Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

On the validation of fine-resolution spatio-temporal agricultural models and the illusion of predictive accuracy

Si Yang Han

siyang.han@sydney.edu.au

The University of Sydney

Thomas Francis Aloysious Bishop

The University of Sydney

Patrick Filippi

The University of Sydney

Research Article

Keywords: wheat yield, yield modelling, data-driven modelling, machine learning, spatio-temporal scales

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11684230/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

On the validation of fine-resolution spatio-temporal agricultural models and the illusion of predictive accuracy

Han S, Bishop TFA, and Filippi P

Precision Agriculture, Hydrology & Geoinformation Sciences Laboratory, Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2015, siyang.han@sydney.edu.au

Abstract

Purpose

A common pitfall in data-driven modelling studies is the assumption that global predictive performance (i.e. across all data points) represents local performance. Agricultural datasets are often inherently structured (e.g. varieties, seasons, fields), where larger between-group differences can mask poor predictions of within-group variability across common prediction metrics. This study showcases and explores the issue with commonly used data-driven models, using crop yield data as the response variable (160 fields, four farms, 10 seasons) and freely accessible spatial products as the predictor variables.

Methods

A series of experiments evaluated the effects of different validation approaches, predictor resolutions and model algorithms. Random-effects models were used understand model performance for different scales by partitioning observed yield and model error between seasons, farms, fields and within-fields.

Results

Most of the observed variance in the dataset occurred between seasons (43.8%), farms (14.8%) and fields (27.3%), with within-field variance accounting for a relatively small 14.1%. When interpolating within a field (Experiment 1), 71.4% of within-field variance was explained by the model which reduced to 37.1% when extrapolating to neighbouring fields (Experiment 2). In Experiment 2, the global NSE was 0.66, yet 59.4% of fields had a negative NSE. Removing coarse-resolution weather predictors or alternative model algorithms did not improve within-field prediction performance.

Conclusion

This study shows that validation needs to be done at the scale of interest for an end-user, e.g. within-field for a grower to understand their yield variability. However, improved validation does not resolve the underlying issue; data-driven models use loss functions which are fit globally across all training data, hence the fitted model appears to prioritise the dominant scales of variation which is between-season, farms and fields.

Keywords: wheat yield, yield modelling, data-driven modelling, machine learning, spatio-temporal scales

Introduction

In the last decade, there has been a rapid increase in the amount of agricultural research that uses data-driven modelling to address spatial and/or temporal problems. This aligns with the vast amount of on-farm data being collected (e.g. crop yield and quality, proximal and in-situ soil sensing, soil samples, crop management), freely accessible spatial data such as satellite imagery (Wulder *et al.,* 2012) and elevation (Van Zyl, 2001), alongside improvements in computational power. Researchers are now applying increasingly complex machine learning and deep learning algorithms to find solutions across larger areas and time periods. Some agricultural and environmental studies primarily vary in space, such as digital soil maps of physical and chemical properties (Benos *et al.*, 2021), whereas others also have a temporal component, e.g. soil carbon and moisture. In the agricultural context, the focus is often on crop modelling and monitoring – yield (Klompenburg *et al.*, 2020), crop type (Orynbaikyzy *et al.*, 2019), weed detection (Adhinata *et al.*, 2024), disease detection (Ngugi *et al.*, 2024) and nitrogen status (Chlingaryan *et al.*, 2018).

The problem with 'global' validation

In many of these studies, variable selection, model fitting and assessment of prediction quality is based on 'global' assessment (i.e. over multiple seasons or a large study area, which could be national, regional, or a farm) of observed and predicted values via a metric (e.g. mean squared error, correlation). Given the spatial and temporal scale of some studies, it is inevitable that there is spatial or spatio-temporal structure in the data, and global assessment masks the predictive capability of these models at more local scales in which they are generally used to make decisions. For example, soil samples collected from a farm would likely be more similar to each other compared to samples collected from a farm in another region. In this context, each farm represents a subgroup. When a dataset contains substantial differences between subgroups, global metrics (e.g. correlation, R^2 and LCCC) can still indicate strong agreement or 'good' fit driven primarily by between-group differences, masking the lack of relationships within groups (Aggarwal and Ranganathan, 2016).

Figure 1 shows observed and manipulated wheat yield data (160 fields across a large farm in Australia) to reflect two extreme scenarios. Each point in Figure 1 represents the crop yield (t ha^{-1}) measured by a yield sensor on a combine harvester which is then aggregated to a 30 by 30 m area in a field. The first example (left) uses the mean yield of each field as the simulated yield, so there is zero spatial variability in yield within each field. In the second example (right), a random amount of bias ($X \sim N(0, 1)$) is added to each field (the same bias across all observations in each field) so the spatial patterns of yield are retained. In both cases, goodness-of-fit is assessed globally (i.e. across all points on the farm, rather than for each individual field) with Pearson's correlation (r), Lin's concordance correlation coefficient (LCCC), Nash-Sutcliffe Efficiency (NSE) and root-mean-squared error (RMSE). All four metrics indicate the fit for the first example (left) is comparable if not stronger than the second example (right) despite having no variability in the predicted value within each field. In the context of precision agriculture predicting variability at within-field resolutions is key for variable rate management, hence this is problematic.

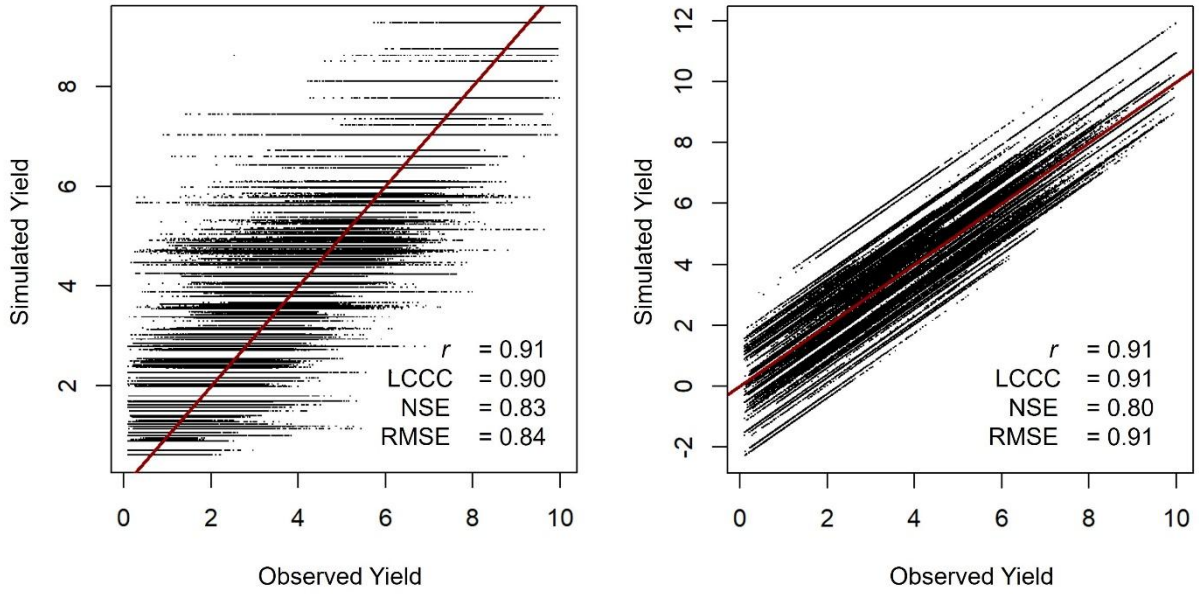


Figure 1. Scatterplots of observed and simulated wheat yield (t ha^{-1}) values. The plot on the left uses field means as the simulated yield, the plot on the right adds random bias to the observed yield each field ($X \sim N(0, 1)$) to form skewed simulations. Global statistics are provided.

Examples in agricultural modelling

The studies below have been included to showcase how often global assessment is solely used in agricultural models and spatial products, although there are clearly spatial or temporal groups in the dataset or figures. The purpose of this paper is not to dismiss or disregard previous work, as many of the methodologies and approaches are still novel and provide valuable insights. In some examples, the authors are limited by what they can show due to the size of their dataset, or the scope of their studies.

Han et al. (2022a) briefly explored the limitations of global validation in publicly available digital soil maps (DSM) with an independent dataset collected across several farms in eastern Australia. The DSM products SoilGrids250m (Hengl *et al.*, 2017) and Soil and Landscape Grid of Australia (Grundy *et al.*, 2015; Viscarra Rossel *et al.*, 2015) were both developed from data-driven models and managed to capture clay and soil organic carbon variability between the farms, but a negligible amount of variability within each farm. As the models were evaluated globally, their published performance metrics indicated moderate to good predictive accuracy. Consequently, end users without extensive independent validation data may be unaware of the models' actual predictive capabilities at the spatial resolution of the products (e.g. 90 m or 250 m) and spatial scales of interest for growers such as within-field variability for precision agriculture. While both DSM products have since been updated with more data as well as improved predictors, models and validation approaches, prediction quality is still presented globally (Poggio *et al.*, 2021; Malone *et al.*, 2025).

Validation typically becomes more difficult with crop models as there is a temporal scale with differences between each growing season. A recent publication in crop phenology modelling by Zhao *et al.* (2025) identifies and tests a curve-derived feature from a timeseries of NDVI for wheat and barley trial sites across Australia (trained on three sites, tested with 183 sites). In several figures, they combined data from wheat and barley crops, multiple sites, and different growing seasons. There is clear clustering around each of the crop types and sites, but global prediction quality metrics (R^2, RMSE) are provided to justify "the significant correlations". Ledvinka *et al*. (2026) is another study on crop phenology where wheat phenology is trained and validated with a large dataset across Australia (169 sites, 19 years), but again, only global statistics (i.e. national) are discussed, and scatterplots show all points available, despite inherent spatio-temporal subgroups (e.g. grain-growing subregions, growing seasons).

The same issues are evident in crop yield modelling, though there is often more data to consider spatio-temporal subgroups. For example, Filippi *et al.* (2020) identifies "fine resolution" (i.e. within-field) predictions as a research gap for cotton yield forecasts. Though the authors briefly show and state "some points were very poorly predicted" at fine 30 m resolution, the discussion instead shifts to field and season averages which had better prediction quality metrics. Another example is a study by Richetti *et al.* (2024) which uses a mechanistic model to simulate wheat yield at trial sites (217 sites, 18 years). Separate observed and predicted scatterplots (e.g. one plot per year) are shown with prediction quality metrics for all points within each plot, and the points are coloured to show independent subgroups (e.g. geographic states). Though some subgroups have predicted values visibly low in variability, subgroup performance is not assessed quantitatively or mentioned. Similarly, Han *et al.* (2022b) hints there is an issue, but does not openly discuss it in the text. In that example, field-scale sugarcane yield forecasts from neighbouring farms are shown on the same plot with different coloured points for each farm but the fit within each farm (different fields) is visibly more mediocre than the global fit suggests.

Showing global validation results does not inform readers or end users how the predictions actually perform at the point or pixel level, and how it represents variability at the scale of interest, in precision agriculture this being within-field scales for a particular season. If a study or product claims the outputs can be used to support precision agriculture or within-field management, 'local' validation (i.e. the performance within spatio-temporal subgroups) through figures or statistics needs to be provided to

justify such claims. As datasets become larger, it becomes more challenging for authors to conduct thorough validation and groundtruthing, particularly with national or global products. Some approaches are explored in this study.

Aims

A prime example to showcase and explore the issue of global and local validation is crop yield monitor data. Yield monitors installed on combine harvesters collect high-resolution, geolocated and time-stamped yield data. The result is a complete, spatial record of a field, rather than a small group of point-based samples per field. Harvesters with yield monitors have been widely adopted across Australian farms (Bramley and Ouzman, 2019), generating large volumes of yield data across regions and seasons. This is also a nested dataset with different spatio-temporal scales of variability to explore: between-season, between-farm, between-field and within-field. Most important of these scales is within-field in terms of prediction quality for a grower interested in precision agriculture.

Filippi *et al.* (2025) used a collection of 160 wheat yield maps across four farms and 10 seasons to address a common misunderstanding among authors about the differences between predicting, interpolating, and forecasting crop yield with data-driven models. The paper showed results for yield predictions only globally and did not address the matter of local validation as not to detract from the focus of the paper. Thus, using the same dataset, this study investigates four experiments to explore potential factors that contribute to poor within-field predictions:

Experiment 1: Assessing the within-field predictive capability of the model and predictors by interpolating within the same field by using 2-fold validation (2FV),

Experiment 2: Comparing the within-field performance when extrapolating to unseen neighbouring fields by using leave-one-field-out cross-validation (LOFOCV),

Experiment 3: Testing whether coarse-resolution predictors cause low prediction variability, as some predictors do not vary at the within-field level, and

Experiment 4: Testing whether poor within-field predictions persist across different model algorithms (tree-based vs neural network), to check that the issue is not specific to one model type.

The focus of this study is not to make the best possible yield prediction but rather showcase how global assessment can mask a model's performance locally and explore potential reasons on why data-driven models fail to capture local variability.

Methods

Study area and yield data

The yield dataset was collated from four neighbouring rainfed cropping farms in northern NSW, Australia. The farms are within 40 km of each other, with fields ranging from 12-496 ha in size. A total of 160 wheat *(Triticum aestivum)* yield maps were collected by yield monitors on combine harvesters between 2014 and 2023, totalling 26,400 ha of yield maps. Yield data was cleaned by a precision agriculture service (e.g. start and end passes, header turns, GPS jitter), and extreme values were then removed if they were more than three times the interquartile range below the first quartile or above the third quartile. Remaining points were then averaged to a 30 m grid. Note 'field' throughout this paper refers to a unique paddock and season, i.e. Paddock A 2015 and Paddock A 2016 are one paddock but two fields.

The farms are sown with an irregular rotation of winter crops (wheat, barley, canola, chickpea) and opportunistic summer crops (cotton, sorghum), but this study focuses solely on wheat yield. Winter grains are sown over 23.5 million ha across Australia, with wheat alone accounting for approximately 60% of all winter grains each season (ABS, 2024). The region is characterised by mild winter temperatures, high summer temperatures, and summer-dominant rainfall (Figure 2). Being a rainfed system, there is high variability in water availability and temperature extremes between growing seasons and hence yield (Feng *et al.*, 2018). In 2015, despite favourable seasonal rainfall totals, low rainfall combined with exceptionally high temperatures and strong winds during the peak growth in September–October period contributed to low and highly variable yields in 2015. Due to drought from 2017 to 2019, no fields were sown in 2018 and just three fields were sown in 2019. In 2022, high rainfall early in the season delayed sowing, and flooding at the end of the season impacted harvest. Within the same season, there is also high spatial variability within and across fields for the same season, due to soil constraints, management, as well as pests and diseases. This kind of spatial and temporal variability reflects the reality for most dryland cropping farms in Australia, and other similar environments around the world.

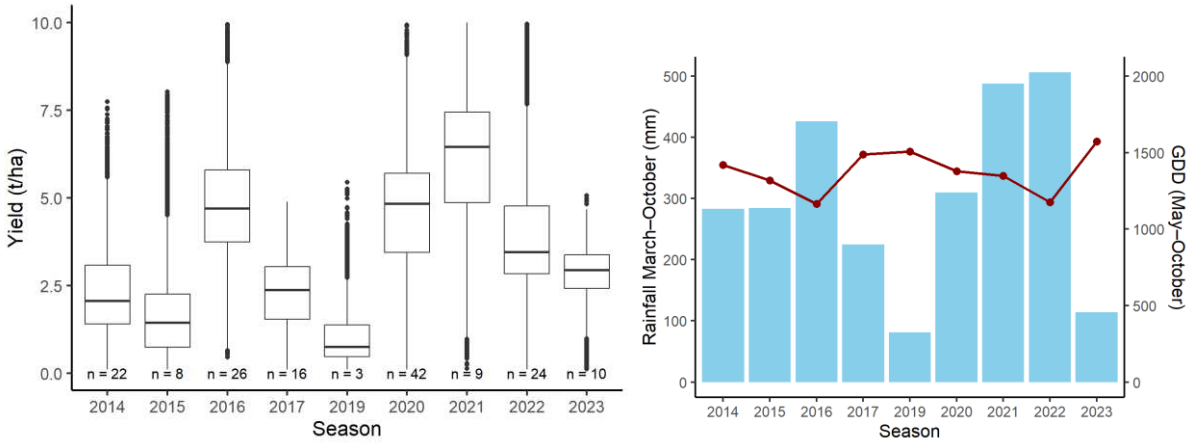


Figure 2. The distribution of wheat yield (30 m resolution) and number of fields (*n*) between each growing season (left), and the seasonal conditions (right) (Jeffrey et al., 2001). The rainfall (1 March – 31 October) (blue bars) and growing degree days (GDD) (1 May – 31 October) (red line) are included.

Predictor variables

A range of freely accessible spatial datasets were extracted and processed into 19 predictors (Table 1). All data cleaning, feature extraction and analysis were completed in R (R Core Team, 2023) unless otherwise specified. All remote sensing data was sourced from Google Earth Engine (Gorelick *et al.*, 2017), or through the 'dataharvester' R package (Haan *et al.*, 2023). The yield data was joined with the respective suite of spatial and spatial-temporal covariates to build the 'datacube' at 30 m resolution to match the resolution of most covariates. The complete details of feature extraction and relevance of each predictor to crop yield are provided in Filippi *et al.* (2025).

There are a few assumptions made when collating the dataset, which reduce the precision of seasonal weather variables. Weather is a coarse resolution modelled product, which has some degree of error. Given the coarse spatial resolution of the pixels (1500 m), the data is averaged to the field to avoid boundary effects between pixels. Additionally, the extracted predictors (i.e. cumulative growing degree days, GDD, and seasonal rainfall) assume constant sowing and harvest dates for the region (e.g. 1 May to 31 October). While there is variability from season to season (e.g. sowing could vary from mid-April to mid-June, harvest from late October to mid-December), the crop is generally expected to be in the

field for most of the window selected, which is adequate to compare between seasons. Seasonal rainfall was calculated from 1 March to 31 October to incorporate stored soil moisture in the months leading up to sowing (in March and April). Summer rainfall was not included as the fraction retained is dependent on many factors, from soil type and properties, to fallow management and crop residue.

Furthermore, another assumption was made regarding remote sensing predictors. Growers in the region aim for the crop to undergo rapid vegetative growth and flower roughly in September, regardless of sowing date, as the risk of frost is reduced, and the crop will experience less heat stress later in the season. Thus, the mean September composite image was chosen, as a composite image caters to differing acquisition dates and cloud cover. While in most cases the mid-September crop response is being captured, it is possible in some seasons the response is either early or late September, and the phenological stages is not as consistent in the image.

Table 1. Source, resolution and names of the spatial datasets used as predictors to model wheat yield.

Type	Source	Data	Resolution
Soil	Geoscience Australia (Wilford and Roberts, 2020)	Sentinel-2 bare earth imagery (red, blue, shortwave-infrared 1)	10-20 m
	(Minty *et al.*, 2009)	Gamma radiometrics (K, total dose)	100 m
	CSIRO (Grundy *et al.*, 2015)	Soil and Landscape Grid of Australia (clay, cation exchange capacity)	90 m
Terrain	CSIRO (Gallant *et al.* 2012; 2016)	Elevation, multi-ridge valley bottom flatness, slope, total shortwave radiation on a sloping surface (January, July)	30 m
Weather	ANUCLIM (Xu and Hutchinson, 2011)	Growing degree days (1 May-31 October), growing season rainfall (1 March-31 October)	1500 m
Imagery	NASA/USGS	Landsat 8 mean September vegetation indices (EVI, MSAVI, NDMI, NDVI)	30 m
	CSIRO (McVicar *et al.,* 2022)	Seasonal CMRSET ET (May-October)	30 m

Modelling

Models

The primary model used in the study is the random forest (RF) regression model from the 'ranger' package (Wright & Ziegler, 2017) with default hyperparameters (Experiment 1, 2, 3). Ensemble, tree-based models are commonly chosen for data-driven crop yield modelling due to their predictive performance, simplicity and robustness (Roßbach, 2018). As an ensemble of decision trees, the prediction is obtained by averaging outputs of hundreds of independent trees, which give a reasonable average prediction by reducing the effect of noisy data and outliers. Variable importance was calculated with permutation importance across the entire dataset (Experiments 1, 2) and with the reduced dataset (Experiment 3).

Experiment 4 instead uses the multi-layer perceptron model (MLP) as it can still incorporate many data layers and represent complex relationships while the model structure is fundamentally different to RF. The MLP model is a relatively simple application of a neural network, where stacked hidden layers allow for complex relationships and recognise subtle difference in inputs. The MLP function from the

'nnet' package (Ripley *et al.*, 2016) was used through the 'tidymodels' interface (Kuhn and Silge, 2022). In a review by Victor *et al.* (2024), MLP models are one of the most used algorithms for pixel-based agricultural predictions, but deep learning models did not perform noticeable better than machine learning models. For the MLP, the dataset was normalised and a 3 × 3 grid search was used for hyperparameter tuning, testing the number of hidden units (i.e. nodes; 10, 20, and 30) and the penalty (i.e. weight decay; 0.0001, 0.001, and 0.01). Detail on data splitting and training is below.

Model validation

Yield was modelled as a late-season hindcast as it performed better than a mid- or late-season yield forecast when assessed globally (Filippi *et al.*, 2025). Predictions were made with one of two validation approaches. Experiment 1 used 2-fold validation (2FV), which only used data from the same field, and subsets alternating strips of 30 m grid cells for training, and the remaining cropping rows for testing. This was intended to mimic a scenario where not all harvesters are equipped with a sensor, and this tested whether a model could interpolate across a field.

Experiment 2, 3 and 4 used leave-one-field-out cross-validation (LOFOCV). While data from the same paddock from a different season could be included in the training dataset, data from the same paddock and season would not be. This tested the ability of the model to extrapolate in space to predict yield in an unsampled field. No hyperparameter tuning was conducted for the RF model, as package default hyperparameters were used.

For Experiment 4, which used the MLP model, LOFOCV was applied slightly differently. To prevent data leakage, each training dataset was randomly split into two equally sized datasets at the field level: one half was used to train the model (train), whereas the other half was reserved for hyperparameter tuning (validation). The withheld field from LOFOCV remained the independent test (test).

Prediction validation

The predictions were assessed globally (all points) as well as by each field (only points from a specific paddock and season) to test local performance. The metrics used include Pearson's correlation coefficient (*r*), Lin's concordance correlation coefficient (LCCC), root-mean-squared error (RMSE) and the Nash-Sutcliffe Efficiency (NSE). Pearson's correlation coefficient (*r*) is a measure of linearity between two variables, regardless of scale. The LCCC incorporates linearity as well as deviance from the 1:1 line, thus incorporating scale. On the other hand, the RMSE is a measure of deviance as the average error across all predictions. These are commonly used statistics to measure prediction quality, and the Pearson's *r* can be squared to estimate the R^2 for comparison with other studies, provided *r* is positive.

The Nash-Sutcliffe Efficiency metric (Nash and Sutcliffe, 1970) can be used as a benchmark to determine whether models have any merit at all. Commonly used in hydrological modelling, NSE (Eq 1) is a measure between individual observations (O_i), the prediction (S_i), and the mean of the observations ($\bar{O}$). If the difference between observations and predictions is greater than the difference between the observations and the mean of observations, then the mean is better than the prediction and the NSE will be less than zero. This is similar to the F-test in simple linear regression. Hence the NSE provides an interpretable metric to determine whether the model should be used.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad \text{[Eq 1]}$$

Random effects analysis of variance

Lastly, a random effects analysis of variance (REANOVA) from the 'lme4' package (Bates *et al.*, 2015) was run on both the wheat yield (t ha^{-1}) observations and model errors. The errors are the difference between observations and predictions. REANOVA partitions the total variability into hierarchical components, allowing the relative contribution of each level to be quantified (Bishop *et al.*, 2015). For observed yield, this separates the variance into spatio-temporal scales of variation (between-season, between-farm, between-field, and within-field), showing where most of the variance lies (Lark, 2005). Comparing the variance components of observed yield and model error then indicates how much variability remains unexplained by the model at each spatio-temporal scale.

Experiments

This study examines four experiments to investigate local (i.e. within-field) predictions and the potential causes for poor prediction performance. Each experiment targets a different aspect of the modelling workflow, including assessment scale, predictor resolution, dataset variation, and model algorithm.

Experiment 1 model and predictor capability (RF, full predictors, 2FV): This experiment fit a separate model for each field. Within each field, alternating rows or columns (of a 30 m grid) were used to train the model, which was used to predict onto the remaining grid to test prediction performance (2FV). This was a benchmark of the upper bound of obtainable goodness-of-fit from the model, full set of predictors, and response. It was also a test of whether there was sufficient complexity in the model and predictors to capture within-field variability at all.

Experiment 2 model validation (RF, full predictors, LOFOCV): A RF model with a full set of predictors was fit using leave-one-field-out cross-validation (LOFOCV). This was a direct comparison with Experiment 1 to test how well the model will perform in neighbouring fields. This is a more challenging prediction problem, as the model must learn relationships from other fields and/or seasons and apply them to the withheld field.

Experiment 3 predictor resolution (RF, reduced predictors, LOFOCV): The most important predictors for crop modelling are often seasonal weather (e.g. rainfall, GDD), especially in Australian grains which are grown in rainfed systems and highly variable from year to year (Dayal *et al.*, 2020). Despite their importance, national weather predictors often have coarse resolution, e.g. this study used 1500 m ANUCLIM (Xu and Hutchinson, 2011) which were averaged to the field-level to avoid pixel edge effects. Coarse-resolution weather products (i.e. seasonal rainfall, growing degree days) were removed to test (against Experiment 2) whether a dataset with only fine-spatial resolution predictors improved local (within-field) yield predictions.

Experiment 4 model algorithm (MLP, full predictors, LOFOCV): Since all experiments above and in Filippi *et al.* (2025) used RF models, a MLP neural network model was tested. This was the same setup as Experiment 2, exchanging the RF model for a MLP model to test whether a different model algorithm improved predictions.

Results

Exploratory data analysis

Exploratory data analysis was undertaken to characterise the spatial and temporal variability in observed wheat yield (t ha^{-1}) as well as its relationship with the predictor variables. Within-field yield variability was characterised using the coefficient of variation (CV) for each field. Additionally, the strength and direction of relationships between yield and each predictor were assessed within individual fields to examine the consistency of predictor-yield relationships across the dataset.

The coefficient of variation (CV) for each field was calculated as the variability in yield relative to the mean yield for each field, which allows variability across fields to be compared. There was considerable within-field variability, with the median field CV of 21% and the interquartile range of 16-30% across the 160 fields. Across most seasons, the median field CV and first and third quartiles were relatively uniform, with only the 2015 and 2019 seasons noticeably more variable (Figure 3). This was primarily due to unusual seasonal conditions and fewer sown fields in those seasons.

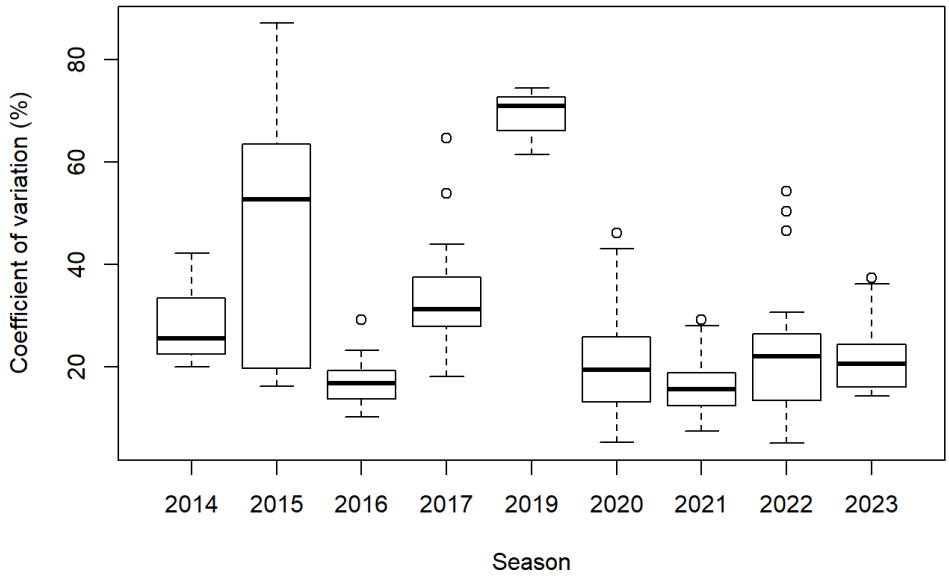


Figure 3. The per-field coefficient of variation for observed wheat yield (t ha^{-1}) from 2014-2023. Each point represents a different field.

The relationships between observed wheat yield and the predictors varied from field to field (Figure 4). In-season vegetation indices were the most consistently moderately to strongly correlated with observed yield. Soil and terrain had weaker relationships, with medians close to zero. In some cases, the field may have relatively uniform landscapes and soil types which are not a large source of within-field variability. In others, there may be interactions with seasonal conditions. For example, areas with greater soil clay content have higher water holding capacity and may support higher yields in dry seasons but are more susceptible to waterlogging in wet seasons. Hence machine learning models such as RF are a suitable choice for this dataset as it can capture potential nonlinear responses and interactions among predictors.

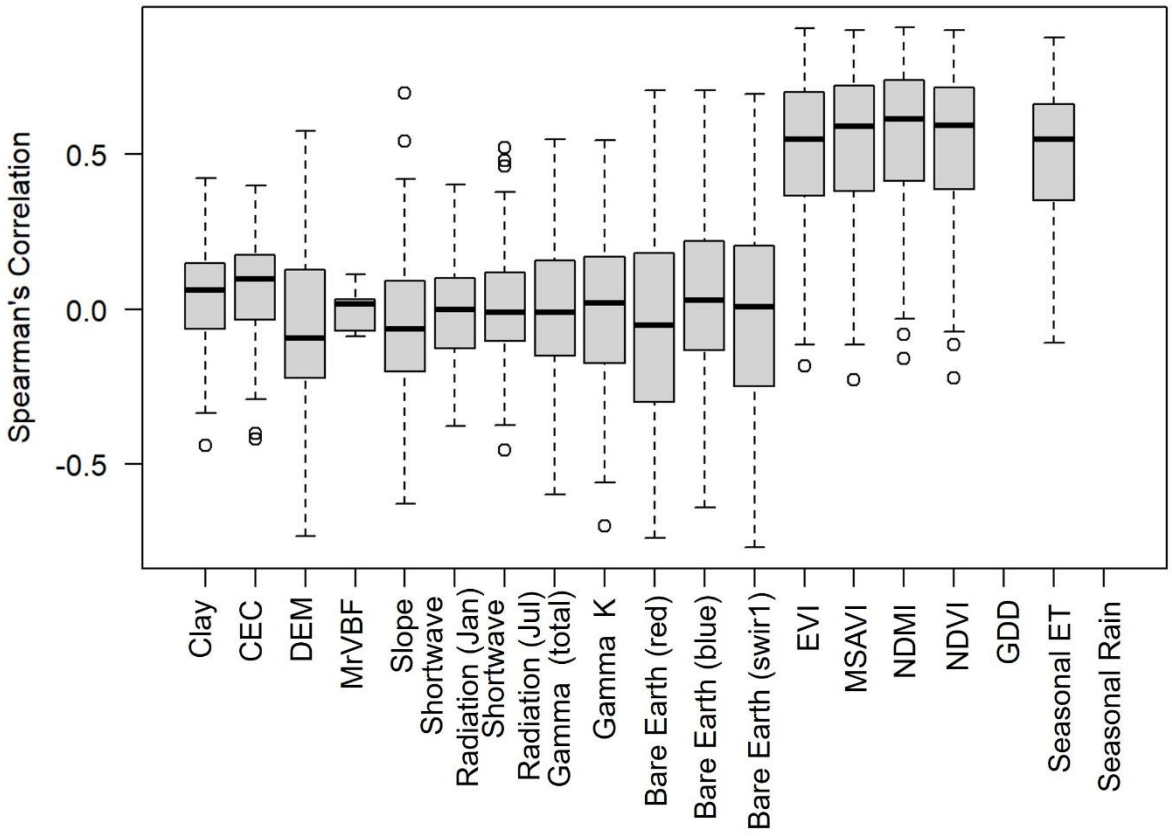


Figure 4. The per-field Spearman's correlation between the observed wheat yield (t ha^{-1}) and each predictor. Note 'GDD' and 'Seasonal Rain' are field-averaged values so per-field correlation cannot be calculated.

Experiment 1: model and predictor capability

Filippi *et al.* (2025) investigated a 2-fold validation approach (2FV) where data from alternating rows or columns of a 30 m grid were used to train the model and then predicted onto the remaining grid. The prediction performed well, as only data from the field (same paddock, same season) were used for training. The scope of the training and test data meant the model was useful for gap-filling or interpolation, but not for a transferable prediction to other fields or seasons.

In this study, the 2FV validation approach serves as the ceiling of the model's ability to capture complex relationships and interpolate, as well as the achievable predictions from the predictors in the model. The RF model with all predictors had a moderately strong median field goodness-of-fit (LCCC = 0.82, RMSE = 0.42 t ha^{-1}), although this was not as strong as the global metrics (LCCC = 0.97, RMSE = 0.45 t ha^{-1}) (Figure 5).

The NSE is similar to the F-test in simple linear regression, where there is a comparison between the predictions and the mean. In the case of the F-test, if the model does not explain more variability than the mean, the model is not statistically significant. Only one of 160 fields had a negative NSE, which indicated for almost all the fields, the model performed better than the mean of observations.

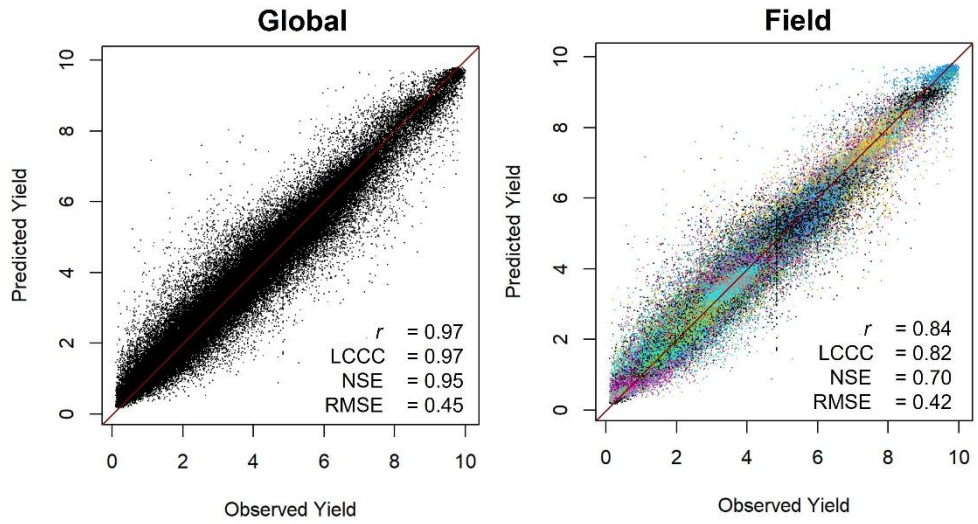


Figure 5. Scatterplots of wheat yield (t ha^{-1}) observations and predictions for Experiment 1 (RF, 2FV, full predictors). Plots on the left show global statistics calculated across all points the farm, plots on the right show median field statistics, calculated for each field (coloured).

Experiment 2: validation approach

A more applicable validation approach in this context is LOFOCV, which estimates the model's ability to predict onto neighbouring fields and farms. The RF model with all predictors appears to predict moderately well with global statistics (r = 0.81, LCCC = 0.80, NSE = 0.66, RMSE = 1.16 t ha^{-1}), particularly in terms of linear association between observations and predictions. An RMSE of 1.16 t ha^{-1} is the only statistic in this case which suggests the fit may not be as strong, and this requires knowledge of this specific dataset and study area to gauge. Even with a scatterplot of observed and predicted (Figure 6, left panel) it is easy to assume there is a good linear trend with some error, as the global statistics suggest. It would be just as easy to assume the global prediction quality at a point (30 m block) resolution holds true across all scales considered in this study, in particular within-fields.

However, once the scatterplot is coloured by individual fields (Figure 6, right panel), it is visibly clear the model does not fit individual fields well. When goodness-of-fit statistics were calculated for each field, there was a moderately strong median field correlation within each field (r = 0.60), but the median field LCCC of 0.32 makes clear that the fit within these fields is not on the 1:1 line. The median field RMSE is lower at 0.85 t ha^{-1}, as the median is less affected by some fields performing very poorly.

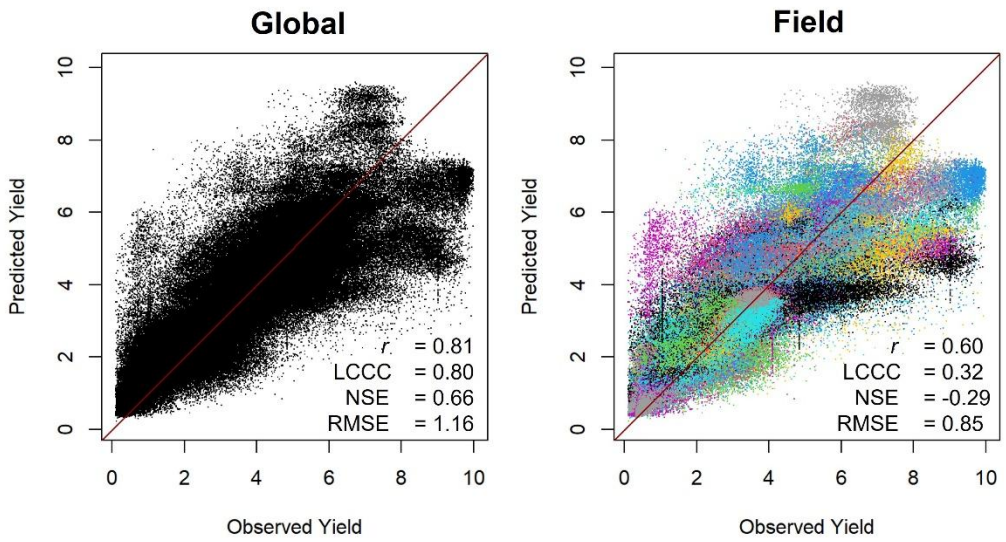


Figure 6. Scatterplots of wheat yield (t ha^{-1}) observations and predictions for Experiment 2 (RF, LOFOCV, full predictors). The left plot shows global statistics calculated across all points in the farm, the right plot shows median field statistics, calculated for each field (coloured).

When global statistics are considered, the predictions have an NSE of 0.66 (Figure 6, left panel), which is well over the benchmark of 0. When the performance of each field is considered independently, a very different account is presented compared to the global metrics (Figure 7). The median field NSE was -0.29, and 59.4% of fields had a negative NSE, indicating that predictions for the majority of fields were less accurate than using their respective mean observed yields. There was a wide range in predictive performance across fields (Figure 7), particularly in *r* and LCCC while RMSE is more consistent, demonstrating that similar magnitudes of prediction error can occur despite substantial differences in the model's ability to capture within-field variability.

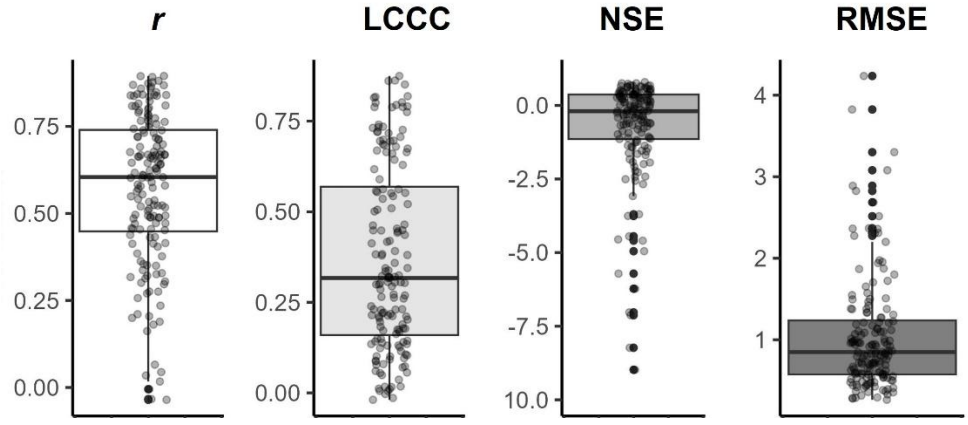


Figure 7. Boxplots of Experiment 2 prediction quality calculated for each field (n = 160), including Pearson's correlation coefficient (r), Lin's concordance correlation coefficient (LCCC), Nash-Sutclife efficiency (NSE) and root-mean-squared error (RMSE, t ha^{-1}).

Figure 8 showcases two fields. Field A (2015) is the best predicted field (r = 0.89, LCCC = 0.87, RMSE = 0.43 t ha^{-1}, NSE = 0.79) with predictions capturing clear spatial patterns in the correct scale. The histogram of yield values shows the predictions do not reach the full range of observed values, but the distribution is fairly representative of the observed data. On the other hand, Field B (2016) (r = 0.20, LCCC = 0.12, RMSE = 1.1 t ha^{-1}, NSE = -0.05) is closer to the median in terms of prediction quality

(top 40th percentile in terms of NSE). The observable spatial patterns are not captured by the model predictions, and the range of predicted values are much more narrow compared to the range of observed values.

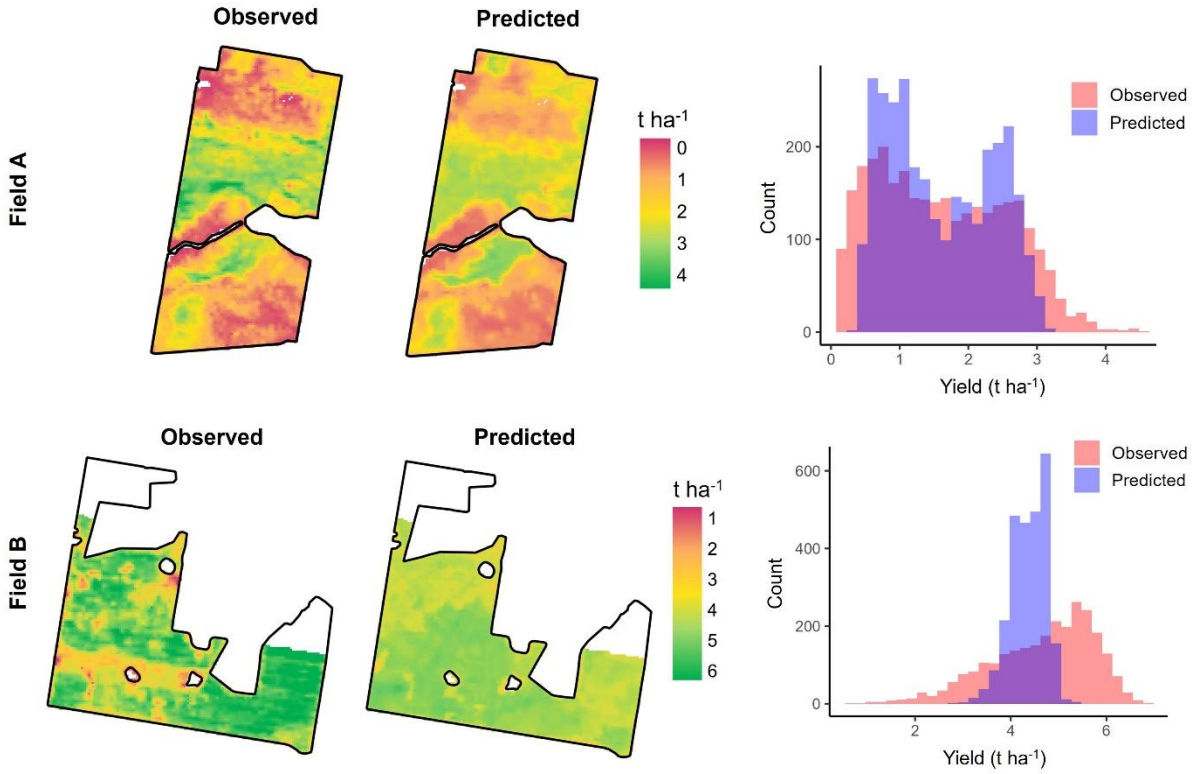


Figure 8. Observed and predicted yield maps (t ha^{-1}) and a histogram of values for the two example fields (Field A, Field B) for Experiment 2. No coordinates or scale are provided to maintain grower privacy.

Variable importance was calculated from the entire dataset, as with leave-one-field-out cross-validation (LOFOCV), most of the training data is common in each iteration, i.e. only one field is left out. The four most important variables were seasonal rainfall, growing degree days (GDD), mean September NDMI, elevation, and mean September EVI (Figure 9). It could be argued that seasonal rainfall and GDD have the most impact on between-season variability, which is where most of the variability is, hence they are being selected. However, the low spatial resolution of weather data (field-averaged) could potentially be responsible for low spatial variability in predictions in each field, given repeated use of the same values.

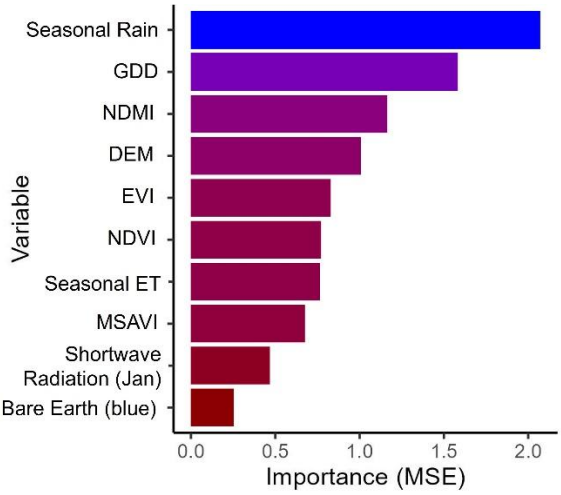


Figure 9. The top ten most important variables (permutation importance) obtained from the random forest model built on the entire dataset and full set of predictors (Experiment 1, Experiment 2).

Experiment 3: predictor resolution

The dataset was simplified to the 17 predictors, removing seasonal rainfall and GDD. Figure 10 (right panel) shows the median field statistics (r = 0.55, LCCC = 0.32, RMSE = 0.92 t ha^{-1}, NSE = -0.59), with a lower Pearson's correlation, LCCC, NSE and higher RMSE than the full model in Experiment 2 (r = 0.60, LCCC = 0.32, RMSE = 0.85 t ha^{-1}, NSE = -0.29) (Figure 6, right panel). With the reduced predictors, however, 70.6% of fields had a negative NSE, as compared to 59.4% with the full model.

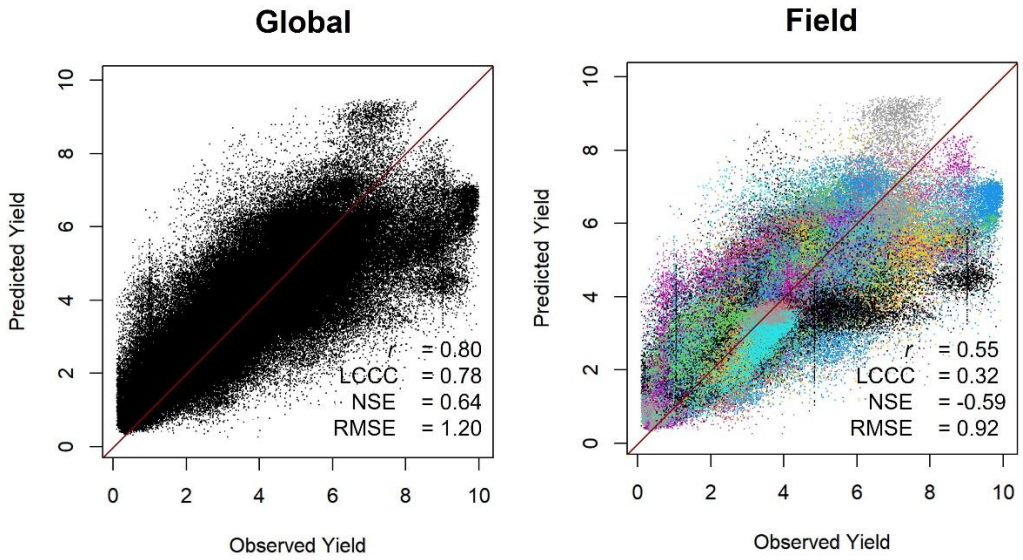


Figure 10. Scatterplot of wheat (t ha^{-1}) yield observations and predictions (Experiment 3, RF, LOFOCV, reduced predictors) with median field statistics, calculated for each field (coloured). A reduced dataset with no coarse-resolution weather predictors was used.

The order of importance is almost identical to that of the full model, except without weather variables. NDMI and DEM rose in rank from the third and fourth most important predictors (Figure 9) to the first and second most important predictors (Figure 11). The only predictor to gain relative importance was

the seasonal ET predictor, given it has information on seasonal water availability, which likely makes up for the removal of seasonal rainfall and growing degree days.

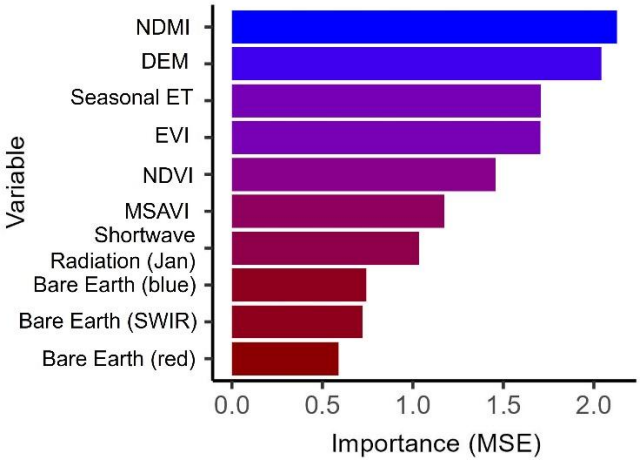


Figure 11. The top ten most important variables (permutation importance) built random forest model using the reduced set of predictors (Experiment 3).

Realistically, many studies would make a similar comparison and choose the model with reduced predictors by principle of parsimony during feature selection, although within-field performance is lower for the reduced model. This could remove predictors that add value at the within-field level, if the performance is only assessed globally. The issue also extends to model selection and hyperparameter tuning, where ultimately the final model is optimised to fit globally, rather than locally.

Experiment 4: model algorithm

All the experiments so far and many of the published studies discussed above have used RF or a tree-based machine learning model. One limitation of tree-based models is that they rely on rule-based splits, resulting in piecewise-constant predictions, thus small changes in input variables may not lead to changes in the predicted output. On the other hand, neural networks rely on continuous functions and hence may be more sensitive to small changes in input variables which are expected within-field. This experiment uses a multilayer perceptron (MLP) neural network model with LOFOCV.

The MLP model performs worse than the RF model with the same full set of 19 predictors (Experiment 2), despite hyperparameter tuning, for both global assessment and local assessment (Figure 12). In terms of median field goodness-of-fit, the MLP model (Figure 12, right panel) has a lower *r,* LCCC and NSE and higher RMSE than the counterpart RF model (r = 0.60, LCCC = 0.32, RMSE = 0.85 t ha^{-1}, NSE = -0.29) (Figure 6, right panel). For the MLP model, approximately 77.5% of the 160 fields had a negative NSE, as compared to 59.4% with the RF model. Even so, there is greater prediction variability in the MLP model predictions compared to the RF models, both with (Experiment 2) and without coarse resolution weather variables (Experiment 3).

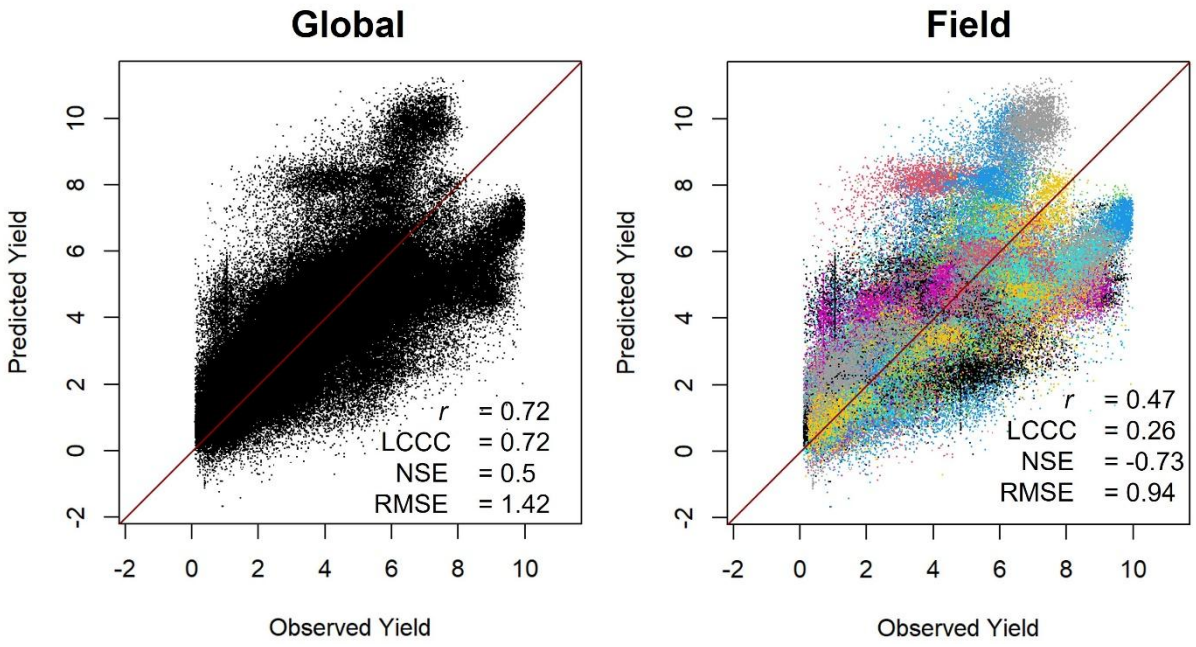


Figure 12. Scatterplot of wheat yield (t ha^{-1}) observations and predictions (Experiment 4, MLP, LOFOCV, full predictors) with median field statistics, calculated for each field (coloured). Leave-one-field-out cross-validation was used to train a multilayer perceptron model with 19 predictors.

Random effects analysis of variance

Variance in yield was examined across different spatial scales to determine how variability was partitioned among and within fields. Table 2 partitions the variance of observed wheat yield into several nested spatio-temporal scales. The first level or variance split is between-season variance, the next is the variance between the four farms (within each season), then the variance between each field (within each farm and season), and lastly the variance within each field. Most of the observed variance occurs between-season (43.8%) and between-fields from the same farm and season (27.3%). The within-field variance accounts for a relatively small 14.1% of total variance in the observed yield (Table 2).

Table 2. Random effects analysis of variance for several nested spatio-temporal scales (levels) of observed wheat yield. The proportion (%) of total observed variance is also included.

Levels	Observed Variance	Proportion Observed (%)
Between-season	2.18	43.8
Between-farm	0.74	14.8
Between-field	1.36	27.3
Within-field	0.70	14.1
Total	4.98	100

Table 3 compares the nested spatio-temporal variance in the observed yield, as well as the remaining error across the experiments. Note with Experiment 1, as a model is fit for each field, there is no between-season, between-farm or between-field variance. At the within-field level, the model still fails to explain 28.6% of observed variance (0.20 error remaining of 0.70 observed variance), despite likely overfitting (Table 3).

In Experiment 2, there is no error remaining for between-season and between-farm which infers the model and predictions have accounted for all variance at these spatio-temporal scales (Table 4). The model, however, captures a relatively small proportion of between-field and within-field variance, failing to explain 80.1% (1.09 error remaining of 1.36 observed variance) and 62.9% (0.44 error remaining of 0.70 observed variance) of variance respectively (Table 3). Thus, the resulting model captures the main and obvious trends in observed yield across seasons and farms (58.6%), but ignores smaller, local trends between and within-field (41.4%) (Table 2). This coincides with the most important predictors being seasonal rainfall and temperature (GDD) (Figure 9), which do not vary spatially across a field, and target between-season variability. As these are the larger sources of variation, assessing predictions globally do capture the major sources of variability.

When the coarse-resolution weather predictors are removed in Experiment 3, the other predictors (NDMI and seasonal ET) can compensate with just 1.8% error remaining between-season (0.04 error remaining of 2.18 observed) and 2.7% error between-farm (0.02 error remaining of 0.74 observed) (Table 3). Compared to Experiment 2 (1.53), Experiment 3 has a slightly higher amount of error remaining collectively at the between- and within-field scales (1.60) which means there is more error in the resulting prediction at finer spatial scales (Table 3). The results indicate the coarse-resolution predictors are not the cause of low prediction variability at these spatial scales.

Experiment 4 uses the full range of predictors in an MLP model. Like Experiment 2 (full predictors, RF model), the between-season and between-farm variability is completely explained by the model and predictors, but most of the between-field and within-field variability is not. There is more variance remaining between-field for Experiment 4 (1.42) compared to the observed yield (1.36) which means the more complex MLP model adds additional noise (Table 3). Within-field, the MLP model does not explain 92.9% of the variance in the observed yield (0.66 error remaining of 0.70 observed variance) (Table 3), which is much lower than Experiment 2 (62%). Thus, the problem of poor within-field predictive variability persists across data-driven model algorithms.

These analyses show that when applied in more realistic scenarios (e.g. predicting neighbouring fields in Experiments 2, 3 and 4) the models are performing well below what the predictors are capable of (i.e. Experiment 1).

Table 3. Variance for several nested spatio-temporal scales obtained from a random effects analysis of variance. The 'error' for the experiments is the variance of the difference between observed and predicted values. Total variance remaining (%) is calculated for each Experiments 2-4 as the proportion of total observed variance (4.98), and Experiment 1 as within-field observed variance (0.70).

	Observed	**Experiment 1 Error**	**Experiment 2 Error**	**Experiment 3 Error**	**Experiment 4 Error**
Between-season	2.18	-	0.00	0.04	0.00
Between-farm	0.74	-	0.00	0.02	0.00
Between-field	1.36	-	1.09	1.01	1.42
Within-field	0.70	0.20	0.44	0.59	0.66
Total Variance	**4.98**	**0.20**	**1.53**	**1.66**	**2.08**
Total Variance Remaining (%)	**100**	**28.6**	**30.7**	**33.3**	**41.8**

Discussion

Using a wheat yield dataset, this study demonstrated that assessing predictions globally can mask local predictive performance particularly if there are spatial and/or temporal scales. Through a series of four experiments, the results showed 1) the RF model and collated predictors had sufficient variability and complexity to represent 71.4% of within-field variability when a per-field model was used, 2) the within-field variability explained declined to 37.1% when the LOFOCV validation approach was applied, 3) removal of coarse resolution predictors did not improve within-field variability in predictions, and 4) poor within-field predictions persisted regardless of the data-driven model algorithm used (RF and MLP). The REANOVA analysis of model error showed the models addressed the scales with the most variance, with negligible variance remaining at the between-season and farm scales which covered 58% of total variance. Consequently, global goodness-of-fit can appear relatively strong although local predictive performance is poor.

To understand how the model performs within a subgroup, in this case a field, individual subgroups need to be assessed. The issue extends beyond validation – feature selection, model selection, and benchmarking or comparison with other studies is all compromised when global statistics are used. The purpose of this paper is not to criticise previous work, but to highlight how widespread the issue is as well as encourage awareness in the community and better practice moving forward. A key reason for validation is to provide end users with a realistic sense of how good or bad a model or predictions are at the resolution and scales it is intended to be used. It is unlikely an end user of a spatial product will have a thorough understanding of how the model was created and shortfalls of validation approaches, much less independent data to verify predictions. From farmers to industry to government, decisions made from inaccurate data have economic, environmental, and cultural consequences.

In recent years, there has been a push for better model training and validation for spatio-temporal data. Many studies have advocated for consideration of spatial autocorrelation and using spatial validation techniques (Kattenborn *et al.*, 2022; Koldasbayeva *et al.*, 2024; Ploton *et al.*, 2020). In soil mapping, further emphasis has been placed on validation of spatial patterns themselves. Rossiter and Poggio (2025) make a strong case for DSM studies to move away from assessing accuracy of point predictions towards general spatial patterns. An example of this occurs in the validation of SoilGrids 2.0, where Poggio *et al.* (2021) uses qualitative checks as a form of ground-truthing, such as "the texture components correctly identify the siltier deltas (e.g. Yellow-Yangtze, Ganges-Brahmaputra)".

Even so, the suggestions to explore local validation would likely only reveal how poorly models are performing, rather than address the cause. An area of concern that needs further investigation are model loss functions. Most data-driven models, including the RF and MLP models used in this study, optimise the fit using statistics (e.g. mean squared error, sum of squared errors) to measure the goodness-of-fit during model fitting. As models are fit iteratively, the loss function is used to determine the ultimate model parameters, or in the case of RF, decision tree splits. Loss functions, however, are also applied globally across the training dataset and thus finer spatio-temporal patterns could also be masked. The default loss function to fit each decision tree in a RF model is the mean squared error, and the same was used for the MLP model in the current study. Given the residual is squared, the mean squared error penalises large errors more heavily, thus favouring predictions closer to the conditional mean. This is a common criticism of machine learning models, frequently observed in crop modelling where high values are underpredicted and low values are overpredicted (Han *et al.,* 2022b; Pang *et al.*, 2022). It is possible that other loss functions (e.g. mean absolute error, Huber-loss) which penalise large outliers to a lesser extent could reduce the impact on predictions.

However, the greater issue may lie in modelling data across different spatial and temporal scales without adequately accounting for their structure. To minimise the structured nature of the data, one option is

to model with a small study area (e.g. a farm) for a single season. This reduces the spatial and temporal structure within the training dataset, and hence what is passed to the model. This has and is still common in practice, as studies are constrained by limited timeframes and funds. However, as shown in Richetti *et al.* (2023) and Filippi *et al.* (2025), these models are not particularly generalisable or transferable in areas or seasons outside the training data. Even with the relatively small scope shown in this study – i.e. several seasons of data from neighbouring farms within a region – the spatial variability between and within-field could not be addressed by the global loss function.

Another option is to consider models that can explicitly account for the structures within the data. For data with high spatial density such as crop yield, there is the opportunity to frame the data as an image-based problem to explicitly learn spatial patterns and local context. U-Net (Ronneberger *et al.*, 2015), for example, is a fully convolutional network whose encoder-decoder architecture and skip connections are designed to retain fine-scale spatial information while learning broader spatial features. Spatial structure could additionally be incorporated into the model objective by adding a spatial loss term that penalises differences between the spatial dependence of observed and predicted values, for example differences in their variograms or local spatial autocorrelation. Even so, increasing the complexity of the loss function reduces the likelihood of model convergence and will likely make optimisation more difficult.

Given many datasets in environmental and agricultural systems are sampled and not spatially 'complete' (e.g. soil pH, soil moisture, crop phenology), not all data is suitable for image-based models. Geostatistical models such as kriging, or hierarchical models, also offer an alternative avenue as they can explicitly account for spatial autocorrelation between observations, which may otherwise leave spatially structured variation unexplained. While the model would not be spatially transferable (e.g. a new location), this may not necessarily be the target of the model, and the results could be transferable across different seasons (Jiang, 2018).

The within-field variance accounted for just 14.1% of total observed variability, which was small relative to the between-season (43.8%), between-field (27.3%) and between-farm (14.9%) components. Given that this study only covers a few farms from one region, it is not known whether these values are similar for the rest of Australia, in other parts of the world, or for other crop types. Poor within-field yield predictions could be less prominent elsewhere, where yield is more uniform across fields, farms or regions, leading to a larger proportion of the variance to be within fields. Model training could then be driven more by within-field variation, and global performance metrics would reflect within-field performance more closely. Future work could explore this further using the REANOVA approach shown here, which requires only yield data. The analysis could also be done for similar types of nested spatio-temporal data (e.g. soil properties, streamflow).

Overall, this study makes it clear that there are issues with the existing modelling and validation approaches for agricultural data, as the inherent spatial and/or temporal structures in these datasets are incompatible with 'global' loss functions. Alternative modelling approaches, such as image-based models and spatially-aware loss functions, involve trade-offs between data requirements, spatial transferability and model complexity that may limit their application across agricultural datasets. A collective effort across the research community is needed to investigate potential solutions and investigate their adoption.

Conclusion

This study highlighted greater consideration for model performance and validation is needed within the research community. Assessing predictions globally when there are spatial or spatio-temporal

subgroups can easily mask local predictive performance. Of greater concern is that similar limitations could extend to loss functions in data-driven models which could contribute to low variability within intrinsic subgroups. This study demonstrates several ways to assess within-subgroup performance, including scatterplots coloured by subgroup, boxplots of per-subgroup performance metrics, per-subgroup NSE as a benchmark against the subgroup mean, and variance partitioning across nested spatio-temporal groups using REANOVA.

With more data, models and tools available by the day, it is vital for modellers to pause and consider how good the outputs truly are: just because a prediction can be made, doesn't mean it is useful. It is not realistic to expect end users to independently validate spatial models and products, and the onus should be on model developers to ensure that the reported predictive accuracy and findings reflect performance at the spatial and temporal scales at which predictions are provided and intended to be used.

Acknowledgements

The authors thank the growers for sharing their data. This work was funded by the University of Sydney, the Thomas Lawrance Pawlett Scholarship through the Sydney Institute of Agriculture, and the Grains Research and Development Corporation under the "Precision ag analytics for frost and heat management" project (UOS2205-006RTX).

Statements and Declarations

The authors report no competing interests.

References

Adhinata, F. D., & Sumiharto, R. (2024). A comprehensive survey on weed and crop classification using machine learning and deep learning. *Artificial intelligence in agriculture*, *13*, 45-63.

Aggarwal, R., & Ranganathan, P. (2016). Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in clinical research*, *7*(4), 187-190.

Al-Shammari, D., Fuentes, I., Whelan, B. M., Wang, C., Filippi, P., & Bishop, T. F. (2024). Combining Sentinel 1, Sentinel 2 and MODIS data for major winter crop type classification over the Murray Darling Basin in Australia. *Remote Sensing Applications: Society and Environment*, *34*, 101200.

Australian Bureau of Statistics (ABS). (2024). *Australian Agriculture: Broadacre Crops*. ABS. https://www.abs.gov.au/statistics/industry/agriculture/australian-agriculture-broadacre-crops/2022-23.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *convergence*, *12*(1), 2.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors*, *21*(11), 3758.

Bishop, T. F. A., Horta, A., & Karunaratne, S. B. (2015). Validation of digital soil maps at different spatial supports. *Geoderma*, *241*, 238-249.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.

Dayal, K., Brown, J. N., Waldner, F., Lawes, R., Hochman, Z., Donohue, R., ... & Chen, Y. (2020). Climate drivers provide valuable insights into late season prediction of Australian wheat yield. *Agricultural and Forest Meteorology*, *295*, 108202.

Feng, P., Wang, B., Liu, D. L., Xing, H., Ji, F., Macadam, I., ... & Yu, Q. (2018). Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern Australia. *Climatic change*, *147*(3), 555-569.

Filippi, P., Han, S. Y., & Bishop, T. F. (2025). On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. *Precision Agriculture*, *26*(1), 8.

Gallant, John; Dowling, Trevor; & Austin, Jenet (2012): Multi-resolution Valley Bottom Flatness (MrVBF). v3. CSIRO. Data Collection. https://doi.org/10.4225/08/5701C885AB4FE

Gallant, J., Austin, J., & Van Niel, T. (2016). Mean monthly total shortwave radiation on a sloping surface modelled using the 1′ DEM-S–3′ mosaic. *Commonwealth Scientific and Industrial Research Organisation (CSIRO)*.

Grundy, M. J., Rossel, R. V., Searle, R. D., Wilson, P. L., Chen, C., & Gregory, L. J. (2015). Soil and landscape grid of Australia. *Soil Research*, *53*(8), 835-844.

Haan, S., Harianto, J., Butterworth, N., & Bishop, T. (2023). Geodata-Harvester: A Python package to jumpstart geospatial data extraction and analysis. *Journal of Open Source Software*, *8*(89), 5205.

Han, S. Y., Filippi, P., Singh, K., Whelan, B. M., & Bishop, T. F. (2022a). Assessment of global, national and regional-level digital soil mapping products at different spatial supports. *European Journal of Soil Science*, *73*(5), e13300.

Han, S. Y., Bishop, T. F. A., & Filippi, P. (2022b). Data-driven, early-season forecasts of block sugarcane yield for precision agriculture. *Field Crops Research*, *276*, 108360.

Jiang, Z. (2018). A survey on spatial prediction methods. *IEEE transactions on knowledge and Data Engineering*, *31*(9), 1645-1664.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *5*, 100018.

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications*, *15*(1), 10700.

Kuhn, M., & Silge, J. (2022). *Tidy modeling with R: A framework for modeling in the tidyverse*. " O'Reilly Media, Inc.".

Lark, R. M. (2005). Exploring scale-dependent correlation of soil properties by nested sampling. *European journal of soil science*, *56*(3), 307-317.

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what?. *PloS one*, *12*(8), e0183250.

Malone, B. P., Searle, R., Stenson, M., Mcjannet, D., Zund, P., Dobarco, M. R., ... & Grundy, M. (2025). Update and expansion of the soil and landscape grid of Australia. *Geoderma*, *455*, 117226.

McPhee, M., Richetti, J., Croke, B., & Walmsley, B. (2024). Model evaluation: The misuse of statistical techniques when evaluating observations versus predictions. *Socio-Environmental Systems Modelling*, *6*, 18758-18758.

McVicar, T., Vleeshouwer, J., Van Niel, T., & Guerschman, J. (2022). Actual Evapotranspiration for Australia using CMRSET algorithm. *Terrestrial Ecosystem Research Network. Dataset*.

Minty, B., Franklin, R., Milligan, P., Richardson, M., & Wilford, J. (2009). The Radiometric Map of AustraliaFN1. *Exploration Geophysics*, *40*(4), 325-333.

Ngugi, H. N., Ezugwu, A. E., Akinyelu, A. A., & Abualigah, L. (2024). Revolutionizing crop disease detection with computational deep learning: a comprehensive review. *Environmental Monitoring and Assessment*, *196*(3), 302.

Orynbaikyzy, A., Gessner, U., & Conrad, C. (2019). Crop type classification using a combination of optical and radar remote sensing data: A review. *International journal of remote sensing*, *40*(17), 6553-6595.

Pang, A., Chang, M. W., & Chen, Y. (2022). Evaluation of random forests (RF) for regional and local-scale wheat yield prediction in southeast Australia. *Sensors*, *22*(3), 717.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, *7*(1), 217-240.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., ... & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, *11*(1), 4540.

Roßbach, P. (2018). Neural networks vs. random forests–does it always have to be deep learning. *Germany: Frankfurt School of Finance and Management*, 1-8.

Richetti, J., Diakogianis, F. I., Bender, A., Colaço, A. F., & Lawes, R. A. (2023). A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield. *Computers and Electronics in Agriculture*, *205*, 107642.

Richetti, J., Lawes, R. A., Whan, A., Gaydon, D. S., & Thorburn, P. J. (2024). How well does APSIM NextGen simulate wheat yields across Australia using gridded input data? Validating Continental-Scale Crop Model Simulations. *European Journal of Agronomy*, *158*, 127212.

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package 'nnet'. *R package version*, *7*(3-12), 700.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

Rossiter, D. G., & Poggio, L. (2025). Representing soil landscapes from digital soil mapping products–helping the map to speak for itself. *Soil*, *11*(2), 849-881.

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture*, *177*, 105709.

Wilford, J., & Roberts, D. (2020, May). Bare Earth-enhanced multi-spectral satellite imagery for mapping soil and exposed rock. In *EGU General Assembly Conference Abstracts* (p. 12371).

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of statistical software*, *77*, 1-17.

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, *122*, 2-10.

Van Zyl, J. J. (2001). The Shuttle Radar Topography Mission (SRTM): a breakthrough in remote sensing of topography. *Acta astronautica*, *48*(5-12), 559-565.

Victor, B., Nibali, A., & He, Z. (2024). A systematic review of the use of Deep Learning in Satellite Imagery for Agriculture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Xu, T., & Hutchinson, M. (2011). ANUCLIM version 6.1 user guide. *The Australian National University, Fenner School of Environment and Society, Canberra*, *90*.

Zhao, Y., Jiang, R., Brider, J., Chapman, S., & Potgieter, A. (2025). Characterizing wheat and barley growth and phenology using multi-spectral remote sensing for site-specific precision agriculture. *in silico Plants*, *7*(2), diaf013.