Agricultural and Forest Meteorology 303 (2021) 108369

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Block-level macadamia yield forecasting using spatio-temporal datasets

James Brinkhoff[*], Andrew J. Robson

Applied Agricultural Remote Sensing Centre, University of New England, Armidale 2350 NSW Australia

ARTICLE INFO

Keywords:
Macadamia
Remote sensing
Yield forecasting
Climate
Statistical models
Machine learning

ABSTRACT

Early crop yield forecasts provide valuable information for growers and industry to base decisions on. This work considers early forecasting of macadamia nut yield at the individual orchard block level with input variables derived from spatio-temporal datasets including remote sensing, weather and elevation. Yield data from 2012–2019, for 101 blocks belonging to 10 orchards, was obtained. We forecast yield on each test year from 2014–2019 using models trained on data from years prior to the test year. Forecasts are generated in January, for the coming harvest in March–September. A linear model using ridge regularized regression produced consistently good predictions compared with other machine learning algorithms including lasso, support vector regression and random forest. Adding meteorological variables offered little improvement over using only remote sensing variables. The 2019 forecast root mean square error at the block level was 0.8 t/ha, and mean absolute percentage error was 20.9%. When block level predictions were aggregated across the multiple orchards per region, production prediction errors were between 0–15% from 2016–2019. The ridge regression model can be easily implemented in GIS platforms to deliver block-level yield forecast maps to end users.

Macadamias are a high value tree nut crop, native to Australia. They are now grown in many countries and the industry is experiencing rapid expansion as demand rises (Stephenson, 2005). In 2016, the major producers were Australia (25% of global production with 46,000 tonnes of nut-in-shell at 10% moisture), South Africa (23%), Kenya (15%) and Hawaii (9%) (Topp et al., 2019). China's production is rapidly increasing. The area of production in Australia in 2017 was estimated to be 23,000 hectares (Brinkhoff and Robson, 2020).

In Australia, most macadamias are harvested between March and September, depending on region and maturity. The mature nuts fall from the trees, and are gathered at regular intervals during the harvest period by finger-wheel harvesters (O'Hare et al., 2004). There is great interest in the industry to have early (January) and accurate forecasts of yield (Mayer et al., 2019). This allows the industry to forecast total crop, and therefore adjust their marketing strategies and logistics planning. Growers are interested in yield predictions as they will aid finance, insurance and logistic decisions. Yield prediction models may also provide information on the drivers of yield variability, and thus offer potential for optimizing yield if variables that can be managed are identified. Spatial yield analysis facilitates precision agriculture applications by adjusting management spatially to achieve optimal yields, considering local variation in tree health, soil, landscape and micro-climate (Felderhof and Gillieson, 2011; Johansen et al., 2020).

Analytic yield prediction methods can be grouped into process-based and statistical models (Lobell and Burke, 2010b). Process-based methods model factors such as light interception, photosynthesis, respiration, carbon assimilation and how this carbon is partitioned into non-harvested and harvested components of a crop (Marcelis et al., 1998). These offer deep insight into the biological and environmental factors driving yield variability. They can therefore be used to predict the impact of factors such as climate change on crop production, and to make predictions in new regions. However, due to the complexity and interaction of biological processes, and the possibility of unforeseen factors impacting these processes in new regions, they are difficult to parameterize and to obtain sufficiently accurate predictions for industry use. Some works use remote sensing data to improve the parameterization of physiological models (Maselli et al., 2012), however no comprehensive process-based model of macadamia trees is currently available, which necessitates using another approach.

On the other hand, statistical models describe yield as a function of combinations of input variables. Such models are parameterized based on historical observations (Zhang et al., 2019). Inputs to such a model may include meteorological variables, remote sensing, soil characteristics, elevation, gradient as well as variables derived from crop factors

[*] Corresponding author.
E-mail address: james.brinkhoff@une.edu.au (J. Brinkhoff).

https://doi.org/10.1016/j.agrformet.2021.108369
Received 4 September 2020; Received in revised form 17 December 2020; Accepted 8 February 2021
Available online 24 February 2021

such as variety, layout and management practices (fertilizer and water application, pruning, pollination). The detail of the physical processes leading to yield outcomes are not described by statistical models, however they may provide some insight into the input factors driving yield variation (Jin et al., 2020). They are typically based on linear methods (for example, ordinary least squares) or machine-learning approaches (for example tree-based models such as random forests (RF), support vector regression (SVR) or artificial neural networks (ANN)) (Zhang et al., 2019). One limitation is that they may fail to predict yield when the input factors, such as weather, fall outside of the range of conditions encountered in the historical data the model was trained on (Deines et al., 2020; Marcelis et al., 1998).

Much research has focused on forecasting yield of annual crops, such as maize (Kang et al., 2020), cotton (Filippi et al., 2020) and rice (Setiyono et al., 2019). Predictor variables for such models typically make use of in-season data. The yield of fruit and nut tree crops however, often have dependence on multiple years of factors such as weather due to their ability to store resources such as carbohydrates (Stephenson et al., 1989), and exhibit complex behavior such as biennial bearing (Huett, 2004). There are relatively few studies reporting tree crop yield forecasting compared with those focussing on annual crops (van Klompenburg et al., 2020).

The impact of particular climatic and management factors on macadamia yield have been the subject of some studies. For example Trochoulias and Lahav (1983) found optimum growth between 20 and 25 °C. Smit et al. (2020) showed that CO_2 assimilation was maximum with leaf-to-air vapour pressure deficits between 1–2.5 kPa. McFadyen et al. (2004) found macadamia yield increases with tree volume up to 43,500 m^3/ha (corresponding to light interception of 94%) and decreases with tree crowding above this value, though no decline in quality with crowding was observed. Stephenson et al. (2000) found optimum yields and quality are obtained at lower nitrogen rates, recommending 355 g nitrogen application in late autumn to early winter and that rainfall during harvest negatively impacted quality. Stephenson et al. (1986a) found that weather is not as important in describing yield variation as leaf nutrients and flushing characteristics and soil zinc levels. A model of 8 parameters (not including weather) was able to account for 58.2% of the observed yield variation, while a model with weather variables described 40.3% of the yield variation.

Currently, yield forecasts for Australian macadamia farms are provided by experienced agronomists and growers using information including weather, flower and nut counts. However, this method is subjective, and is limited in the amount of temporal and spatial (considering only selected trees within an orchard) information it can utilise.

At a larger scale, work on predicting yield over large regions of the macadamia industry in Australia using statistical models is ongoing (Mayer et al., 2019). These models take data on total production per region per year, and fit models using input variables including weather, market price, tree age and total area. Recent work has used satellite data to improve accuracies of the estimates of total planted area and tree age per region (Brinkhoff and Robson, 2020) to aid regional yield predictions.

The potential of using high-resolution remote sensing imagery to predict macadamia yield variability has been examined by Robson et al. 2017. Models were calibrated using total nut weight measured from trees selected from three vigour zones within each study block. While a positive relationship between tree vigour and yield was identified across a number of locations and seasons, the remotely sensed vegetation index that best predicted yield varied between orchard blocks. For each site and season, the optimal vegetation indices were able to describe between 69 and 86% of the spatial variation of yield. Similarly, Johansen et al. (2020) assessed macadamia tree condition using high-resolution imagery and showed that it is difficult to generalise a model developed from one location or variety to another. These studies did not aim to produce a yield forecast model, rather they analyzed spatial variability using high-resolution imagery.

Macadamia yields are highly variable from year-to-year (McFadyen et al., 2004), with suggested causes being climatic variation (Mayer et al., 2019) and carbohydrate cycling (Huett, 2004). However, the precise causes and methods of predicting and reducing this variation are still unknown, which necessitates accumulation of high quality yield data from many years and the use of controlled experiments to establish causal links between yield and management or environmental factors (Huett, 2004).

In this study, we aim to produce a macadamia yield forecast model at the orchard block level. The forecasts are produced in a timeframe useful for growers, at least two months before harvest begins. The predictor variables come from publicly available spatio-temporal datasets, so that predictions can be generalized to new orchards and areas. First, we investigated the importance of the possible yield predictor variables from remote sensing, meteorological and spatial datasets. Then, forecast models of a range of complexities were trained using historical block level data and machine learning approaches. Forecasts were assessed, before selecting and implementing a final model.

1. Study area

The study included 8 years of yield data (2012–2019) from 101 orchard blocks belonging to 10 orchards across 3 significant Australian macadamia growing regions. The locations are shown in Fig. 1 and encompass a range of climate conditions and orchard management strategies. We define a block as an area in an orchard that has had yield data recorded, typically with uniform management and plant year.

The growers supplied maps of their blocks, which were digitized. They also provided production data, with values reported as kilograms (kg) of nut-in-shell (NIS) at 10% moisture (O’Hare et al., 2004), per block and per year. A summary of data from each region is shown in Table 1. The range of planting dates spans 25 years for the Ballina blocks, 4 years for the Bundaberg blocks and 19 years for the Macksville blocks. In the Macksville region, trees were older on average, and yields

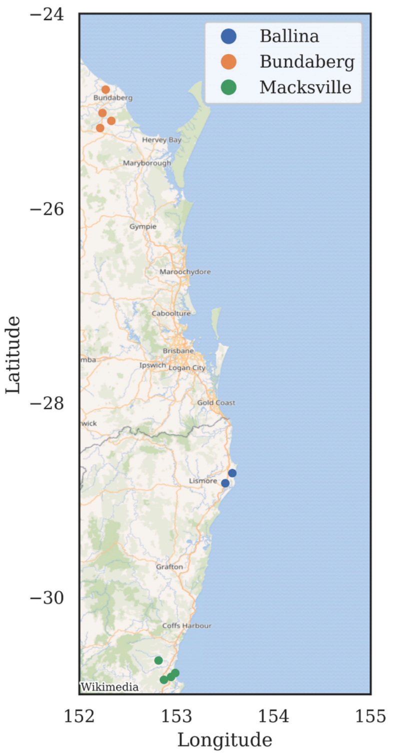


Fig. 1. Location of study orchards from each of the growing regions.

Table 1
Number of blocks in each region, data points (sum of blocks by years of recorded yield data), and the average block areas, planting years and yields.

Region	Blocks	Datapoints	Mean area (ha)	Rangeplant years	Mean yield (t/ha)
Ballina	39	125	4.2	1989–2014	2.9
Bundaberg	31	203	9.1	2004–2008	3.3
Macksville	31	147	7.0	1991–2010	2.1
All	101	475	7.1	2003 (mean)	2.8

lower compared with the northern Bundaberg region. Most Bundaberg orchards are irrigated, whereas Ballina and Macksville orchards mainly rely on rainfall.

The number of block yield data points per region and year is shown in Fig. 2b. There was no data recorded for Ballina before 2014, and no data for Macksville in 2014. One of the Ballina orchards suffered a hailstorm which affected 2018 yields, so that data was omitted. We have also omitted Ballina data from 2017, because of the severe rainstorms during harvest that washed much of the fallen nut away. Future work could include rolling updates to forecasts to capture adverse affects such as this during the long harvest period, however the current models are aimed at providing estimates of potential yield in January, before harvest starts. The minimum number of annual data points was 37 in 2012, and a maximum of 93 in 2019.

The wide spatio-temporal variation in yields between years, regions and blocks is shown in Fig. 2a. Block yields range from close to 0 t/ha to almost 7 t/ha. Some of the yield variability may be explained by tree age and climatic differences. Four key meteorological variables for each region are shown in Fig. 3. There is significant variation in rainfall from year-to-year across all regions. 2019 was in the midst of a severe drought with low rainfall and high temperatures and evapotranspiration. In general, Bundaberg experiences lower rainfall and higher evapotranspiration than other regions, which necessitates irrigation. There are significant differences in maximum and minimum temperatures between regions. Thus, Fig. 3 shows that our dataset encompasses a wide range of climatic conditions, due to both season and region.

Fig. 3e shows the green normalized difference vegetation index (GNDVI, Gitelson et al. 1996), computed from the remote sensing data. This shows the variability of tree reflectance from year-to-year, which may be caused by a combination of factors such as climate, management, tree age and canopy cover.

2. Methods

2.1. Spatio-temporal data

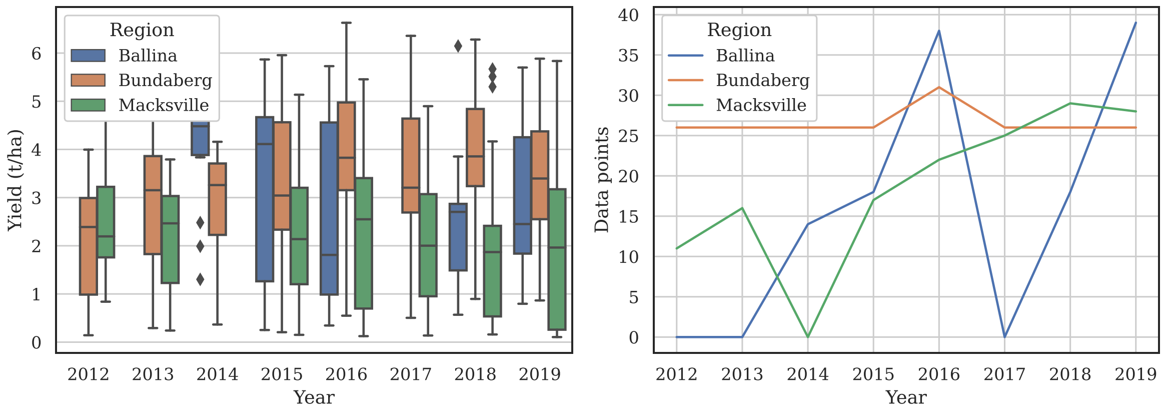
Spatio-temporal datasets were accessed and processed in Google Earth Engine (GEE) (Gorelick et al., 2017). Spatial data aggregated per block included:

- Smoothed elevation model (from Geoscience Australia, derived from LiDAR 5m grid, as available in GEE), from which the slope, north-facing slope and east-facing slope were calculated. The elevation difference between orchards was not significant, they were all close to sea-level.
- Tree planting year model. This was generated from Landsat data using the methods in (Brinkhoff, Robson, 2020), from which the median tree age per block per year was calculated. We used this model instead of grower-supplied tree ages so the yield forecast model could be applied to orchards for which we don't have access to grower-supplied tree age data.

Two spatio-temporal datasets were used, also available in GEE:

- Landsat 5, 7 and 8 tier 1 scenes, from 1988–2019 at 30 meter resolution. The scenes are available as surface reflectance in GEE (atmospherically corrected using the respective USGS procedures). The images in GEE also contain a cloud mask produced using USGS CFMASK, which was applied to all images. We investigated harmonizing the reflectances measured by the TM, ETM+ and OLI sensors using the equations proposed by Roy et al., 2016, however these did not result in improved model predictions, so we omitted this step. All normalized difference spectral indices (NDSIs) were calculated from combinations of input bands (blue=b, green=g, red=r, near infra-red=nir, shortwave infra-red 1=swir1 and shortwave infra-red 2=swir2). This yields 15 linearly-independent NDSIs. For example, ND(NIR,R)=(NIR-R)/(NIR+R) corresponds to NDVI.
- SILO (Jeffrey et al., 2001), a 5 km resolution daily meteorological dataset, interpolated from weather station observations. Variables include minimum and maximum temperatures (Tmin and Tmax), solar radiation (SolarRad), vapour pressure deficit (VPD), reference evapotranspiration (ETo) and rainfall (Rain). The SILO dataset is updated regularly in GEE, and so can be used to produce yield forecasts in the required timeframe.

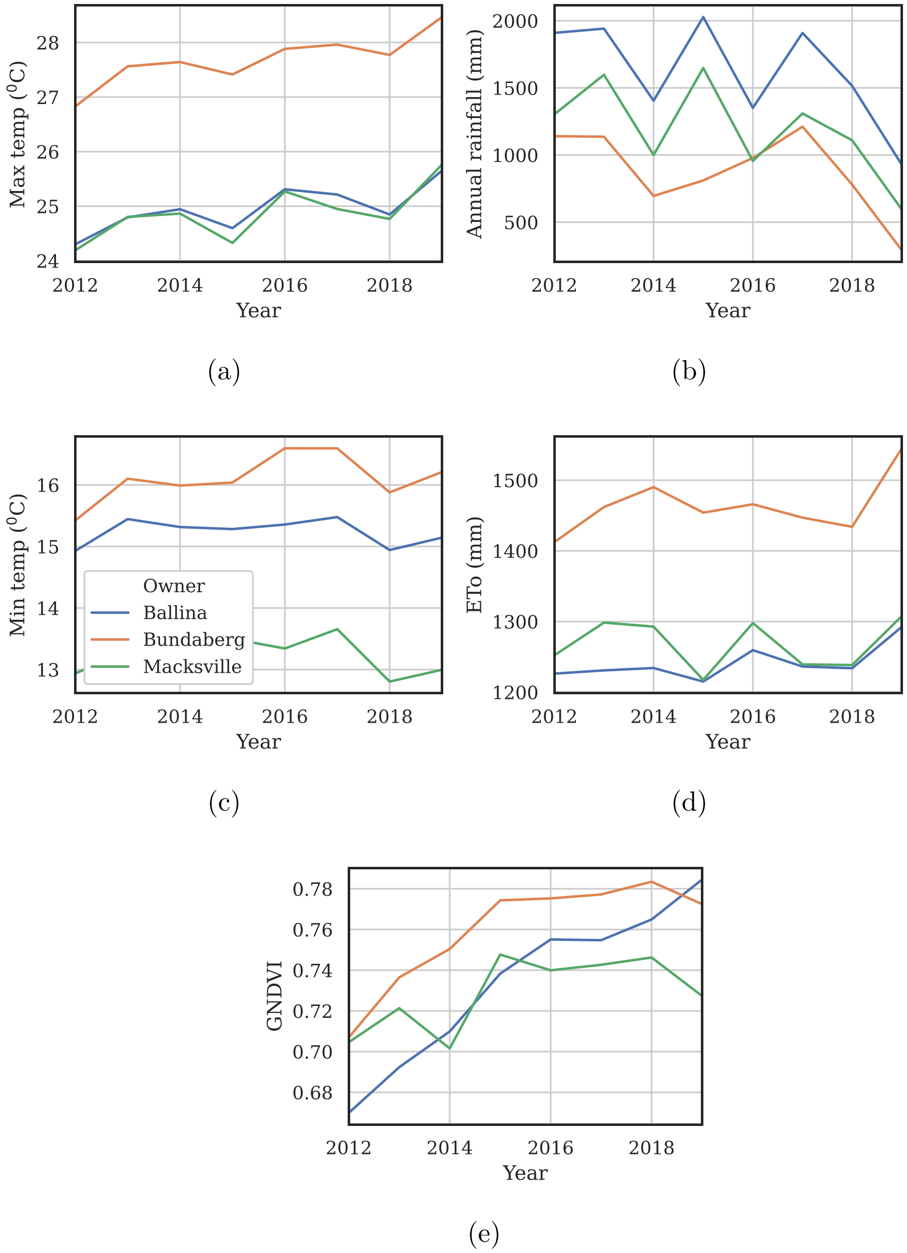
Landsat satellite data was selected because of its relatively high resolution compared with other data that covers a similar historical timeframe. SILO is used because of its ability to capture spatial variation



(a)

(b)

Fig. 2. (a) Distribution of all block yields per year and region. (b) The number of block yield data points per year and region.



(a)

(b)

(c)

(d)

(e)

Fig. 3. Annual aggregate meteorological and remote sensing variables for each region. (a) Average maximum temperature. (b) Total rainfall. (c) Average minimum temperature. (d) Total reference evapotranspiration. (e) Average GNDVI.

at a reasonable scale and its availability within GEE, which facilitates its use in generation of annual yield prediction maps.

Considering the different temporal frequency of the Landsat and SILO data, and the fact that cloud sometimes caused monthly Landsat mosaics to be totally masked over a number of blocks, we aggregated each of the spatio-temporal datasets at two temporal scales. Firstly, per year and secondly per quarter. Quarters were defined as s1=January-March, s2=April-June, s3=July-September and s4=October-December. Macadamia phenology varies with climate (Stephenson et al., 1986b), but s1 approximately encapsulates summer leaf flush, s2 flower initiation, s3 spring flush and peak flowering and s4 nut growth and beginning of oil accumulation (Schaffer and Andersen, 2018). Other temporal aggregations were also assessed before settling on quarterly aggregation. For example, using only three four-month periods produced less accurate forecasts. Using five two-month periods had the disadvantage that many blocks had no valid image data for some periods due to cloud.

The SILO variables were aggregated at the two temporal scales (annual and quarterly) using the mean operation. The Landsat variables were aggregated using the median operation, as this avoided the possibility of outlying pixels in the time-series stack (such as from unmasked clouds and shadows) skewing the aggregated values.

The squared values for all these spatio-temporal variables were calculated and added to the set of predictor variables, as shown in Fig. 4. Including these nonlinear terms improved the yield forecasts.

The spatial mean per block and per year of each of the variables was computed. The resulting table was then widened, so that each row included the aggregated spatio-temporal values for the two years previous to the yield prediction year (y1 and y2). This was merged with the tables of the recorded yields for the current yield year (y0) and block areas. The resulting table was used in the training and testing of the yield prediction models.

Importantly, the models are true forecast models in that yield in the current year is forecast without using any data from the current year. It is based only on variables aggregated from two previous years (y1 and

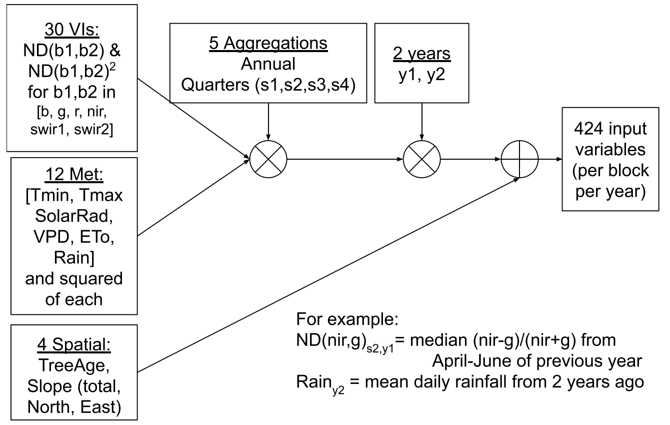


Fig. 4. Model variables created from spatial and spatio-temporal datasets. Abbreviations are defined in Section 2.1.

y2). Data from the two previous years was used as macadamia yield depends on previous management and climatic conditions. For example, topping (pruning the tops of trees) reduces yield and can take several years to recover from, and carbohydrate storage from previous years can affect current yield (Huett (2004)).

Fig. 4 shows the input variables and the temporal aggregations. The result was 424 variables in total, consisting of 4 spatial variables, 300 variables derived from Landsat observations and 120 from the SILO data.

2.2. Modeling methods

2.2.1. Model evaluation and optimization metrics

The actual macadamia production (P) in tonnes (t) for a given area is defined as:

$$P = \sum_{n=1}^{N} A_n Y_n \tag{1}$$

where A_n is the area of individual units, such as pixels, blocks, farms or regions, measured in hectares (ha). Y_n is the yield in tonnes/ha for each of those units. The goal is to find a predictive model f given input variables X (from data prior to the yield prediction year) to estimate the yield:

$$\widehat{Y_n} = f(X) \tag{2}$$

If the yield is described by a linear combination of the input variables, the relationship can be written as:

$$\widehat{Y_n} = \widehat{\beta_0} + \sum_{j=1}^{p} \widehat{\beta_j} X_j \tag{3}$$

where $\widehat{\beta_0}$ is the intercept, X_j are the p predictor variables, and $\widehat{\beta_j}$ are the fitted coefficients. In our case, $p <= 424$. Similarly, the total predicted production $\widehat{P}$ over a number of units (for example, orchard blocks), can be found by substituting $\widehat{Y_n}$ in (1).

To examine the degree to which each of the predictor variables can explain the variation in yields, we used the coefficient of determination R^2, defined as the square of Pearson's r.

We use a number of accuracy/error metrics to compare, assess and select forecast models. Lin's Concordance Correlation Coefficient (LCCC) (Lin, 1989) assesses the degree to which predicted verses observed yields lie along the 1:1 line, and as such is a useful metric to compare model performance between different studies and crops. It is defined as:

$$LCCC = \frac{2 s_{\widehat{Y}Y}}{s^2_{\widehat{Y}} + s^2_Y + \left(\overline{\widehat{Y}} - \overline{Y}\right)^2} \tag{4}$$

where $s_{\widehat{Y}Y}$ is the covariance between predicted and measured yields, $s^2_{\widehat{Y}}$ and s^2_Y are the variances of the predicted and measured yields respectively, and $\overline{\widehat{Y}}$ and $\overline{Y}$ are the means of the predicted and measured yields.

The mean absolute percentage error gives an easily interpretable assessment of average prediction accuracy relative to average yields (note, we use the weighted definition of MAPE throughout):

$$MAPE = \frac{1}{N\overline{Y}} \sum_{n=1}^{N} \left| Y_n - \widehat{Y}_n \right| \times 100 \tag{5}$$

The root-mean squared error penalizes larger errors more than mean absolute error, and is therefore used to select between models and as the scoring metric for optimizing model tuning parameters:

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(Y_n - \widehat{Y}_n \right)^2} \tag{6}$$

2.2.2. Cross validation and model testing

The model selection, training and testing procedure is illustrated in Fig. 5. Grid-search cross-validation (CV) in the SciKit-Learn Python package (Pedregosa et al., 2011) was used to select model tuning parameters, such as α of lasso and ridge regression, or the number of trees for random forest. We note that the number of predictor variables (424) is similar to the total number of samples available for training models ($<$ 475, Table 1), making proper cross validation procedures coupled with machine learning algorithms that are able to deal with this $p \approx n$ scenario crucial to avoid over-fitting to training data.

Cross validation was performed using the leave-one-group-out method of SciKit-Learn, with the groups split by year, which we call leave-one-year-out (LOYO). This LOYO CV method produces a model that generalizes to an unseen year better than randomly splitting data into training and validation sets. The latter method often results in an overfit model, because the model tuning parameters are optimized for predictions for the unrealistic case of data from the test year being available for training (Brinkhoff et al., 2019). For each test year, only data from years prior to the test year were used to train the models. So

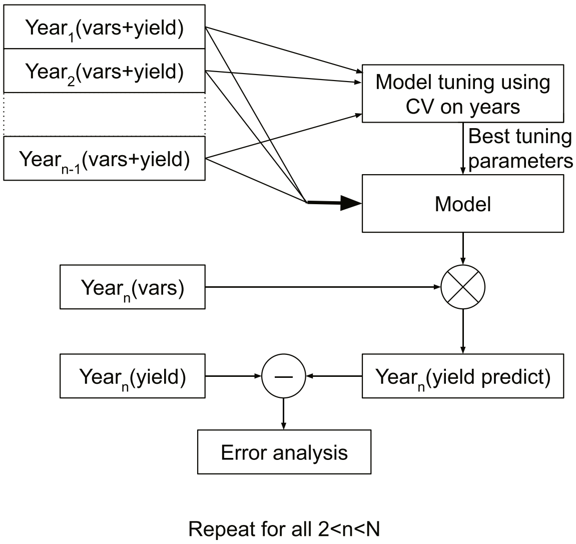


Fig. 5. Model training and testing process. Leave-one-year-out cross validation was used to optimize model tuning parameters, where the groups are years. Predictions for a given test year (n) only use model training data from previous years.

for example, 2014 predictions used cross validation on 2012 and 2013 data (2 groups). 2019 predictions used cross validation on 2012–2018 (7 groups).

After the CV procedure selected the optimized tuning parameters, the final model for each test year was fit using all training data. Finally, yield predictions were generated for the unseen test year, which were then compared with the measured yields. This process was repeated for all test years from 2014–2019.

2.2.3. Model algorithms

We compared a number of algorithms to provide inference and prediction of macadamia yield. These algorithms are available in the Scikit-Learn (Pedregosa et al., 2011) and Statsmodels (Seabold and Perktold, 2010) packages for Python:

- Ordinary least squares with forward-backward (hybrid) selection. At each step in the variable selection process, the variable that provides the biggest decrease in the Bayesian Information Criterion (BIC) was chosen, and then if removal of a previously-selected variable further decreases BIC, it was removed. BIC results in a simpler model than the Akaike Information Criterion (AIC), and also has the property that if a large number of samples are available, the process will select the correct model, making it useful for inference (Hastie et al., 2009).
- Ridge regression, which uses the L2 penalty to shrink the regression coefficients, without shrinking them to 0, thus retaining all the coefficients. The α parameter was tuned using GridSearch CV in Scikit-Learn (Pedregosa et al., 2011), which was also used to optimize tuning parameters for the following algorithms. Larger α shrinks coefficients more, leading to increased bias and reduced variance and vice versa (Hastie et al., 2009).
- Lasso, which uses the L1 penalty to shrink coefficients, some of which will become zero, and thus provides variable selection resulting in a more compact model than Ridge regression. The α parameter was tuned.
- Random forest (RF), a nonlinear method. The tuning parameters optimized were n_estimators, max_depth, min_samples_split, min_samples_leaf and max_features.
- Support vector regression (SVR), using the nonlinear radial basis function kernel, with tuning parameters cost and gamma.

Overfitting is avoided by tuning the algorithm parameters (such as α for ridge) using leave-one-year-out cross validation as noted above.

3. Results

We first examined which predictor variables best explain the variation in the observed yields (inference) in Section 3.1. We then built predictive forecast models, described in Section 3.2.

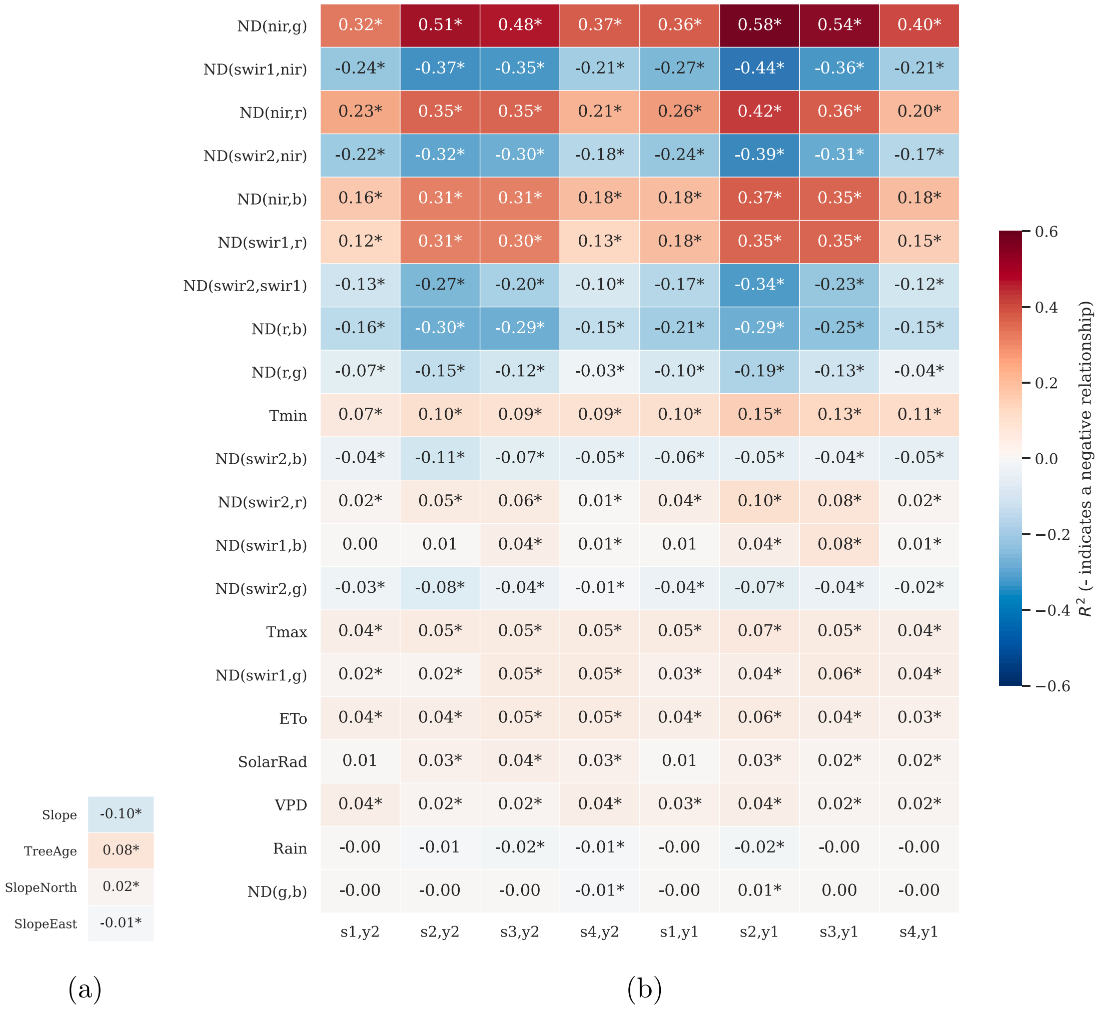


Fig. 6. Coefficient of determination between yield and each of the predictor variables, sorted by the strength of the correlation. Significant relationships at $p < 0.05$ are indicated by a *. Negative relationships are indicated with a - sign.

3.1. Inference: important predictors

3.1.1. Correlation analysis

Correlation analysis was performed, to determine which variables are most related with yield. The results for the quarterly aggregated spatio-temporal and spatial variables (excluding the squared variables for brevity) are shown in Fig. 6. We observe:

- The best predictor of macadamia yield at the block level was $\mathrm{ND(nir, g)}_{\mathrm{s2,y1}}$, otherwise known as the green normalized difference vegetation index (GNDVI), measured during late autumn-winter of the year preceding the harvest year, with $R^2 = 0.58$. This index is sensitive to chlorophyll concentration, and has a wider dynamic range than the normalized difference vegetation index (NDVI) Gitelson et al. (1996). It has been found to be a good predictor of yield in other crops, for example sugar cane in Rahman and Robson (2020).
- The next best predictors were ND(swir1,nir)= - NDWI (normalized difference water index) and ND(nir,r)=NDVI (normalized difference vegetation index) (Gao, 1996).
- Meteorological variables were less important than the remote-sensing variables. The most important among these are the minimum temperature during winter, which is positively correlated with yield, with $R^2 = 0.15$, followed by maximum temperature and evapotranspiration.
- Of the spatial variables, land slope was negatively correlated with yield. However, slope towards the north is positively correlated with yield. Tree age is positively correlated with yield.

3.1.2. The single best predictor

The correlation analysis above revealed the most important variable is $\mathrm{ND(nir, g)}_{\mathrm{s2,y1}}$ (winter GNDVI). Further investigation revealed the square of this variable is even more correlated with yield. We therefore performed a linear regression against these two variables. The coefficients and intercept were significant at $p < 0.001$, with the equation being:

$$Y_p = 82.3 \times \mathrm{ND(nir, g)}_{\mathrm{s2,y1}}{}^2 - 99.4 \times \mathrm{ND(nir, g)}_{\mathrm{s2,y1}} + 30.2 \tag{7}$$

This can be approximately factorised into:

$$Y_p = \beta\left(\mathrm{ND(nir, g)}_{\mathrm{s2,y1}} - \gamma\right)^2 = \beta \times \mathrm{GNDVIN} \tag{8}$$

Interestingly, γ is close to the minimum $\mathrm{ND(nir, g)}_{\mathrm{s2,y1}}$ in the dataset, which is somewhat expected, given that this value corresponds to a yield close to zero. We selected $\gamma = 0.56$, which minimized the intercept of the regression of (8). As shown in Fig. 7, the ordinary least squares (OLS) regression using only the linear term $\mathrm{ND(nir, g)}_{\mathrm{s2,y1}}$ gave $R^2 = 0.59$, whereas the regression using the transformed variable GNDVIN gave an improved fit to the data with $R^2 = 0.63$.

We evaluated how stable this predictor is with respect to region and year. When each region was analyzed separately, R^2 varied between 0.56 (Macksville) and 0.79 (Ballina). When each year from 2014–2019 was analyzed separately, R^2 varied between 0.64 (2015) and 0.82 (2014). This demonstrates GNDVIN is a good predictor of yield, describing both spatial and temporal variability.

We also computed the coefficient β for the linear regression between yield and GNDVIN (Eq. 8), for each year-region combination separately. The results are shown in Fig. 8. There is a significant relationship in all regions and years ($p < 0.001$). The slope of the relationship varies from year-to-year and between regions, which motivates searching for a more complex model that can describe more of the spatio-temporal yield variability.

3.1.3. Inferential models describing yield variability using OLS with forward-backward variable selection

To find the most important predictors using a multi-variable linear model, and to assess how much variation can be explained by such a model, we used OLS with forward-backward selection, trained on the

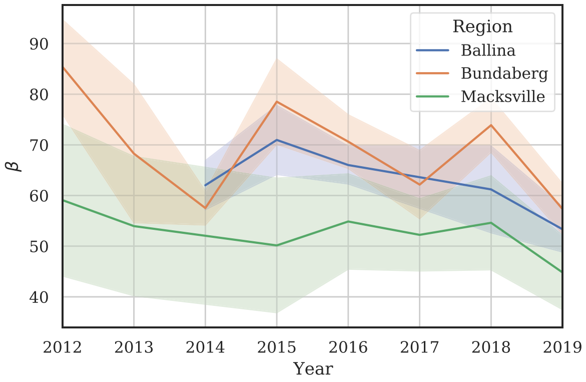


Fig. 8. Slope (β) of the relationship between yield (Y_p) and GNDVIN for each year and region, for the relationship $Y_p = \beta \times$ GNDVIN. The shaded area shows the 95% confidence interval of the coefficient.

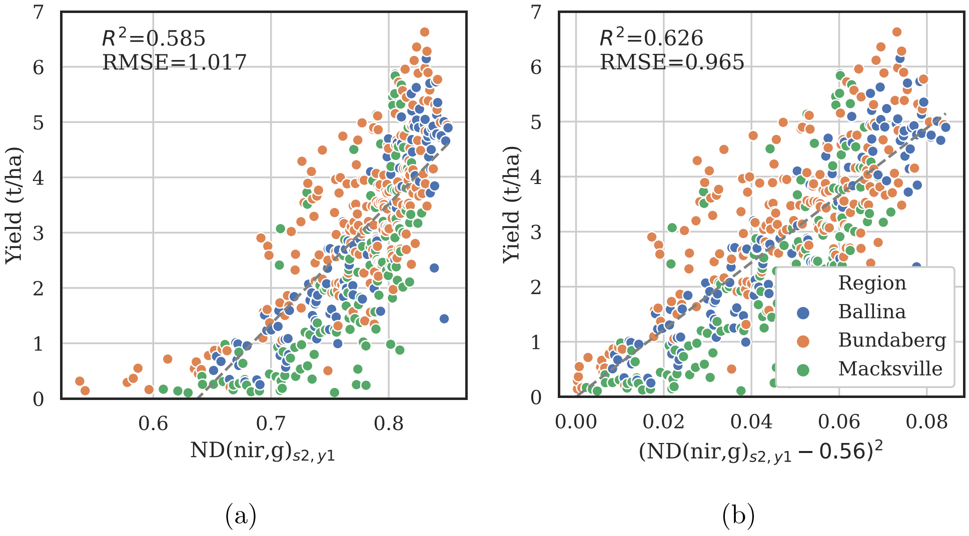


Fig. 7. Correlation between yield and (a) $\mathrm{ND(nir,g)}_{\mathrm{s2,y1}}$, and (b) the transformed predictor GNDVIN=$(\mathrm{ND(nir, g)}_{\mathrm{s2,y1}} - 0.56)^2$.

entire dataset. Note, the intention of this model is inference about important variables (similar to Stephenson et al. 1986a), rather than to forecast yield. Forecast models are covered below in Section 3.2.

Table 2
Variables selected by the OLS forward-backward selection method fit to all observations. Significance at $p < 0.05$ *, $p < 0.01$ ** and $p < 0.001$ ***. The variables are listed in the order they were selected (greatest reduction in AIC/BIC first). The sign of the coefficient is also indicated.

Met only	NDSIs only	All variables
$\text{-Tmin}_{s2,y1}{}^{2}$ *	$\text{ND}(\text{nir},\text{g})_{s2,y1}{}^{2}$ ***	$\text{ND}(\text{nir},\text{g})_{s2,y1}{}^{2}$ ***
$\text{-Tmax}_{s4,y1}$ ***	$\text{-ND}(\text{nir},\text{b})_{s2,y1}$ ***	$\text{-ND}(\text{nir},\text{b})_{s2,y1}$ ***
$\text{SolarRad}_{s4,y1}{}^{2}$ ***	$\text{ND}(\text{swir1},\text{b})_{s3,y1}{}^{2}$ ***	$\text{Tmin}_{s3,y2}{}^{2}$ ***
Tmin_{y2} ***	$\text{-ND}(\text{swir1},\text{nir})_{s4,y1}$ ***	$\text{ND}(\text{nir},\text{g})_{s4,y1}{}^{2}$ ***
$\text{-Rain}_{y2}{}^{2}$ ***	$\text{-ND(r,g)}_{s4,y2}{}^{2}$ ***	$\text{-ND}(\text{nir},\text{g})_{s4,y1}$ ***
$\text{-Tmin}_{s2,y2}{}^{2}$ ***	$\text{ND}(\text{swir2},\text{g})_{y2}{}^{2}$ ***	SlopeNorth ***
$\text{ETo}_{s1,y2}{}^{2}$ ***	$\text{ND}(\text{nir},\text{g})_{s2,y2}{}^{2}$ ***	$\text{-ND}(\text{swir1},\text{r})_{s4,y2}{}^{2}$ ***
$\text{Tmin}_{s2,y1}$ *	$\text{-ND}(\text{nir},\text{b})_{y1}$ ***	$\text{ND}(\text{nir},\text{g})_{s4,y2}$ ***
$\text{-Rain}_{s3,y1}$	$\text{-ND}(\text{swir2},\text{b})_{s4,y2}{}^{2}$ ***	$\text{ND}(\text{swir2},\text{r})_{s3,y1}$ **
$\text{Rain}_{s2,y1}{}^{2}$ ***	$\text{-ND}(\text{g},\text{b})_{s4,y1}$ ***	-SlopeEast **
	$\text{ND}(\text{swir2},\text{b})_{s3,y2}{}^{2}$ ***	$\text{-ND}(\text{r},\text{b})_{s1,y1}{}^{2}$ **
	$\text{ND}(\text{swir1},\text{nir})_{y2}{}^{2}$ *	$\text{ND}(\text{r},\text{b})_{s3,y2}{}^{2}$ ***
	$\text{-ND}(\text{g},\text{b})_{s4,y2}{}^{2}$ **	
$R^2 = 0.28$	$R^2 = 0.79$	$R^2 = 0.80$
LCCC=0.43	LCCC=0.88	LCCC=0.89
RMSE=1.34 t/ha	RMSE=0.72 t/ha	RMSE=0.71 t/ha

The results are shown in Table 2, including three models built with (i) meteorological variables only, (ii) with remote sensing variables only, and (iii) with all variables. The more selective BIC criterion (Hastie et al., 2009) was used for the remote sensing and all variable cases, and the more inclusive AIC criterion was used for the meteorological variable case (as BIC resulted in a model with only one retained variable). We made the following observations:

- The meteorological variables can explain 28% of the variation in the dataset, while remote sensing variables can explain 79%. Adding meteorological variables to remote sensing only improves the explanatory power of the model by 1%, to 80%.
- Of the remotely sensed variables, $\text{ND}(\text{nir},\text{g})_{s2,y1}{}^{2}$ is the most important (as also noted in the correlation analysis in the previous section).
- For the all variable model, of the meteorological variables, only the minimum temperature during s3 was selected.
- Of the spatial variables, the north-facing slope is positively related with yield. We expected tree age would also be important. However, we found that many NDSIs were strongly correlated (R^2 >0.65) with tree age, which could explain why tree age was not explicitly selected by the models. Another reason could be that the yields of most blocks in this study are greater than 5 years old, and so their yield vs tree age has plateaued (Mayer et al., 2006).

The model fit using the model considering all variables (Table 2) is shown in Fig. 9. It will be shown below in Section 3.2.3 that this forward-backward variable selection method is not as good as other CV-based methods at producing an accurate forecast model. However, this method is useful to find a minimal set of variables that best explains the variability in the whole dataset (80% of the yield variation is described using only 12 variables). The residuals are greatest in the Macksville region, where the coefficient of variation of yields is the greatest. However, the residuals are relatively evenly distributed, indicating the ability of the model to describe a large proportion of the variability in the dataset.

3.2. Prediction: forecast models

Next, we investigated models to forecast future macadamia yield. We started with (1) a simple model based on the average yield of previous years, (2) a model using the best predictor GNDVIN, discussed above in Section 3.1.2, (3) comparison of different multi-variable model algorithms and combinations of predictors, and (4) evaluation of the final selected model.

3.2.1. Predictions based on previous years yield

To establish a baseline for prediction accuracy assessment, we used the null model (Deines et al., 2020), which simply predicts future yield based on the mean of all previous years yields. The observed yields varied between 2.4 t/ha (average of years prior to 2014), to 2.9 t/ha (average of years prior to 2017).

The results are shown in Fig. 10a-b. Of course for this null model, there is very little variability in the inter-annual predictions and the predictions for all regions are the same. The block yield RMSE when test results from all years are pooled was 1.65 t/ha, LCCC was -0.02 and the MAPE was 47.7%. The poor performance of this null model is expected, as no predictors that could account for spatio-temporal variability are included. The total production errors are also high. However, for the mid-yielding region of Ballina the errors are less than 15% for test years after 2014, due to the fact that this simple model under-predicts some block yields, and over-predicts others, so the aggregated yield errors cancel to some degree.

3.2.2. Model using the most important predictor

Section 3.1.2 showed that $\text{GNDVIN}=(\text{ND}(\text{nir},\text{g})_{s2,y1} - 0.56)^2$ is the best single predictor. We fit an OLS model using this variable to previous

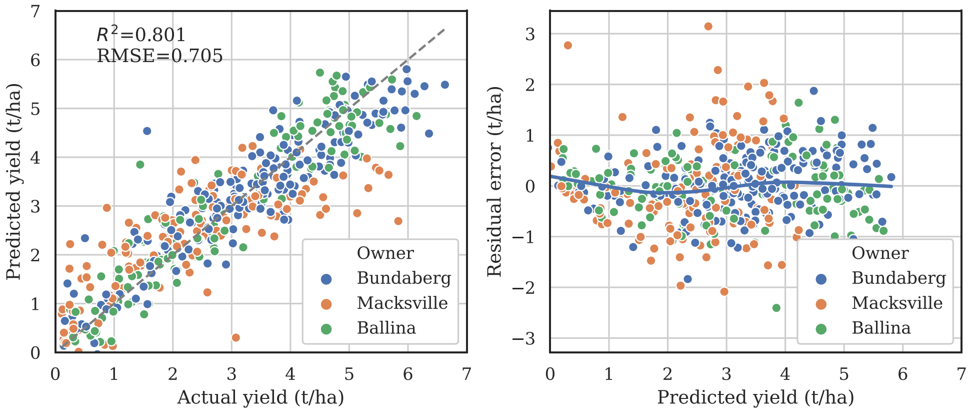
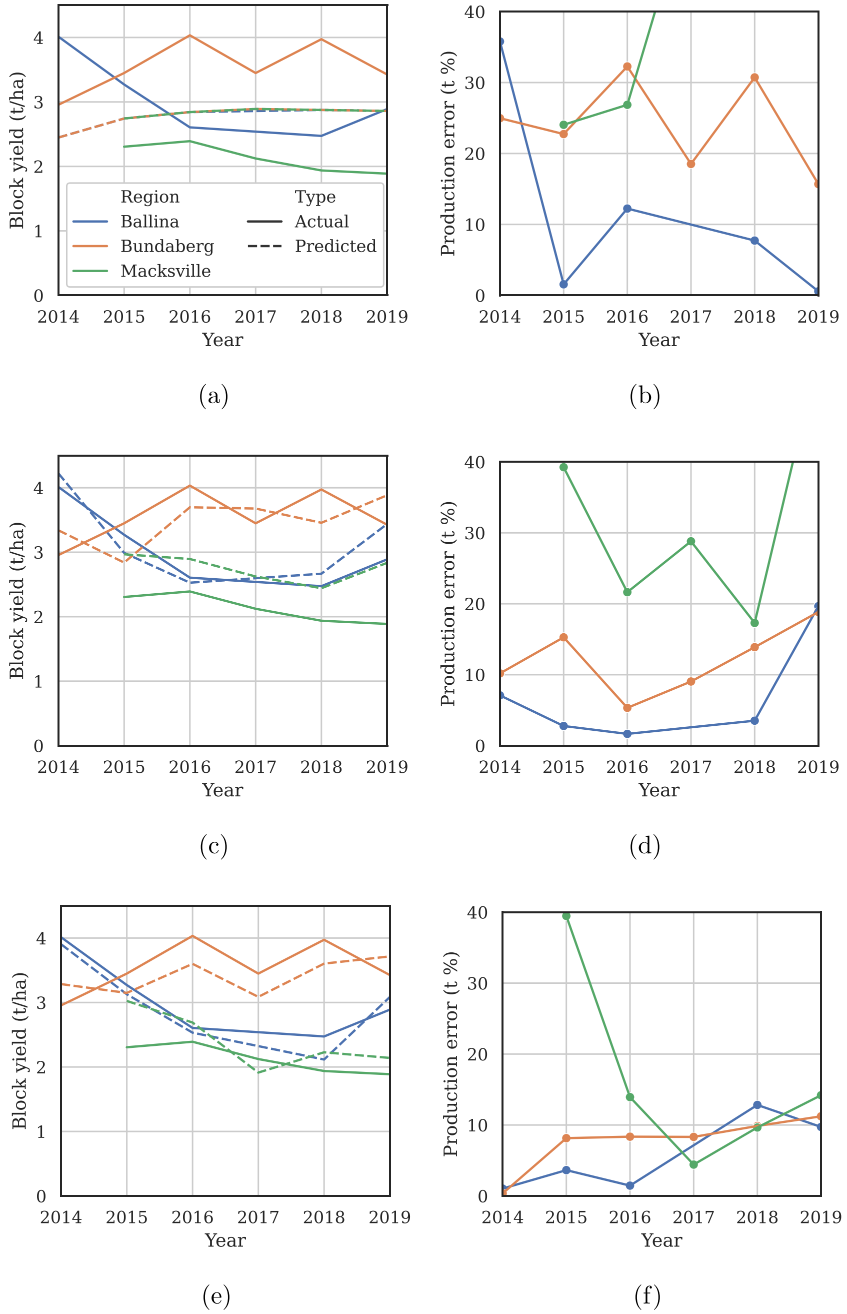


Fig. 9. Model fit and residuals using the entire dataset and the OLS forward-backward variable selection algorithm, with all 424 variables considered. Twelve variables were selected by the algorithm, shown in Table 2.



(a)

(b)

(c)

(d)

(e)

(f)

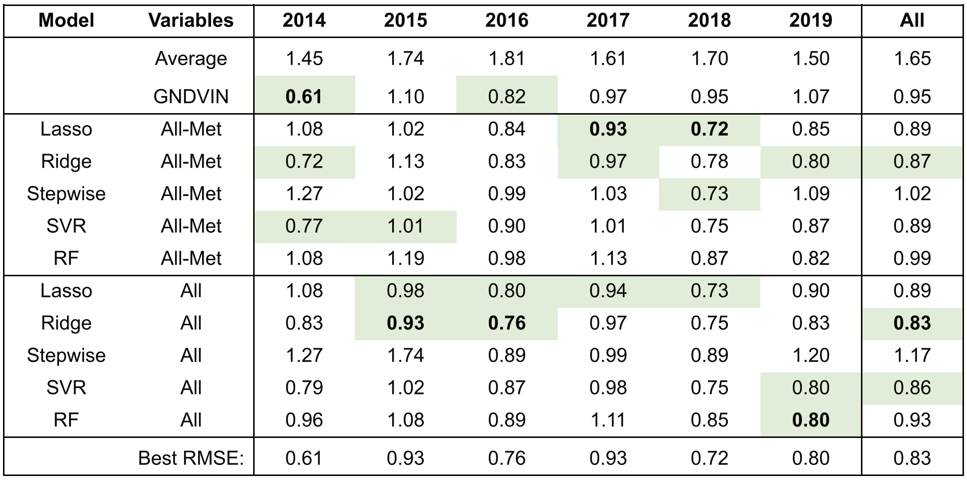
Fig. 10. Average block yields per region (left) and production error per region (right) for three different forecast models. (a-b) Simple model based on average of previous yields. (c-d) Single-variable model based on the best single predictor GNDVIN. (e-f) Multi-variable model using ridge regression with all NDSI and spatial variables.

years yields, and assessed predictions on the following test year, where the test year was varied from 2014 to 2019. The results are shown in Fig. 10c-d. This simple model is able to describe some of the variability between regions, and between years. The RMSE for this model across all test years was 0.94 t/ha, LCCC=0.79 and MAPE=24.8%. Errors are lowest for the Ballina region. On average, the model over-predicts Macksville yields, and slightly under-predicts Bundaberg yields. Model predictions using this simple model are poor for 2019 perhaps due to drought conditions, which began in 2018.

3.2.3. *Comparison of multi-variable prediction algorithms*

We then compared a number of model algorithms, and combinations of variables, with the goal of making a selection for a practical forecast model given the current dataset. We note that the conclusions of this comparison may change as more data becomes available, and the strengths of different algorithms can be utilized. The same methodology used in previous sections was followed, in that each model was trained with previous data to predict the test year yield, for test years between 2014–2019. The results are shown in Fig. 11. We made the following observations:

- In earlier years, where relatively less training data is available, the best models were relatively simple. The single variable GNDVIN model was best at predicting 2014 yields (trained on only 2012–2013 data), and in the top-3 models in 2016.
- Addition of meteorological variables (the ‘All’ models in Fig. 11) did reduce prediction RMSE in some cases, particularly for the nonlinear



Model	Variables	2014	2015	2016	2017	2018	2019	All
	Average	1.45	1.74	1.81	1.61	1.70	1.50	1.65
	GNDVIN	**0.61**	1.10	0.82	0.97	0.95	1.07	0.95
Lasso	All-Met	1.08	1.02	0.84	**0.93**	**0.72**	0.85	0.89
Ridge	All-Met	0.72	1.13	0.83	0.97	0.78	0.80	0.87
Stepwise	All-Met	1.27	1.02	0.99	1.03	0.73	1.09	1.02
SVR	All-Met	0.77	1.01	0.90	1.01	0.75	0.87	0.89
RF	All-Met	1.08	1.19	0.98	1.13	0.87	0.82	0.99
Lasso	All	1.08	0.98	0.80	0.94	0.73	0.90	0.89
Ridge	All	0.83	**0.93**	**0.76**	0.97	0.75	0.83	**0.83**
Stepwise	All	1.27	1.74	0.89	0.99	0.89	1.20	1.17
SVR	All	0.79	1.02	0.87	0.98	0.75	0.80	0.86
RF	All	0.96	1.08	0.89	1.11	0.85	**0.80**	0.93
	Best RMSE:	0.61	0.93	0.76	0.93	0.72	0.80	0.83

Fig. 11. Comparison of forecast model algorithms and input variables based on prediction RMSE (t/ha). The top 3 models for each test year is highlighted, with the best model in bold. For the variables, 'Average' refers to the null model (Section 3.2.1), 'GNDVIN' refers to the single variable model (Section 3.2.2), 'All-Met' uses remote sensing and spatial variables, and 'All' adds the meteorological variables.

SVR and RF algorithms. However, there was not a consistent and significant advantage in terms of lower prediction RMSE.

- The lasso and ridge linear models generally offered similar performance to the more complex RF and SVR models.

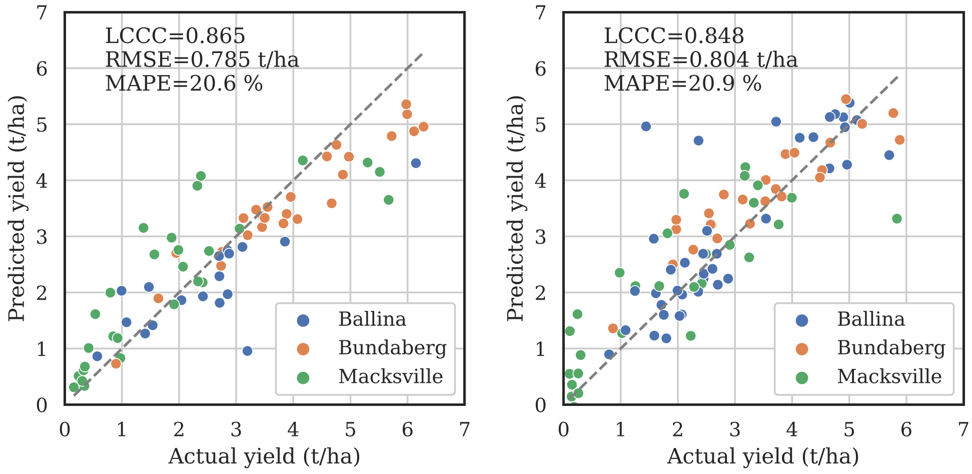
Based on these observations, we made a number of choices regarding the final model. Firstly, we chose to use the ridge algorithm. RF and SVR produced comparable results to ridge when meteorological variables were included, however ridge specifies the forecast model as a simple linear equation that can be easily ported to multiple GIS platforms for industry delivery (which is not the case for RF and SVR). Secondly, we chose to exclude meteorological variables, as it was not clear they offered a consistent advantage. We chose instead to use the 304 All-Met variables. It is possible that as more years data are accumulated, the inclusion of meteorological variables will improve model performance, so this choice will be re-evaluated.

3.2.4. *Final model performance*

Finally, we assessed the accuracy of the chosen algorithm (ridge) and variables (All-Met). The results are shown in Fig. 10e-ff. This model is able to describe much of the variation in yields between regions, and between years. It correctly predicts higher yields in Bundaberg and lower yields in Macksville. Temporal fatures such as the lower yields in Bundaberg in 2015 and 2017, and higher yields in 2016 and 2018 are captured. With data from all test years, the RMSE was 0.87 t/ha, LCCC was 0.82 and MAPE was 22.9%. Again, the regionally aggregated total production predictions (Fig. 10f) are lower than the the block-level predictions, because of the tendency of over- and under-predictions at the block-level to cancel. The production forecast errors are less than 15% from 2016–2019, with an average error for all test years of 9.8%. In contrast to the simpler models, this model gives good performance across all regions.

The model predictions for all blocks for 2018 (trained on 2012–2017 data) and 2019 (trained on 2012–2018 data) are shown in Fig. 12. LCCC indicates excellent agreement between forecast model predictions and actual yields, with values of 0.87 and 0.85 respectively. Over all test years, Macksville yields tend to be over-predicted (average 0.24 t/ha), and Bundaberg yields tend to be under-predicted (average 0.14 t/ha).

Fig. 13 shows an example comparison of block level yield predictions and measurements from 2019 for one of the Ballina orchards that includes 18 blocks. The yield prediction RMSE for this orchard and year is 0.5 t/ha. The spatial pattern of high and low yielding blocks is described well by the model.



(a) (b)

Fig. 12. Comparison of actual and predicted yields for (a) 2018 and (b) 2019, using the ridge regression model with All-Met variables.

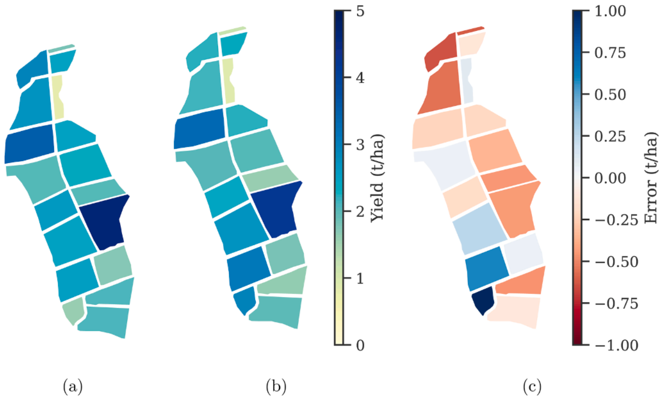


Fig. 13. Block-level yields from one of the Ballina orchards in 2019. (a) Measured. (b) Predictions (using model trained on data previous to 2019). (c) Prediction error.

4. Discussion

Macadamia yields are notoriously difficult to predict, as yields are highly variable and many of the causes of this variability are still unknown (Huett, 2004; Topp et al., 2019). There is no standard practice for tree, block or orchard scale macadamia yield prediction, and current methods rely on estimates based on visual inspection and weather conditions. These estimates are prone to error. This work has developed a method for forecasting macadamia yield at the block-level, using model variables derived from public spatio-temporal datasets including land elevation, remote sensing imagery and spatially interpolated meteorological observations. Forecasts in successive years were tested for accuracy, with forecast models were trained on data from previous years.

4.1. Model design and algorithms

Given the significant differences of yields between the regions used in this study, it is reasonable to question why a separate model was not generated for each region. Alternatively, a model parameter for region that encodes the average yield difference between regions (similar to the panel model in Lobell and Burke, 2010a) could be used. We decided against these methodologies, generating instead a global model that attempts to capture the spatial variation simply using the spatial and spatio-temporal variables described, as other studies have done (Donohue et al., 2018; Filippi et al., 2019). The reasons are two fold. Firstly, the aim was to generate a model that could provide predictions for orchards for which we currently don't have tree or farm data (Fig. 1), using only publicly available datasets. Additionally, it is possible that the yields of the orchards for which we have data do not necessarily represent all orchards in the regions to which they belong. Comparison with the industry benchmark report (Queensland-Government, 2020) suggests many of the orchards used in this study produce higher than average yields. Therefore adding a regional adjustment to model the specific characteristics of these orchards is not likely to produce a model that correctly describes the true variation of the average performance of orchards across regions.

We found that for our dataset, the linear ridge regression model gave competitive performance compared with more complex algorithms such as SVR and RF. The latter algorithms may provide benefit as a more extensive training data set is built, so that nonlinear effects and interactions between predictors can be confidently modeled. However, linear formulations provide the important practical advantage that models can easily be ported between software packages and platforms, as the model is a simple linear summation of input variables. For example, we trained the models in Python using the Scikit-Learn library, copied the resulting linear model equations to Google Earth Engine, and deployed predictions as web applications for growers to view. The predictions could easily be scaled over whole growing regions and countries using macadamia maps such as that of Shephard and McKechnie 2017, potentially complementing predictions based on regional-scale models (Mayer et al., 2019). However, as noted above, there is great variability in grower practices and harvest efficiencies, so care would need to be taken to train models on a representative sample of farms over the whole area predictions are required.

We noted, as also discussed by Filippi et al. 2020 and Deines et al. 2020, that applying a yield prediction model developed at finer scales (at the block-level in our case), to predicting yield at coarser scales (farm or region) tends to reduce errors due to cancellation. Therefore predictions at coarser scales have greater accuracy. This strengthens the value of building block-level models to provide production forecasts at regional or industry-wide scales. Conversely, as noted in Deines et al., 2020, assessing model accuracy at a coarse scale does not guarantee similar performance at finer scales.

4.2. Predictor variables

Similar to studies on other crops (Deines et al., 2020; Kang et al., 2020), we found meteorological variables did not add significant predictive value over using only remote-sensing variables. This does not imply that meteorological variables are unimportant. Rather, remote sensing variables are able to capture the effects of weather on crop parameters, such as LAI, as other studies have shown (Cai et al., 2019). Remote sensing variables may describe many physical characteristics (such as leaf area and light interception, leaf nutrients, water stress) that depend on a range of factors including weather, management and soil, even though these relationships are not explicit in the statistical models (Cai et al., 2019; Zhang et al., 2019).

With variations in climate currently being experienced in Australia (Fig. 3), it is likely that the range of variation seen over a year will not be captured by the current yield dataset, thus limiting the ability of meteorological variables in predicting yield in new years (Deines et al., 2020; Marcelis et al., 1998). Possibly, a dataset incorporating more years, and thus more climatic variation, will make meteorological variables more useful. We also note that our model was built on block-level data, so models are selected based on their ability to describe variation between block yields, as well between regional and annual yields. If data was aggregated at the farm level, thus eliminating block-block variability, it is possible that meteorological variables would become more important as the variation in yield would be dominated by seasonal and regional differences, rather than block differences.

Of the most important meteorological variables, minimum temperature during winter was the most correlated with yield (positive relationship), followed by maximum temperature and evapotranspiration. We noted some correlations between yield and spatial variables. Land slope was negatively correlated with yield, perhaps due to difficulty in harvesting non-flat blocks. However, slope towards the north is positively correlated with yield, perhaps due to increased solar exposure in the southern hemisphere (higher nut set generally occurs on the northern side of trees in Australia, Huett 2004). The most important remote sensing variable for predicting yield was the average GNDVI from April–June, which had a coefficient of determination $R^2 = 0.58$.

Our final model used ridge regression, with 304 variables, with variables derived from public remote sensing and spatial datasets. Even though no meteorological variables were included, and a single model that covered all regions, the model was able to predict inter-annual variation as well as spatial variation between blocks and regions.

Unfortunately, yield forecast studies that report relative accuracy metrics that can be directly compared (such as LCCC or MAPE) are rare (van Klompenburg et al., 2020), particularly so for tree crops. However, our macadamia forecast models compare well with similar work on predicting grain (LCCC=0.89–0.94 Filippi et al. 2019), canola and wheat (RMSE=32-33% (Donohue, Lawes, Mata, Gobbett, Ouzman, 2018)), rice (RMSE=15% for block level data in Setiyono et al. 2019) and cotton (LCCC=0.63 Filippi et al. 2020).

4.3. Limitations

One of the sources of irreducible errors in this methodology is the uncertainties in measured yield data, due to variability in harvest efficiency and accuracy. Growers have noted the variability in the proportion of nuts being successfully gathered, due to equipment and weather conditions (for example, rain causes issues in some areas with nuts being washed away before being swept up). There is also estimation involved in deriving block yields from farm yields, with growers using different methods to calculate these.

Macadamia yield is dependent on variety (Stephenson et al., 1986a), and on cross-pollination between varieties (Howlett et al., 2015). However, our models did not include macadamia variety as a variable. There are two reasons for this. Firstly, many blocks have multiple varieties interleaved (to promote cross-pollination), and the number of rows for each variety are often smaller than the 30 meter Landsat pixels. Secondly, an important goal of this work is to produce a forecast model that can be applied to orchards for which we have no information other than the public remote sensing, climate and landscape data. Requiring tree variety as a model variable would make predictions over these areas impossible (unless variety could be estimated from remote sensing data). Future work could involve using higher resolution remote sensing data to investigate relationships between variety, reflectance and yield.

Management factors that may affect yield were not directly modeled. These include pruning, fertilizer application, irrigation, mulching, and control of weeds, pests and diseases (Jin et al., 2020). However, remote sensing variables may capture some of the effects of these factors on tree health (Zhang et al., 2019). Irrigation mitigates the effects of dry weather to some extent, so future work could involve adding an irrigation variable to allow different model coefficients for irrigated and non-irrigated orchards.

Despite these sources of error, the models were able to predict yield variability between regions and years, and total production with errors less than 15%.

4.4. Deployment and future prospects

The yield forecasts are currently being delivered to growers through a web application in January, giving sufficient time for decision making before the March–September harvest season. This provides a tool that growers are able to use to support their decisions about harvest planning, contracts and marketing. It also provides information about spatial variability that can aid management decisions (Robson et al., 2017). It is expected that as more data is added, predictions will become more accurate and further insights into the drivers of yield variability will be possible.

5. Conclusion

Accurate yield forecasting within many horticultural tree crop industries remains elusive, and macadamia is no exception. This work examined a range of model algorithms and input variables to forecast macadamia yield. The methodology allowed forecasts to be generated in January, predicting the harvest in March–September, which is the timeframe useful for growers and industry to base important decisions on. Remotely sensed variables were found to be the most important predictors of yield, when compared with meteorological variables. The ridge regularized linear model was selected, and was able to predict block-level yield over three regions and years from 2014–2019 with an RMSE=0.87 t/ha, LCCC=0.82 and MAPE= 22.9%. This is a significant improvement over simply using the average of historical yields, which produced RMSE=1.65 t/ha LCCC=0 and MAPE=47.7%. When block-level predictions were aggregated at the regional scale to generate total production predictions, the errors were between 0–15% from 2016–2019.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This project is being delivered by Hort Innovation – with support from the Australian Government Department of Agriculture and Water Resources as part of its Rural R&D for Profit program – and UNE as the co-investor for ST19008 and ST19015.

The authors are grateful for the support of the Australian Macadamia Society, and for helpful discussions with David Mayer, Chris Searle, and data from collaborating macadamia growers including Macadamias Australia, Chris Cook of Dymocks Farm and Graham Wessling of LNL Australia Pty Ltd.

References

Brinkhoff, J., Dunn, B.W., Robson, A.J., Dunn, T.S., Dehaan, R.L., 2019. Modeling mid-season rice nitrogen uptake using multispectral satellite data. Remote Sens. 11 (15), 1837. https://doi.org/10.3390/rs11151837.

Brinkhoff, J., Robson, A.J., 2020. Macadamia orchard planting year and area estimation at a national scale. Remote Sens. 12 (14), 2245. https://doi.org/10.3390/rs12142245.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. Forest Meteorol. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Deines, J.M., Patel, R., Liang, S.-Z., Dado, W., Lobell, D.B., 2020. A million kernels of truth: insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sens. Environ. 112174. https://doi.org/10.1016/j.rse.2020.112174.

Donohue, R.J., Lawes, R.A., Mata, G., Gobbett, D., Ouzman, J., 2018. Towards a national, remote-sensing-based model for predicting field-scale crop yield. Field Crops Res. 227, 79–90. https://doi.org/10.1016/j.fcr.2018.08.005.

Felderhof, L., Gillieson, D., 2011. Near-infrared imagery from unmanned aerial systems and satellites can be used to specify fertilizer application rates in tree crops. Can. J. Remote Sens. 37 (4), 376–386. https://doi.org/10.5589/m11-046.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S. U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F.A., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20 (5), 1015–1029. https://doi.org/10.1007/s11119-018-09628-4.

Filippi, P., Whelan, B.M., Vervoort, R.W., Bishop, T.F.A., 2020. Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. Agric. Syst. 184, 102894. https://doi.org/10.1016/j.agsy.2020.102894.

Gao, B.-c., 1996. NDWIA normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58 (3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58 (3), 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.

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

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. In: Springer Series in Statistics. Springer-Verlag, New York.

Howlett, B.G., Nelson, W.R., Pattemore, D.E., Gee, M., 2015. Pollination of macadamia: review and opportunities for improving yields. Sci. Hortic. 197, 411–419. https://doi.org/10.1016/j.scienta.2015.09.057.

Huett, D.O., 2004. Macadamia physiology review: a canopy light response study and literature review. Aust. J. Agric. Res. 55 (6), 609. https://doi.org/10.1071/AR03180.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Modell. Softw. 16 (4), 309–330. https://doi.org/10.1016/S1364-8152(01)00008-1.

Jin, Y., Chen, B., Lampinen, B.D., Brown, P.H., 2020. Advancing agricultural production with machine learning analytics: yield determinants for California's almond orchards. Front. Plant Sci. 11, 290. https://doi.org/10.3389/fpls.2020.00290.

Johansen, K., Duan, Q., Tu, Y.-H., Searle, C., Wu, D., Phinn, S., Robson, A., McCabe, M.F., 2020. Mapping the condition of macadamia tree crops using multi-spectral UAV and

WorldView-3 imagery. ISPRS J. Photogramm. Remote Sens. 165, 28–40. https://doi.org/10.1016/j.isprsjprs.2020.04.017.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environ. Res. Lett. 15 (6), 064005. https://doi.org/10.1088/1748-9326/ab7df9.

Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268. https://doi.org/10.2307/2532051.

Climate Change and Food Security: Adapting Agriculture to a Warmer World. In: Lobell, D.B., Burke, M. (Eds.), 2010, Advances in Global Change Research. Springer, Netherlands. https://doi.org/10.1007/978-90-481-2953-9.

Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agric. Forest Meteorol. 150 (11), 1443–1452. https://doi.org/10.1016/j.agrformet.2010.07.008.

Marcelis, L.F.M., Heuvelink, E., Goudriaan, J., 1998. Modelling biomass production and yield of horticultural crops: a review. Sci. Hortic. 74 (1), 83–111. https://doi.org/10.1016/S0304-4238(98)00083-1.

Maselli, F., Chiesi, M., Brilli, L., Moriondo, M., 2012. Simulation of olive fruit yield in Tuscany through the integration of remote sensing and ground data. Ecol. Modell. 244, 1–12. https://doi.org/10.1016/j.ecolmodel.2012.06.028.

Mayer, D.G., Chandra, K.A., Burnett, J.R., 2019. Improved crop forecasts for the Australian macadamia industry from ensemble models. Agric. Syst. 173, 519–523. https://doi.org/10.1016/j.agsy.2019.03.018.

Mayer, D.G., Stephenson, R.A., Jones, K.H., Wilson, K.J., Bell, D.J.D., Wilkie, J., Lovatt, J.L., Delaney, K.E., 2006. Annual forecasting of the Australian macadamia crop integrating tree census data with statistical climate-adjustment models. Agric. Syst. 91 (3), 159–170. https://doi.org/10.1016/j.agsy.2006.02.004.

McFadyen, L.M., Morris, S.G., Oldham, M.A., Huett, D.O., Meyers, N.M., Wood, J., McConchie, C.A., 2004. The relationship between orchard crowding, light interception, and productivity in macadamia. Aust. J. Agric. Res. 55 (10), 1029–1038. https://doi.org/10.1071/AR04069.

O'Hare, P., Quinlan, K., Stephenson, R., Vock, N., Drew, H., Ekman, J., Firth, D., Gallagher, E., O'Farrell, P., Rigden, P., Searle, C., Vimpany, I., Waite, G., 2004. Macadamia Information Kit. Agrilink, your growing guide to better farming guide. Library Catalog: era.daf.qld.gov.au Place: Queensland Horticulture Institute. Brisbane, Queensland Publisher: Agrilink Series Q103052. Department of Primary Industries.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, A., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

Queensland-Government, 2020. Macadamia industry benchmark report - 2009 to 2019 seasons.

Rahman, M.M., Robson, A., 2020. Integrating Landsat-8 and Sentinel-2 time series data for yield prediction of sugarcane crops at the block level. Remote Sens. 12 (8), 1313. https://doi.org/10.3390/rs12081313.

Robson, A., Rahman, M.M., Muir, J., Saint, A., Simpson, C., Searle, C., 2017. Evaluating satellite remote sensing as a method for measuring yield variability in Avocado and Macadamia tree crops. Adv. Anim. Biosci. 8, 498–504. https://doi.org/10.1017/S2040470017000954.

Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70. https://doi.org/10.1016/j.rse.2015.12.024.

Schaffer, B., Andersen, P.C., 2018. Handbook of environmental physiology of fruit crops. CRC Press.

Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, Vol. 57, p. 61.

Setiyono, T.D., Quicho, E.D., Holecz, F.H., Khan, N.I., Romuga, G., Maunahan, A., Garcia, C., Rala, A., Raviz, J., Collivignarelli, F., Gatti, L., Barbieri, M., Phuong, D.M., Minh, V.Q., Vo, Q.T., Intrman, A., Rakwatin, P., Sothy, M., Veasna, T., Pazhanivelan, S., Mabalay, M.R.O., 2019. Rice yield estimation using synthetic aperture radar (SAR) and the ORYZA crop growth model: development and application of the system in South and South-east Asian countries. Int. J. Remote Sens. 40 (21), 8093–8124. https://doi.org/10.1080/01431161.2018.1547457.

Shephard, C., McKechnie, J., 2017. Australian Tree Crop Rapid Response Map. State of Queensland (Department of Science, Information Technology and Innovation).

Smit, T.G., Taylor, N.J., Midgley, S.J., 2020. The seasonal regulation of gas exchange and water relations of field grown macadamia. Sci. Hortic. 267, 109346. https://doi.org/10.1016/j.scienta.2020.109346.

Stephenson, R., 2005. Macadamia: Domestication and commercialization. Chronica Horticulture 45 (2), 11–15.

Stephenson, R.A., Cull, B.W., Mayer, D.G., 1986. Effects of site, climate, cultivar, flushing, and soil and leaf nutrient status on yields of macadamia in South East Queensland. Sci. Hortic. 30 (3), 227–235. https://doi.org/10.1016/0304-4238(86)90101-9.

Stephenson, R.A., Cull, B.W., Stock, J., 1986. Vegetative flushing patterns of macadamia trees in South East Queensland. Sci. Hortic. 30 (1), 53–62. https://doi.org/10.1016/0304-4238(86)90081-6.

Stephenson, R.A., Gallagher, E.C., Doogan, V.J., Mayer, D.G., 2000. Nitrogen and environmental factors influencing macadamia quality. Aust. J. Exp. Agric. 40 (8), 1145. https://doi.org/10.1071/EA99077.

Stephenson, R.A., Gallagher, E.C., Rasmussen, T.S., 1989. Effects of growth manipulation on carbohydrate reserves of macadamia trees. Sci. Hortic. 40 (3), 227–235. https://doi.org/10.1016/0304-4238(89)90115-5.

Topp, B.L., Nock, C.J., Hardner, C.M., Alam, M., O'Connor, K.M., 2019. Macadamia (Macadamia spp.) Breeding. In: Al-Khayri, J.M., Jain, S.M., Johnson, D.V. (Eds.), Advances in Plant Breeding Strategies: Nut and Beverage Crops: Volume 4. Springer International Publishing, Cham, pp. 221–251. https://doi.org/10.1007/978-3-030-23112-5_7.

Trochoulias, T., Lahav, E., 1983. The effect of temperature on growth and dry-matter production of macadamia. Sci. Hortic. 19 (1–2), 167–176. https://doi.org/10.1016/0304-4238(83)90058-4.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709.

Zhang, Z., Jin, Y., Chen, B., Brown, P., 2019. California almond yield prediction at the orchard level with a machine learning approach. Front. Plant Sci. 10 https://doi.org/10.3389/fpls.2019.00809.