26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Comparing machine learning and process-based models for predicting wheat flowering timing

Harris D Ledvinka [a], Jonathan Richetti [b], Thomas Bishop [a], Patrick Filippi [a]

[a] *Sydney Institute of Agriculture, School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Eveleigh, Sydney, New South Wales, 2015, Australia*
[b] *CSIRO Agriculture & Food, 147 Underwood Avenue, Floreat, Perth, Western Australia, 6014, Australia*
Email: harris.ledvinka@sydney.edu.au

Abstract: Accurate predictions of flowering timing in wheat (*Triticum aestivum* L.) are an important decision-support tool for crop management. Process-based models such as the Agricultural Production Systems sIMimulator Next Generation (APSIM-NG) are widely used, but recent advances in machine learning (ML) may offer more accurate and flexible alternatives. We present a tree-based ML model for predicting wheat flowering timing and compare its accuracy with APSIM-NG using an independent dataset to assess if ML can outperform a traditional process-based approach.

Keywords: *Machine learning, XGBoost, APSIM, anthesis, phenology, predict*

1. INTRODUCTION

The flowering period in wheat (*Triticum aestivum* L.) is particularly sensitive to adverse environmental conditions such as frost and heat, which can diminish yield by reducing both grain size and number (Barlow et al. 2015). However, being able to accurately predict when flowering will occur can support growers to select the optimal sowing dates and varieties for the growing environment.

Process-based models such as APSIM-NG (Holzworth et al. 2018) are widely used for this purpose. These models use parameterised, mechanistic equations to represent crop physiological processes in response to environmental conditions such as temperature and solar radiation. While they can offer accurate predictions that are grounded in biology, some limitations include their extensive input data requirements and their need for recalibration for new environments (George et al. 2024).

Unlike process-based approaches, ML models are data-driven and can capture complex, nonlinear relationships between inputs and outputs, without the need for recalibration. However, ML models require large and representative training datasets to be both accurate and scalable. The training dataset used in this research was the largest among known existing ML studies for predicting wheat flowering timing. Over 30,000 wheat flowering observations were provided by the Grains Research and Development Corporation (GRDC)'s National Variety Trials (NVT) program (https://nvt.grdc.com.au), which is the world's largest coordinated variety trial network. The observations were collected across 18 seasons (2005-2022) from sites across the Australian wheat belt.

The aim was to compare process-based and ML methods for predicting wheat flowering timing in response to key climate, environmental and crop management factors. To achieve this aim, the following objectives were addressed: (1) develop and validate an eXtreme Gradient Boosting (XGBoost) model for predicting wheat flowering timing, (2) use an independent dataset to simulate wheat flowering timing with APSIM-NG, and (3) compare the predictive accuracy of the XGBoost model with APSIM-NG using the independent dataset.

2. METHODOLOGY AND RESULTS

The first objective was to develop and validate an XGBoost model, a powerful and efficient tree-based ensemble method. Over 30,000 wheat flowering observations from the NVT program, spanning 18 growing seasons (2005-2022) across the Australian wheat belt, served as the primary dataset. This comprehensive dataset, which included sowing date, flowering date, GPS coordinates, and variety information, was the largest of its kind for predicting wheat flowering timing in known existing ML studies.

Features for the model were derived from the original NVT data and external sources, including cumulative weather variables (e.g., growing degree days, heat stress), environmental variables (e.g., daylength), soil variables (e.g., plant available water capacity), and management and genetic factors (e.g., variety, sowing day) (**Table 1**).

Table 1 Features used in the ML model and data sources

Feature type	Feature(s)	Data source
Cumulative weather	Growing degree days, degree days <2ºC (frost), degree days <10ºC (vernalisation), degree days >32ºC (heat stress), rainfall, vapour-pressure deficit, solar radiation	Scientific Information for Land Owners (SILO) database
Cumulative environmental	Daylength	'geosphere' package in R
Soil variable	Plant available water capacity (0-1 m)	Soil and Landscape Grid of Australia
Management and genetic	Variety maturity class	University of Sydney Plant Breeding Institute classifications, based on variety from NVT data
	Sowing day	NVT

To assess model performance, we conducted three validation tests. An initial 70:30 training-test split was used for general performance evaluation, ensuring that year-site combinations were allocated to prevent data leakage. To test the model's ability to generalise over time and space, we performed leave-one-year-out cross-validation and leave-one-site-out cross-validation. The predictive accuracy of the model was evaluated using a suite of metrics: Lin's concordance correlation coefficient (LCCC), R^2, root mean squared error (RMSE, in days), and accuracy within ±3, 5, and 7 days.

To compare with the process-based model, APSIM-NG, we used an independent test dataset comprising over 3,000 wheat flowering observations, also from the NVT program. The APSIM-NG wheat flowering simulations were performed using the 'apsimx' R package. The XGBoost model was then trained using data from year-site combinations not represented in the independent test set, to prevent data leakage, to make predictions using the test dataset. This enabled fair comparison between APSIM-NG and the XGBoost model. The results of this final comparison are still being processed and will be presented at the conference.

3. CONCLUSIONS

The aim of this research was to compare the predictive accuracy of a data-driven ML model with a traditional process-based model for predicting wheat flowering timing, which is a critical tool for precision agriculture. While process-based models such as APSIM-NG provide predictions grounded in biology, they often require extensive input data and recalibration for new environments. In contrast, the data-driven ML model developed in this research used a comprehensive dataset of over 30,000 observations to capture complex, nonlinear relationships, potentially offering a more flexible and scalable alternative approach for predicting wheat flowering timing.

The preliminary results, which will be presented at the conference, suggest that the XGBoost model shows strong spatiotemporal generalisability, demonstrating its ability to accurately predict flowering dates in new sites and years beyond the training data. The comparison on an independent test set will be particularly revealing, providing quantitative evidence of whether a data-driven ML approach can outperform a process-based one in this domain.

ACKNOWLEDGEMENTS

This research was enabled by GRDC through the provision of data resources generated as part of the NVT program and provided through the NVT Resource Sharing pathway. We respectfully acknowledge the Traditional Owners of the lands where the NVT data were collected and acknowledge their enduring custodianship of these lands.

REFERENCES

Barlow KM, Christy BP, O'Leary GJ, Riffkin PA and Nuttall JG (2015) Simulating the impact of extreme heat and frost events on wheat crop production: A review. Field Crops Research 171, 109-119. Doi: https://doi.org/10.1016/j.fcr.2014.11.010.

George NA, de Jesus Pedro Cuamba H, Lundy ME and Bennett SJ (2024) Evaluating the Agricultural Production Systems sIMulator (APSIM) wheat module for California. Crop and Pasture Science 75(3), -. Doi: https://doi.org/10.1071/CP23046.

Holzworth D, Huth NI, Fainges J, Brown H, Zurcher E, Cichota R, Verrall S, Herrmann NI, Zheng B and Snow V (2018) APSIM Next Generation: Overcoming challenges in modernising a farming systems model. Environmental Modelling & Software 103, 43-51. Doi: https://doi.org/10.1016/j.envsoft.2018.02.002.

