is particularly advantageous when observational datasets are limited, as it enables ML to benefit from mechanistic knowledge encoded in PBMs. However, its effectiveness is contingent on the accuracy of PBM simulations. If the synthetic data are biased or poorly calibrated, they may mislead the ML model during pretraining, thereby introducing structural errors. Moreover, Maestrini et al. (2022) do not provide detailed performance metrics, which makes it challenging to evaluate the quantitative contribution of PBM pretraining. Nevertheless, the study illustrates a promising pathway for embedding physiological realism into data-driven models, potentially enhancing prediction robustness in environments where field data is sparse or variable.

2.3 Learning Structure and Choices

Across hybrid phenology modelling studies, the choice of process-based models is mostly coupled within land surface models. At the same time, the ML algorithm, on the other hand, is closely tied to the quantity and structure of available data and the complexity of the modelled phenological processes. Widely adopted algorithms included XGBoost (Zhang et al., 2024; Droutsas et al., 2022), Random Forests (Yu et al., 2025), neural networks (Worrall et al., 2022), ensemble learners (McCormick et al., 2021), and advanced deep learning architectures such as transformers (Garnot et al., 2025). Simpler regression-based models have also been employed for relatively straightforward predictive tasks (Kawakita et al., 2023; Da Veiga Grubert, 2023).

Engineered features typically reflect climatic and crop-specific variables, including temperature, precipitation, solar radiation, vegetation indices (e.g., NDVI), photoperiod, cultivar identity, and soil properties. For instance, Zhang et al. (2024) incorporated rice varietal data for genotypic differences. Garnot et al. (2025) included vapour pressure deficit to capture moisture stress dynamics in temperate forest systems. The quality and granularity of these features often depend on the study region's data infrastructure.

Supervised learning remains the dominant paradigm, with most models trained to predict continuous outcomes such as flowering dates or yield levels. Ensemble strategies, like those used by McCormick et al. (2021), offer robustness by aggregating multiple base learners, while transformer models (Garnot et al., 2025) leverage temporal dependencies in high-resolution time series datasets.

Critically, algorithm selection reflects both data richness and environmental complexity. Tree-based models (e.g., XGBoost, Random Forest) perform well on structured datasets with moderate complexity and are resilient to missing data. Deep learning models excel where data volume supports parameter learning, but risk overfitting and computational inefficiency in sparse or noisy datasets. The absence of methodological transparency in some studies (e.g., Maestrini et al., 2022), such as missing architectural or training details, limits reproducibility and transferability. Inconsistent feature selection, such as omission of soil parameters in Zhang et al. (2024), can further reduce model generalizability, particularly in environmentally heterogeneous contexts.

2.4 Current Scale of Reviewed Research

Reviewed studies spanned diverse spatial and temporal domains, yet tend to cluster around well-instrumented, agriculturally important, and climatically stable regions. At the global level, McCormick et al. (2021) evaluate soybean phenology across multiple continents, benefiting from extensive satellite and ground-truth datasets. Regional studies are more prevalent: Zhang et al. (2024) focus on 337 sites across China's rice-growing zones, Da Veiga Grubert (2023) on sugarcane in Brazil, and Garnot et al. (2025) on deciduous forest phenology using 70,000 records from Switzerland. Even studies that appear localized often reflect data-abundant, climatically stable zones. For example, Worrall et al. (2022) utilize long-term remote sensing and weather data from the U.S. Corn Belt. Maestrini et al. (2022) build a transfer learning framework for the Netherlands, using broader Northern European data as the pretraining domain. These contexts offer consistent phenological cycles, dense sensor coverage, and rich agronomic records, enabling sophisticated modelling approaches.

Underrepresented regions such as arid zones, mountainous areas, or tropical systems with complex multi-cropping receive little attention, despite their exposure to climate variability and phenological instability. This geographic imbalance introduces a contextual gap: the assumption that models trained in data-rich, stable systems generalise to data-sparse, highly variable contexts remains largely untested. Temporally, most studies aim at intra-seasonal or annual phenological predictions (e.g., flowering, maturity). This narrows the potential for hybrid models to inform long-term adaptation planning under climate change.

2.5 Research Applicability

Hybrid models demonstrate substantial promise across agricultural and ecological domains. High predictive performance is consistently reported in well-studied systems: Zhang et al. (2024) report RMSEs of 4.65 and 5.72 days for rice phenology, with R^2 values above 0.90; Droutsas et al. (2022) explain 98% of the variance in wheat biomass and yield using feature-augmented PBMs; and Garnot et al. (2025) reduce RMSE by 1.1 days for tree phenology predictions with transformer-based residual learning.

These applications align with practical goals: improved planting and harvesting windows (Zhang et al., 2024; Kawakita et al., 2023), enhanced yield forecasting (Maestrini et al., 2022), and in-season monitoring (Worrall et al., 2022). Notably, these studies operate in data-rich settings with access to long-term meteorological and phenological archives, conditions often absent in low-income or climatically extreme regions.

However, limitations persist. Several studies fail to report full evaluation metrics (e.g., Yu et al., 2025; Worrall et al., 2022), undermining cross-study comparability and real-world validation. High-end data requirements (e.g., satellite indices, genomic inputs) constrain model transfer to regions with limited infrastructure. Likewise, the computational demands of deep learning architectures (e.g., PhenoFormer in Garnot et al., 2025) limit scalability, particularly in operational or resource-constrained environments. In sum, while hybrid models yield strong performance where data availability and environmental stability align, their applicability in low-resource or ecologically dynamic regions remains underexplored. Addressing this discrepancy is critical for ensuring equitable access to phenology forecasting tools under global climate variability.

2.6 Gaps in Current Research

Despite the progress made in hybrid phenology modelling, several critical gaps and limitations persist, both contextually and operationally. Sequential hybrid approaches, while effective in refining PBM outputs, are susceptible to the propagation of errors originating from poorly calibrated or oversimplified PBMs. This dependence on the mechanistic model's baseline accuracy may limit its applicability in heterogeneous or data-sparse environments. In addition, there remains a lack of standardised evaluation frameworks. Studies often report different combinations of metrics (e.g., RMSE, R^2, MAE) or omit them entirely, making it difficult to benchmark model performance objectively. The geographical focus of existing literature is heavily skewed toward temperate and tropical systems, with limited attention paid to phenological modelling in arid, semi-arid, or underrepresented agroecologies.

Another significant gap lies in the limited crop and vegetation diversity addressed in current hybrid phenology modelling studies. The literature is heavily skewed toward globally dominant cereals such as rice, wheat, and soybean, reflecting data availability and global economic importance. However, this narrow focus overlooks regionally vital or climate-resilient crops, such as millets, pulses, sorghum, and indigenous varieties, that play a critical role in the food security and nutritional diversity of vulnerable populations, particularly in arid, semi-arid, and resource-constrained regions. Furthermore, non-crop plant species, such as native forest trees, grasslands, and shrubs, are largely excluded from hybrid modelling efforts, despite their ecological importance. These species underpin key ecosystem services, including carbon sequestration, water regulation, pollination support, and biodiversity maintenance. Their phenological responses to climate variability offer vital indicators of ecosystem health and resilience, especially under climate change (Lindborg et al., 2023; Qiu & Mitchell, 2024; Kattel, 2022). The absence of such vegetation types from modelling efforts limits the applicability of current models for conservation planning, biodiversity monitoring, and ecosystem management. Finally, most studies remain limited to seasonal phenology or short-term yield prediction. Few models explicitly incorporate real-time sensing or long-term forecasting capabilities necessary to assess phenological responses under climate change scenarios. In addition, no existing hybrid phenological modelling framework explicitly incorporates or accounts for the fuzzy and uncertain nature of phenological transitions. The field currently lacks formal treatment of phenophases as fuzzy intervals, probabilistic transitions, or uncertain windows, despite clear biological and observational justification for doing so and instead focuses on predicting a rigid DOY.

Table 1Categorization of Hybrid Modelling Approaches

Study Title	***Hybrid Coupling Method***	***PBM Used***	***ML Algorithm Used***	**Hybrid** Modelling **Specifics**
Predicting Rice Phenology Across China (Zhang et al., 2024)	Sequential Hybrid - PBM output used as ML input	Crop phenology model	XGBoost	The crop phenology model provides initial predictions, and ML use additional climatic and varietal data to refine initial predictions.
Dual Ensemble Approach for Rice Heading Date (Kawakita et al., 2023)	Sequential Hybrid - ML predicts genotype-specific priors for PBM	Multiple rice phenology models	Genetic parameter regression (ML-based)	ML estimates genotype-specific parameters fed into PBMs; final predictions are averaged from multiple models.
Comparing Physics-Based, Data-Driven, and Hybrid Models for Rice Phenology (Yu et al., 2025)	Sequential Hybrid - PBM outputs serve as ML features	ORYZA	Distributed Random Forest (DRF)	ORYZA predicts phenological stages, and DRF refines predictions using learned climate-phenology interactions.
Integration of ML into Process-Based Modelling for Complex Crop Responses (Droutsas et al., 2022)	Feature Augmentation - ML refines key PBM variables	GLAM-Parti	Random Forest, XGBoost	ML predicts key physiological parameters (Radiation Use Efficiency - RUE, rate of change of harvest index - dHI/dt), which are then used as inputs in the PBM.
In-Season Crop Phenology Using Remote Sensing & Model-Guided ML (Worrall et al., 2022)	ML guided by PBM constraints	DSSAT	Neural Networks (NN)	DSSAT guides NN training using phenological stage outputs, ensuring physiological plausibility.
Intercontinental Prediction of Soybean Phenology (McCormick et al., 2021)	Parallel Hybrid -	PBM for soybean phenology	Ensemble Super Learner	PBMs and ML models generate independent predictions, which are then weighted and combined into a final ensemble prediction.
PhenoFormer: Deep Learning Meets Tree Phenology Modelling Garnot et al., 2025)	Residual Learning - PBM output used as prior, ML refines it	Multiple PBMs	Transformer-based deep learning (ResNet-152)	PBM outputs serve as priors, and PhenoFormer refines them dynamically by learning from real-world observations.
Mixing process-based and data-driven approaches in yield prediction (Maestrini et al., 2022)	PBM- generated synthetic data used to pre-train the ML model	PBM for crop growth	ML models (not specified)	PBM simulations generate synthetic training data, which ML models use for pre-training before fine-tuning on observational data.

3. STUDY AREA, DATASETS AND METHODOLOGY

3.1 Study Area

This study focused on 229 points spread across Germany, a temperate country located in Central Europe, spanning latitudes 47°N to 55°N and longitudes 5°E to 15°E. The region experiences a predominantly temperate seasonal climate, with moderate winters, warm summers, and substantial interannual variability in temperature and precipitation (Zhu & Siebert, 2024; Deutscher Wetterdienst [DWD], 2023). Germany hosts a range of forest ecosystems, with *Fagus sylvatica* (European beech) being one of the dominant and ecologically significant deciduous species across its natural range. The species' sensitivity to thermal cues makes it an ideal candidate for phenological modelling under changing climatic conditions (Hacket-Pain et al., 2025; Vitasse et al., 2021).

3.2 Datasets and Sources

This study integrated phenological and climate datasets to develop and evaluate hybrid budburst prediction models. Table 2 summarises the key sources and datasets used, including phenological observations of *Fagus sylvatica* and associated climate variables.

3.2.1 Phenological Observations

Phenological records for *Fagus sylvatica* were obtained from the PEP725 database, aggregating harmonized phenological observations from national European networks. The dataset used in this study spans 2000 to 2015, covering multiple years of spring budburst observations across numerous German monitoring sites from which 229 were isolated for this research. All records were quality-checked and pre-filtered to exclude missing, inconsistent, or outlier values based on established phenological criteria (Templ et al., 2018).

3.2.2 Climate Data

Daily climate data for each of the 229 points were derived from the ERA5-Land Hourly Reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and accessed via the Google Earth Engine (GEE) platform from the Copernicus Climate Data Store. The dataset provides hourly climate variables at approximately 9 km spatial resolution and includes core surface meteorological drivers relevant to phenology. For this study, the data were aggregated to daily means or totals for 1969–2015 using the GEE JavaScript API (Muñoz Sabater, 2019; Gorelick et al., 2017).

Table 2: Datasets and Sources

Dataset	***Source***	***Temporal Coverage***	***Spatial Resolution***	***Key Variables***	***Purpose***
ERA5-Land	ECMWF Copernicus Climate Data Store	1969 - 2015	~9 km (0.1°)	Daily air temperature, dewpoint temperature, precipitation, soil moisture, pressure, solar and thermal radiation, latent heat flux, wind, VPD	Environmental drivers for phenology modelling
PEP725 Phenology	Pan European Phenology Network (PEP725)	1969 - 2015	Point-based (site-level)	Budburst dates (BBCH 11) for *Fagus sylvatica*, with site ID and geographic coordinates	Ground truth for model training and evaluation

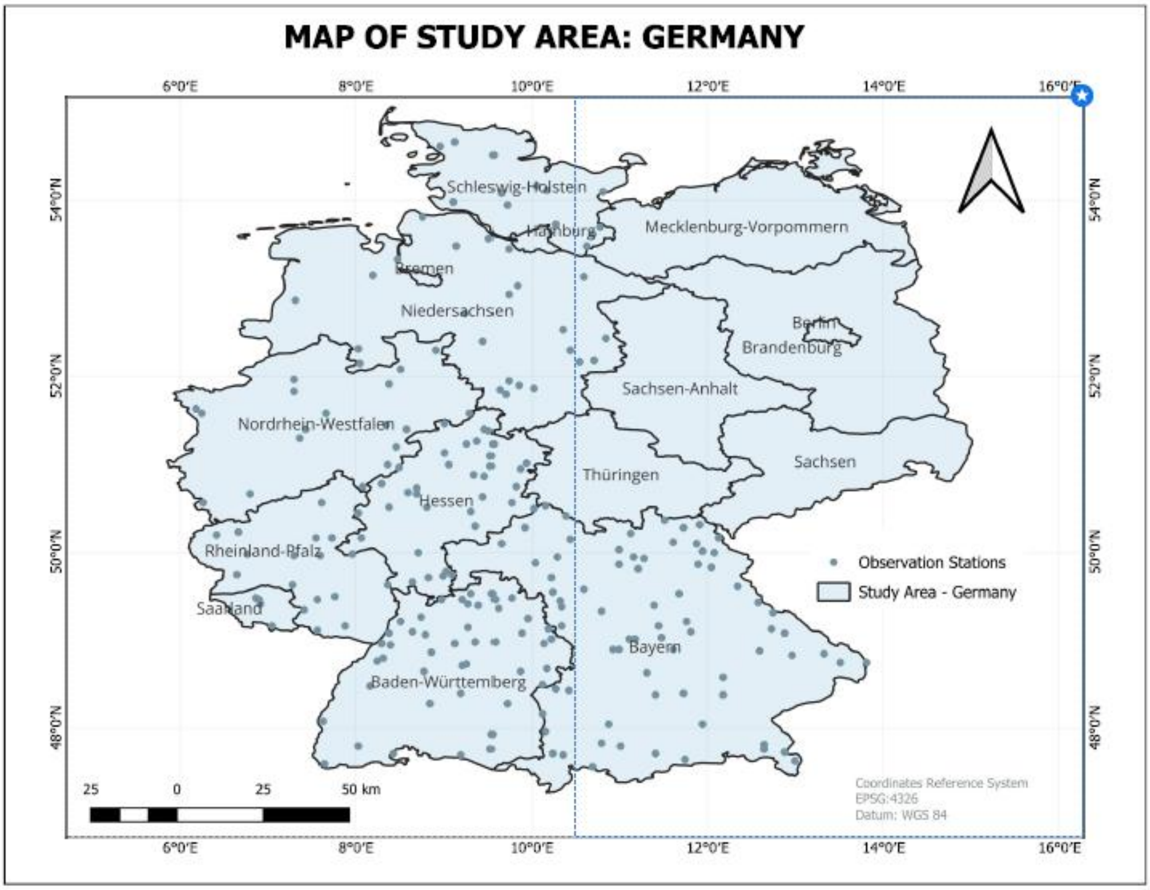


Figure 2: Study Area Map

3.3 Exploratory Data Analysis and Preprocessing

Prior to model development, the structure and quality of the phenology and climate datasets were assessed. This process involved data cleaning, outlier detection, temporal alignment, and standardization to ensure all variables were coherent and appropriately scaled for machine learning and process-based modelling.

Phenological data, obtained from the PEP725 database, consisted of ground-based observations of budburst (BBCH Stage 11) for *Fagus sylvatica* across Germany. From the initial dataset, 230 sites were selected based on their temporal completeness and spatial coverage. These sites represented those with consistent and uninterrupted observations from 1969 to 2015. Observations reflecting data entry errors or extreme departures from long-term trends were removed after manual inspection. The goal was to retain only phenological records that reflect natural, biologically realistic interannual variability.

The corresponding climate data were extracted from the ERA5-Land reanalysis archive via the Google Earth Engine platform. Daily values were computed by aggregating hourly measurements for each variable throughout the study period. Climate data were extracted from each phenology site's nearest ERA5-Land grid cell, offering an approximate spatial resolution of 9 km. Climate variables included daily mean air temperature, dewpoint temperature, surface pressure, precipitation, soil moisture, wind speed and direction, shortwave and longwave surface radiation, and latent heat flux. All variables were converted into standard units. For example, temperature values were converted from Kelvin to degrees Celsius, precipitation from meters to millimetres, and pressure from hectopascals to kilopascals. Energy and radiation fluxes were standardized to daily average W/m².

Data completeness was verified, and years with extensive missing values were excluded from further analysis. Derived metrics, including growing degree days (GDD), chilling hours, and vapour pressure deficit, were calculated from the cleaned variables to characterize temperature and energy dynamics relevant to plant development. These variables were selected due to their physiological relevance to dormancy break and spring phenology in temperate deciduous species (Basler & Körner, 2023; Zohner et al., 2023).

No further standardisation was applied as the ERA5-Land climate variables were pre-normalized and aggregated in Google Earth Engine using consistent methods and units. Finally, the phenological and climate datasets were temporally aligned site-yearly. Climate windows were extracted for each phenological event, ensuring that daily records preceding budburst were retained for feature extraction. The final dataset consisted of matched, cleaned, and unit-standardized records for 229 sites over 47 years, ready for use in the hybrid modelling frameworks. In Figure 3, a line plot shows the yearly average temperature distribution.

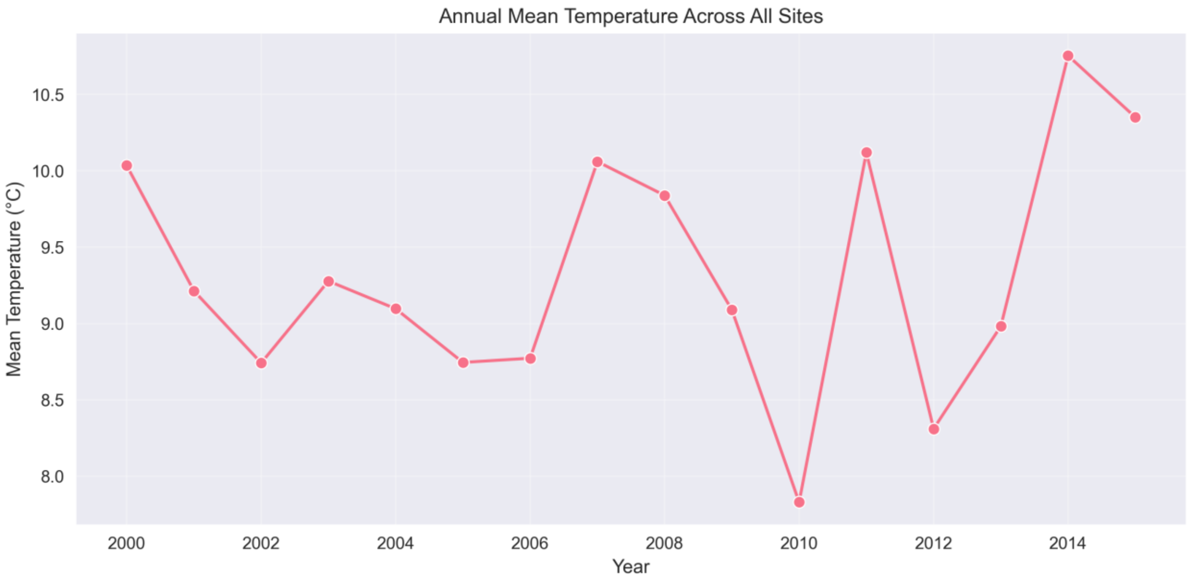


Figure 3Time Series of Mean Annual Temperature distribution across observed stations

The line plot illustrates annual mean temperature trends across all sites from 2000 to 2015. Considerable interannual variability can be seen, with temperatures fluctuating between just under 8.0°C and over 10.5°C. The lowest point occurs in 2010, followed by a sharp rebound and a noticeable warming trend leading up to 2014, the warmest year in the series. Although the pattern is not uniformly linear, the latter years, particularly from 2012 onward, suggest a possible shift toward higher average temperatures compared to the early 2000s. This may reflect a broader trend of recent warming.

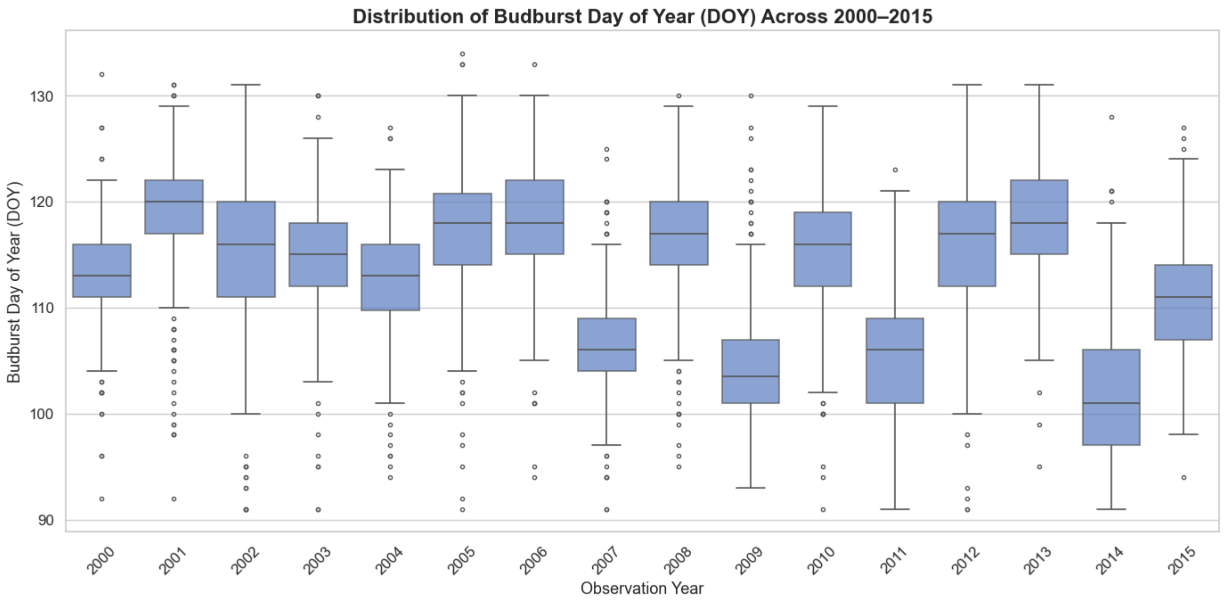


Figure 4Box plots of Budburst DOY distributions over the years

In Figure 4, the boxplot reveals notable budburst timing shifts corresponding to interannual temperature variation. In 2010, which recorded the lowest average temperatures, the distribution of budburst days is relatively delayed, with a higher median and a concentration of values later in the season. Conversely, 2014, identified as the warmest year, shows the earliest budburst timing in the entire period, with a markedly lower median DOY and a compressed distribution. This contrast underscores the sensitivity of budburst phenology to temperature, with warmer conditions generally advancing the onset of budburst and colder years delaying it.

3.4 Methodology

This chapter outlines the methodologies employed in this study. Data was structured using a sliding window approach and partitioned using a year-forward chaining strategy. Two hybrid modelling pipelines were developed. The first hybrid model adopts a data-driven parameterization strategy, where an ML model predicted PBM (UNIFORC) parameters, which were then used to simulate DOY. This pipeline included a feedback loop that triggered re-optimization of the parameters and re-simulation when certain conditions were met. The second combined the same PBM as the first, (UNIFORC) with five fixed parameter sets running simultaneously with three ML regressors (Random Forest, XGBoost, and Support Vector Regression), each independently predicting budburst DOY. Predictions from each base model were fused using a linear regression ensemblemodel to produce a single stacked output.

3.4.1 Temporal Data Splitting

A year-forward chaining strategy was employed for data splitting to simulate realistic forecasting conditions (Roberts et al., 2017). For each test year $y_{test} \in \{2001, 2002, \ldots, 2015\}$, the model was trained on all preceding years $\{y: y < y_{test}\}$. Formally, for each fold:

$$\text{Train}_y = \{(x_i, y_i) \mid \text{Year}_i < y_{test}\}, \text{Test}_y = \{(x_i, y_i) \mid \text{Year}_i = y_{test}\} \quad (1)$$

In equation (1), x is the covariates (climate variables) used to model phenology, while y is the Target variable (Budburst Day of year (DOY)).

This approach resulted in 15 temporally ordered folds, where training windows expanded over time (e.g., training on 2000 - 2004 to predict 2005). Unlike random or stratified splits, this method avoids data leakage by respecting the temporal sequence of climate phenology interactions. It also enables an evaluation of temporal generalizability, assessing how well models trained on historical data perform under potentially novel climatic conditions (evaluation procedure is explained in section 3.4.6). While this strategy may introduce class imbalance across folds, it remains consistent with the seasonal and interannual variability inherent to ecological time series.

3.4.2 Data Preparation and Feature Engineering

Cleaned and split climate and phenology data, including daily temperature, precipitation, latitude, photoperiod, and Budburst DOYs, were aggregated into 14-day sliding windows (Dai et al., 2022). These windows preserve the temporal dynamics leading up to budburst and allow the models to learn context-specific responses. Each sliding window comprised a multivariate time series that included mean temperature, precipitation, soil moisture, growing degree days (GDD), and chilling units. Features were explicitly labelled from day 1 to day 14 (e.g., tmean_day1, precipitation_day5, etc.), resulting in a high-dimensional feature space. Additionally, static variables such as latitude and photoperiod were appended to provide location-specific context. In total, the feature space comprised 72 features. The dependent variable for all modelling tasks was budburst's observed day-of-year (DOY), drawn from validated phenological records. This target variable served as the ground truth for model training and evaluation.

3.4.3 Process-Based Modelling Using the UNIFORC Model

The UNIFORC model defined in equation (2) served as the foundation for the process-based component of the hybrid frameworks. It simulates budburst timing by accumulating daily forcing units derived from a sigmoid function of mean temperature. For each day *t*, the forcing unit is defined as:

$$FU_t = \frac{1}{1+\exp\left(d\cdot(T_t-e)\right)} \tag{2}$$

where T_t is the daily mean temperature, d determines the steepness of the response curve, and e is the inflexion point representing the critical temperature at which bud development accelerates.

Budburst is assumed to occur on the earliest day when the cumulative forcing units exceed a species-specific threshold, f_{crit}. (Chuine, 1999). This threshold is accumulated from a fixed starting day-of-year t_0, leading to the condition modelled by equation (3):

$$\sum_{t=t_0}^{DOY} FU_t \geq f_{crit} \xrightarrow{yields} Budburst\ on\ DOY \tag{3}$$

The UNIFORC model requires a continuous daily temperature series for each site-year combination. A custom function consolidates temperature data from sliding windows into a full-year series (DOY 1 to 366). The resulting temperature series, $T = [T_1, T_2, \ldots, T_{366}]$, is input to the UNIFORC model. The model's sigmoid temperature response offers a biologically grounded tool for modelling Fagus sylvatica budburst in the study area, where spring temperatures between 2000 and 2015 showed considerable variability, but chill accumulation rarely fell below critical thresholds. The key parameters, slope (d), inflexion (e), and forcing threshold (fcrit), allow flexible response across cooler years like in 2010 and warmer springs such as in 2014. Although UNIFORC omits chilling, studies show that its forcing-only structure often outperforms more complex models in European beech phenology (Roberts et al., 2015). Additionally, field observations within the study area confirmed that temperature strongly governs *Fagus sylvatica* budburst: a long-term analysis across 405 German sites found that years with slower spring warming, such as 2010, corresponded with delayed and more variable budburst timing (Malyshev et al., 2022).

While the model's sensitivity to parameter estimates and its neglect of photoperiod or drought cues can limit predictions under novel climates, its integration into a hybrid framework, complemented by ML in this study, offers a promising approach to validating UNIFORC as a physiologically grounded backbone for reconstructing and forecasting budburst under historically observed temperature regimes.

3.4.4 The Parallel (Ensemble) Hybrid Model Architecture

The model architecture described in Figure 5 has the ML component running independently of the PBM, with its outputs stacked by a Linear Regression ensemble model (Dai, et al., 2022). In the following sections, the distinct components of the strategy are explained.

3.4.4.1 The Process-Based Component

Five fixed parameter set options were tested in this strategy (See Table 3). They varied in terms of temperature sensitivity (via d) and thermal threshold (via f_{crit}), although a common accumulation onset date (t_0) was set. Based on literature and prior modelling work, each parameterization was chosen to represent a plausible ecophysiological regime. The UNIFORC model was run deterministically for each site-year, yielding multiple candidate DOY predictions per parameter set.

Table 3Parallel Hybrid- Process-based Parameters

Set	***d***	***e***	***fcrit***	***t0 (start DOY)***
UN_1	-0.01	5.96	41.0	26.5
UN_2	-0.0	5.98	42.5	26.5
UN_3	-0.01	6.0	44.0	26.5
UN_4	-0.01	6.02	45.0	26.5
UN_5	-0.02	6.04	46.5	26.5

The selection of the five UNIFORC parameter sets was guided by the need to capture a range of phenological responses to interannual climate variability, particularly as reflected in the fluctuating temperature patterns observed between 2000 and 2015. As illustrated in Figure 3, the period featured considerable year-to-year variation, with colder anomalies such as 2010 and significantly warmer episodes in 2007, 2011, and 2014. The parameter sets were therefore not arbitrarily defined but were intentionally constructed to span a gradient of thermal sensitivity, each bringing distinct interpretive value to the ensemble framework.

UN_1, for instance, was defined with a low forcing threshold (fcrit = 41.0) and a relatively early inflexion point (e = 5.96°C), making it highly responsive to limited thermal accumulation.

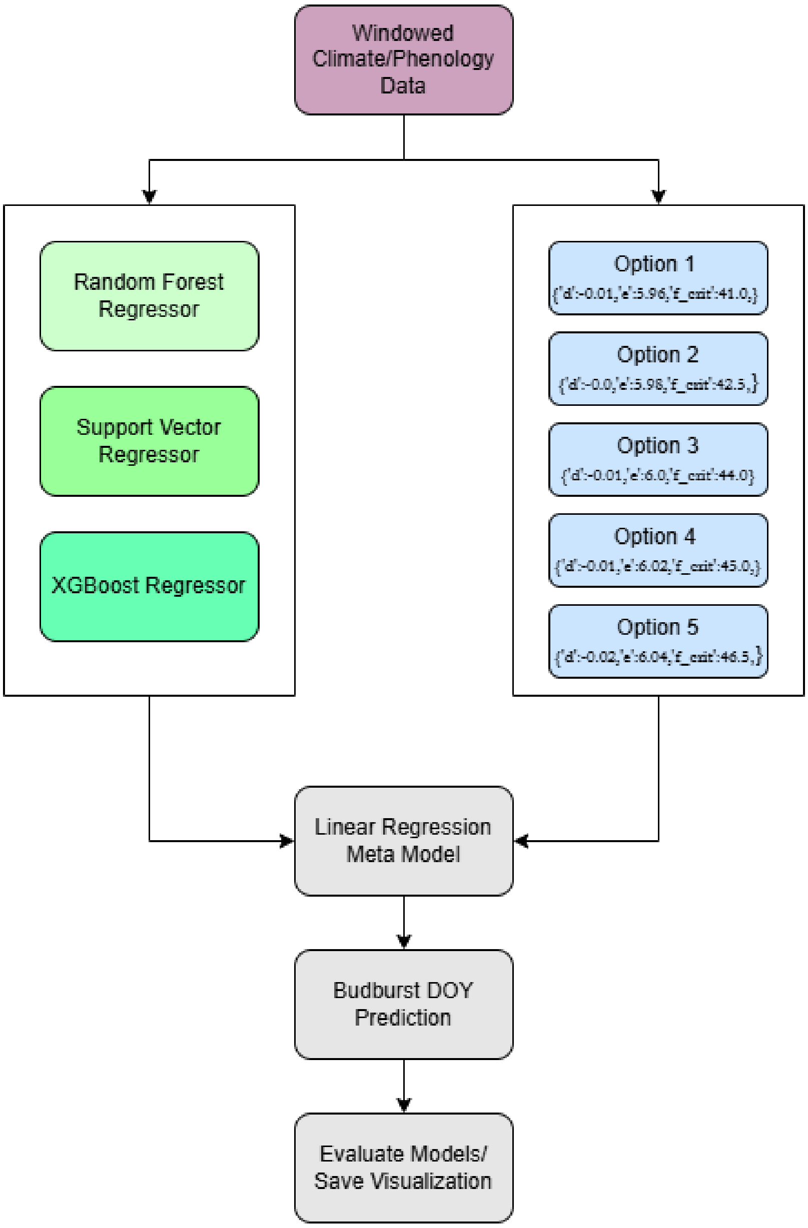


Figure 5Conceptual diagram of the Ensemble Model

This responsiveness is particularly useful in colder years, such as 2010, where temperatures may not consistently exceed higher activation thresholds; it allows the model to still register budburst events under constrained conditions, albeit with a risk of premature triggering under transient warming. Option 2, featuring a near-zero slope (d = –0.00), was included to represent a biologically smoother, less reactive forcing response. Though less precise, this variant helps the ensemble model to avoid erratic or overly aggressive predictions in noisy, ambiguous, or ecologically diffuse systems. It adds stability and realism where phenological cues are weak, indirect, or broadly distributed over time. Options 3 and 4 reflect a gradual tightening of the thermal requirements, with increasing fcrit and e values designed to delay budburst until more consistent warming is achieved. These configurations were calibrated to align with warmer years, such as 2007 and 2011, where their conservative settings improve fit by avoiding premature predictions. Option 5 represents the upper bound of thermal selectivity, with a steeper slope (d = –0.02) and the highest fcrit (46.5), making it particularly suitable for capturing delayed phenological responses in high-temperature years like 2014

Its sharp activation profile ensures that budburst is simulated only when warming is sustained and sufficient, thus avoiding overestimation in scenarios where early thermal spikes are not followed by continued heat accumulation.

Collectively, these parameter sets were intended to simulate plausible biological responses across a climatic gradient and strategically populate the ensemble model space with functionally diverse behaviours. In doing so, they enhance the parallel model's ability to generalize across species and years, allowing it to accommodate both early and late budburst tendencies and to capture the full range of ecological responses to shifting climate signals.

3.4.4.2 The Machine Learning Component

Three regression algorithms, Random Forest, XGBoost, and Support Vector Regression (SVR), were selected to form a diverse ensemble of machine learning models, each offering distinct advantages for modelling budburst under variable environmental conditions. These models, selected through GridSearchCV and trained on the same 72-feature input matrix, serve complementary roles in the hybrid pipeline. Hyperparameter configurations for each model ensured that each was optimally tuned for the training dataset rather than arbitrarily or heuristically defined. This methodological rigour enhances not only predictive performance but also comparability across algorithms.

The Random Forest Regressor, optimized with 150 estimators, a maximum depth of 15, and constrained splitting criteria (minimum samples split = 5, leaf = 3), was selected for its robustness in handling noisy, nonlinear ecological datasets. Its capacity for ensemble averaging enables it to generalize well in the presence of missing or collinear features, a common challenge when modelling climate-driven phenology using many interacting predictors (Li et al., 2023; Zhang et al., 2023). This model serves as the ensemble's stabilizer, reliably capturing broader trends while buffering against outliers and irregular seasonal signals.

In contrast, the XGBoost Regressor, fine-tuned with 200 estimators, a learning rate of 0.05, maximum depth of 6, and subsampling ratios of 0.7, contributes both bias reduction and high-resolution pattern detection through its gradient-based boosting mechanism (Ma et al., 2023). These hyperparameters were selected not only for optimal fit but also to enforce regularization and stable convergence, critical in modelling phenological events, which are often temporally clustered and governed by interacting environmental thresholds that require fine-scale discrimination without overfitting.

The SVR, configured with an RBF kernel, penalty parameter C = 10.0, and ε = 0.2, was included for its capacity to map non-linearly separable input-output relationships in high-dimensional feature spaces. While more computationally intensive and sensitive to hyperparameter tuning, SVR offers flexibility in scenarios where phenological responses follow curved or diffuse relationships, particularly in marginal climate zones where linear models or decision-tree-based splits may struggle (Rahimi et al., 2023). Its kernel approach allows the ensemble to account for latent interactions that might otherwise go undetected.

3.4.4.3 Hybrid Modelling via Stacked Linear Regression

The hybrid component of this methodology integrated the predictions from both the process-based and machine learning models through a stacked linear regression model. The meta-model takes as input the outputs of each base model, including all UNIFORC variants and ML regressors, and learns an optimal weighted combination that best approximates the true budburst DOY.

Formally, let y_i^j denote the prediction for instance i from model j. The hybrid model predicts DOY using equation (4):

$$y_i^{Hybrid} = \beta_0 + \sum_{j=1}^{p} \beta_j y_i^j \tag{4}$$

Where β_j are the coefficients learned through ordinary least squares regression, and β_0 is the intercept. The training of the meta-model was restricted to only instances where all component predictions and actual DOY values were available, ensuring data integrity during fusion. This approach offers two main advantages. First, it allowed for interpretability; the learned coefficients β_j reveal how much each base model contributed to the final prediction. A higher weight indicates that the meta-model relied more heavily on that model's output, offering insight into model relevance and reliability (Zhang & Ma, 2022). By combining multiple predictive sources, the stacked model can buffer against poor performance by any single base model. If one model underperforms in certain conditions, the meta-model can downweigh its influence, thereby reducing prediction error and improving generalization to new, unseen data.

3.4.5 ML-Derived Parameterization with Feedback Loop

This hybrid strategy frames phenological modelling as a parameter-learning task. ML models (LightGBM) are trained to predict physiological parameters (d, e, fcrit, t0) of the UNIFORC model from environmental features. Rather than directly predicting budburst DOY, the ML model infers the internal structure of the process-based model PBM, which is then used to simulate budburst DOY. The prediction error between the simulated and actual DOY is minimized through numerical optimization using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) (Zhu et al., 1997) algorithm, and the improved parameters are used to retrain the ML model. This feedback loop iterates until error convergence, thereby gradually aligning ML-inferred parameters with biologically plausible outputs. Figure 6 shows the conceptual flow of the strategy. This hybrid architecture estimates the internal parameters of the UNIFORC process model, specifically d, e, and $f$crit from environmental features using ML. This reframes phenological modelling as a parameter inference task, wherein machine learning predicts the biological response parameters that control the timing of budburst. Although L-BFGS-B is used in both stages, its roles are distinct: initial refinement to generate training targets vs. feedback correction of ML-predicted parameters.

3.4.5.1 Initial Parameter Derivation

Initial UNIFORC parameters (d, e, f_crit, t0) are derived heuristically from literature. These initial values are loosely optimized using the L-BFGS-B algorithm (Zhu et al., 1997) for up to 20 iterations, constrained within biologically plausible bounds. The goal is not fine-tuning but rather to generate reasonable targets for supervised learning. This step avoids arbitrary assignments and ensures empirical observations inform the starting parameters.

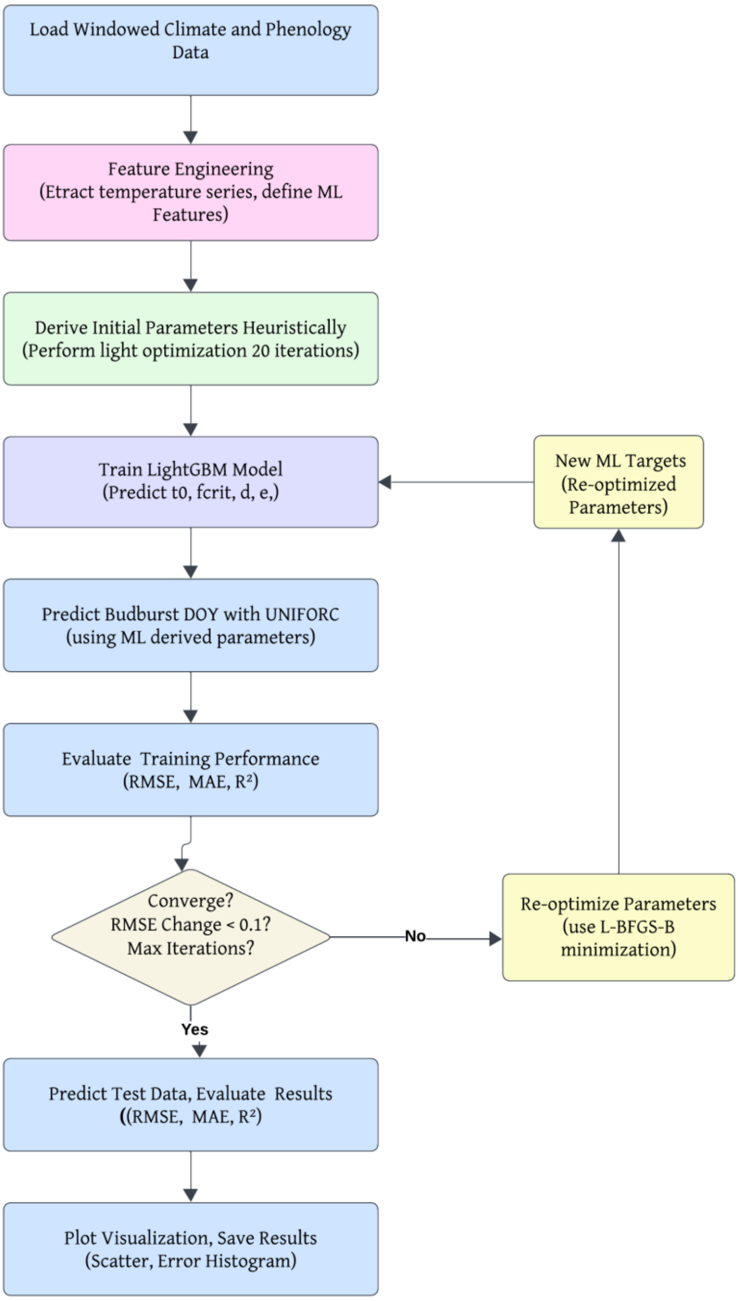


Figure 6Conceptual diagram of the ML-derived parameterization hybrid model strategy

3.4.5.2 Machine Learning Parameter Prediction

LightGBM regressors are trained to map climate features (mean temperature, photoperiod, GDD, etc.) to the initial parameter estimates. Separate models are trained for each parameter. These regressors are then used to predict site-specific UNIFORC parameters, which are input into the PBM to simulate Budburst DOY. This indirect modelling ensures predictions remain constrained by the mechanistic structure of UNIFORC, grounding them in plant physiology. LGBM was selected for predicting UNIFORC parameters due to its speed and scalability, which are critical in the iterative feedback process involving multiple rounds of re-training and optimization.

Its leaf-wise tree growth enables fine-grained learning of site- and climate-specific interactions influencing UNIFORC's parameters, while maintaining low computational cost compared to XGBoost or Random Forest. This efficiency ensures faster convergence in the looped structure of the hybrid phenology model without sacrificing predictive accuracy (Ke et al., 2017).

3.4.5.3 Parameter Re-optimisation

After predicting DOY using the ML-derived parameters, the model evaluates the difference between simulated and observed budburst dates. If the error exceeds acceptable limits, the predicted parameters are again refined via numerical optimization using L-BFGS-B, minimizing the squared error between predicted and actual DOY. This optimization uses the ML predictions as initial guesses and is constrained by clipping them within biologically reasonable parameter bounds ($d \in [-0.13, 0.13]$, $e \in [4.5, 9.5]$, $f\_crit \in [20.0, 100.0]$). This step corrects ML-induced errors by aligning predictions with actual plant responses.

3.4.5.4 Iterative Refinement

The newly optimized parameters are used as updated targets to retrain the ML regressors. This closes the feedback loop, where each iteration incrementally improves the ML models' ability to infer biologically meaningful parameters. In each repeat cycle:

1. Initial parameters are heuristically generated and lightly optimized
2. ML predicts parameters using lightly optimized parameters.
3. UNIFORC simulates DOY using these parameters.
4. Prediction errors are minimized via re-optimization.
5. Updated parameters serve as new targets and are passed back to ML.

This loop continues until one of the following stopping criteria is met:

- A maximum of five iterations was reached. (Five is selected for convenience, since a larger part of the parameter is optimized in the first iteration.)
- Change in RMSE falls below a threshold ($\Delta RMSE < 0.1$)
- Training RMSE reaches a predefined target ($RMSE \leq 7.0$)

This iterative design ensures convergence by reducing error while preserving physiological realism through repeated PBM grounding.

3.4.5.5 Prediction and evaluation of DOY on Test data

Once convergence is achieved, or a stopping criterion is met, the final iteration's ML models are used to predict parameters on held-out test data. These parameters are input into UNIFORC to simulate budburst DOY for unseen samples. Performance is assessed using RMSE, MAE, and R^2, evaluating both the accuracy and generalizability of the final hybrid model. By merging ML's predictive flexibility with PBM's mechanistic structure, this architecture yields interpretable and adaptive phenological forecasts suited to varying environmental conditions and long-term climate shifts.

3.4.6 Model Evaluation and Visualisation

To assess model performance, three standard regression metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) (See equations (5), (6) and (7) respectively). These metrics were computed year by year, allowing performance evaluation across temporal variability, after which values were averaged across years to obtain an overall summary metric for each model. This ensured that no year disproportionately influenced the final assessment and supported fair comparison across hybrid strategies.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}i| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} \tag{7}$$

The Mean Absolute Error (MAE) estimates the average deviation in day-of-year (DOY) between predicted and true budburst DOY. It is robust to moderate outliers, making it suitable for datasets with local variability (Karunasingha, 2022). In contrast, the Root Mean Squared Error (RMSE) penalizes larger errors more heavily, thus highlighting the impact of substantial prediction deviations, critical for phenology models used in time-sensitive decisions like planting or harvest (Hodson, 2022). The coefficient of determination (R^2) quantifies the model's ability to explain interannual and temporal variability, with higher values indicating better generalization (Fu et al., 2014).

In addition to these metrics, scatter plots of predicted versus observed DOY were plotted and saved for visual inspection, enabling detection of systematic bias (deviation from the 1:1 line), clustered errors, or residual patterns not evident in aggregate statistics (Zhang et al., 2023). These visual diagnostics complement numerical evaluation by revealing spatial or temporal inconsistencies that may inform further model refinement.

4. RESULTS

4.1 Results Overview

This section presents the outputs and performances of the two hybrid phenological modelling approaches explored in this research: (1) an ML-parameterized UNIFORC model, which dynamically tunes physiological parameters, and (2) a parallel hybrid ensemble, integrating multiple fixed-parameter PBM sets with multiple ML regressors. Predictive accuracy, generalization, parameter interpretability, and ensemble component contributions across 2001 - 2015 were assessed, using year-forward validation, with root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) as metrics.

4.1.1 Predictive Accuracy Across Years

To assess the temporal robustness of both hybrid phenology modelling strategies, we evaluated their predictive accuracy across the 15-year (2001–2015) study period using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). These metrics, summarized in Table 4, offer insight into annual model behaviour under realistic forecasting conditions using year-forward validation.

Table 4Regression metrics of both modelling approaches

S/No	*Year*	*RMSE(days)*	*MAE(days)*	*R^2*	*RMSE(days)*	*MAE(days)*	*R^2*
		ML-derived Parametrisation			**Metamodel**		
1	2001	8.62	7.39	-0.67	8.74	7.84	-0.72
2	2002	9.38	7.37	-0.41	7.91	5.70	0.00
3	2003	8.51	6.53	-1.02	6.52	4.86	-0.19
4	2004	8.81	7.4	-1.34	5.91	4.60	-0.05
5	2005	8.35	7.01	-0.6	11.71	10.31	-2.14
6	2006	11.05	9.56	-2.85	6.83	5.37	-0.47
7	2007	6.71	4.95	-0.55	6.83	5.54	-0.60
8	2008	7.25	5.6	-0.45	6.15	4.91	-0.04
9	2009	10.16	8.64	-1.56	13.86	12.89	-3.78
10	2010	8.88	7.37	-0.99	7.34	5.42	-0.36
11	2011	12.81	11.26	-3.41	11.94	10.72	-2.83
12	2012	7.54	5.49	-0.11	8.07	6.54	-0.28
13	2013	7.45	5.94	-0.86	8.23	6.40	-1.27
14	2014	12.19	12.19	-2.02	10.56	8.92	-1.27
15	2015	6.58	5.28	-0.47	6.46	5.07	-0.42
Average		**8.95**	**7.47**	**-1.15**	**8.47**	**7.01**	**-0.96**

Model performance dynamics, illustrated in Figure 7, provide a visual expression of each model's trend. Contrary to a uniform trend, neither model exhibited consistent dominance across all years. The parallel ensemble model outperformed 8 of the 15 years, while the ML-derived parameterisation model led in 7. For instance, in 2004, a year of relatively stable climatic conditions, the ensemble model achieved an RMSE of 5.91 compared to 8.81 from the ML-derived model, suggesting superior responsiveness to low-noise signals. Similarly, in 2010 and 2012, the ensemble model showed lower MAE values, although the ML-derived model slightly outperformed in RMSE during 2012 (7.54 vs. 8.07). In 2013, the ML-derived model again posted better RMSE and MAE, indicating better adaptation during moderate variability.

In contrast, during highly anomalous years, performance varied. In 2011, the ML-derived model performed worse across all metrics (RMSE: 12.81 vs. 11.94; MAE: 11.26 vs. 10.72), while in 2014, both models struggled (negative R^2 values), but the ensemble model still outperformed the ML-derived model in RMSE (10.56 vs. 12.19) and MAE (8.92 vs. 12.19), suggesting better resilience to extreme conditions.

Interestingly, 2015 marked a convergence point. Both models recorded their lowest RMSEs (ML: 6.58; Ensemble: 6.46) and their least negative R^2 values (-0.47 and -0.42, respectively), suggesting favourable environmental signals that both approaches could interpret effectively.

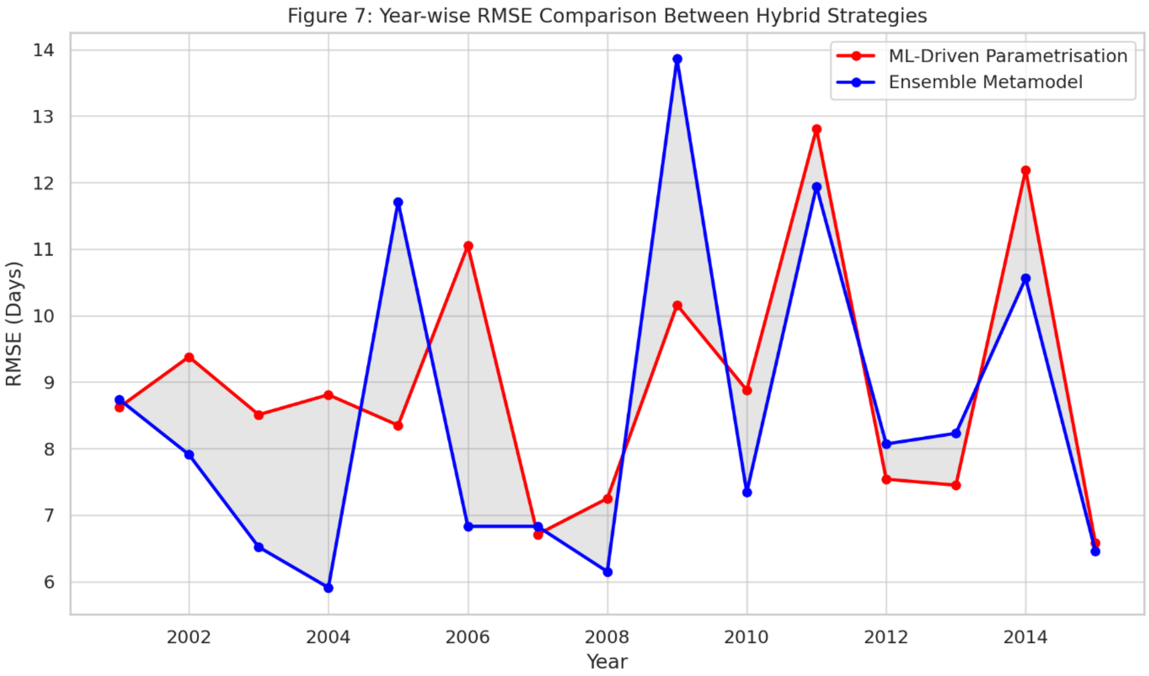


Figure 7: Year-wise RMSE comparison between both hybrid strategies

On average, the ensemble model achieved a marginally better RMSE (8.47 days vs. 8.95) and MAE (7.01 vs. 7.47) than the ML model. Despite this, both models recorded negative averageR^2 values, suggesting they generally explained less variance than a baseline climatology, as illustrated in Figure 8.

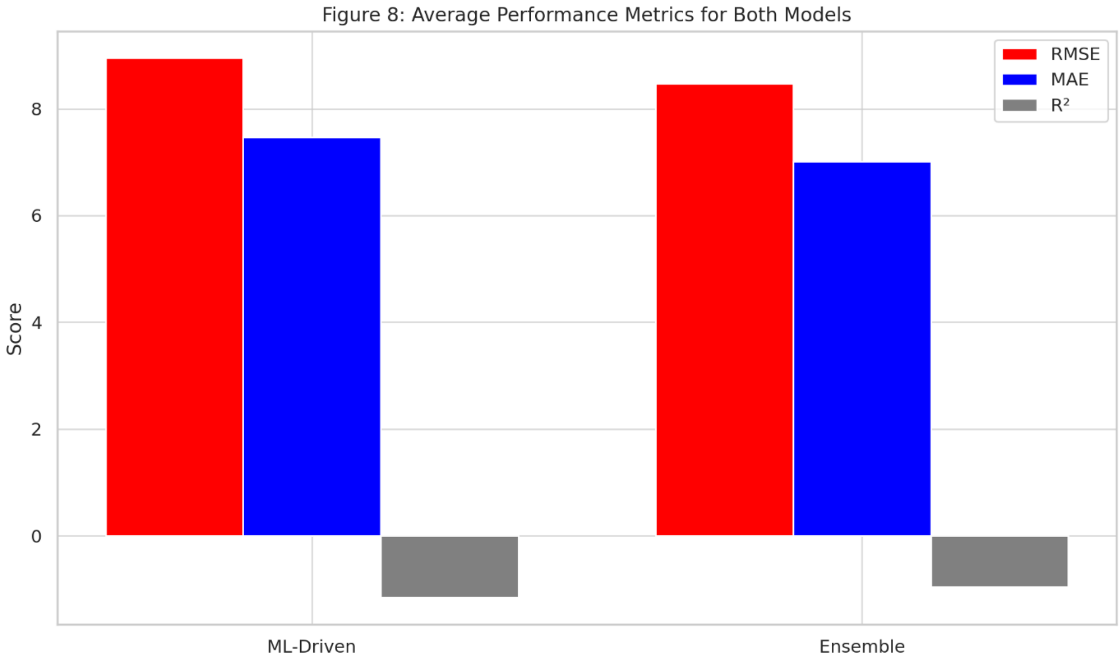


Figure 8: Average performance metrics for both models

To assess statistical significance, a Wilcoxon signed-rank test (See Table 5) was conducted comparing year-wise RMSE and MAE values. The resulting p-values 0.30 for RMSE and 0.42 for MAE indicate that performance differences between the ML-derived parameterisation model and the ensemble model are not statistically significant at the 5% or even 10% level. Although the ensemble model shows slightly lower error metrics, the differences are not strong enough to rule out chance variation. The Wilcoxon test was selected because it makes no assumptions about the normality of the error distribution and is more robust to outliers and small sample sizes, conditions common in interannual phenological evaluations (Wilcoxon, 1945; Demšar, 2006). This reinforces the interpretation that both models perform comparably across years, with observed differences likely reflecting specific temporal or climatic contexts rather than consistent structural superiority.

Table 5: Wilcoxon test results and confidence intervals for RMSE/MAE averages

Metric	***ML Mean***	***ML 95% CI***	***En-model Mean***	***En-model 95% CI***	***Wilcoxon p-value***
RMSE	8.95	(8.00, 9.91)	8.47	(7.24, 9.70)	0.30
MAE	7.47	(6.38, 8.55)	7.01	(5.71, 8.30)	0.42

To complement this year-wise evaluation, average RMSE and MAE were computed across all years and visualised with 95% confidence intervals (Figure 9). The ensemble model showed slightly lower mean RMSE (8.47 days) and MAE (7.01 days) compared to the ML-derived parameterisation model (RMSE: 8.95 days, MAE: 7.47 days). However, overlapping confidence intervals and the non-significant p-values confirm that these differences lack statistical robustness. Notably, the ML model exhibits greater interannual variability, suggesting it is more sensitive to specific environmental conditions, whereas the ensemble model displays more stable performance. This complementary behaviour strengthens the case for hybrid or model-switching strategies in operational phenology forecasting.

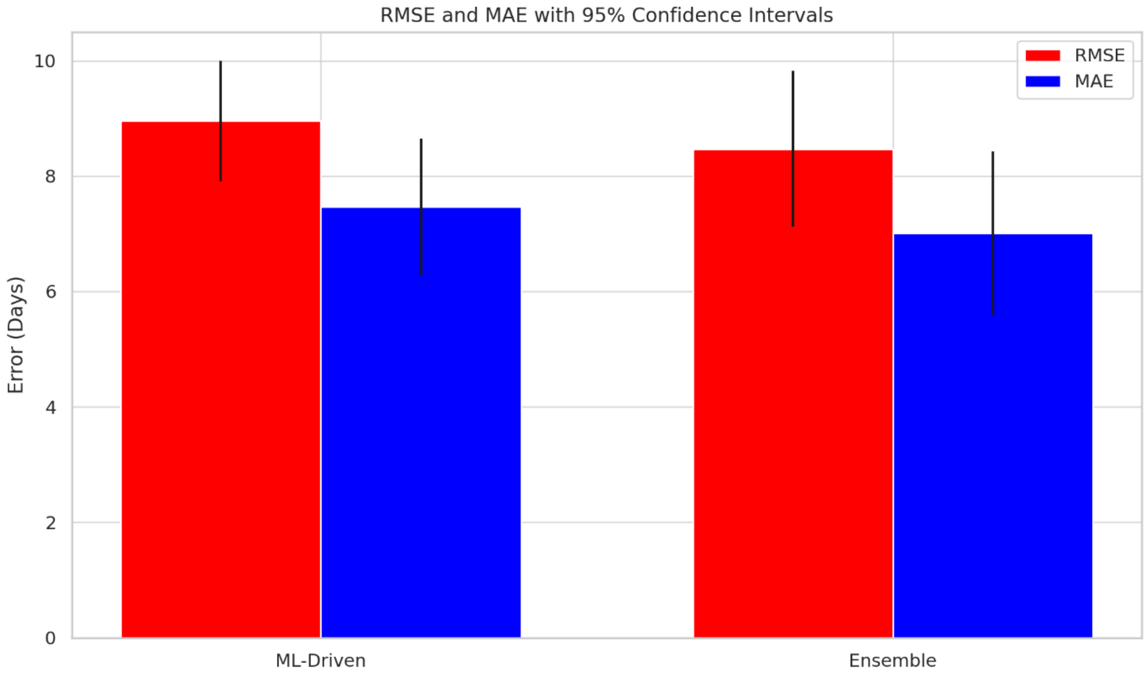


Figure 9: RMSE vs MAE error bars; 95% CI overlay

4.1.2 Prediction Error Distributions and Structural Biases

To gain deeper insight into model performance beyond summary statistics, this section evaluates the distribution and structural characteristics of prediction errors across all site-year combinations. Specifically, we examine the error spread, skewness, temporal patterns, and alignment of predicted vs. observed DOY values.

4.1.2.1 Histogram and Density of Residuals

Figure 10 displays histograms and kernel density estimates (KDEs) of residuals (predicted DOY minus observed DOY), aggregated across all years for both the ML-derived and ensemble models. Table 6 shows the summary statistics of the residual analysis. While both distributions centre loosely around zero, the ML model shows a visibly wider and right-skewed error spread, reflecting its positive bias (mean error = 3.10). In contrast, the ensemble model exhibits a more peaked distribution near zero (mean error = 0.53), with narrower tails.

Though skewness values (ML = 0.14, Ensemble = 0.28) suggest only slight asymmetry in both models, the KDE curves clearly show that the ML-derived model produces more extreme overpredictions. This visual evidence overrides the minor statistical skew difference, highlighting the importance of combining quantitative and graphical diagnostics.

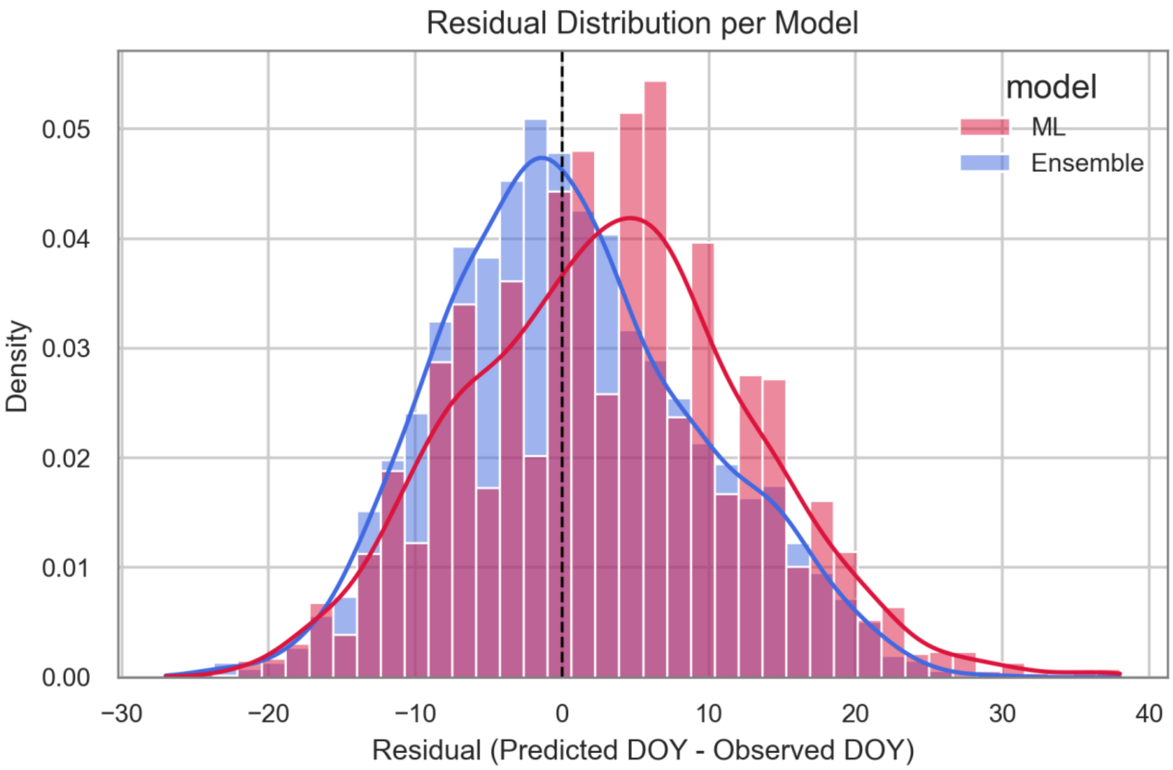


Figure 10: Histogram and KDE plots of residuals per model

Table 6: Residual Summary Statistics

Model	*Count*	*Mean Error(days)*	*Median Error (days)*	*Std Dev (days)*	*Skewness*	*Kurtosis*	*Min Error (days)*	*Max Error (days)*
Ensemble	3166	0.53	-0.25	8.79	0.28	-0.22	-26.31	29.89
ML-derived	3166	3.1	3	9.58	0.14	-0.04	-27	38

Overall, the ML-derived model exhibits broader error dispersion and a heavier upper tail, while the ensemble model is more concentrated around the mean with fewer extreme deviations.

4.1.2.2 Scatterplots and Structural Deviations

Figures 11 and 12 present combined scatterplots of predicted versus observed budburst DOY values for the ML-derived and ensemble models, respectively. Notably, these scatterplots aggregate predictions across a 15-year period, encompassing substantial inter-annual climatic variability. While this aggregation enables a holistic view of model behaviour, it can also obscure temporal nuances, such as year-specific strengths or weaknesses in prediction accuracy.

For the ML-derived model (Figure 11), the R^2 value of 0.04 indicates that only 4% of the variance in observed budburst DOY is explained by the fitted linear relationship over the aggregated dataset. The associated trendline, with a slope of 0.18, is markedly flatter than the ideal 1:1 line, revealing a strong regression-to-the-mean effect. This implies that the model consistently predicts within a narrow DOY range and struggles to reflect interannual extremes, particularly early or delayed phenological events that deviate from climatological norms.

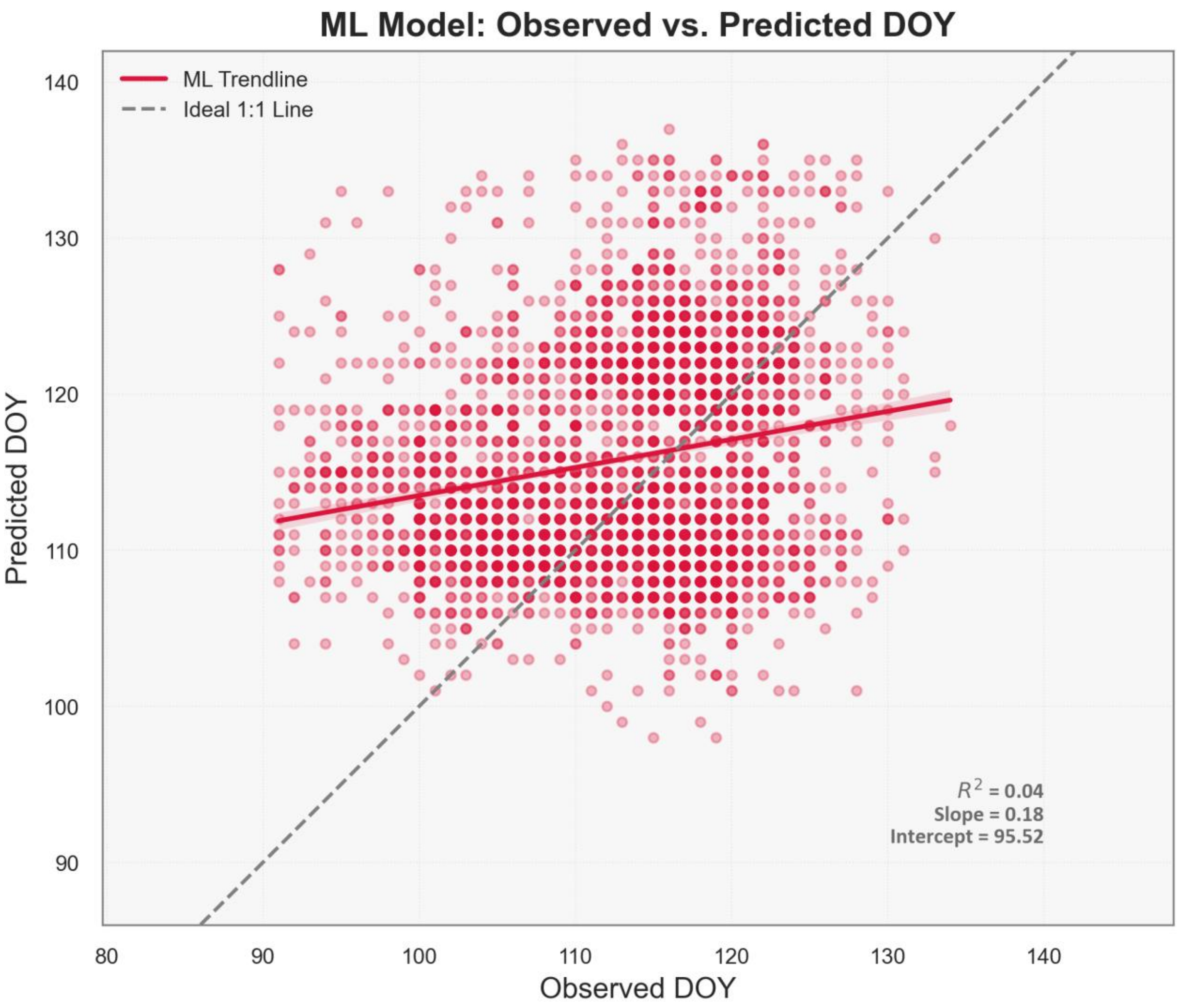


Figure 11: ML-Parameterised Scatter plots of predicted vs observed Budburst DOY across 2001 - 2015

The ensemble model (Figure 12) performs even more weakly in structural terms, with an R^2 of just 0.02 and an even flatter slope of 0.08. The dense vertical clustering of predictions indicates a strong central bias, suggesting the model emphasizes stability over responsiveness. This behaviour is consistent with ensemble averaging, which tends to dampen predictive variance, potentially at the cost of reduced sensitivity to environmental anomalies.

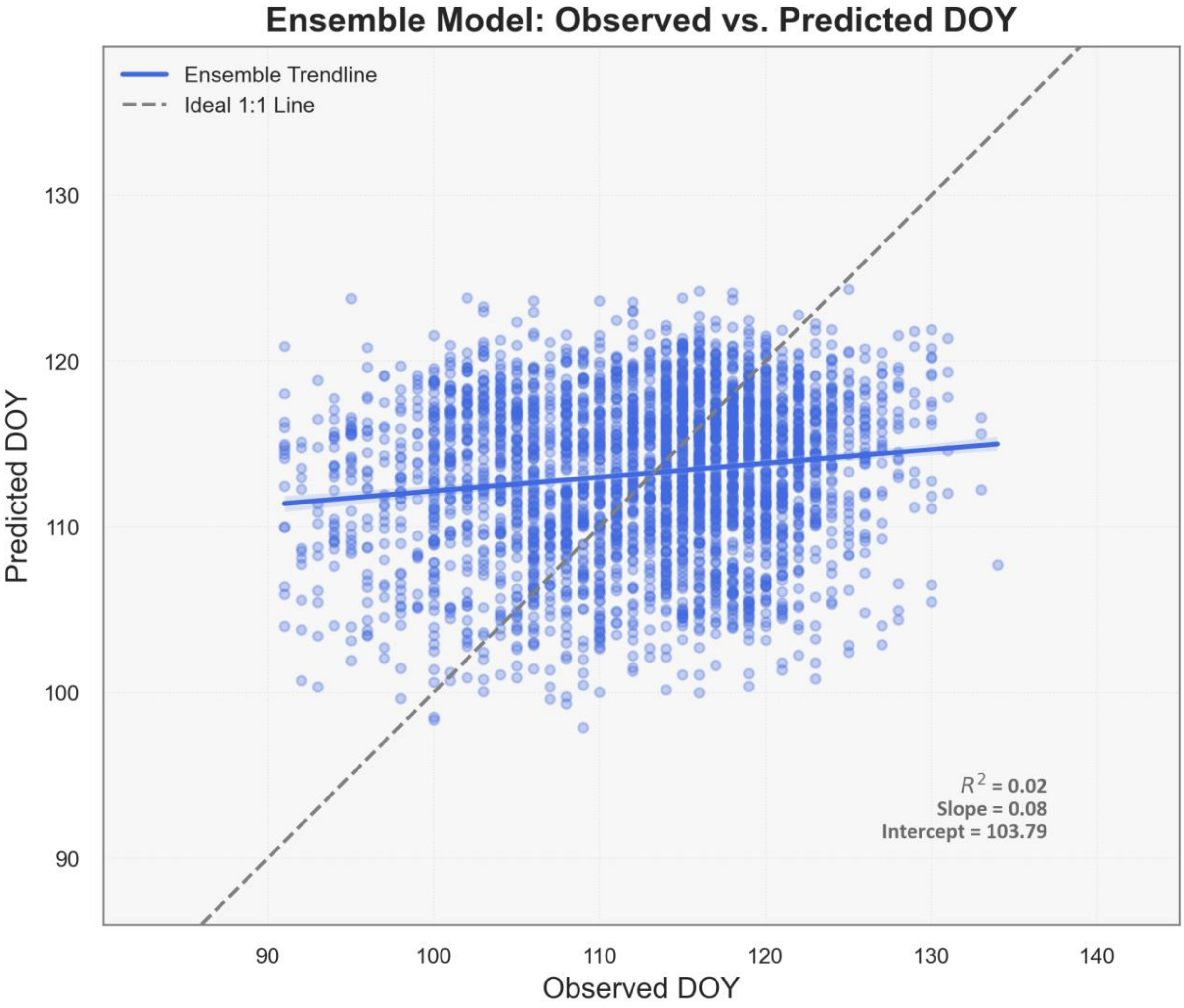


Figure 12: Ensemble Model Scatter plots of predicted vs observed Budburst DOY across 2001 - 2015

Both models exhibit clear limitations in capturing phenological variability at the distribution tails. The observed flattening of trendlines and predictive compression at early and late DOY extremes highlight their underperformance under atypical climatic conditions. This is especially concerning in the context of long-term climate change, where the ability to track shifting phenological baselines is critical.

When interpreted in the context of temporal aggregation, the ML-derived model's slightly steeper slope suggests marginally better responsiveness to observed variation compared to the ensemble. However, both trendlines ultimately underscore the limited discriminatory power of these models across a diverse climatic range.

While these aggregated scatterplots offer a useful overview of average performance, they mask temporal dynamics and evolving model biases. Therefore, they should be interpreted alongside complementary diagnostics such as temporal residual trajectories or year-wise error decomposition, which more effectively expose sensitivity to interannual variability and systematic drift.

4.2 Regression Fit and Granularity

While Figures 11 and 12 offer model-specific performance assessments, Figure 13 presents a critical overlaid perspective, enabling direct visual and quantitative comparison of the ML-derived and ensemble models across the 15-year aggregated dataset.

This comparative view immediately highlights fundamental structural differences in predictive behaviour. The ML-derived model (red) retains its previously observed characteristics, with a regression slope of 0.18, intercept of 95.52, and an R^2 of 0.043. Its red data points form a more dispersed and continuously distributed cloud, suggesting higher flexibility in capturing site-year DOY variability. This reflects the adaptive nature of its parameter inference, which allows it to adjust predictions across a broader phenological range.

In contrast, the ensemble model (blue) exhibits a markedly flatter trendline (slope = 0.084, intercept = 103.79) and a lower R^2 of 0.020, confirming a stronger regression-to-the-mean tendency. A notable artefact in this overlaid view is the horizontal banding of blue points, indicative of output quantisation. This likely arises from its fusion methodology, where combining predictions from base learners yields a limited set of discrete outcomes, reducing resolution and limiting the model's ability to respond to fine-scale phenological signals or capture extremes.

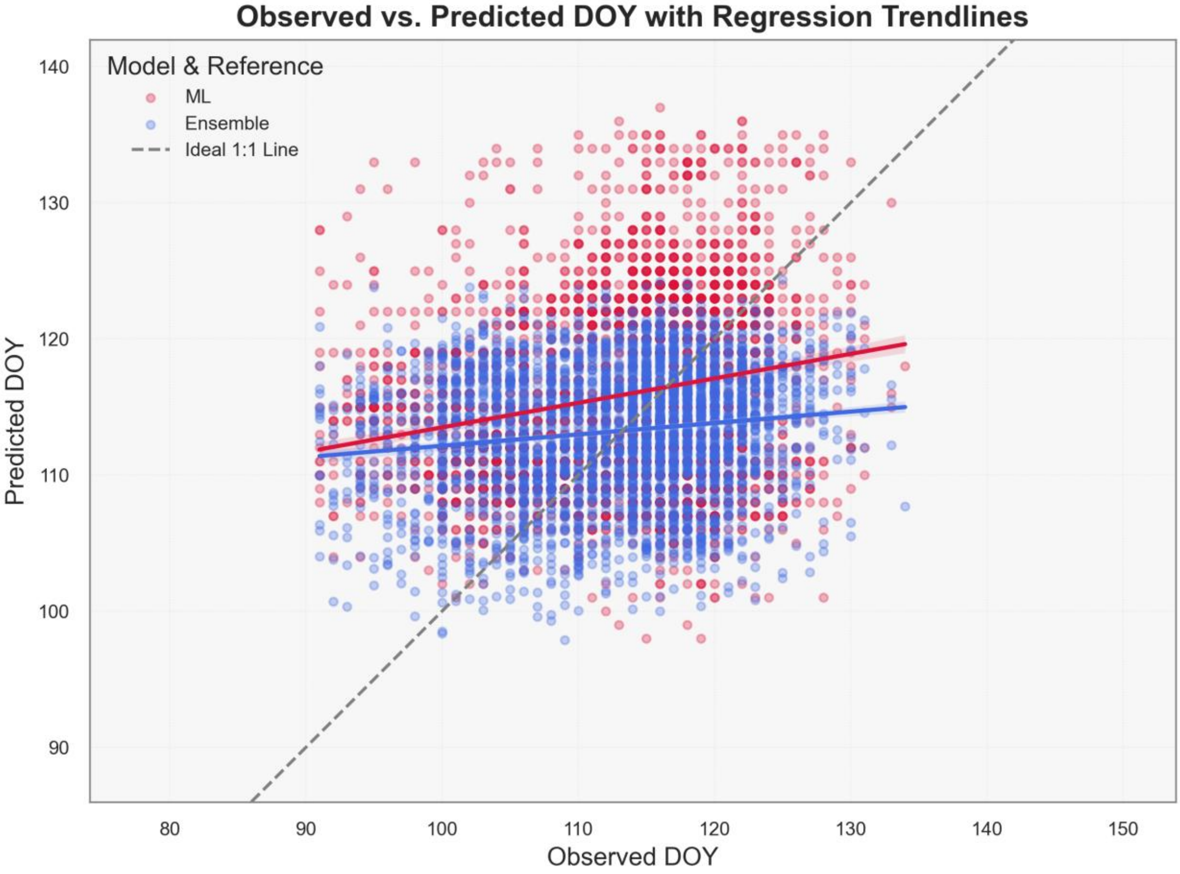


Figure 13: Predicted vs. Observed DOY with regression lines per model

Both models deviate substantially from the ideal 1:1 line yet do so in distinct ways. The ML-derived model, while noisier, demonstrates a more flexible alignment with the phenological gradient. The ensemble model, by contrast, sacrifices responsiveness for stability, leading to predictive compression around climatological norms. This structural comparison underscores the inherent trade-offs between the dynamic expressiveness of ML-derived parameterisation and the conservative averaging behaviour typical of ensemble approaches.

4.3 Model Robustness Under Climatic Extremes

This section evaluates how each hybrid modelling approach performs under extreme climatic conditions, using years with marked deviations from the climatological mean, specifically 2002, 2010, and 2014, as case studies. These years were selected based on annual temperature dynamics and the resulting challenge they pose to phenological prediction systems.

Figure 14 shows the RMSE values for both the ML-derived parameterisation model and the ensemble model in these years. In all three cases, the ensemble model demonstrated lower RMSEs: 7.91 vs. 9.38 in 2002, 7.34 vs. 8.88 in 2010, and 10.56 vs. 12.19 in 2014. These consistent outcomes indicate that the ensemble model had a performance stability under both warm and cool anomalies.

The ensemble's structure, based on fixed physiological variants and model fusion, appears to have conferred robustness by preventing overreaction to anomalous inputs. In contrast, the ML-derived model, though theoretically more adaptive, may have miscalibrated its parameter estimates in response to extreme or non-stationary thermal cues. This underlines a potential limitation of data-driven models in extrapolative scenarios where input-output relationships deviate from historical norms.

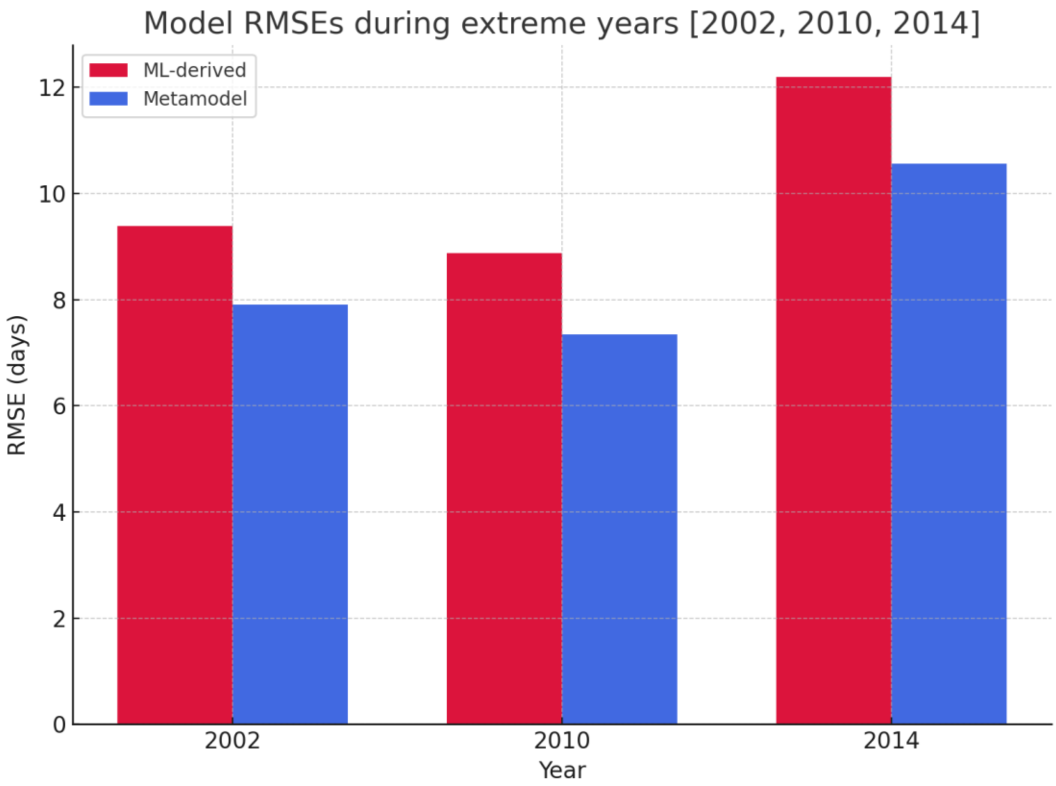


Figure 14: Model RMSEs during extreme years [2002, 2010, 2014]

These findings indicate that the ensemble model is comparatively stabilising during climatic volatility in years where climatic conditions diverge from the norm.

4.4 Model Interpretability and Structural Transparency

This section evaluates the potential of each model to produce interpretable outputs, a critical criterion in ecological modelling, The ML-derived parameterisation model generates predictions by learning site-specific values for core phenological parameters, including the base growth rate (d), shape factor (e), forcing threshold (f_crit), and thermal accumulation start (t_0). In contrast, the ensemble model produces its predictions through a weighted combination of multiple machine learning models. These weights, learned via a linear regressor meta-learner, reflect the statistical influence of each base model in the final prediction. The following sections analyse the temporal dynamics of both the learned phenological parameters and the ensemble model weights, with the aim of evaluating their stability, interpretive value, and implications for model transparency over time.

4.4.1 Comparative Dynamics of Parameter Behaviour

In Table 7, the mean and standard deviation of each ML-derived parameter are presented, summarising the distribution of site-specific parameter estimates.

Table 7: Mean and standard deviation of d, e, fcrit across models and selected years

Year	*d_mean*	*d_std*	*e_mean*	*e_std*	*f_crit_mean*	*f_crit_std*	*t0_mean*	*t0_std*
2001	-0.01	0.01	6.01	0.02	43.62	1.34	26.24	0.37
2002	0	0.01	6	0.07	42.1	1.61	28.79	0.53
2003	0.01	0	5.98	0.03	46.03	1.86	25.56	0.47
2004	-0.01	0	6.05	0.01	46.2	1.94	26.27	1.04
2005	0	0.01	6.04	0.03	43.44	1.86	26.62	0.6
2006	0.01	0.01	6.02	0.03	43.21	1.96	26.31	0.47
2007	0.01	0.01	5.99	0.04	41.41	2.81	25.65	0.78
2008	-0.02	0.01	6.02	0.03	43.74	1.65	26.82	0.91
2009	0.01	0.01	5.95	0.02	45.42	1.31	27.82	0.35
2010	0	0.01	6.01	0.01	43.3	1.85	26.28	0.25
2011	0	0	6.03	0.01	45.47	0.72	26.5	0.26
2012	-0.02	0.01	6.04	0.02	44.96	1.19	26.15	0.55
2013	0.01	0.01	5.97	0.02	45.24	2.03	26.44	0.5
2014	0	0	5.99	0.02	43.45	0.69	25.97	0.3
2015	0.01	0.01	6	0.01	44.67	0.96	26.29	0.36

Figure 15 complements this by illustrating the interannual dynamics of the ML-derived model's predicted UNIFORC parameters (d, e, f_crit, t_0). These trends provide insight into the model's internal decision logic and its responsiveness to climatic variability. The slope parameter d fluctuates around zero, with steeper negative values in years such as 2008 and 2012, potentially reflecting accelerated forcing responses under cooler spring conditions. The inflexion parameter e remains tightly clustered around 6.0, aligning with established thermal responsiveness in temperate species.

Meanwhile, variation in f_crit and t0 across years points to shifts in forcing thresholds and the onset of accumulation, respectively, suggesting that the model dynamically adjusts to interannual environmental signals. Collectively, these parameter trajectories support the ecological plausibility and interpretability of the ML-derived approach.

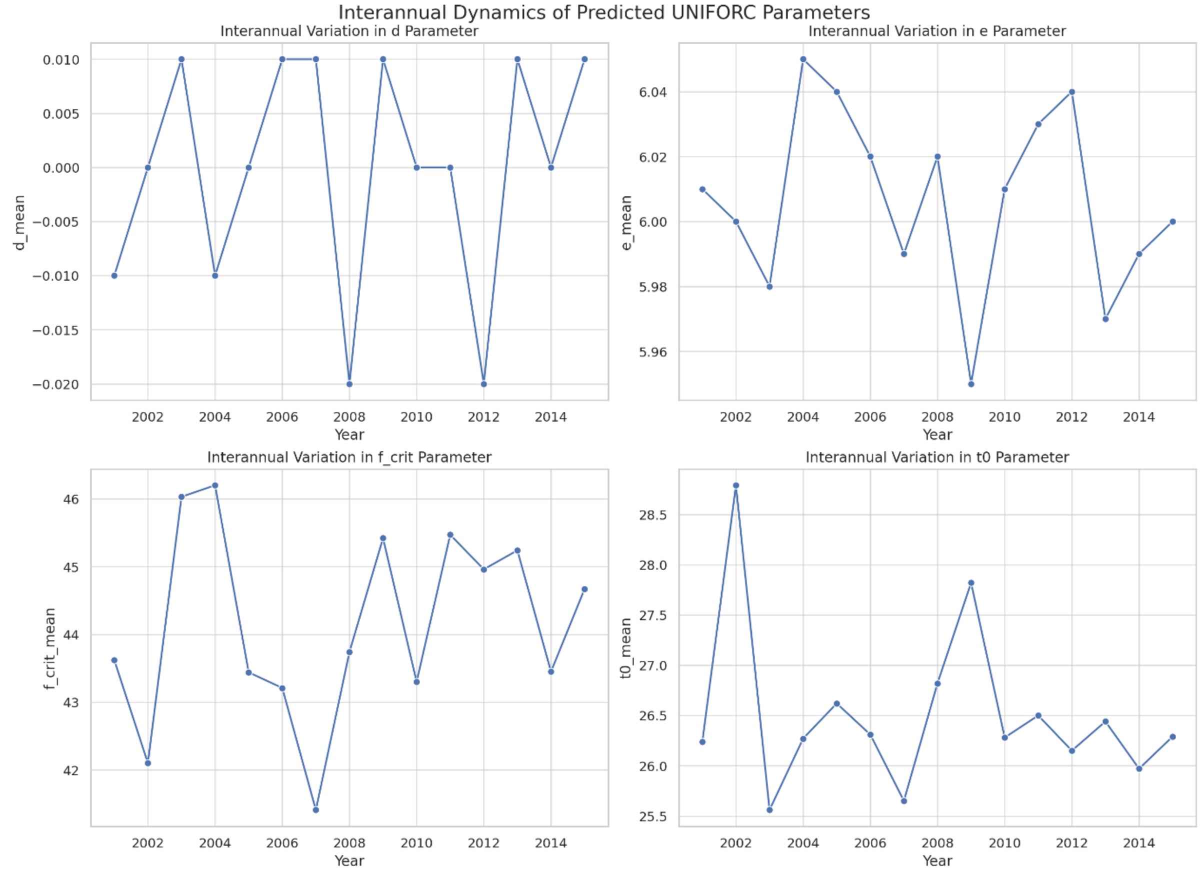


Figure 15: Line plot of parameter distributions for d, e, fcrit across both models

Yet, as shown in Table 8, these changes do not strongly correlate with prediction error, reinforcing that the ML-derived model maintains accuracy without sacrificing transparency.

Table 8: Parameter-to-prediction correlations

Parameter	***RMSE_corr***	***RMSE_p***	***MAE_corr***	***MAE_p***
d_mean	0.12	0.66	0.12	0.66
e_mean	0.02	0.94	0.00	1.00
f_crit_mean	0.10	0.71	0.08	0.79
t0_mean	0.16	0.58	0.09	0.76

4.4.2 Ensemble Composition and Weighting Dynamics

The Ensemble model integrates predictions from three ML regressors (XGBoost, Random Forest, SVR) and five UNIFORC variants using a linear regression-based weighting. Weights assigned to the individual models by the Linear Regression ensemble learner are illustrated in Figure 16.

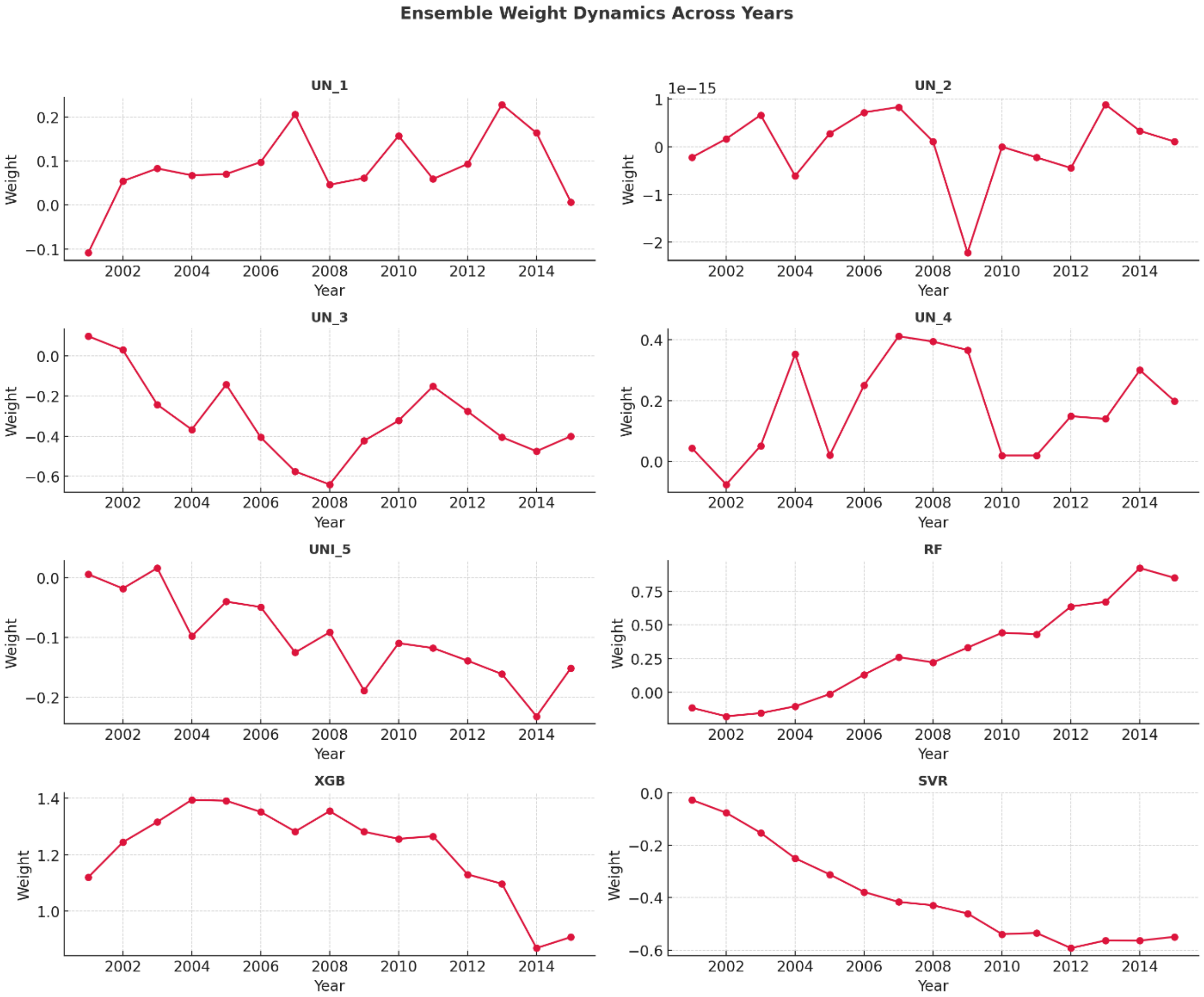
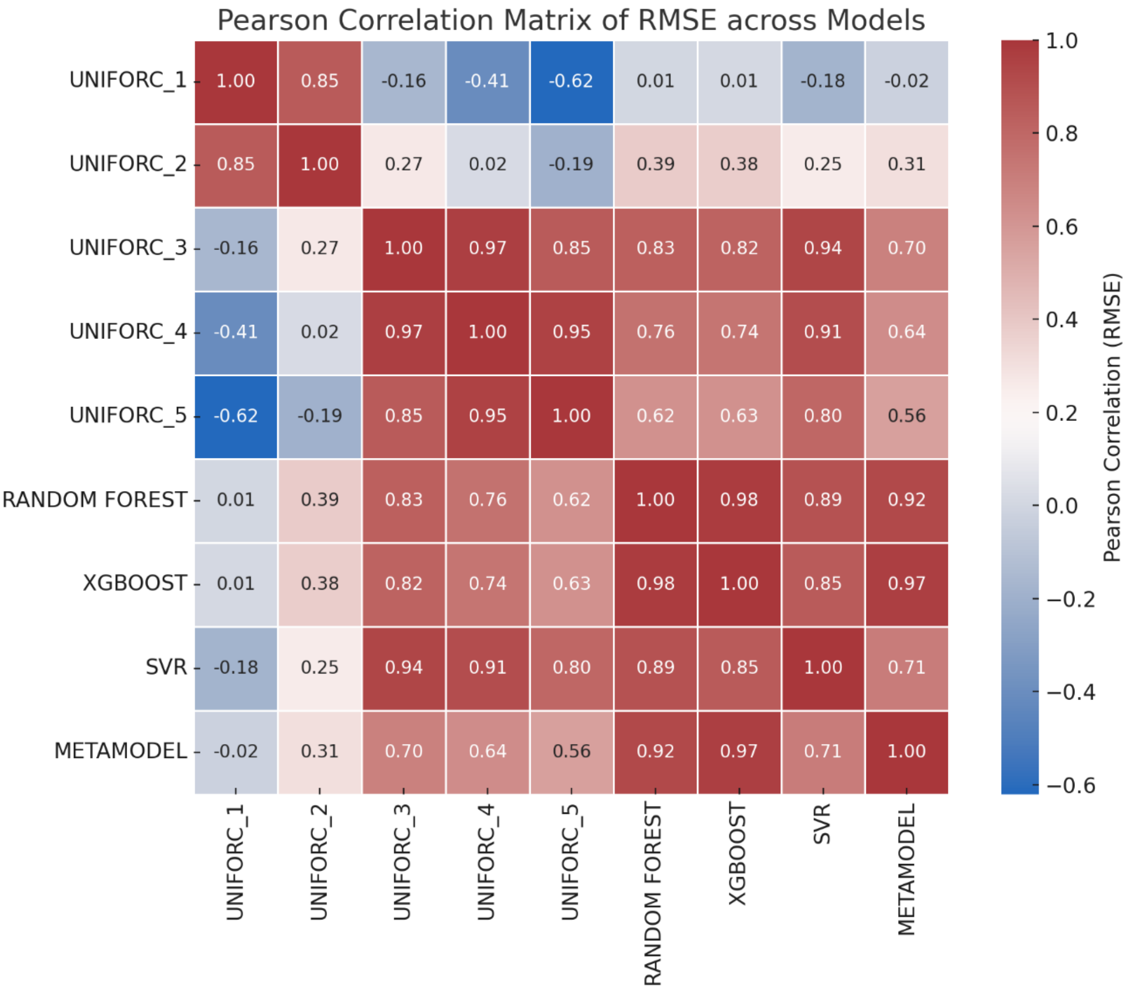


Figure 16: Line plot of model weights across cross-validation folds

XGBoost consistently received the highest weight (>0.5), reflecting the ensemble model's reliance on its strong performance. However, several UNIFORC variants, despite performing better than some ML models in certain years, received negligible or negative weights. This suggests the ensemble model prioritised variance reduction over physiological interpretability.

Such bias may compromise long-term generalisability, particularly under novel climates. Future work should explore constrained or regularised weighting to preserve mechanistic contributions within ensemble learning.

To understand the basis of the ensemble model's weighting behaviour, we examined the Pearson correlation between RMSE values of all constituent models (five UNIFORC variants and three machine learning models) and the final metamodel across all test years. The resulting correlation matrix (Figure 17) reveals three major insights. First, UNIFORC models 3 and 4 show near-perfect mutual correlation (r = 0.97) and exhibit strong correlation with UNIFORC_5 (r = 0.85 and r = 0.95, respectively). This cluster of high mutual similarity suggests that these models often carry overlapping signal content. Consequently, the ensemble model may suppress one or more of them to avoid redundancy even when their standalone performance is strong. This aligns with ensemble learning principles that prioritize diversity over raw accuracy to reduce correlated errors and enhance generalization (Ganaie et al., 2021).



Pearson Correlation Matrix of RMSE across Models

	UNIFORC_1	UNIFORC_2	UNIFORC_3	UNIFORC_4	UNIFORC_5	RANDOM FOREST	XGBOOST	SVR	METAMODEL
UNIFORC_1	1.00	0.85	-0.16	-0.41	-0.62	0.01	0.01	-0.18	-0.02
UNIFORC_2	0.85	1.00	0.27	0.02	-0.19	0.39	0.38	0.25	0.31
UNIFORC_3	-0.16	0.27	1.00	0.97	0.85	0.83	0.82	0.94	0.70
UNIFORC_4	-0.41	0.02	0.97	1.00	0.95	0.76	0.74	0.91	0.64
UNIFORC_5	-0.62	-0.19	0.85	0.95	1.00	0.62	0.63	0.80	0.56
RANDOM FOREST	0.01	0.39	0.83	0.76	0.62	1.00	0.98	0.89	0.92
XGBOOST	0.01	0.38	0.82	0.74	0.63	0.98	1.00	0.85	0.97
SVR	-0.18	0.25	0.94	0.91	0.80	0.89	0.85	1.00	0.71
METAMODEL	-0.02	0.31	0.70	0.64	0.56	0.92	0.97	0.71	1.00

Figure 17: Pearson correlation heatmap of all base model predictions across 2001–2015

Second, the machine learning models (Random Forest, XGBoost, and SVR) display strong correlations with each other (r > 0.85) and with the metamodel (r = 0.92 for Random Forest, r = 0.97 for XGBoost). Their generally superior RMSE values (relative to most UNIFORC variants) and higher orthogonality to the UNIFORC models likely increase their utility as independent sources of predictive power, thereby securing them higher weights in the ensemble. Lastly, while UNIFORC 1 and 2 showed only moderate correlations with other models and the metamodel (r ≈ 0.0–0.3), they sometimes received modest weighting due to offering slightly distinct error distributions. However, UNIFORC 1, which frequently exhibited the worst RMSE, remained downweighted across nearly all years. Taken together, the metamodel's weighting strategy appears performance-sensitive but also redundancy-aware.

The foregoing showed that the ensemble model's weights only represent the statistical contribution of each base learner to the final prediction. However, these weights do not encode any biologically interpretable or mechanistically grounded information. As such, despite the ensemble approach's competitive predictive performance in certain years, it lacks interpretability. Its internal decision process remains relatively opaque. It remains difficult to trace or decompose the model's predictive behaviour in terms of phenological mechanisms or physiological parameters. This limitation hinders its utility for ecological inference, where understanding the why behind a prediction is often as critical as the prediction itself.

4.5 Performance Trends across Base Models

The RMSE trends across 2001-2015 for all the representative models (both stand-alone and hybrid) illustrated in Figure 18 highlight nuanced distinctions in how each model responds to varying climatic regimes. The ML-derived model shows high flexibility, performing well in many years but faltering in anomalous ones like 2011 and 2014, likely due to over-adjustment to noisy forcing cues. Its dynamic parameter learning enables adaptability but may suffer from overfitting in climatically volatile years.

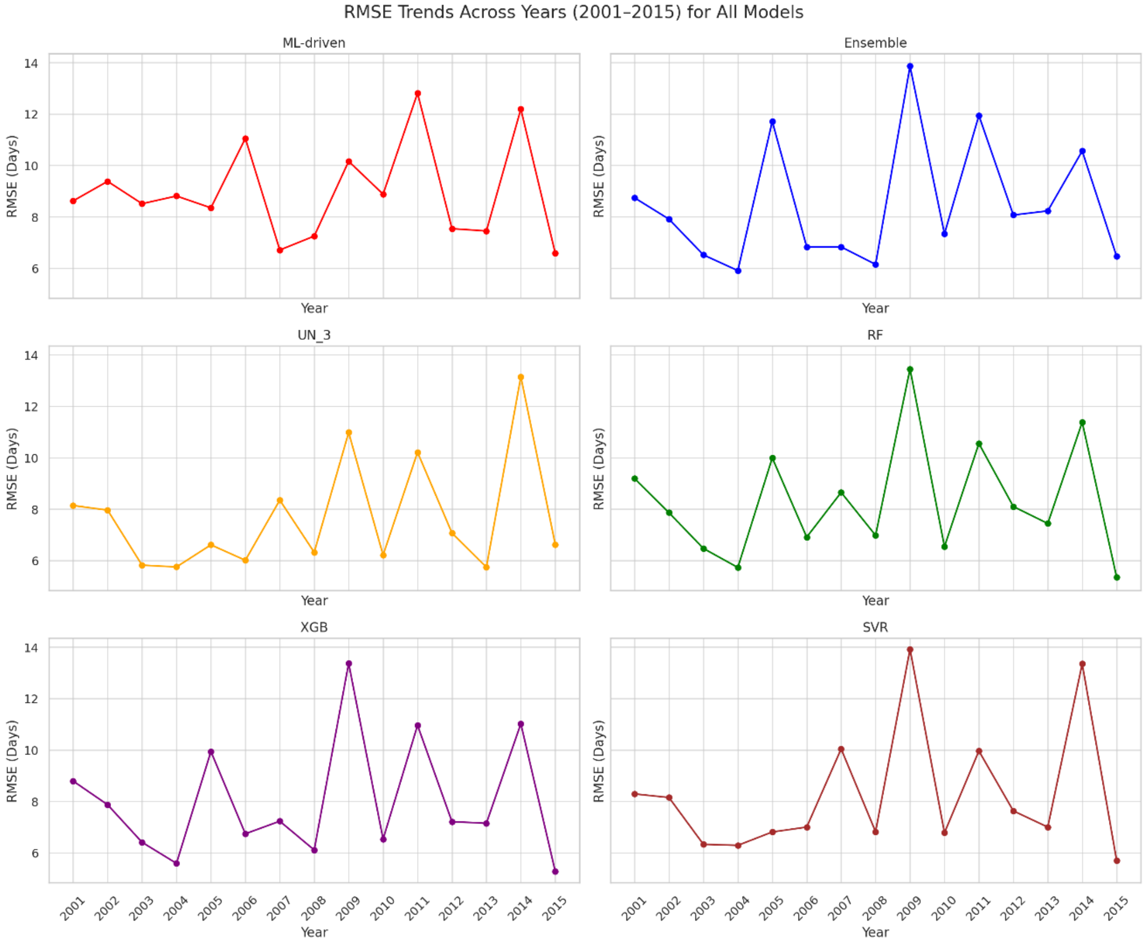


Figure 18: Performance metrics of the stand-alone models

The ensemble model offers greater stability, avoiding extreme errors through conservative averaging. However, its underperformance in certain years, such as 2009, suggests a limited capacity to adjust when sharp shifts in climate occur, revealing the trade-off between robustness and responsiveness.

The UNIFORC model (UN_3) was selected as the best-performing variant among five physiological configurations, consistently yielding competitive RMSEs in stable years. This underscores the potential of well-tuned process-based models to achieve high accuracy under typical climatic conditions, despite their relative rigidity.

Among the machine learning baselines, XGBoost maintained strong overall performance across years, likely due to its capacity to model complex interactions without excessive variance. Random Forest showed more erratic behaviour, possibly due to overfitting from unpruned trees, while SVR struggled in highly variable years, reflecting its limitations in capturing non-linear seasonal patterns.

Overall, these results confirm that no model is universally optimal. Each demonstrates context-specific strengths, supporting the rationale for hybrid approaches that combine physiological interpretability with statistical flexibility.

An interesting performance paradox is, however, observed between the ML-derived model and the UN_3: although the UN_3 model merely uses fixed parameters inferred from the ML-derived pipeline, it consistently outperforms the ML model across most years (See Figure 19). The ML-derived model, despite its adaptability, exhibits erratic RMSE behaviour, peaking in years like 2006 and 2011, where it underperforms by margins exceeding 4–5 days.

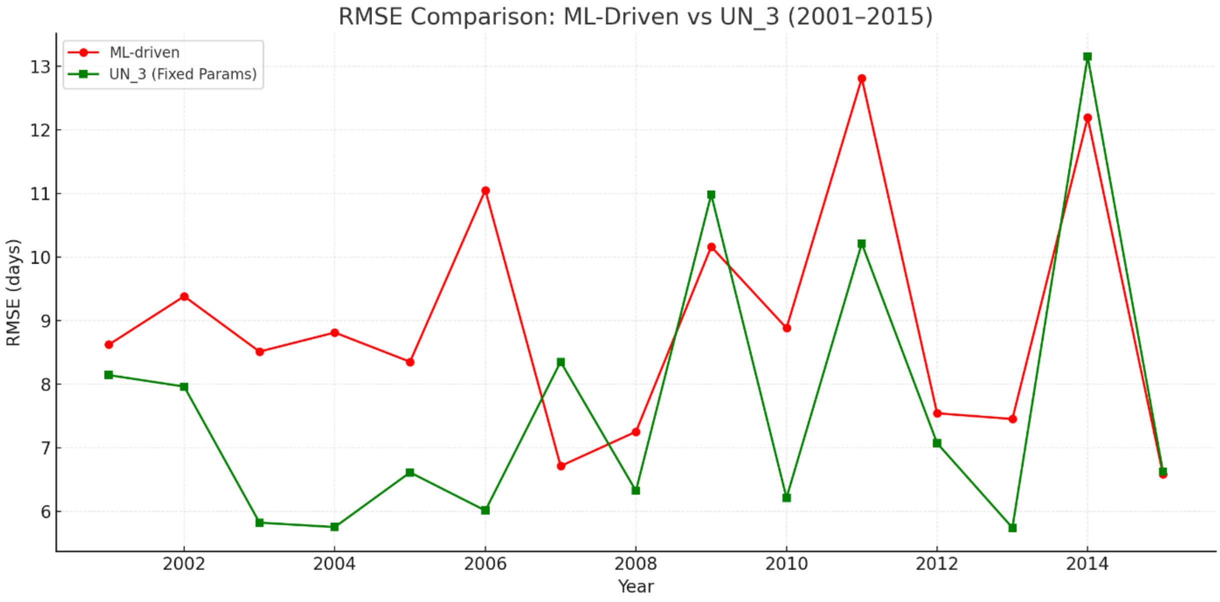


Figure 19: RMSE comparison between ML-derived parameterization and UN_3

What this suggests is that the iterative feedback loop in the ML model may be introducing instability rather than improving generalization. The retraining steps, while intended to refine parameter estimation, may in fact be reinforcing noise or overfitting to subtle variations in the training labels, especially in climatically variable years.

In contrast, the UN_3 model's static formulation appears to regularise performance, acting as a stabiliser that preserves the signal from ML-derived parameters without accumulating feedback error. The fact that UN_3 was not retrained but still outperformed its dynamic source model implies that the value lies more in stabilising the learning loop than in the parameter estimates themselves.

This insight challenges common assumptions about ML superiority in hybrid modelling. It underlines that carefully constrained, biologically grounded process models, even when seeded by data-driven approaches, can outperform ML when iterative feedback is poorly managed. Therefore, future improvements might focus on stabilising the feedback mechanism or adopting hybrid decoupling strategies that retain ML's flexibility without compromising PBM consistency.

4.7: Base Models' Response to Erratic Temperature Dynamics

The bar chart in Figure 20 reveals contrasting behaviours of the models under thermal extremes. In the coldest year (2010), the UNIFORC model achieved the lowest RMSE, closely followed by the

RF and XGB models, suggesting that fixed, well-calibrated physiological thresholds can offer stability in cooler conditions.

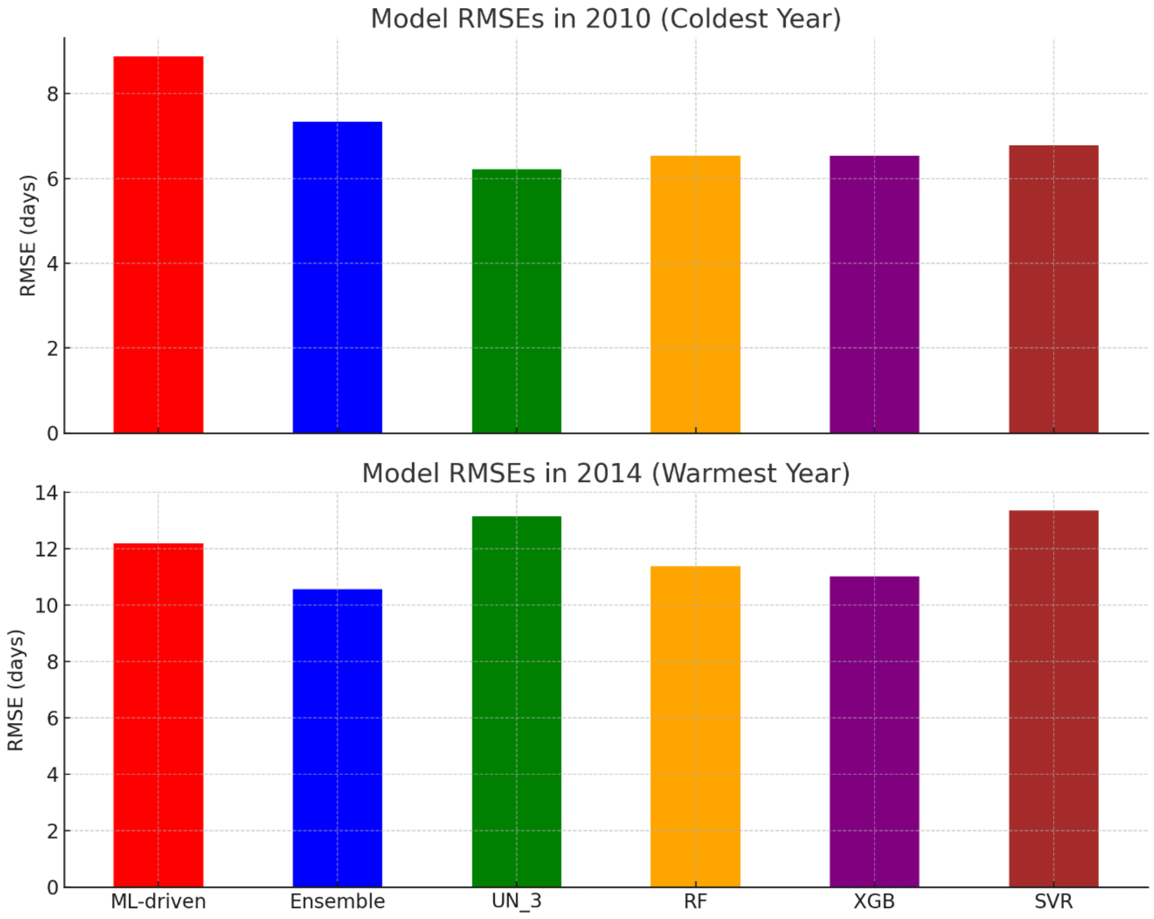


Figure 20: Stand-alone model adaptability to temperature extremes

Conversely, in the warmest year (2014), ML-derived models like RF and XGB outperformed both the UNIFORC and ensemble approaches, indicating that adaptive parameterization may better capture accelerated phenological responses. The ensemble model performed moderately in both years, reinforcing its role as a conservative average rather than a specialist under extremes. These patterns highlight the complementary strengths of static and adaptive modelling strategies.

4.8 Convergence of the ML-derived Parameterisation Loops

The ML-derived parameterisation model adopts an iterative hybrid modelling framework in which machine learning predicts site-specific PBM parameters, and these are used to simulate budburst DOY. To assess the stability and efficiency of this iterative process, RMSE trajectories from four representative years (2002, 2006, 2010, and 2015) were examined.

As shown in Figure 21, a consistent pattern emerges: RMSE decreases sharply after the first iteration, followed by a plateau in subsequent steps. This behaviour indicates that most of the predictive error is corrected early in the loop, as the ML models rapidly adjust to PBM-informed parameter feedback. In many cases, just one or two iterations are sufficient to achieve near-optimal performance, thereby enhancing computational efficiency.

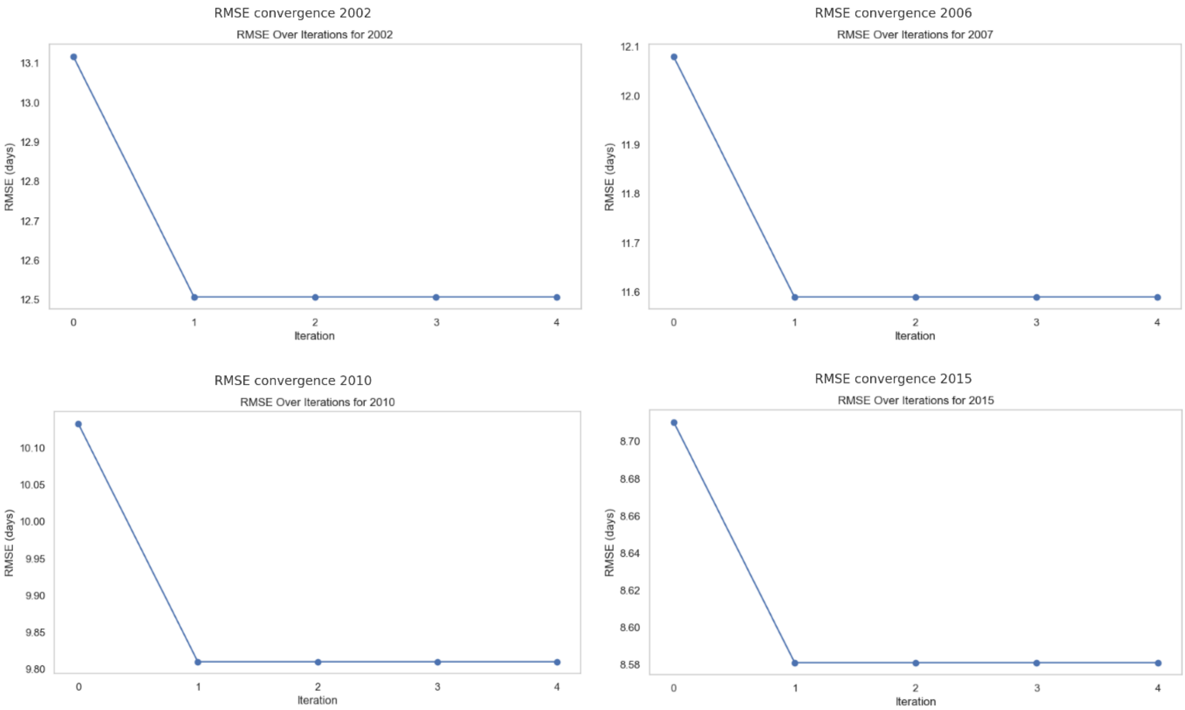


Figure 21: RMSE trajectory by iteration of the ML-derived parameterisation model

While minor improvements persist beyond the first iteration, they tend to yield diminishing returns. This highlights the importance of stabilising the learning loop, as later iterations may introduce noise or overfitting rather than meaningful refinement. Consequently, the current stopping criteria, though functionally correct, may be overly permissive, especially regarding the number of iterations. Tightening them could reduce redundant computation without sacrificing model accuracy.

5. DISCUSSION AND CONCLUSIONS

5.1 Overview

This discussion addresses the core research objectives of this study: (1) to review existing hybrid phenological frameworks, (2) to implement and compare two modelling strategies, and (3) to evaluate model adaptability using year-forward chaining. In doing so, it responds directly to the research questions posed in Section 1.5

Drawing from the two hybrid strategies, this section reflects on methodological strengths, trade-offs, and structural limitations inherent in each approach. It situates these findings within the broader context of hybrid phenological modelling literature, highlighting key gaps, such as the lack of dynamic parameter adaptation, limited model interpretability, and insufficient validation under realistic ecological variability and future climate conditions.

5.1 Common Modelling Strategies, Their Drawbacks, and Unaddressed Gaps

Hybrid phenological modelling has emerged as a promising avenue to leverage the mechanistic understanding of process-based models (PBMs) with the predictive power of ML (Basler & Körner, 2023; Chuine & Régnière, 2017). This study contributes to this growing field by implementing and evaluating two distinct hybrid frameworks: an ML-derived Parameterization Strategy for the UNIFORC model, and a Parallel Hybrid Model with Metamodel Fusion.

Common strategies in the literature include sequential hybrids (PBM predictions feed into ML); parallel hybrids (ML and PBM predictions are combined); data-driven parameterization (ML predicts PBM parameters); ML guided by PBM constraints; residual learning (ML corrects PBM errors); and PBM-generated synthetic data (for ML pre-training). This study implemented the parallel hybrid approach (fusion of fixed-parameter UNIFORC and ML regressors using a linear metamodel) and data-driven parameterization (ML-predicted UNIFORC parameters).

Each method presents trade-offs. Sequential hybrids risk error propagation. Parallel hybrids demand careful calibration of weighting schemes and can be computationally heavy. Data-driven parameterization relies on a robust PBM structure and feature engineering. ML constrained by PBMs may sacrifice flexibility. Residual learning and synthetic data generation require high data quality and strong PBM realism.

Across the literature and this study, several gaps persist. Research is biased toward well-studied crops and temperate systems, neglecting less-monitored ecosystems. Evaluation frameworks remain inconsistent, complicating cross-study comparison. Studies do not incorporate the fuzziness that indicates the gradual nature of climate-driven events like phenology. There is also limited support for long-term or real-time forecasting under climate change scenarios, which constrains the practical utility of hybrid models.

5.2 Hybrid Modelling for Dynamic Parameter Adjustment, and Feedback

Two modelling strategies were implemented, each offering unique fusion characteristics. Between the two strategies explored in this research, ML-derived parameterization proved most effective in enabling feedback and adaptability. In this study, LightGBM was used to predict four UNIFORC parameters (d, e, fcrit, t0), which were iteratively optimised via a feedback loop. This approach helped align model predictions with observed DOY, ensuring adaptability to site- and year-specific conditions.

For example, t0, typically fixed, was modelled dynamically, with values ranging from DOY 25 to 29. This better accounted for interannual variation in chilling completion. Temporal parameter summaries (Figure 15) confirmed biologically plausible interannual variability, enhancing realism and responsiveness to climatic anomalies.

In contrast, the Ensemble Model fused predictions from static PBMs and ML regressors using a linear metamodel. While this improved aggregate performance, it did not allow physiological parameters to adapt annually. Coefficient plots (Figure 16) showed that XGBoost often dominated the metamodel, with weights exceeding 1.0 (e.g., 1.395 in 2004). However, these weights offered no insight into biological mechanisms and were primarily statistical corrections.

This distinction highlights a core limitation of fusion strategies using fixed PBMs: they may correct for biases statistically, but do not dynamically adjust in biologically meaningful ways. However, when properly tuned as seen in Figure 19, they have the potential to challenge common assumptions about ML superiority in hybrid modelling

5.3 Validation for Robustness, Adaptability, and Interpretability

Validation under realistic conditions is crucial. While the year-forward chaining strategy adopted in this study emulates realistic forecasting by leveraging past observations to predict future years, it does not constitute true scenario-based forecasting. Specifically, no simulations were conducted under externally defined climate projections such as CMIP6 or Representative Concentration Pathways (RCPs). Future work should incorporate downscaled climate scenario data to evaluate model robustness under long-term warming trends and extreme event regimes.

However, despite overall robust performance, all base models struggled in 2014, a year with unusually high error (ML RMSE: 12.19, Metamodel RMSE: 10.56; R^2: -2.02 and -1.27, respectively). This suggests shared limitations, possibly due to data quality or climatic anomalies, such as insufficient chilling followed by rapid warming (Zohner et al., 2023).

Interpretability was another key focus. The ML-derived parameterization offered high interpretability by estimating ecologically meaningful parameters (e.g., more negative d in colder years, stable e near 6.0). These patterns directly tied physiological understanding to model outputs.

The ensemble model, by contrast, provided statistical transparency (via weights) but limited physiological insight. Correlation matrices showed UNIFORC variants (e.g., UNIFORC 3 and 4) were downweighed despite low RMSE due to high mutual correlation ($r = 0.97$), suggesting the ensemble model prioritized diversity over raw performance (Ganaie et al., 2021).

To improve interpretability, future hybrid designs should integrate parameterized PBMs within ensemble frameworks and explore nonlinear meta-learners (e.g., gradient boosting or neural nets). Climate-aware regularization could prevent over-reliance on dominant models. Lastly, robust diagnostic tools and visualization methods should be developed to facilitate a clearer understanding of model mechanics and error sources.

5.4 Study Limitations and Implications

Despite the methodological rigour and comparative breadth of this study, several limitations must be acknowledged.

First, the dataset was geographically and climatically constrained to temperate environments in Central Europe, characterized by moderate thermal variability. While this choice enabled focused model calibration, it limits the generalizability of results to more extreme or variable biomes such as Mediterranean, boreal, or arid systems (Piao et al., 2022). Expanding spatial coverage would be necessary to ensure broader ecological relevance.

Second, although the ML-derived parameterization enabled dynamic adjustment of UNIFORC parameters, it depended heavily on the quality and completeness of engineered environmental covariates, particularly thermal indices. Any omission, redundancy, or collinearity among features may have influenced parameter plausibility, convergence, and model stability (Dormann et al., 2012). The absence of chilling and photoperiod variables also restricts physiological completeness, particularly in years with decoupled thermal signals.

Third, the ensemble model fusion relied on a linear regression framework, selected for transparency. However, this structure may not have adequately captured non-linear dependencies or synergistic interactions among model outputs and environmental drivers. As a result, high-performing yet correlated models like UNIFORC 3 and 4 were occasionally assigned negligible or negative weights due to multicollinearity, undermining their ecological relevance (Dai et al., 2022). Incorporating more flexible fusion strategies (e.g., kernel-based or tree-based metamodels) could mitigate this issue.

Fourth, while the year-forward chaining validation was realistic and temporally robust, the models were not tested for spatial transferability. This represents a key limitation for ecological forecasting, where models must generalize across heterogeneous landscapes. Without spatial cross-validation, the models' extrapolation capacity remains uncertain (Tang et al., 2023).

Lastly, this study did not explicitly assess computational scalability. The iterative feedback loop in the ML-derived parameterization model and the multi-model architecture of the metamodel pose non-trivial runtime demands. For real-time applications such as operational forecasting or resource-constrained decision systems, optimizing computational efficiency remains an open priority.

5.5 Conclusions

5.5.1 Summary of Findings

This research explored two distinct hybrid phenological modelling strategies: an ML-derived Parameterization Model and a Parallel Hybrid Ensemble Model. The ML-derived model dynamically estimated physiological parameters of the UNIFORC process-based model using LightGBM, integrating iterative feedback to enhance adaptive learning. The ensemble model fused outputs from five fixed-parameter UNIFORC variants and three ML regressors through a linear regression meta learner, aiming to balance predictive diversity.

Key findings showed that the ML-parametrised model demonstrated greater physiological interpretability and temporal adaptability, with parameters such as t0, d, e, and $f_{(crit)}$ fluctuating within biologically plausible ranges in response to interannual climatic variability. The ensemble model excelled in some statistical accuracy metrics, especially in years where certain base learners (notably XGBoost) dominated. However, it lacked the capacity to adjust process parameters, leading to rigidity in years with anomalous climate behaviour.

A comparative analysis revealed no consistent superiority across all years. For instance, in 2011 and 2014, years marked by phenological anomalies, both models struggled, underscoring the limitations of each framework under stress conditions. The correlation analysis revealed a weak relationship between prediction error (RMSE) and mean annual temperature (Pearson's r ≈ 0.08 - 0.16), suggesting that performance is governed more by other complex interactions and model architecture than by warming trends alone. The ensemble model exhibited selective weighting behaviour, often suppressing well-performing mechanistic models (e.g., UNIFORC 3 and 4) due to high collinearity with other base learners, favouring statistical decorrelation over ecological relevance.

5.5.2 Conclusion on Research Questions

This section revisits the core research questions posed at the outset, drawing on empirical evidence, modelling outputs, and theoretical insights to provide concise, evidence-based conclusions.

RQ1: What are the prevailing strategies in hybrid phenological modelling, and what methodological limitations and knowledge gaps remain?

This study identified six dominant hybrid strategies: Sequential coupling, Parallel ensemble fusion, Data-driven parameterisation, Constraint-guided ML, Residual learning, and PBM-generated synthetic data. Each varies in its balance between interpretability, adaptability, and computational cost. Despite progress, the literature remains skewed toward temperate systems and lacks standardised evaluation protocols. There is also limited support for dynamic parameter tuning or real-time simulation under future climate scenarios.

RQ2: How do different hybrid strategies, such as parallel ensemble modelling and machine learning-driven parameter estimation with feedback, compare in their ability to integrate PBM and ML components for accurate and interpretable budburst prediction?

The ML-derived parameterisation strategy, which predicts and iteratively refines process-based parameters (e.g., *d*, *e*, *f_crit*, t_0), outperformed the parallel ensemble in terms of interpretability and adaptability. It allowed the model to maintain biological transparency while dynamically adjusting to climatic variability. Conversely, the ensemble model—though occasionally more accurate—relied on metamodel weights that lacked ecological meaning, offering limited insight into the underlying biological processes.

RQ3: How effectively can hybrid models simulate future phenological responses under evolving climate conditions when evaluated using a year-forward chaining strategy and windowed climate data?

The year-forward chaining framework approximated real-world forecasting by simulating prediction under non-stationary temporal conditions. Both models showed resilience across most years, but struggled during climatically extreme periods (e.g., 2014). Notably, only the ML-derived model retained explanatory clarity through biologically interpretable parameters. While no external climate scenario simulations (e.g., RCPs) were conducted, this study lays a foundation for future integration of scenario-based projections to test long-term model robustness under changing climates.

RQ3: *How effectively can hybrid models simulate future phenological responses under evolving climate conditions when evaluated using a year-forward chaining strategy and windowed climate data?*

5.6 Recommendations for Future Work

To advance the robustness, adaptability, and interpretability of hybrid phenology models, several strategic directions are recommended:

1. **Incorporate Climate Scenario-Based Testing (e.g., Representative Concentration Pathways [RCPs] or Shared Socioeconomic Pathways [SSPs]):** Future research should

evaluate hybrid model robustness under long-term climate change scenarios by simulating responses to projected shifts in chilling requirements, spring warming patterns, and extreme event frequencies.

2. **Advance Hybrid Models with Interpretable and Adaptive Structures:** Improve the ML-derived parameterisation approach through multi-objective optimisation and cross-site generalisability testing. Simultaneously, embed biologically meaningful constraints into ensemble meta-models to balance predictive power with ecological interpretability.
3. **Expand Hybrid Modelling to Underrepresented Ecosystems, Phenophases, and Species:** To increase generalizability and ecological applicability, future work should apply hybrid frameworks to less-monitored ecosystems (e.g., tropical forests, drylands), phenophases beyond budburst (e.g., senescence, flowering), and a wider range of plant or crop species.

5.7 Ethical Considerations

This study was conducted in accordance with the Research Ethics Policy of the ITC, University of Twente and adhered to the highest standards of academic and professional integrity. The following ethical considerations were addressed

Data Privacy and Security: All data used in this study were obtained from publicly available, open-access sources. Phenological observations were sourced from the PEP725 database, and climate data were acquired from the ERA5-Land reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). No personal, identifiable, or sensitive information was required or used at any point in this study.

Data acquisition and use complied fully with relevant legal and ethical standards. All datasets were evaluated for representativeness and completeness, and preprocessing steps, including data cleaning, handling missing values, and removing outliers, were applied to ensure data quality and minimize bias. Regular validation checks were performed to maintain data integrity throughout the modelling process.

Acknowledgement of Prior Research: All previous research and data sources used to inform this study have been appropriately cited and acknowledged in the main text and reference list. This approach reflects a commitment to academic transparency, intellectual honesty, and respect for the contributions of other researchers.

Use of AI tools: AI tools were used to improve text clarity, grammar, formatting, and to assist in code development. Their use was limited to enhancing language precision and programming efficiency, without influencing the design, outcomes, or interpretation of the research. All AI applications adhered to ethical guidelines and academic standards to ensure the integrity of the work.

All AI-generated outputs were critically reviewed, edited, and validated by the author to ensure accuracy, originality, and scholarly rigour. The author takes full responsibility for the final content presented in this thesis.

Overall, the study was conducted with full ethical responsibility, ensuring proper use of open data, maintaining data privacy, and upholding the integrity of the scientific process.

LIST OF REFERENCES

Akhavizadegan, F., Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). A time-dependent parameter estimation framework for crop models. Scientific Reports, 11(1), 11437. https://doi.org/10.1038/s41598-021-90820-2

Arcomano, T., Szunyogh, I., Wikner, A., Pathak, J., Hunt, B. R., & Ott, E. (2022). A hybrid approach to atmospheric modeling that combines machine learning with a physics-based numerical model. Journal of Advances in Modeling Earth Systems, 14(3), e2021MS002712. https://doi.org/10.1029/2021MS002712

Asse, D., Randin, C. F., Bonhomme, M., Delestrade, A., & Chuine, I. (2020). Process-based models outcompete correlative models in projecting spring phenology of trees in a future warmer climate. Agricultural and Forest Meteorology, 285–286, 107931. https://doi.org/10.1016/j.agrformet.2020.107931

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology, 249(1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Chuine, I., Cour, P., & Rousseau, D. D. (1999). Selecting models to predict the timing of flowering of temperate trees: implications for tree phenology modelling. Plant, Cell & Environment, 22(1), 1–13. https://doi.org/10.1046/j.1365-3040.1999.00395.x

Chuine, I. (2000). A unified model for budburst of trees. Journal of Theoretical Biology, 207(3), 337–347. https://doi.org/10.1006/jtbi.2000.2178

Chuine, I., Bonhomme, M., Legave, J. M., García de Cortázar-Atauri, I., Charrier, G., Lacointe, A., & Améglio, T. (2016). Can phenological models predict tree phenology accurately in the future? The unrevealed hurdle of endodormancy break. Global Change Biology, 22(10), 3444–3460. https://doi.org/10.1111/gcb.13383

Chuine, I., & Régnière, J. (2017). Process-based models of phenology for trees and shrubs of temperate and boreal regions. Global Ecology and Biogeography, 26(5), 535–548. https://doi.org/10.1111/geb.12567

Currier, C. M., & Sala, O. E. (2022). Precipitation versus temperature as phenology controls in drylands. Ecology, 103(10), e3772. https://doi.org/10.1002/ecy.3772

Da Veiga Grubert, D. A. (2023). A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling (Doctoral dissertation, University of São Paulo). https://doi.org/10.11606/t.11.2023.tde-03112023-103445

Dai, W., Jin, H., Liu, T., Jin, G., Zhang, Y., & Zhou, Z. (2022). Applying ensemble learning in ecophysiological models to predict spring phenology. Forest Ecology and Management, 505, 119911. https://doi.org/10.1016/j.foreco.2021.119911

Dai, W., Jin, H., Zhang, Y., Liu, T., & Zhou, Z. (2019). Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: Application of machine learning in phenological model. Agricultural and Forest Meteorology, 279, 107702. https://doi.org/10.1016/j.agrformet.2019.107702

Dai, W., Jin, H., Zhou, L., Liu, T., Zhang, Y., Zhou, Z., Fu, Y. H., & Jin, G. (2022). Testing machine learning algorithms on a binary classification phenological model. Global Ecology and Biogeography, 32(1), 178–190. https://doi.org/10.1111/geb.13612

Deng, G., Zhang, H., Hong, Y., Guo, X., Yi, Z., & Biniyaz, E. (2024). Modeling of spring phenology of boreal forest by coupling machine learning and diurnal temperature indicators. Chinese Geographical Science, 35(1), 38–54. https://doi.org/10.1007/s11769-024-1478-x

Deutscher Wetterdienst (DWD). (2023, December 29). The weather in Germany in the year 2023. Retrieved from https://www.dwd.de/EN/press/press_release/EN/2023/20231229_the_weather_in_germany_in_year_2023_news.html

Devkota, K. P., Devkota, M., Moussadek, R., & Nangia, V. (2023). Genotype × environment × agronomic management interaction to enhance wheat yield in the Mediterranean rainfed environments of Morocco: II. Process-based modeling. European Journal of Agronomy, 151, 126973. https://doi.org/10.1016/j.eja.2023.126973

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dronova, I., & Taddeo, S. (2022). Remote sensing of phenology: Towards the comprehensive indicators of plant community dynamics from species to regional scales. Journal of Ecology, 110(7), 1430–1444. https://doi.org/10.1111/1365-2745.13902

Droutsas, I., Challinor, A. J., Deva, C. R., & Wang, E. (2022). Integration of machine learning into process-based modelling to improve simulation of complex crop responses. In Silico Plants, 4(2). https://doi.org/10.1093/insilicoplants/diac017

European Centre for Medium-Range Weather Forecasts (ECMWF). (2023). ERA5 hourly data on single levels from 1940 to present [Data set]. Copernicus Climate Data Store. https://doi.org/10.24381/cds.adbb2d47

Forrest, J., & Miller-Rushing, A. J. (2010). Toward a synthetic understanding of the role of phenology in ecology and evolution. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1555), 3101–3112. https://doi.org/10.1098/rstb.2010.0145

Fu, Y., Zhang, H., Dong, W., & Yuan, W. (2014). Comparison of phenology models for predicting the onset of growing season over the Northern Hemisphere. PLoS ONE, 9(10), e109544. https://doi.org/10.1371/journal.pone.0109544

Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151. https://doi.org/10.1016/j.engappai.2022.105151

Gao, J., Yang, B., Mura, C., Boucher, Y., & Rossi, S. (2023). Increased inter-annual variability in budburst dates towards the northern range edge of black spruce. Agricultural and Forest Meteorology, 333, 109410. https://doi.org/10.1016/j.agrformet.2023.109410

Garnot, V. S. F., Spafford, L., Lever, J., Sigg, C., Pietragalla, B., Vitasse, Y., Gessler, A., & Wegner, J. D. (in press). Deep learning meets tree phenology modelling: PhenoFormer versus process-based models. Methods in Ecology and Evolution. [DOI pending publication]

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hacket-Pain, A., Szymkowiak, J., Journé, V., Barczyk, M. K., Thomas, P. A., Lageard, J. G. A., Kelly, D., & Bogdziewicz, M. (2025). Growth decline in European beech associated with temperature-driven increase in reproductive allocation. Proceedings of the National Academy of Sciences, 122(5). https://doi.org/10.1073/pnas.2423181122

Hänninen, H. (1990). Modelling dormancy release in trees from cool and temperate regions. Acta Forestalia Fennica, 213, 1–47. https://doi.org/10.14214/aff.7651

Hänninen, H., Kramer, K., Tanino, K., Zhang, R., Wu, J., & Fu, Y. H. (2019). Experiments are necessary in process-based tree phenology modelling. Trends in Plant Science, 24(3), 199–209. https://doi.org/10.1016/j.tplants.2018.11.006

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 15(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Inouye, D. W. (2022). Climate change and phenology. Wiley Interdisciplinary Reviews: Climate Change, 13(3), e764. https://doi.org/10.1002/wcc.764

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., & Ritchie, J. T. (2003). The DSSAT cropping system model. European Journal of Agronomy, 18(3–4), 235–265. https://doi.org/10.1016/S1161-0301(02)00107-7

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. Geoscientific Model Development, 15(13), 5147–5162. https://doi.org/10.5194/gmd-15-5147-2022

Kattel, G. R. (2022). Climate warming in the Himalayas threatens biodiversity, ecosystem functioning and ecosystem services in the 21st century: Is there a better solution? Biodiversity and Conservation, 31(7), 1827–1849. https://doi.org/10.1007/s10531-022-02429-8

Kawakita, S., Yamasaki, M., Teratani, R., Yabe, S., Kajiya-Kanegae, H., Yoshida, H., Fushimi, E., & Nakagawa, H. (2023). Dual ensemble approach to predict rice heading date by integrating multiple rice phenology models and machine learning-based genetic parameter regression models. Agricultural and Forest Meteorology, 344, 109821. https://doi.org/10.1016/j.agrformet.2023.109821

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N. G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J. P., Silburn, M., Wang, E., Brown, S., Bristow, K. L., Asseng, S., … Smith, C. J. (2003). An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy, 18(3–4), 267–288. https://doi.org/10.1016/S1161-0301(02)00108-9

Keenan, T. F., Gray, J., Friedl, M. A., Toomey, M., Bohrer, G., Hollinger, D. Y., Munger, J. W., O'Keefe, J., Schmid, H. P., Wing, I. S., Yang, B., & Richardson, A. D. (2014). Terrestrial biosphere model performance for inter-annual variability of land-atmosphere CO2 exchange. Global Change Biology, 20(6), 1971–1984. https://doi.org/10.1111/gcb.12477

Körner, C., & Basler, D. (2010). Phenology under global warming. Science, 327(5972), 1461–1462. https://doi.org/10.1126/science.1186473

Leolini, L., Costafreda-Aumedes, S., Santos, J. A., Menz, C., Fraga, H., Molitor, D., Merbold, L., Buchmann, N., & Moriondo, M. (2023). Spring phenology in subtropical trees: Developing process-based models with leaf differentiation and budburst prediction. Agricultural and Forest Meteorology, 341, 109678. https://doi.org/10.1016/j.agrformet.2023.109678

Li, Y., Wang, X., & Ren, Z. (2023). Random Forest-based ensemble models for ecological forecasting under uncertainty. Ecological Informatics, 74, 101914. https://doi.org/10.1016/j.ecoinf.2023.101914

Lindborg, R., Hartel, T., Helm, A., Prangel, E., Reitalu, T., & Ripoll-Bosch, R. (2023). Ecosystem services provided by semi-natural and intensified grasslands: Synergies, trade-offs and linkages to plant traits and functional richness. Applied Vegetation Science, 26(2), e12729. https://doi.org/10.1111/avsc.12729

Lisboa, P., Saralajew, S., Vellido, A., Fernández-Domenech, R., & Villmann, T. (2023). The coming of age of interpretable and explainable machine learning models. Neurocomputing, 535, 25–39. https://doi.org/10.1016/j.neucom.2023.02.040

Luedeling, E., Schiffers, K., Fohrmann, T., & Urbach, C. (2021). PhenoFlex - an integrated model to predict spring phenology in temperate fruit trees. Agricultural and Forest Meteorology, 307, 108491. https://doi.org/10.1016/j.agrformet.2021.108491

Malyshev, A. V., Van Der Maaten, E., Garthen, A., Maß, D., Schwabe, M., & Kreyling, J. (2022). Inter-individual budburst variation in Fagus sylvatica is driven by warming rate. Frontiers in Plant Science, 13. https://doi.org/10.3389/fpls.2022.853521

Ma, L., Fu, T., & Song, X. (2023). A comprehensive review of XGBoost and its applications in environmental modelling. Environmental Modelling & Assessment, 28(1), 13–29. https://doi.org/10.1007/s10666-022-09830-1

Maestrini, B., Mimić, G., van Oort, P. A., Jindo, K., Brdar, S., Athanasiadis, I. N., & van Evert, F. K. (2022). Mixing process-based and data-driven approaches in yield prediction. European Journal of Agronomy, 139, 126569. https://doi.org/10.1016/j.eja.2022.126569

Masago, Y., & Lian, M. (2022). Estimating the first flowering and full blossom dates of Yoshino cherry (Cerasus × yedoensis 'Somei-yoshino') in Japan using machine learning algorithms. Ecological Informatics, 71, 101835. https://doi.org/10.1016/j.ecoinf.2022.101835

McCormick, R. F., Truong, S. K., Rotundo, J., Liu, L., Kyle, A., Wang, Y., & Messina, C. D. (2021). Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. In Silico Plants, 3(1). https://doi.org/10.1093/insilicoplants/diab004

Mcmaster, J. (1997). Growing conditions: One equation, two interpretations. Agricultural and Forest Meteorology, 87(4), 291–300. https://doi.org/10.1016/s0168-1923(97)00027-0

Menzel, A., Yuan, Y., Matiu, M., Sparks, T., Scheifinger, H., Gehrig, R., & Estrella, N. (2020). Climate change fingerprints in recent European plant phenology. Global Change Biology, 26(4), 2599–2612. https://doi.org/10.1111/gcb.15000

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. Information Systems, 102549. https://doi.org/10.1016/j.is.2025.102549

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac

Ogle, K., Barber, J. J., Barron-Gafford, G. A., Bentley, L. P., Young, J. M., Huxman, T. E., Loik, M. E., & Tissue, D. T. (2014). Quantifying ecological memory in plant and ecosystem processes. Ecology Letters, 18(3), 221–235. https://doi.org/10.1111/ele.12399

Olsson, C., & Jönsson, A. M. (2014). Process-based models not always better than empirical models for phenology predictions. Agricultural and Forest Meteorology, 184, 1–8. https://doi.org/10.1016/j.agrformet.2013.09.001

Ou, J., Chen, F., Zhang, M., Batchelor, W. D., Wang, B., Wu, D., & Feng, P. (2025). A continuous learning framework based on physics-guided deep learning for crop phenology simulation. Agricultural and Forest Meteorology, 368, 110562. https://doi.org/10.1016/j.agrformet.2024.110562

Pan European Phenology Network (PEP725). (n.d.). PEP725 phenological database. Zentralanstalt für Meteorologie und Geodynamik (ZAMG). Retrieved from https://www.pep725.eu/

Patidar, N., & Behera, M. D. (2024). Role of phenology as discriminant in vegetation type mapping using Sentinel-2 data. Remote Sensing Applications: Society and Environment, 36, 101255. https://doi.org/10.1016/j.rsase.2024.101255

Pei, J., Tan, S., Zou, Y., Liao, C., He, Y., Wang, J., Huang, H., Wang, T., Tian, H., Fang, H., Wang, L., & Huang, J. (2024). The role of phenology in crop yield prediction: Comparison of ground-based phenology and remotely sensed phenology. Agricultural and Forest Meteorology, 361, 110340. https://doi.org/10.1016/j.agrformet.2024.110340

Piao, S., Liu, Q., Chen, A., Janssens, I. A., Fu, Y., Dai, J., Liu, L., Lian, X., Shen, M., & Zhu, X. (2019). Plant phenology and global climate change: Current progresses and challenges. Global Change Biology, 25(6), 1922–1940. https://doi.org/10.1111/gcb.14619

Piao, S., Wang, J., Li, X., Xu, H., & Zhang, Y. (2022). Spatio-temporal changes in the speed of canopy development and senescence in temperate China. Global Change Biology, 28(24), 7366–7375. https://doi.org/10.1111/gcb.16408

Piña-Rey, A., Ribeiro, H., Fernández-González, M., Abreu, I., & Rodríguez-Rajo, F. J. (2023). Data mining methods to detect airborne pollen of spring flowering arboreal taxa. Environmental Monitoring and Assessment, 195(10), 1234. https://doi.org/10.1007/s10661-023-11799-7

Ponti, R., & Sannolo, M. (2023). The importance of including phenology when modelling species ecological niche. Ecography, 2023(8), e06758. https://doi.org/10.1111/ecog.06758

Qiu, J., & Mitchell, M. (2024). Understanding biodiversity–ecosystem service linkages in real landscapes. Landscape Ecology, 39(3), 58. https://doi.org/10.1007/s10980-024-01824-8

Rahimi, M., Soleimani, H., & Liaghat, A. (2023). Support vector regression with RBF kernel for high-dimensional agricultural phenology prediction. Computers and Electronics in Agriculture, 208, 107785. https://doi.org/10.1016/j.compag.2023.107785

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Richardson, A. D., Keenan, T. F., Migliavacca, M., Ryu, Y., Sonnentag, O., & Toomey, M. (2012). Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. Agricultural and Forest Meteorology, 169, 156–173. https://doi.org/10.1016/j.agrformet.2012.09.012

Richardson, E. A., Seeley, S. D., & Walker, D. R. (1974). A model for estimating the completion of rest for 'Redhaven' and 'Elberta' peach trees. HortScience, 9(4), 331–332. https://doi.org/10.21273/hortsci.9.4.331

Roberts, A. M., Tansey, C., Smithers, R. J., & Phillimore, A. B. (2015). Predicting a change in the order of spring phenology in temperate forests. Global Change Biology, 21(7), 2603–2611. https://doi.org/10.1111/gcb.12896

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

Stuble, K. L., Bennion, L. D., & Kuebbing, S. E. (2021). Plant phenological responses to experimental warming—A synthesis. Global Change Biology, 27(17), 4110–4124. https://doi.org/10.1111/gcb.15685

Tang, Y., Wang, W., Fu, B., Liu, Y., & Zhang, X. (2023). Evaluating ecological model transferability using cross-regional and temporal predictions. Ecological Modelling, 482, 110005. https://doi.org/10.1016/j.ecolmodel.2023.110005

Templ, B., Koch, E., Bolmgren, K., Ungersböck, M., Paul, A., Scheifinger, H., & Zust, A. (2018). Pan European Phenological database (PEP725): A single point of access for European data. International Journal of Biometeorology, 62(7), 1109–1113. https://doi.org/10.1007/s00484-018-1512-8

Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. Electronics, 12(8), 1789. https://doi.org/10.3390/electronics12081789

Vitasse, Y., Signarbieux, C., & Fu, Y. H. (2021). Global warming leads to more uniform spring phenology across elevations. Nature Climate Change, 11(5), 396–401. https://doi.org/10.1038/s41558-021-01008-5

Way, D. A., & Montgomery, R. A. (2015). Photoperiod constraints on tree phenology, performance and migration in a warming world. Plant, Cell & Environment, 38(9), 1725–1736. https://doi.org/10.1111/pce.12431

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys, 55. https://doi.org/10.1145/3514228

Worrall, G., Judge, J., Boote, K., & Rangarajan, A. (2022). In-season crop phenology using remote sensing and model-guided machine learning. Agronomy Journal, 115(3), 1214–1236. https://doi.org/10.1002/agj2.21230

Yu, J., Zhao, Y., Lei, G., & Zeng, W. (2025). A comparison of physics-based, data-driven, and hybrid modeling approaches for rice phenology prediction. Agronomy Journal, 117(1). https://doi.org/10.1002/agj2.70010

Zhang, C., & Ma, Y. (Eds.). (2022). Ensemble machine learning: Methods and applications. Springer. https://doi.org/10.1007/978-3-030-92575-3

Zhang, J., Lin, X., Jiang, C., Hu, X., Liu, B., Liu, L., Xiao, L., Zhu, Y., Cao, W., & Tang, L. (2024). Predicting rice phenology across China by integrating crop phenology model and machine learning. The Science of the Total Environment, 951, 175585. https://doi.org/10.1016/j.scitotenv.2024.175585

Zhang, Q., Zhao, X., Han, Y., Yang, F., Pan, S., Liu, Z., Wang, K., & Zhao, C. (2023). Maize yield prediction using federated random forest. Computers and Electronics in Agriculture, 210, 107930. https://doi.org/10.1016/j.compag.2023.107930

Zhang, X., Wang, J., & Li, Y. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS ONE, 18(2), e0281234. https://doi.org/10.1371/journal.pone.0281234

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software, 23(4), 550–560. https://doi.org/10.1145/279232.279236

Zhu, M., Dai, J., Wang, H., Alatalo, J. M., Liu, W., Hao, Y., & Ge, Q. (2024). Mapping 24 woody plant species phenology and ground forest phenology over China from 1951 to 2020. Earth System Science Data, 16(1), 277–293. https://doi.org/10.5194/essd-16-277-2024

Zhu, W., & Siebert, S. (2024). Climate-driven interannual variability in subnational irrigation areas across Europe. Communications Earth & Environment, 5(1). https://doi.org/10.1038/s43247-024-01721-z

Zohner, C. M., Mirzagholi, L., Renner, S. S., Mo, L., Rebindaine, D., Bucher, R., Palouš, D., Vitasse, Y., Fu, Y. H., Stocker, B. D., & Crowther, T. W. (2023). Effect of climate warming on the timing of autumn leaf senescence reverses after the summer solstice. Science, 381(6653). https://doi.org/10.1126/science.adf5098

Zeng, X., Wang, X., Xue, Q., & Du, Y. (2024). Time interval between budburst and leaf-out of subtropical woody species is shortened by increases in both seasonal and diurnal temperatures. Agricultural and Forest Meteorology, 358, 110254. https://doi.org/10.1016/j.agrformet.2024.110254

APPENDICES

APPENDIX 1

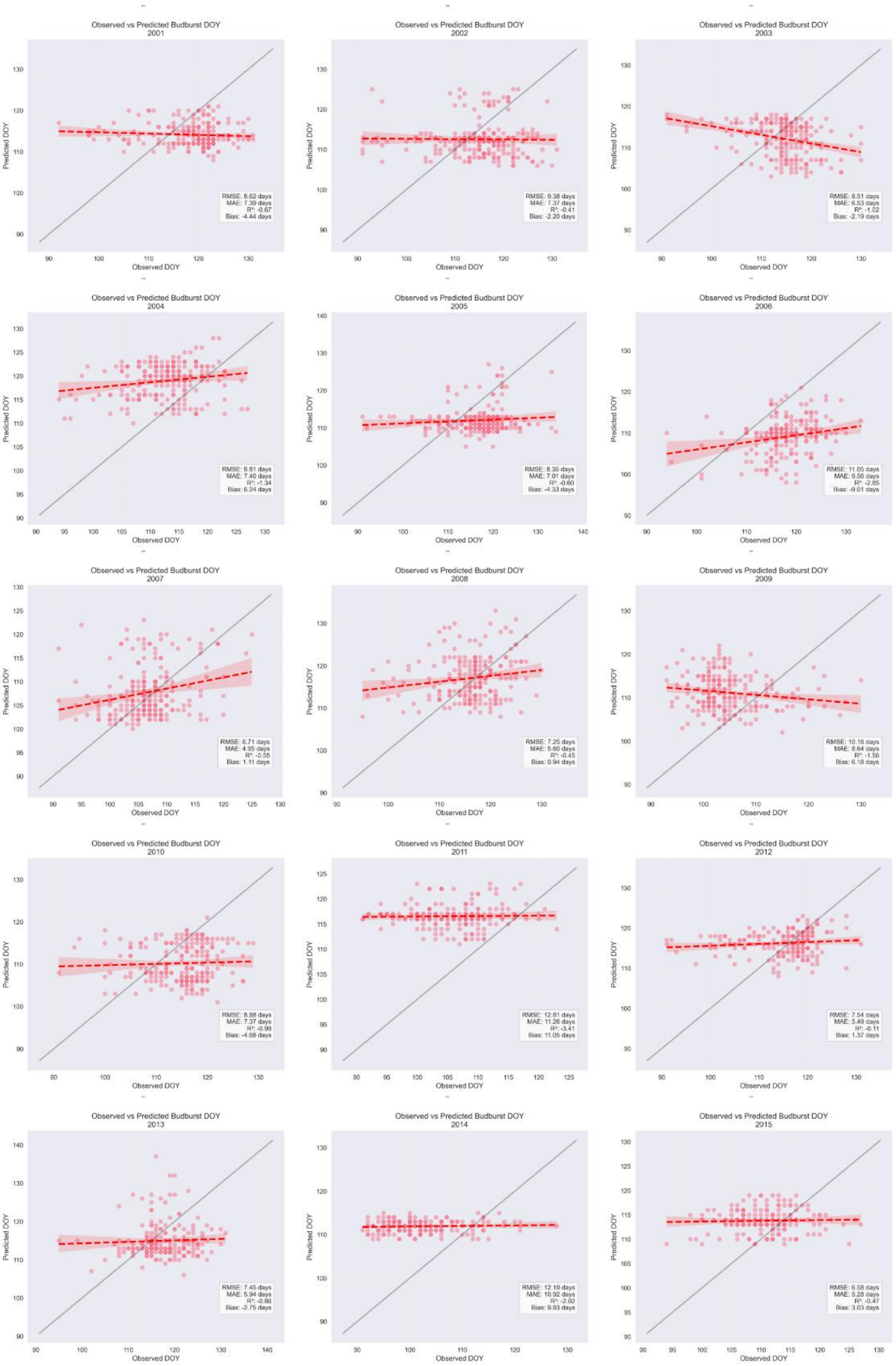


Figure 22: Yearly scatterplots of Predicted vs Observed Budburst DOY using ML-derived Model

APPENDIX 2

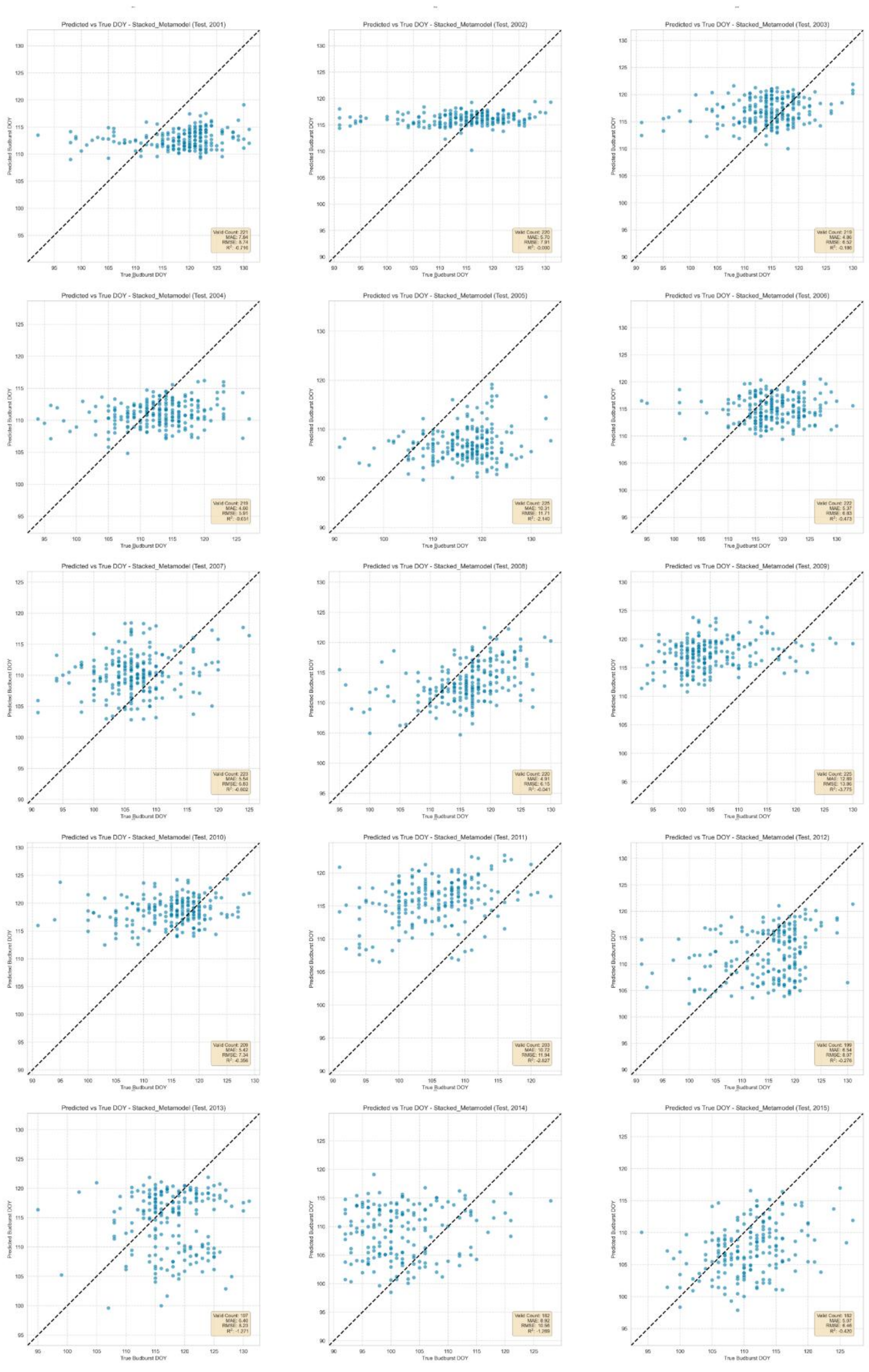


Figure 23: Yearly scatterplots of Predicted vs Observed Budburst DOY using Ensemble Model

APPENDIX 3: Site specific parameters predicted by the ML-derived Model for Year 2001

Table 9: Site specific parameters predicted by the ML-derived Model for Year 2001

site_id	year	d	e	f_crit	t0
94	2001	-0.01	6.00	42.19	26.25
103	2001	-0.01	6.01	42.10	26.24
106	2001	-0.01	6.01	42.38	26.30
124	2001	-0.01	5.98	42.89	26.83
152	2001	-0.01	6.00	42.59	26.76
199	2001	-0.01	6.00	42.95	26.83
213	2001	-0.01	5.98	42.49	26.69
244	2001	-0.01	5.99	42.99	26.83
280	2001	-0.01	5.98	42.49	26.55
291	2001	-0.01	5.99	42.24	26.23
317	2001	-0.01	5.98	43.01	26.61
427	2001	-0.01	6.00	43.25	26.40
450	2001	-0.01	5.98	43.08	26.09
502	2001	-0.01	6.02	42.40	26.56
531	2001	-0.01	6.02	42.78	26.58
532	2001	-0.01	6.02	42.78	26.58
638	2001	-0.01	6.00	42.76	26.16
656	2001	-0.02	5.99	42.81	26.25
680	2001	-0.01	5.98	42.24	26.39
765	2001	-0.01	6.00	43.01	26.21
766	2001	-0.01	6.01	42.45	26.33
799	2001	-0.01	5.98	43.05	26.22
810	2001	0.01	6.08	41.69	26.71
855	2001	-0.01	5.98	42.18	26.42
911	2001	-0.01	5.97	44.60	26.21
975	2001	-0.01	5.98	44.89	25.95
977	2001	-0.01	5.98	44.60	25.87
1072	2001	-0.02	5.99	42.68	26.24
1074	2001	-0.01	6.03	42.39	26.55
1076	2001	-0.02	6.02	42.42	26.46
1111	2001	-0.02	5.99	42.68	26.23
1119	2001	-0.02	6.00	42.53	26.19
1127	2001	-0.02	5.99	42.52	26.21
1129	2001	-0.01	5.97	44.41	25.96
1203	2001	-0.01	5.98	44.04	26.00
1324	2001	-0.01	5.99	44.92	25.94
1326	2001	-0.01	5.99	44.93	26.03
1358	2001	-0.02	5.99	44.47	25.94
1469	2001	-0.01	6.02	42.05	26.39
1537	2001	-0.01	5.98	44.98	25.95
1539	2001	-0.01	5.98	44.98	25.95

1619	2001	-0.01	6.00	44.09	26.45
1690	2001	-0.01	6.00	43.26	26.31
1713	2001	-0.02	6.01	42.40	26.59
1723	2001	-0.01	5.99	44.10	26.46
1758	2001	-0.01	6.01	43.07	26.63
1759	2001	0.00	6.04	43.14	26.69
1771	2001	0.00	6.04	42.23	26.81
1821	2001	-0.01	6.03	41.81	26.74
1851	2001	0.00	6.04	41.99	26.82
1872	2001	-0.01	5.99	42.19	26.71
1876	2001	-0.01	5.98	43.13	26.68
1887	2001	-0.01	5.99	44.23	26.49
1920	2001	-0.02	5.98	43.91	26.03
1969	2001	-0.02	5.97	44.46	25.64
1974	2001	-0.02	5.98	44.52	25.64
1980	2001	-0.02	5.98	43.15	25.64
1981	2001	-0.02	5.98	44.52	25.64
2009	2001	-0.02	5.99	43.14	26.03
2023	2001	-0.02	6.03	42.38	25.85
2024	2001	-0.01	6.04	42.78	26.18
2038	2001	-0.02	5.98	44.33	25.57
2075	2001	0.00	6.04	41.97	26.59
2112	2001	-0.02	6.04	42.40	26.13
2126	2001	0.00	6.04	43.93	25.90
2140	2001	-0.02	6.02	42.39	25.95
2153	2001	-0.02	6.03	42.39	25.87
2164	2001	-0.02	6.02	42.41	25.93
2168	2001	-0.02	6.02	42.39	25.95
2170	2001	-0.02	6.03	42.38	25.86
2182	2001	-0.02	6.01	42.41	26.15
2203	2001	-0.02	6.02	42.38	25.87
2216	2001	-0.02	6.03	42.39	25.84
2217	2001	-0.02	6.02	42.40	25.77
2225	2001	-0.02	6.02	42.41	25.73
2241	2001	-0.01	6.03	42.41	26.24
2250	2001	-0.01	6.03	42.48	26.57
2286	2001	-0.02	6.02	42.39	25.94
2294	2001	-0.02	6.03	42.40	25.92
2295	2001	-0.01	6.03	42.52	25.93
2296	2001	-0.01	6.02	42.46	26.15
2327	2001	-0.01	6.03	41.73	26.70
2360	2001	-0.02	6.00	42.17	26.21
2390	2001	-0.02	5.98	44.26	25.84
2415	2001	-0.02	5.99	42.96	26.26
2432	2001	-0.02	5.99	42.81	26.19
2444	2001	-0.02	6.00	42.40	26.13
2489	2001	-0.01	5.99	43.53	26.35

2545	2001	-0.02	6.00	42.46	26.32
2552	2001	-0.02	5.98	44.07	26.04
2561	2001	-0.01	5.99	43.81	25.77
2566	2001	-0.01	5.98	43.32	26.32
2602	2001	-0.01	6.01	42.43	26.23
2615	2001	-0.01	6.02	42.47	26.54
2674	2001	-0.02	5.98	44.07	26.10
2698	2001	-0.02	6.08	42.99	25.95
2739	2001	-0.02	5.98	44.14	25.66
2761	2001	-0.02	5.98	43.55	26.01
2763	2001	-0.02	5.98	43.60	26.00
2808	2001	-0.02	5.98	46.45	25.64
2831	2001	-0.02	5.97	43.74	26.16
2842	2001	-0.02	6.01	46.08	26.00
2852	2001	-0.01	6.01	46.33	26.07
2853	2001	-0.01	6.01	44.78	26.35
2867	2001	-0.02	5.99	44.60	26.10
2869	2001	-0.01	6.02	46.20	25.97
2886	2001	-0.01	6.06	41.49	26.22
2893	2001	-0.01	6.07	41.70	26.46
2899	2001	-0.02	6.00	46.06	26.04
2909	2001	0.00	6.07	43.67	26.62
2937	2001	-0.02	5.98	43.28	25.79
2953	2001	-0.02	5.99	46.52	25.62
2958	2001	-0.02	5.98	46.56	25.70
2976	2001	-0.02	5.98	42.54	25.67
3014	2001	-0.01	6.06	41.53	26.37
3018	2001	-0.01	6.01	44.70	26.56
3023	2001	-0.02	5.98	46.75	25.51
3053	2001	-0.02	5.98	44.75	25.59
3063	2001	-0.02	5.98	44.74	25.58
3065	2001	-0.02	5.98	42.58	25.71
3068	2001	-0.02	5.98	42.54	25.62
3069	2001	-0.02	5.98	42.62	25.86
3076	2001	-0.02	5.98	42.58	25.67
3082	2001	-0.02	5.98	43.32	25.68
3091	2001	-0.02	5.97	44.01	25.58
3099	2001	-0.02	5.98	46.58	25.74
3100	2001	-0.02	5.99	47.01	25.53
3105	2001	-0.02	5.99	45.25	25.89
3124	2001	-0.02	5.98	42.71	25.67
3127	2001	-0.02	5.98	42.57	25.67
3173	2001	-0.02	5.99	45.02	25.92
3213	2001	-0.02	6.08	41.72	26.15
3230	2001	0.00	6.01	44.36	26.88
3248	2001	-0.01	6.02	41.61	26.62
3264	2001	0.00	6.01	42.81	26.89

3284	2001	-0.01	6.01	44.67	26.05
3290	2001	-0.02	5.99	45.02	25.92
3332	2001	0.00	6.01	42.63	26.93
3346	2001	0.00	6.07	41.82	26.88
3366	2001	-0.01	6.02	46.05	25.99
3378	2001	0.00	6.02	42.73	26.77
3417	2001	0.00	6.01	42.99	26.71
3422	2001	0.00	6.07	43.62	26.94
3438	2001	0.00	6.02	43.10	26.90
3482	2001	0.00	5.99	44.90	26.68
3484	2001	0.00	6.01	43.08	26.68
3486	2001	0.00	6.02	44.90	26.30
3515	2001	0.01	6.03	44.03	26.74
3520	2001	0.00	5.99	44.88	26.68
3526	2001	0.00	6.01	44.90	26.36
3534	2001	0.00	6.02	45.06	26.53
3542	2001	0.00	6.02	43.14	26.74
3589	2001	0.00	6.03	44.83	26.64
3628	2001	0.01	6.03	44.92	26.52
3630	2001	0.01	6.03	44.91	26.52
3633	2001	0.00	6.01	44.46	26.54
3656	2001	0.00	5.99	44.85	26.72
3659	2001	0.00	5.99	44.95	26.49
3693	2001	0.00	6.03	45.27	26.85
3694	2001	0.00	6.03	44.87	26.80
3723	2001	0.00	6.03	45.30	26.63
3744	2001	0.00	6.03	45.01	26.86
3750	2001	0.00	6.02	45.03	26.55
3798	2001	0.00	6.03	44.90	26.88
3824	2001	0.00	6.03	44.73	26.38
3827	2001	0.00	6.03	44.57	26.44
3852	2001	0.00	6.03	46.10	26.42
3876	2001	-0.01	6.03	44.18	26.19
3884	2001	-0.01	6.06	41.85	26.66
3893	2001	-0.01	6.03	44.80	26.33
3911	2001	-0.01	6.03	45.95	26.61
3918	2001	-0.01	6.03	45.89	26.60
3945	2001	0.00	6.02	46.08	26.55
3959	2001	-0.01	6.03	46.06	26.31
3983	2001	-0.01	6.03	44.83	26.43
3984	2001	0.00	6.03	44.79	26.80
3986	2001	0.00	6.04	45.89	26.53
3993	2001	-0.01	6.02	44.32	25.80
3995	2001	-0.01	6.03	44.77	26.28
3997	2001	-0.01	6.03	44.74	26.07
4015	2001	-0.02	6.02	43.46	25.89
4046	2001	-0.02	6.03	42.61	25.92

4054	2001	-0.01	6.03	43.77	26.08
4055	2001	-0.01	6.03	42.26	26.41
4057	2001	-0.01	6.02	42.32	26.36
4080	2001	-0.01	6.03	44.25	25.93
4098	2001	-0.01	6.03	44.08	25.98
4121	2001	-0.01	6.03	44.47	26.05
4146	2001	-0.02	6.02	41.95	26.38
4156	2001	-0.01	6.03	44.56	26.26
4157	2001	-0.01	6.03	44.39	26.13
4172	2001	-0.02	6.01	43.22	25.85
4198	2001	-0.01	6.08	41.51	26.45
4223	2001	-0.01	6.03	44.72	26.79
4230	2001	-0.02	6.01	42.03	26.20
4239	2001	-0.01	6.01	42.88	26.50
4240	2001	-0.02	6.01	41.93	26.27
4256	2001	-0.02	6.01	41.93	26.27
4260	2001	-0.02	6.00	42.54	25.68
4263	2001	-0.01	6.02	46.26	26.43
4266	2001	-0.01	6.02	44.78	26.79
4268	2001	-0.01	6.02	46.26	26.43
4317	2001	-0.02	6.03	42.64	25.91
4333	2001	-0.02	6.02	42.57	25.80
4340	2001	-0.02	6.00	42.53	25.67
4364	2001	-0.02	6.02	41.77	26.18
4386	2001	-0.02	5.99	42.53	25.65
4415	2001	-0.01	6.03	43.54	25.88
4419	2001	-0.02	6.03	42.86	25.80
4445	2001	-0.02	5.98	42.52	25.68
4456	2001	-0.01	6.01	46.13	26.28
4492	2001	0.00	6.03	42.98	26.57
4495	2001	-0.01	6.01	46.12	26.28
4506	2001	0.00	6.02	44.10	26.42
4517	2001	0.01	6.03	42.88	26.61
4519	2001	0.01	6.03	43.04	26.56
4539	2001	0.00	6.02	43.01	26.47
4615	2001	-0.02	5.98	44.74	26.31
4632	2001	-0.02	5.97	43.31	26.19
4641	2001	-0.02	5.97	45.10	26.30
4643	2001	-0.02	5.97	45.10	26.30

APPENDIX 4 : Budburst DOY predictions for year 2001 using the Ensemble Model

Table 10: Site-specific Budburst DOY Predictions for 2001 using Ensemble Model

site_id	year	true_budburst_doy	Ens_model_predicted_doy
94	2001	123.0	112.0
103	2001	123.0	112.4
106	2001	119.0	112.1
124	2001	130.0	112.8
152	2001	127.0	114.9
199	2001	130.0	119.1
213	2001	123.0	117.5
244	2001	129.0	116.3
280	2001	120.0	117.5
291	2001	123.0	115.5
317	2001	125.0	110.8
427	2001	98.0	114.1
450	2001	123.0	112.1
502	2001	116.0	113.6
531	2001	124.0	115.2
532	2001	123.0	115.1
638	2001	122.0	113.6
656	2001	105.0	112.8
680	2001	115.0	113.8
765	2001	123.0	113.2
766	2001	117.0	110.8
799	2001	122.0	114.3
810	2001	123.0	110.3
855	2001	121.0	113.8
911	2001	112.0	115.9
975	2001	122.0	114.3
977	2001	123.0	115.3
1072	2001	121.0	112.6
1074	2001	122.0	114.4
1076	2001	121.0	114.1
1111	2001	120.0	112.5
1119	2001	120.0	112.2
1127	2001	121.0	111.8
1129	2001	122.0	114.4
1203	2001	124.0	115.3
1324	2001	118.0	112.5
1326	2001	117.0	112.9
1358	2001	107.0	112.0
1469	2001	128.0	114.1

1537	2001	123.0	111.5
1539	2001	120.0	110.6
1619	2001	122.0	113.8
1690	2001	122.0	114.5
1713	2001	126.0	114.5
1723	2001	119.0	113.8
1758	2001	120.0	115.2
1759	2001	131.0	114.6
1771	2001	122.0	117.1
1821	2001	124.0	116.2
1851	2001	124.0	116.0
1872	2001	122.0	115.1
1876	2001	128.0	114.0
1887	2001	119.0	110.8
1920	2001	101.0	111.7
1969	2001	115.0	112.3
1974	2001	112.0	111.7
1980	2001	99.0	113.2
1981	2001	114.0	111.7
2009	2001	125.0	111.5
2023	2001	123.0	114.4
2024	2001	123.0	114.5
2038	2001	117.0	111.5
2075	2001	121.0	114.9
2112	2001	126.0	114.2
2126	2001	126.0	113.8
2140	2001	122.0	114.9
2153	2001	122.0	113.8
2164	2001	122.0	114.5
2168	2001	106.0	114.1
2170	2001	123.0	114.0
2182	2001	122.0	114.5
2203	2001	119.0	114.4
2216	2001	121.0	114.7
2217	2001	120.0	114.6
2225	2001	121.0	113.9
2241	2001	120.0	114.7
2250	2001	121.0	115.2
2286	2001	121.0	114.2
2294	2001	117.0	114.7
2295	2001	122.0	113.7
2296	2001	120.0	114.4
2327	2001	122.0	115.9
2360	2001	114.0	115.6
2390	2001	98.0	112.1
2415	2001	122.0	113.8
2432	2001	126.0	114.1

2444	2001	120.0	114.3
2489	2001	120.0	114.0
2545	2001	105.0	114.9
2552	2001	118.0	111.9
2561	2001	128.0	113.0
2566	2001	122.0	115.2
2602	2001	118.0	114.7
2615	2001	118.0	114.8
2674	2001	122.0	113.2
2698	2001	113.0	111.8
2739	2001	118.0	113.7
2761	2001	121.0	113.4
2763	2001	120.0	111.9
2808	2001	123.0	114.4
2831	2001	99.0	112.9
2842	2001	118.0	113.2
2852	2001	122.0	111.2
2853	2001	120.0	111.4
2867	2001	115.0	113.4
2869	2001	120.0	112.9
2886	2001	118.0	113.0
2893	2001	121.0	111.8
2899	2001	106.0	113.4
2909	2001	122.0	110.9
2937	2001	104.0	112.8
2953	2001	110.0	112.4
2958	2001	121.0	113.3
2976	2001	122.0	112.5
3014	2001	114.0	113.6
3018	2001	126.0	110.8
3023	2001	116.0	110.5
3053	2001	100.0	110.6
3063	2001	119.0	111.4
3065	2001	116.0	112.8
3068	2001	115.0	112.8
3069	2001	117.0	114.1
3076	2001	118.0	113.3
3082	2001	118.0	113.2
3091	2001	131.0	112.0
3099	2001	115.0	111.5
3100	2001	110.0	110.6
3105	2001	119.0	111.6
3124	2001	119.0	112.7
3127	2001	115.0	112.4
3173	2001	92.0	113.5
3213	2001	118.0	112.4
3230	2001	98.0	109.0

3248	2001	111.0	109.6
3264	2001	122.0	109.3
3284	2001	110.0	112.6
3290	2001	103.0	113.1
3332	2001	122.0	109.6
3346	2001	121.0	111.2
3366	2001	122.0	112.9
3378	2001	121.0	110.6
3417	2001	114.0	109.6
3422	2001	126.0	109.5
3438	2001	120.0	111.4
3482	2001	108.0	112.1
3484	2001	121.0	111.9
3486	2001	120.0	110.6
3515	2001	121.0	111.2
3520	2001	120.0	111.6
3526	2001	115.0	112.1
3534	2001	116.0	111.5
3542	2001	118.0	110.1
3589	2001	120.0	111.5
3628	2001	121.0	111.4
3630	2001	120.0	111.2
3633	2001	118.0	110.9
3656	2001	119.0	112.2
3659	2001	120.0	111.8
3693	2001	118.0	110.2
3694	2001	116.0	110.9
3723	2001	114.0	111.3
3744	2001	119.0	112.2
3750	2001	117.0	111.5
3798	2001	124.0	111.6
3824	2001	115.0	111.9
3827	2001	117.0	111.6
3852	2001	108.0	112.1
3876	2001	121.0	111.5
3884	2001	115.0	111.8
3893	2001	130.0	111.1
3911	2001	116.0	111.3
3918	2001	111.0	110.6
3945	2001	121.0	111.5
3959	2001	120.0	111.3
3983	2001	119.0	111.2
3984	2001	119.0	111.5
3986	2001	122.0	111.5
3993	2001	119.0	111.5
3995	2001	123.0	110.4
3997	2001	126.0	112.1

4015	2001	129.0	113.3
4046	2001	121.0	115.8
4054	2001	118.0	113.1
4055	2001	123.0	114.0
4057	2001	120.0	113.9
4080	2001	122.0	112.4
4098	2001	120.0	113.2
4121	2001	124.0	112.7
4146	2001	106.0	114.7
4156	2001	122.0	110.8
4157	2001	128.0	113.2
4172	2001	119.0	112.0
4198	2001	121.0	112.6
4223	2001	124.0	112.1
4230	2001	112.0	113.4
4239	2001	117.0	111.0
4240	2001	112.0	113.1
4256	2001	116.0	113.9
4260	2001	108.0	112.6
4263	2001	122.0	110.8
4266	2001	117.0	112.4
4268	2001	124.0	110.8
4317	2001	120.0	113.3
4333	2001	102.0	112.5
4340	2001	109.0	112.3
4364	2001	119.0	113.8
4386	2001	106.0	112.3
4415	2001	122.0	114.0
4419	2001	120.0	113.9
4445	2001	117.0	112.5
4456	2001	119.0	110.9
4492	2001	105.0	109.2
4495	2001	121.0	111.5
4506	2001	118.0	111.0
4517	2001	123.0	110.8
4519	2001	122.0	110.2
4539	2001	123.0	111.0
4615	2001	124.0	113.5
4632	2001	111.0	112.0
4641	2001	129.0	112.6
4643	2001	122.0	113.1

APPENDIX 5: Budburst DOY predictions for year 2001 using the ML-derived Model

Table 11: Site specific Budburst DOY Predictions for 2001 using ML-derived Parameterization Model

site_id	***year***	***true_budburst_doy***	***predicted_budburst_doy***
94	2001	123	119
103	2001	123	119
106	2001	119	120
124	2001	130	130
152	2001	127	118
199	2001	130	121
213	2001	123	122
244	2001	129	125
280	2001	120	122
291	2001	123	120
317	2001	125	120
427	2001	98	123
450	2001	123	121
502	2001	116	124
531	2001	124	125
532	2001	123	125
638	2001	122	124
656	2001	105	123
680	2001	115	124
765	2001	123	122
766	2001	117	122
799	2001	122	120
810	2001	123	124
855	2001	121	124
911	2001	112	128
975	2001	122	124
977	2001	123	124
1072	2001	121	119
1074	2001	122	123
1076	2001	121	123
1111	2001	120	119
1119	2001	120	121
1127	2001	121	121
1129	2001	122	128
1203	2001	124	123
1324	2001	118	126
1326	2001	117	126
1358	2001	107	126
1469	2001	128	123
1537	2001	123	127
1539	2001	120	127
1619	2001	122	123

1690	2001	122	124
1713	2001	126	123
1723	2001	119	124
1758	2001	120	124
1759	2001	131	120
1771	2001	122	121
1821	2001	124	124
1851	2001	124	122
1872	2001	122	129
1876	2001	128	127
1887	2001	119	124
1920	2001	101	117
1969	2001	115	116
1974	2001	112	115
1980	2001	99	116
1981	2001	114	115
2009	2001	125	122
2023	2001	123	110
2024	2001	123	113
2038	2001	117	113
2075	2001	121	123
2112	2001	126	114
2126	2001	126	113
2140	2001	122	123
2153	2001	122	111
2164	2001	122	113
2168	2001	106	123
2170	2001	123	112
2182	2001	122	122
2203	2001	119	111
2216	2001	121	112
2217	2001	120	112
2225	2001	121	112
2241	2001	120	122
2250	2001	121	125
2286	2001	121	113
2294	2001	117	110
2295	2001	122	112
2296	2001	120	120
2327	2001	122	124
2360	2001	114	123
2390	2001	98	124
2415	2001	122	123
2432	2001	126	123
2444	2001	120	121
2489	2001	120	129
2545	2001	105	124

2552	2001	118	120
2561	2001	128	124
2566	2001	122	128
2602	2001	118	123
2615	2001	118	127
2674	2001	122	120
2698	2001	113	117
2739	2001	118	113
2761	2001	121	118
2763	2001	120	117
2808	2001	123	113
2831	2001	99	129
2842	2001	118	120
2852	2001	122	120
2853	2001	120	121
2867	2001	115	124
2869	2001	120	119
2886	2001	118	121
2893	2001	121	122
2899	2001	106	121
2909	2001	122	121
2937	2001	104	121
2953	2001	110	122
2958	2001	121	121
2976	2001	122	118
3014	2001	114	124
3018	2001	126	122
3023	2001	116	122
3053	2001	100	120
3063	2001	119	117
3065	2001	116	120
3068	2001	115	119
3069	2001	117	119
3076	2001	118	118
3082	2001	118	119
3091	2001	131	117
3099	2001	115	119
3100	2001	110	121
3105	2001	119	124
3124	2001	119	121
3127	2001	115	121
3173	2001	92	122
3213	2001	118	123
3230	2001	98	117
3248	2001	111	125
3264	2001	122	119
3284	2001	110	122

3290	2001	103	122
3332	2001	122	120
3346	2001	121	120
3366	2001	122	120
3378	2001	121	122
3417	2001	114	117
3422	2001	126	121
3438	2001	120	116
3482	2001	108	108
3484	2001	121	122
3486	2001	120	115
3515	2001	121	114
3520	2001	120	108
3526	2001	115	119
3534	2001	116	118
3542	2001	118	123
3589	2001	120	112
3628	2001	121	114
3630	2001	120	114
3633	2001	118	119
3656	2001	119	108
3659	2001	120	110
3693	2001	118	115
3694	2001	116	116
3723	2001	114	114
3744	2001	119	115
3750	2001	117	119
3798	2001	124	114
3824	2001	115	111
3827	2001	117	113
3852	2001	108	118
3876	2001	121	116
3884	2001	115	123
3893	2001	130	116
3911	2001	116	117
3918	2001	111	116
3945	2001	121	118
3959	2001	120	116
3983	2001	119	115
3984	2001	119	117
3986	2001	122	112
3993	2001	119	112
3995	2001	123	115
3997	2001	126	115
4015	2001	129	117
4046	2001	121	111
4054	2001	118	119

4055	2001	123	123
4057	2001	120	121
4080	2001	122	112
4098	2001	120	116
4121	2001	124	114
4146	2001	106	120
4156	2001	122	114
4157	2001	128	117
4172	2001	119	118
4198	2001	121	122
4223	2001	124	122
4230	2001	112	123
4239	2001	117	120
4240	2001	112	121
4256	2001	116	121
4260	2001	108	118
4263	2001	122	118
4266	2001	117	122
4268	2001	124	118
4317	2001	120	111
4333	2001	102	111
4340	2001	109	113
4364	2001	119	116
4386	2001	106	119
4415	2001	122	113
4419	2001	120	112
4445	2001	117	118
4456	2001	119	120
4492	2001	105	112
4495	2001	121	120
4506	2001	118	120
4517	2001	123	115
4519	2001	122	114
4539	2001	123	120
4615	2001	124	132
4632	2001	111	128
4641	2001	129	128
4643	2001	122	128