A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool

Kefei Chen[1,2], Rebecca A. O'Leary[1], Fiona H. Evans[1,2,3,†]

[1] Department of Primary Industries and Regional Development, 3 Baron-Hay Court, South Perth, WA 6151, Australia

[2] Faculty of Science and Engineering, Curtin University, Bentley, WA 6102, Australia

[3] Big Data in Agriculture, Murdoch University, Murdoch, WA 6150, Australia

[†] Corresponding author

Phone: +61 (0) 427 871 776

Fax: +61 8 9368 3082

Email: Fiona.Evans@murdoch.edu.au

Abstract

Yield prediction is a major determinant of many management decisions for crop production. Farmers and their advisors want user-friendly decision support tools for predicting yield. Simulation models can be used to accurately predict yield, but they are complex and difficult to parameterise. The goal of this study is to build a simple and parsimonious model for predicting wheat yields that can be implemented in a decision tool to be used by farmers at a paddock level.

A large yield data set accumulated from trials on commonly grown varieties in Western Australia is used to build and validate a generalised additive model (GAM) for predicting wheat yield. Explanatory variables tested included weather data and derivatives, geolocation, soil type, land capability, and wheat varieties. Model selection followed a forward stepwise approach in combination with cross-validation to select the smallest set of explanatory variables. The predictive performance is also evaluated using independent data.

The final model uses seasonal water availability, location and year to predict wheat yield. Because the GAM model has minimal inputs, it can be easily employed in a decision tool to predict yield throughout the growing season using rainfall data up to the prediction date and either climatological averages or seasonal forecasts of rainfall for the remainder of the growing season. It also has the potential to be used as an input to agronomic models that predict the effect on yield of various management choices for fertilizer, pest, weed and disease management.

Keywords: Yield prediction; Waterlogging; Precision farming; Crop modelling; Crop water relations; Decision support

1. Introduction

Grain production in Western Australia (WA) is located in the south-west of Australia and wheat is the main crop grown. The climate is Mediterranean, with broadacre cropping reliant on winter rainfall in a dryland system. Climate variability explains around 40% of total wheat yield variability in the Australian wheat belt, and in parts of WA that figure can be greater than 60% (Ray et al., 2015).

The WA grainbelt has experienced a 20% decline in winter rainfall since the 1970s due to southward shifts in rain-bearing synoptic systems (Bates et al., 2008; Hope et al., 2006). Despite the challenge of declining rainfall, grain growers have improved productivity; largely through better use of existing technologies, including decision support tools and precision agriculture technologies (Kingwell et al., 2013).

Decision support tools help farmers assess different tactical management choices by predicting their effects on crop yield and profit. Many incorporate agronomic models that apply Mitscherlich's law of diminishing returns, which states that the increase in crop yield as a result of increasing a single growth factor is proportional to the decrease from the maximum yield that takes into account all other limiting factors (Mitscherlich, 1909). For instance, this approach is used in decision support tools for nitrogen application in common use in WA such as 'NPdecide', 'Select Your Nitrogen' and its derivative the N-Broadacre mobile application (Bowden, 2003; Burgess et al., 1991). The user of these decision tools is required to input an estimate of maximum yield when nitrogen is not a limiting factor. In the absence of better information, many farmers and / or agricultural advisors using decision tools use an average estimate of yield. Given that wheat yields can vary considerably from year to year depending on the amount, frequency and timing of seasonal rainfall, the benefits of using decision support tools for on-farm management can be greatly improved with better prediction of yield to use in the tools.

With further reduction in rainfall likely for WA (IOCI, 2012), improvements in the performance of decision support tools via better yield prediction can help grain growers maintain and / or continue to improve wheat yields despite a drying climate.

In addition, yield predictions can be applied regionally to support strategic decision-making and planning processes in agribusiness in the context of the increasing pressure on food security and sustainability. Predictions of crop yield before harvest can assist transport and storage planning, allowing grain handling organisations to best allocate resources in any given season. Crop yield predictions have been used as an aid in land use planning, determining crop insurance premiums (Abbaspour et al., 1992; Choudhury and Jones, 2014) and food supply considerations (Rosengrant et al., 2002), and considering future production in a changing climate (Ludwig and Asseng, 2006). Regional yield predictions can also be used to understand where and why gaps between potential and actual yield occur (van Ittersum et al., 2013).

Prediction of crop yield is usually by one of two methods: (1) crop simulation modelling or (2) statistical crop modelling. Crop simulation models, such as APSIM (Ahmed et al., 2016; Brown et al., 2018; Keating et al., 2003; McCown et al., 1996; Yang et al., 2014), CERES (Lv et al., 2017; Ritchie et al., 1988), SWAP (Mokhtari et al., 2018; van Lier et al., 2015) and WOFOST (Boogaard et al., 1998; Ma et al., 2013), use fundamental mechanisms of crop growth and development, soil and water processes to simulate plant growth in different scenarios. They are mathematical representations of the real-world situation and usually include sub-models for crop growth, soil water movement, fertiliser uptake and dissolution and more. Crop simulation models give accurate predictions of yield, but require extensive parameter inputs that describing characteristics of the modelled situation (including crop, soil, climate and crop management). Use of crop simulation models for on-farm decision making typical requires soil testing and meticulous calibration, often with the assistance of a specialised consultant (Hunt et al., 2006). For regional yield predictions, a generic set of typical model parameters are used, but these parameters will give inaccurate predictions for any individual location and are therefore not of use in decision support tools.

Statistical crop models encode relationships between environmental and/or management factors and crop yield to make predictions. The most commonly used methods in WA are French and Schultz (F&S) type equations that relate the total amount of water available during the growing season with the potential non-water-limited yield possible in Australian dryland cropping systems (French and Schultz, 1984). They developed the following equation: *Potential yield (kg/ha) = (Wavail – Evaporation) x WUE,* where *Wavail* is the amount of water that is available during the growing season, defined as $Wavail = RF_S/3 + RF_{GS}$, where RF_S is summer rainfall (November-March) and RF_{GS} is growing season rainfall (April-October). French and Schultz estimated values of 110 mm for *Evaporation* and 20 kg/ha/mm for the water use efficiency (*WUE*).

A recent modification frequently used in WA is the broken stick method that specifies an upper limit to potential yield dependent on the plant available water capacity of the soil (Oliver et al., 2009). This has been shown to better account for waterlogging, which has adverse effects on the development and growth of wheat in Mediterranean-type environments (Turner, 1992). Both the F&S and broken stick methods have the advantage of being simple, easy to understand, and require minimal inputs and have been employed in decision support tools such as PYCAL (Tennant and Tennant, 2000), CliMate (https://climateapp.net.au) and the Department of Primary Industries and Regional Development (DPIRD) ‘Potential yield tool’ (https://www.agric.wa.gov.au/climate-weather/potential-yield-tool). However, these equations can be perceived as being too simple and they only provide a prediction of potential non-water-limited yield not actual yield.

Outside of WA, the most common statistical method used for predicting crop yields is multiple linear regression (linear models) which has been used with covariates including seasonal climate data (Gornott and Wechsung, 2016; Landau et al., 2000; Lobell and Burke, 2010), remote sensing vegetation indices (Kern et al., 2018; Qader et al., 2018; Zhang et al., 2017) and crop water stress indices (Schut et al., 2009), using either simulated data from crop models or real data to train the models. Bayesian implementations of linear models have also been applied (Bornn and Zidek, 2012;

Chipanshi et al., 2015), and partial least squares regression has been used to identify patterns in daily climate records that have use for predicting yield (Ceglar et al., 2016). Machine learning approaches to yield prediction from seasonal climate data include artificial neural network (ANN) models (Çakır et al., 2014; Das et al., 2018); random forests, which have been used at global and / or regional scales (Folberth et al., 2019; Jeong et al., 2016); and genetic programming algorithms (Ali et al., 2018).

The use of time series models, including auto-regressive integrated moving average (ARIMA) models, has been investigated for forecasting crop yields (Ali et al., 2015; Choudhury and Jones, 2014; Craparo et al., 2015; Debnath et al., 2013). However, these models rely on the assumption that yield in any one year is related to that of the previous year and because wheat yields in WA vary considerably from year to year with seasonal rainfall, they are unsuitable for this study.

Motivated by the need for seasonal yield predictions for use in decision support tools for farmers, the objective of this study is to determine whether a large yield data set can be used to build a statistical model that combines the advantages of simple F&S type equations with added complexity, as required, to produce results akin to those of crop simulation models. The goal is to produce a model that is more accurate than simple linear equations while still having few, easily obtained inputs so that it can be implemented in a decision tool. For this reason, we limit climate inputs to seasonal summaries rather than daily data, and compare model outputs only with F&S and not with crop simulation models that require daily climate data as input.

A generalized additive model (GAM) is a generalized linear model in which the predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. GAMs were originally developed to blend properties of generalized linear models with additive models (Hastie and Tibshirani, 1990). GAMs allow for flexible specification of the dependence of the response (yield) on covariates, by specifying the model in terms of smooth functions or 'smooths', rather than as detailed parametric relationships (Wood, 2006). GAMs have a number of advantages that suggest their use for crop yield modelling – most importantly that they capture nonlinear relationships as compared to multivariate linear regression. GAMs are easy to interpret by plotting fitted smooths against yield. Relationships between smooths and yield can be considered in light of existing knowledge to ensure that they are sensible. They can also identify hidden patterns in the data, potentially improving our understanding of crop-soil-weather interactions.

This study uses a large yield data set accumulated from variety trials on commonly grown and well-adapted wheat varieties in WA. The data encompass a wide range of locations, soils and rainfall zones. Explanatory variables tested in the study were selected using existing knowledge and following consultation with local agronomists. They include weather data and derivatives, geolocation (latitude and longitude), soil type, differences in wheat variety and various classifications of wheat variety (eg. into classes of maturity type).

Forward stepwise model selection is used to build a GAM that has the smallest possible set of inputs required to predict yield. The forward selection process starts with the null model (i.e. intercept only model) and adds variables to the model one at a time. At each step, the ability of each additional explanatory variable to add predictive performance to the model is assessed using statistics for model selection that balance the 'goodness of fit' of the model with its complexity. The variable that provides the most improvement is then added to the model, until no further improvement is possible. At each step, cross-validation is applied to test the ability of the model to perform on unseen data and thus avoid overfitting. The final model is also tested using previously unseen data from a separate source.

This approach is used to build a simple and parsimonious GAM that reflects current understanding of how a wheat crop responds to seasonal rainfall, and how yields vary spatially and through time in WA. Because the GAM has minimal inputs, it can be easily employed in a decision tool to predict yield throughout the growing season using rainfall data up to the prediction date and either climatological averages or seasonal forecasts of rainfall for the remainder of the growing season. It also has the potential to be used as an input to agronomic models that predict the effect on yield of various management choices for farm inputs, pest, weed and disease management.

2. Data

2.1. Yield data

Variety trial data: A total of 25,505 observations from variety trials, including WA-based Crop Variety Trials (CVT) and National Variety Trials (NVT, http://www.nvtonline.com.au) in a 40-year period of 1975 – 2014 were considered in this study. The data include 109 varieties and 775 unique locations; however, the majority of varieties are not commonly grown and/or not well-adapted in Western Australia (WA). A panel of 26 varieties commonly grown and well-adapted in WA with a total of 17,701 observations were selected for the base model construction. Figure 1 shows the distribution of 775 locations in variety trials conducted in the WA grainbelt, overlaid on 30-year average rainfall (Garlinge 2005).

Focus paddock data: A total of 428 observations from 164 different Focus Paddocks (Harries et al., 2015) over the period of 2010 – 2013 were used to test the predictive performance of the base model constructed using the variety trial data. Unlike the variety trial data which was specifically collected to determine performance differences between wheat varieties, the Focus Paddocks included data from operational wheat-growing farm paddocks. They were mainly distributed in the medium rainfall zones (see Figure S1) and were used in this study as an independent test of the fitted yield prediction model to help understand how the model might predict previously unseen, real paddock data.

2.2. Weather and derived data

Weather data for the nearest weather station to each trial location were extracted from the Patched Point Database (https://www.longpaddock.qld.gov.au/silo). Data derived from the daily weather include:

Growing season available water (Wavail): The amount of water available for crop use in any growing season is defined to be one third of summer rainfall plus growing season rainfall.
30-year average rainfall: Average rainfall for years 1971 to 2000 (*avgrf30*) was used as a continuous surrogate for rainfall zone.

Early season rainfall: Early season rainfall (*sumrf2m*) is defined to be the sum of the first two months of rainfall occurring after germination.

Germination time: Germination time (*gdoy*) is estimated from daily rainfall using a common germination rule where germination is assumed to occur if there is 25mm of rainfall over three days after 25 April, or 5mm of rainfall over three days after 5 June.

2.3. Soil data

Soil type: The dominant soil types for each location were extracted from the soils database managed by the Department of Primary Industries and Regional Development (DPIRD). The soil groups were first classified according to their soil-landscape mapping (Purdie et al., 2004), and then simplified to six general functional types (*soils*) including clays, duplex, gravel, loamy, sandy and wet to reduce the number of categories for modelling. The classification of the agricultural soil groups to functionally simplified soil classes is shown in Table S1.

Land capability and qualities: A total of 19 variables related to land capability and qualities were derived from the Natural resource information (NRInfo) database maintained by DPIRD and other government agencies (https://www.agric.wa.gov.au/resource-assessment/nrinfo-western-australia). The *in-silico* land capability and qualities include variables of land evaluation for percentage of map unit for soil pH value, phosphorus export hazard, surface salinity, drainage potential, subsurface compaction susceptibility, water erosion hazard, waterlogging susceptibility, water repellence susceptibility, soil water storage, wind erosion hazard, and dryland cropping (van Gool et al., 2005).

A detailed description of the 19 variables related to the land capability and qualities is presented in Table S2.

Other explanatory variables of interest for the yield prediction model were also assessed, including wheat variety (*variety*) and maturity type (*mattype*).

3. Methodology

3.1. Modelling using GAMs

A generalised additive modelling (GAM) approach (Hastie and Tibshirani, 1990) is used for the yield prediction. In particular, we used the `'gam'` function from R package `'mgcv'` (Wood, 2006).

3.2. Model selection

A forward stepwise model selection is used to determine the optimal set of explanatory variables to use in the model. The forward selection process starts with the null model (i.e. intercept only model) and adds variables to the model one at a time. At each step, variables that provide the most information to the model are calculated using various statistics for model selection.

3.3. Statistics for model performance

Several model performance statistics were used for model selection and to assess the goodness-of-fit of the models including the Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978), root mean square error (RMSE), correlation coefficient (*r*) and coefficient of determination (R^2).

3.4. Cross validation

When fitting statistical models, the goodness-of-fit of the model must be balanced against model complexity in order to avoid overfitting. That is, to avoid building models that describe the data used to fit them very well but predict poorly on previously unseen data. To this end, the `'gam'` function in

R package 'mgcv' can be used to fit smooth terms by generalised cross validation (GCV). GCV is based on a leave-one-out cross validation, where only one datum from the dataset is omitted from model fitting (Craven and Wahba, 1978). For small data sets, this results in reasonable smooths; however, for a large data set such as ours, this resulted in smooths that appeared to be over-fitted. We therefore applied a 5-fold-cross-validation (Geisser, 1993), to divide the variety trial data into 5 testing and training sets with a 80:20 split of training to test GAMs at each step of the forwards model selection procedure.

3.5. Model checking with deviance residuals

The final fitted GAM was checked with deviance residuals. A variety of residual plots were examined, including normal Q-Q plot, residuals versus fitted values, histogram of residuals and response versus fitted values.

3.6. Bayesian analysis

Bayesian inference using Gibbs sampling approach was used for comparison to the GAM approach. The 'jagam' function in the 'mgcv' package in R was used to convert the best GAM model from the frequentist approach into a Bayesian 'JAGS' (Just Another Gibbs Sampler) model, following the approach outlined in Wood (2016). Applying the 'jagam' function converts the GAM into a Bayesian graphical model for simulation with 'JAGS', which can then be passed to 'JAGS' *via* the 'rjags' package in R. The following model diagnostics were applied to test the convergence of the Markov chain Monte Carlo (MCMC) simulations: Geweke's Z-score (Geweke, 1991), Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992) and Heidelberger and Welch's convergence diagnostic (Heidelberger and Welch, 1983). In addition, the convergence was also checked using a variety of plots, including a trace density plot, Gelman-Rubin-Brooks plot and Geweke-Brooks plot.

4. **Results**

4.1. Exploratory data investigation

In the variety trial data, there are a large number of observations with the same location, variety and year, but with large within-cluster variation (see Figure S2). To reduce the within-cluster variance component, the data were aggregated over location, variety, and year. Thus, the average yield within the same location, year and variety was used for model construction. After aggregation, there were a total of 9,116 observations for the variety trial data set.

Figure 2 shows changes in mean yield and *Wavail* through time. The mean yield and *Wavail* with standard error of the mean (SEM) are shown in Table S3. There were obvious fluctuations of the mean yield and *Wavail* through the 40-year period, and the overall trends of the mean yield and *Wavail* differ markedly. The correlation between yield and *Wavail* in the early years of the variety trial data is much lower compared to that in the late years. This suggests that seasonal rainfall has played a more important role in crop production as the *Wavail* has gradually declined in WA. Regardless of the declining trend of *Wavail*, yield has gradually increased over these years. This could possibly be attributed to the improvement in crop management techniques and technologies, including crop breeding for better selection of varieties and timely sowing. For this reason, we have tested the year in which the trial was conducted as a potential variable in the model.

Figure 3 shows the distribution of yield, and its relationship with explanatory variables *Wavail*, *sumrf2m*, *avgrf30*, germination time *(gdoy)*, longitude (*longi*), latitude (*lat*), year, variety and soils. Because the data are from variety trials, a range of different varieties are planted simultaneously at each trial location causing a wide range in recorded yields shown in the y-axes. The correlation of yield with each continuous variable is marked on the plots in Figure 3, showing that seasonal available water, *Wavail*, has the highest correlation with yield.

As may be expected, rainfall in the first two months after germination and 30-year average rainfall are correlated with *Wavail*, with correlations of 0.73 and 0.69 respectively. Both show similar, but weaker, relationships with yield as for *Wavail*. Germination time *(gdoy)* has a significant correlation with yield (p-*value* $< 2.2 \times 10^{-16}$) (Figure 3), which reflects regional knowledge that longer growing periods (i.e. earlier sowing) results in higher yields in the absence of other constraints such as frost damage.

Analysis of variance shows that inter-variety variation in yield is significant (p-*value* $< 2.2 \times 10^{-16}$); however, the amount of variability in yield explained by variety is only 2.5%. The same is true for soils. Similarly, the five maturity types explained less than 0.5% of variability in yield.

4.2. Model selection

A selected list of models for water-limited yield with the best performance at each level is presented in Table 1, with the model formulae as specified in R. A full list of models for yield compared in this study, including model performance statistics, is presented as supplementary material Table S4. Seasonal water availability (*Wavail*) was selected as the first variable added by stepwise forward selection using a linear model. It had the greatest coefficient of determination (R^2) compared to all of the univariate analyses, showing that more variation can be explained by *Wavail* than any of the other explanatory variables. This supports our prior knowledge that seasonal rainfall is the main driver of yield in water-limited environments. Use of a GAM with the single explanatory variable *Wavail*, showed that using a smooth term provided a better fit and resulted in a lower AIC and RMSE than the linear model, and higher correlation coefficient and R^2. Thus, the GAM with *Wavail* as a smooth (s) term was selected as a base model. The degrees of freedom of the smooth are dictated by the dimension, *k*, of the spline basis used. We optimised *k* between numbers of 3 to 10 using 5-fold cross-validation. According to the test statistics and examination of the smooth plots for *Wavail*, increasing *k* beyond 3 was not required because higher values only increased the complexity of the smooth while resulting in greater values of the AIC (see Table S4). Therefore, after the first step of forward model selection, the selected model had the form: *yield ~ s(Wavail, k=3)* (see Figure 4). The plot of the fitted smooth against yield shows that the partial effect of *Wavail* on yield reaches a maximum around 400 mm, after which it starts to have a negative effect on yield (Figure 5a).

The second step in the forward model selection tested the addition of smooth terms for each explanatory variable, as well as a two-dimensional smooth term fitted to capture the spatial correlated effect of geolocation, *s(lat, longi)*. Adding the smooth term *s(lat, longi)* to the base model provided the greatest improvement, and it was therefore selected as the second covariate to add into the model. Five-fold cross validation showed that setting *k* equal to 6 was optimal. The contour plot for *s(lat, longi)* is similar to the grainbelt rainfall contours, and is a realistic representation of prior knowledge that yield decreases with both rainfall and distance from the Indian and Southern oceans (Figure 5b). After adding the second term, the model after performing the step 2 of forward selection is *yield ~ s(Wavail, k=3) + s(lat, longi, k=6)*.

Following the same process, the third covariate to be added to the model was the variable year (Table 1). There was no benefit gained by adding year as a smooth term, and therefore year was fitted in model as a linear effect.

Adding soil types to the model could provide slightly better model performance based on 5-fold cross-validation (Figure 4), but the model does not inherently possess higher predictive power based on a test with out-of-sample data (i.e. Focus Paddocks data). Since this model is built for a predictive purpose, a simple and parsimonious model with few inputs was selected as the final fitted model: *yield ~ s(Wavail, k=3) + s(lat, longi, k=6) + year.*

4.3. Model checking with deviance residuals

The residual plots (Figure S3) indicate that the GAM model adequately fits the data, but there is a large amount of variability in the observed data that is not captured by the model. Because this model is not accounting for any kind of farm or paddock management (e.g. fertiliser applications, soil applications or weed, disease and pest management), this variability is expected.

4.4. Spatial and temporal model diagnostics

Figure 6 shows the GAM model predicted yields compared to the observed yield for each year in the variety trial data. The model clearly performs better in later years, most obviously 2000. This is likely due to the change in the relationship between *Wavail* and yield (see Figure 2). Figure S4 confirms this, and also shows that model performance is better in low to medium seasonal rainfall.

4.5. Model performance using unseen data

The Focus Paddocks data were used as an independent test of the fitted yield prediction model. They were not used to fit the model. The correlation between observed and predicted yields was 0.549 and the RMSE was 0.948 tonnes/ha. Compared with the results of the 5-fold cross-validation, which showed a correlation of 0.655 and RMSE of 0.833 tonnes/ha calculate over the test sets. Figure 7 shows the annual plots of observed versus predicted yield for the Focus Paddock data. The GAM model tends to under-predict yield for the Focus Paddocks, particularly in years 2011 and 2012. These years are both examples of types of years that have already been identified as years in which the model performs poorly: in 2011 there was a poor correlation between yield and *Wavail* in the variety trial data; and 2012 was a wetter year.

The plot for the smoothing term *s(Wavail)* in the GAM model is displayed over the Focus Paddock data in Figure S5a. This suggests the effect of *Wavail* on yield is approximately linear over this data set. Similarly the smoothing term *s(lat, longi)* using Focus Paddocks data is displayed in Figure S5b.

4.6. Bayesian analysis

The selected "best" GAM model was used to create a Bayesian(`'JAGS)` model code and data, which was then run using `'JAGS'` to make inference about the model via Gibbs sampling. The trace and density plot, Gelman-Rubin-Brooks plot, and the Geweke-Brooks plot for all parameters in the Gibbs simulations are presented in Figure S6, Figure S7 and Figure S8, respectively. The Geweke convergence diagnostic suggests that there is a good equity of the means of the first (10%) and last part (50%) of a Markov chain. The Gelman-Rubin convergence diagnostic indicates that the output from all chains is indistinguishable. The Heidelberger-Welch convergence diagnostic suggests that there is little evidence to reject the null hypothesis i.e. the Gibbs sampled values come from a stationary distribution. The results of these convergence diagnostics are presented in supplementary Table S5. Together, these diagnostics indicate that the fitted Bayesian `'JAGS'` model adequately describes the variety trial data.

The resulting `'JAGS'` predictions were similar to those of the original GAM model with (r = 0.999), suggesting that no benefit is gained by use of a full `'JAGS'` implementation of this model.

4.7. Comparison with French & Schultz method

The comparisons between the (traditional) GAM predicted yield and French & Schultz (F&S) potential yield using variety trial data and Focus Paddock data are presented in Figure S9 and Figure S10, respectively. The potential yield was estimated using the F&S approach with 120 mm as evaporation and 20 kg/ha.mm as the water use efficiency. In general, predicted yields are lower than potential yields, as would be expected. When *Wavail* is low, the GAM predicted yield is slightly higher than the potential yield estimated by F&S method (Figure S9d).

5. Discussion

The model selection process used in this study ensured that goodness-of-fit was balanced against model complexity by adding explanatory variables one at a time, only if they added to the predictive ability of the model. This process showed that water availability is the major determinant of wheat yield in Western Australia. Based on a historical trial series from 1975 to 2014, the partial effect of *Wavail* on yield reaches a plateau around 400 mm, and then it starts to have a negative effect on yield (Figure 5a).

The second most important predictor is location. The partial effect of the combination of latitude and longitude is similar to the contour map of the rainfall zones. The addition of location to the model adds to its predictive ability, indicating that the spatial effect is caused by more than rainfall patterns. The visualised contour plot of the effect of latitude and longitude on the GAM yield prediction is shown in Figure 8, in which the partial effect of latitudes and longitudes on yield are classified into six different contours i.e. 1.5, 2, 2.5, 3, 3.5 and 4 (see Figure 8).

The third most important predictor is the year in which the crop is sown. This predictor is most likely capturing the increasing ability of WA grain growers to grow higher yielding crops despite decreasing rainfall, either by new technology or improved knowledge. Adding soil and variety covariates did not improve yield predictions, so the final fitted GAM predicts a mean yield across all varieties and soil types.

The simple and parsimonious GAM model developed in this study can be used as a base model for predicting yield that can potentially be integrated with agronomic models for predicting further effects of farm management for nutrition, pest, disease and weed management. The GAM model can be used to predict yields throughout the growing season using rainfall data up to the prediction date and either climatological averages or seasonal forecasts of rainfall for the remainder of the growing season, similar to the 'Rainfall to Date' tool that predicts growing season rainfall , developed by DPIRD in WA . The tool provides cumulative growing seasonal rainfall and projected seasonal rainfall finishes at a given date based on climatological history (https://www.agric.wa.gov.au/climate-weather/rainfall-date).

Model assessment showed better performance in more recent years, most obviously from 2000, that is likely due to changes in the relationship between available water and wheat yield. The model also performs better in years with low to medium rainfall, with a tendency to under-predict in high rainfall years. This pattern was supported by model validation using the independent Focus Paddocks data set.

Model validation suggests that the effect of *Wavail* on yield is approximately linear (Figure S5a) across the domain of the Focus Paddocks data. This is probably because the Focus Paddocks data are mainly distributed in the medium rainfall zone, with few of the high *Wavail* values evident in the variety trial data.

Comparison of the GAM predictions with the French & Schultz (F&S) potential yield showed that GAM predicted yields were generally lower than the F&S potential yield (Figure S9d and Figure S10d), as is expected because F&S aims to predict non-water-limited yield potential whereas the GAM is aiming to predict actual yield taking into account other constraints.

6. Conclusions

Variation in wheat yield in Western Australia can be described using a simple GAM model with minimal inputs derived from rainfall, latitude/longitude and year. This model can easily be developed into a user-friendly online tool for use by grain growers and their consultants. The model can be used to predict yields throughout the growing season using rainfall data up to the prediction date and either climatological averages (Hunt et al., 2006), or seasonal forecasts of rainfall as used by (Brown et al., 2018) for the remainder of the growing season. It also has the potential to be used as an input to agronomic models that predict the effect on yield of various management choices for farm inputs, pest, weed and disease management.

7. Acknowledgments

We are grateful to many colleagues from the Department of Primary Industries and Regional Development (DPIRD) and grain growers for providing various data and help. This project is made possible by Royalties for Regions. The authors would like to thank Dennis Van Gool, Karen Holmes, Ted Griffin, Tony Leeming, Avril Russell-Brown and Phil Goulding from the GIS soil group at DPIRD for providing the soil type and land capability data; many other colleagues at DPIRD for helpful discussions and suggestions.

8. Conflicts of interest

The authors declare no conflicts of interest.

9. References

Abbaspour, K., Hall, J., Moon, D., 1992. A yield model for use in determining crop insurance premiums. Agricultural and Forest Meteorology 60, 33-51.

Ahmed, M., Akram, M.N., Asim, M., Aslam, M., Hassan, F.-u., Higgins, S., Stöckle, C.O., Hoogenboom, G., 2016. Calibration and validation of APSIM-Wheat and CERES-Wheat for spring wheat under rainfed conditions: Models evaluation and application. Computers and Electronics in Agriculture 123, 384-401.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory, 267-281.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018. Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach. Agricultural and Forest Meteorology 263, 428-448.

Ali, S., Badar, N., Fatima, H., 2015. Forecasting production and yield of sugarcane and cotton crops of Pakistan for 2013-2030. Sarhad Journal of Agriculture 31, 1-10.

Bates, B.C., Hope, P., Ryan, B., Smith, I., Charles, S., 2008. Key findings from the Indian Ocean Climate Initiative and their impact on policy development in Australia. Climatic Change 89, 339-354.

Boogaard, H., Van Diepen, C., Rotter, R., Cabrera, J., Van Laar, H., 1998. WOFOST 7.1; user's guide for the WOFOST 7.1 crop growth simulation model and WOFOST Control Center 1.5. SC-DLO.

Bornn, L., Zidek, J.V., 2012. Efficient stabilization of crop yield prediction in the Canadian Prairies. Agricultural and forest meteorology 152, 223-232.

Bowden, J.W., 2003. Select your Nitrogen. A decision tool for quantifying nitrogen availability and crop response in broad-acre farming systems. . WA Department of Agriculture Bulletin.

Brown, J.N., Hochman, Z., Holzworth, D., Horan, H., 2018. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. Agricultural and Forest Meteorology 260, 247-254.

Burgess, S.J., Bowden, J.W., Diggle, A.J., 1991. NPDECIDE - User's Guide. A computer program to help with nitrogen and phosphorus decisions for cereals. Department of Agriculture, Government of Western Australia.

Çakır, Y., Kırcı, M., Güneş, E.O., 2014. Yield prediction of wheat in south-east region of Turkey by using artificial neural networks, Agro-geoinformatics (Agro-geoinformatics 2014), Third International Conference on. IEEE, pp. 1-4.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., Dentener, F., 2016. Impact of meteorological drivers on regional inter-annual crop yield variability in France. Agricultural and forest meteorology 216, 58-67.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F., 2015. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agricultural and Forest Meteorology 206, 137-150.

Choudhury, A., Jones, J., 2014. Crop yield prediction using time series models. Journal of Economics and Economic Education Research 15, 53.

Craparo, A., Van Asten, P., Läderach, P., Jassogne, L., Grab, S., 2015. Coffea arabica yields decline in Tanzania due to climate change: Global implications. Agricultural and Forest Meteorology 207, 1-10.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numerische mathematik 31, 377-403.

Das, B., Nair, B., Reddy, V.K., Venkatesh, P., 2018. Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. Int J Biometeorol 62, 1809-1822.

Debnath, M., Bera, K., Mishra, P., 2013. Forecasting area, production and yield of cotton in India using ARIMA model. Research & Reviews: Journal of Space Science & Technology 2, 16-20.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., Obersteiner, M., 2019. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. Agricultural and Forest Meteorology 264, 1-15.

French, R.J., Schultz, J.E., 1984. Water use efficiency of wheat in a Mediterranean-type environment. I. The relation between yield, water use and climate. Australian Journal of Agricultural Research 35, 743-764.

Geisser, S., 1993. Predictive inference. Chapman and Hall, New York, NY.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical science, 457-472.

Geweke, J., 1991. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Gornott, C., Wechsung, F., 2016. Statistical regression models for assessing climate impacts on crop yields: A validation study for winter wheat and silage maize in Germany. Agricultural and Forest Meteorology 217, 89-100.

Harries, M., Anderson, G.C., Hüberli, D., 2015. Crop sequences in Western Australia: what are they and are they sustainable? Findings of a four-year survey. Crop and Pasture Science 66, 634-647.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. CRC Press, Boca Raton, London, New York, Washington, D.C.

Heidelberger, P., Welch, P.D., 1983. Simulation run length control in the presence of an initial transient. Operations research 31, 1109-1144.

Hope, P.K., Drosdowsky, W., Nicholls, N., 2006. Shifts in the synoptic systems influencing southwest Western Australia. Climate Dynamics 26, 751-764.

Hunt, J., van Rees, H., Hochman, Z., Carberry, P.S., Holzworth, D., Dalgliesh, N.P., Brennan, L.E., Poutlon, P.L., van Rees, S., Huth, N.I., Peake, A., 2006. Yield Prophet®: An online crop simulation service, 13th Australian Agronomy Conference.

IOCI, 2012. Project 1.2 South-west Western Australia's Regional Surface Climate ad Weather Systems, IOCI Milestone Report

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., Kim, S.H., 2016. Random forests for global and regional crop yield predictions. PLoS One 11, e0156571.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., McLean, G., Verburg, K., Snow, V., Dimes, J.P., Silburn, M., Wang, E., Brown, S., Bristow, K.L., Asseng, S., Chapman, S., McCown, R.L., Freebairn, D.M., Smith, C.J., 2003. An overview of APSIM, a model designed for farming systems simulation. European Journal of Agronomy 18, 267-288.

Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., Lichtenberger, J., 2018. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. Agricultural and forest meteorology 260, 300-320.

Kingwell, R., Anderton, L., Islam, N., Xayavong, V., Wardell-Johnson, A., Feldman, D., Speijers, J., 2013. Broadacre farmers adapting to a changing climate, in: Facility, N.C.C.A.R. (Ed.), p. 171.

Landau, S., Mitchell, R., Barnett, V., Colls, J., Craigon, J., Payne, R., 2000. A parsimonious, multiple-regression model of wheat yield response to environment. Agricultural and forest meteorology 101, 151-166.

Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology 150, 1443-1452.

Ludwig, F., Asseng, S., 2006. Climate change impacts on wheat production in a Mediterranean environment in Western Australia. Agricultural Systems 90, 159-179.

Lv, Z., Liu, X., Cao, W., Zhu, Y., 2017. A model-based estimate of regional wheat yield gaps and water use efficiency in main winter wheat production regions of china. Scientific reports 7, 6081.

Ma, H., Huang, J., Zhu, D., Liu, J., Su, W., Zhang, C., Fan, J., 2013. Estimating regional winter wheat yield by assimilation of time series of HJ-1 CCD NDVI into WOFOST–ACRM model with Ensemble Kalman Filter. Mathematical and Computer Modelling 58, 759-770.

McCown, R., Hammer, G., Hargreaves, J., Holzworth, D., Freebairn, D., 1996. APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. Agricultural systems 50, 255-271.

Mitscherlich, E.A., 1909. Das gesetz des minimums und das Gesetz des abnehmenden Bodenertrages (Eng: The law of the minimum and the law of diminishing soil productivity). Landwirtschaftliche Jahrbücher 38, 537-552.

Mokhtari, A., Noory, H., Vazifedoust, M., 2018. Improving crop yield estimation by assimilating LAI and inputting satellite-based surface incoming solar radiation into SWAP model. Agricultural and Forest Meteorology 250, 159-170.

Oliver, Y.M., Robertson, M.J., Stone, P.J., Whitbread, A., 2009. Improving estimates of water-limited yield of wheat by accounting for soil type and within-season rainfall. Crop and Pasture Science 60, 1137-1146.

Purdie, B.R., Tille, P.J., Schoknecht, N.R., 2004. Soil-landscape mapping in south-western Australia: an overview of methodology and outputs., Resource Management Technical Reports. Department of Agriculture and Food Western Australia.

Qader, S.H., Dash, J., Atkinson, P.M., 2018. Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: A case study in Iraq. Science of the Total Environment 613, 250-262.

Ray, D.K., Gerber, J.S., MacDonald, G.K., West, P.C., 2015. Climate variation explains a third of global crop yield variability. Nat Commun 6, 5989.

Ritchie, J., Godwin, D., Otter-Nacke, S., 1988. CERES-Wheat. A simulation model of wheat growth and development. Univ. of Tex. Press, Austin.

Rosengrant, M., Cai, X., Cline, S.A., 2002. World water and food to 2025. International Food Policy Research Institute, Washington, DC.

Schut, A.G.T., Stephens, D.J., Stovold, R.G.H., Adams, M., Craig, R.L., 2009. Improved wheat yield and production forecasting with a moisture stress index, AVHRR and MODIS data. Crop and Pasture Science 60, 60-70.

Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics 6, 461-464.

Tennant, D., Tennant, S., 2000. Potential Yield Calculator, software package. Department of Agriculture and Food Western Australia.

Turner, N.C., 1992. Crop production on duplex soils: an introduction. Animal Production Science 32, 797-800.

van Gool, D., Tille, P.J., Moore, G.A., 2005. Land evaluation standards for land resource mapping : assessing land qualities and determining land capability in south-western Australia, in: Department of Agriculture and Food, W.A. (Ed.), Perth, Western Australia.

van Ittersum, M.K., Cassman, K.G., Grassini, P., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance—A review. Field Crops Research 143, 4-17.

van Lier, Q.d.J., Wendroth, O., van Dam, J.C., 2015. Prediction of winter wheat yield with the SWAP model using pedotransfer functions: An evaluation of sensitivity, parameterization and prediction accuracy. Agricultural Water Management 154, 29-42.

Wood, S., 2006. Generalized additive models: an introduction with R. CRC press, Boca Raton, London, New York.

Wood, S.N., 2016. Just Another Gibbs Additive Modeller: Interfacing JAGS and mgcv. arXiv preprint arXiv:1602.02539.

Yang, Y., Yang, Y., Han, S., Macadam, I., Li Liu, D., 2014. Prediction of cotton yield and water demand under climate change and future adaptation measures. Agricultural water management 144, 42-53.

Zhang, N., Zhao, C., Quiring, S.M., Li, J., 2017. Winter Wheat Yield Prediction Using Normalized Difference Vegetative Index and Agro-Climatic Parameters in Oklahoma. Agronomy Journal 109, 2700-2713.

10. Tables

Table 1: Selected list of models for water-limited yield tested with cross validation in this study.

Model	Model_Formula	Level	Corr (train)	Corr (test)	RMSE (train)	RMSE (test)	R^2 (train)	R^2 (test)	AIC (all)	BIC (all)
1	lm(yield~Wavail)	1	0.526	0.525	0.939	0.939	0.276	0.276	24723	24744
2	gam(yield~s(Wavail, k = 3))	1	0.583	0.582	0.896	0.897	0.34	0.338	23884	23913
12	gam(yield~s(Wavail, k = 3) + s(longi, lat, k = 6))	2	0.631	0.63	0.856	0.857	0.397	0.394	23049	23113
22	gam(yield~s(Wavail, k = 3) + s(longi, lat, k = 6) + year)	3	0.656	0.655	0.832	0.833	0.43	0.427	22546	22617
24	gam(yield~s(Wavail, k = 3) + s(longi, lat, k = 6) + s(year, k = 4))	3	0.658	0.656	0.831	0.833	0.431	0.427	22522	22607
32	gam(yield~s(Wavail, k = 3) + s(longi, lat, k = 6) + year + soils)	4	0.663	0.661	0.826	0.828	0.437	0.433	22425	22532
33	gam(yield~s(Wavail, k = 3, by = soils) + s(longi, lat, k = 6) + year)	4	0.661	0.659	0.828	0.83	0.434	0.428	22473	22614

k refers to the degree of freedom of the smooth function; Corr: correlation coefficient;

†: A full list of models for yield tested with cross validation in this study is provided as supplementary information (Table S4).

11. Figures

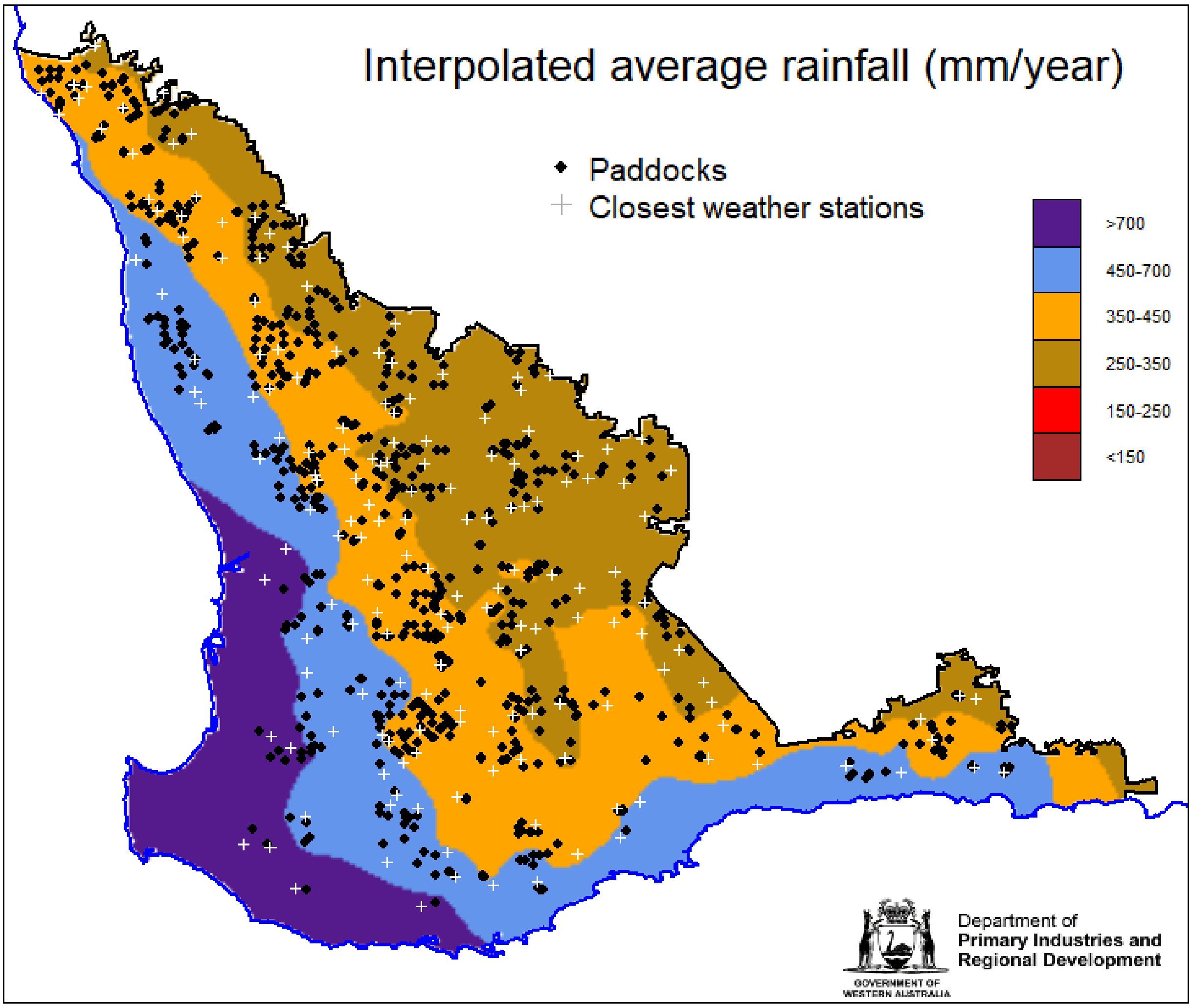


Figure 1: Map of the paddocks and patched point weather stations in variety trials with interpolation of *avgrf30.*

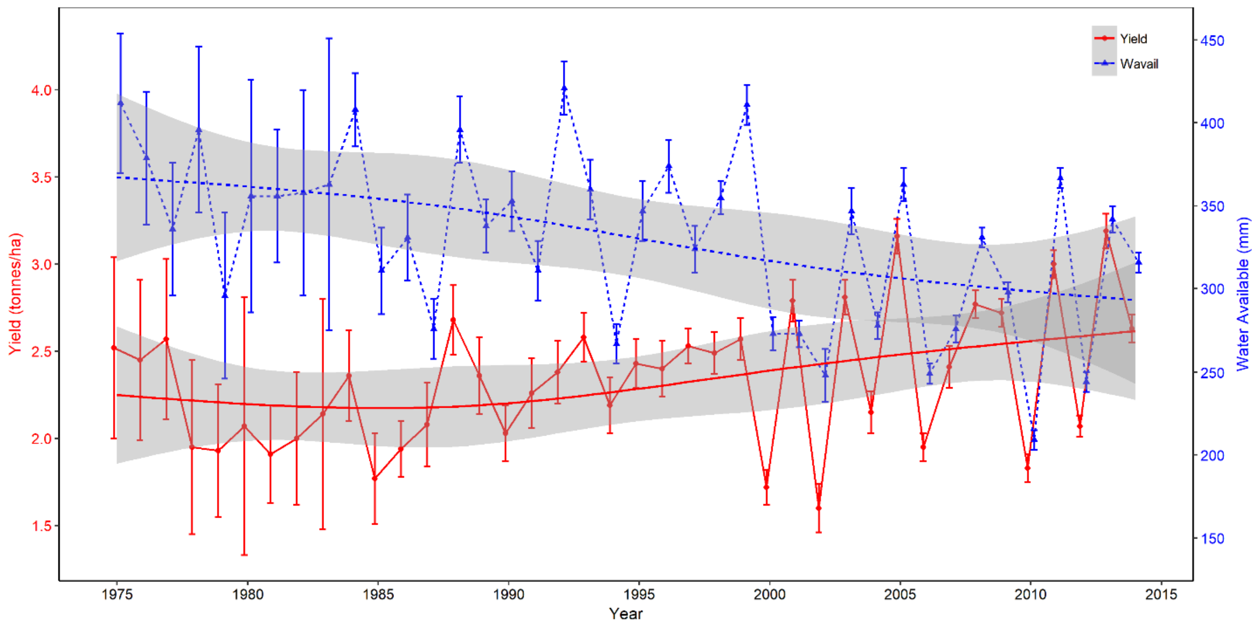


Figure 2: Plots of yield and *Wavail* through time. Note: The error bars represent 2×SEM (standard error of the mean). The non-linear curves were fitted with natural cubic spline function with knot number of 3. The shaded region represents 95% confidence intervals for the smooth curves.

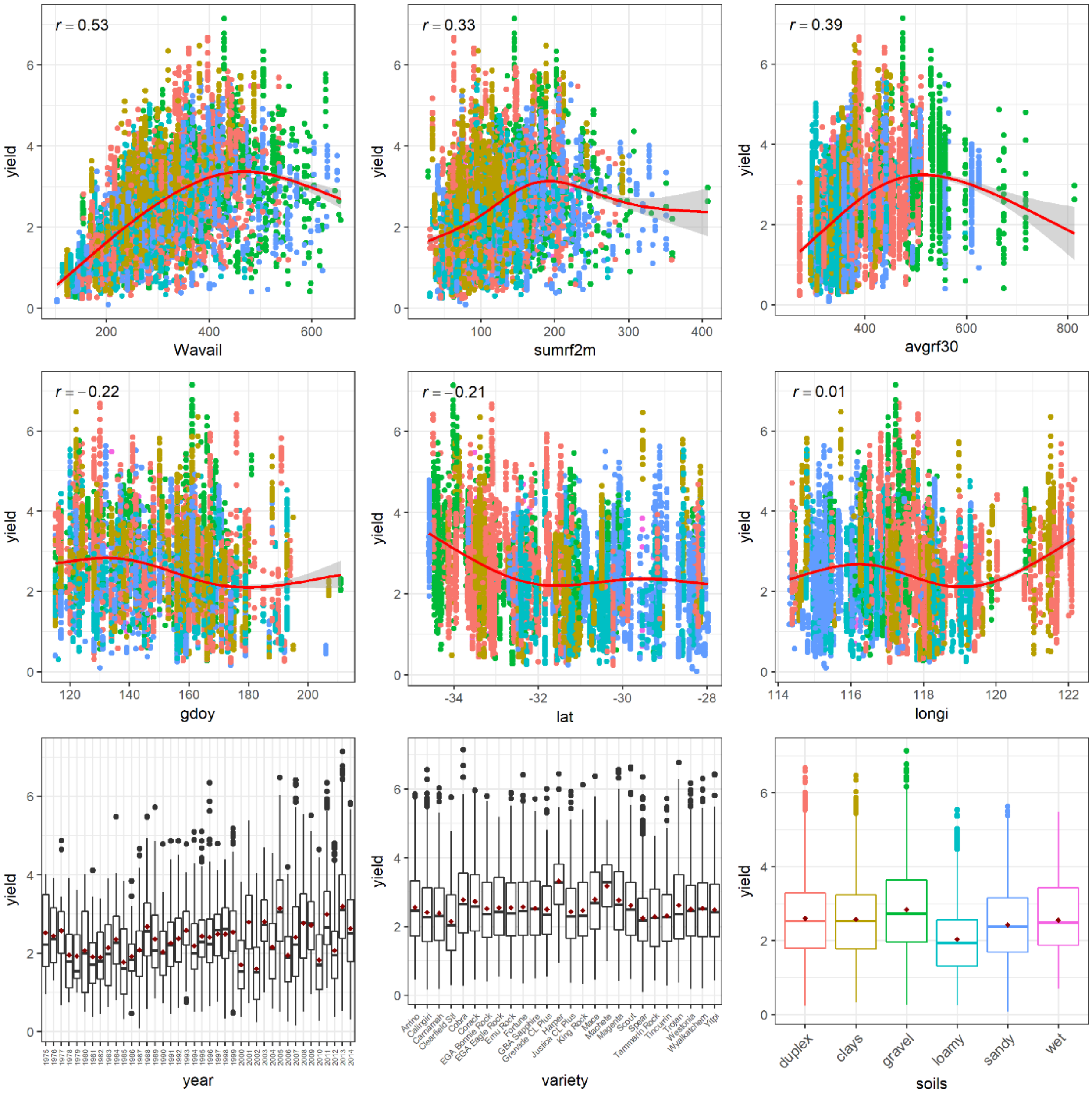


Figure 3: Scatter plot / boxplot matrix for the response variable (yield) and some explanatory variables including *Wavail*, *sumrf2m*, *avgrf30*, day of germination *(gdoy)*, latitude (*lat*), longitude (*longi*), year, variety and soils. Scatter plots are shown for the continuous variables and boxplots for the categorical variables. The points on the scatter plots are observed yield (from the aggregated data based on year, location (latitude/longitude), variety and soil type (n=9068)) and are coloured according to soil types (red=duplex, brown=clays, green=gravel, light blue=loamy, dark blue=sandy, purple=wet). Because the data are from variety trials, a range of different varieties are planted simultaneously at each location causing a wide range in recorded yields shown in the y-axes. The non-linear curves were fitted with natural cubic spline function with knot number of 3. The shaded region on the continuous variable plots represents the 95% confidence intervals for the smooth curves. For the continuous variables, the correlation of yield with each variable is displayed in the top left-hand corner.

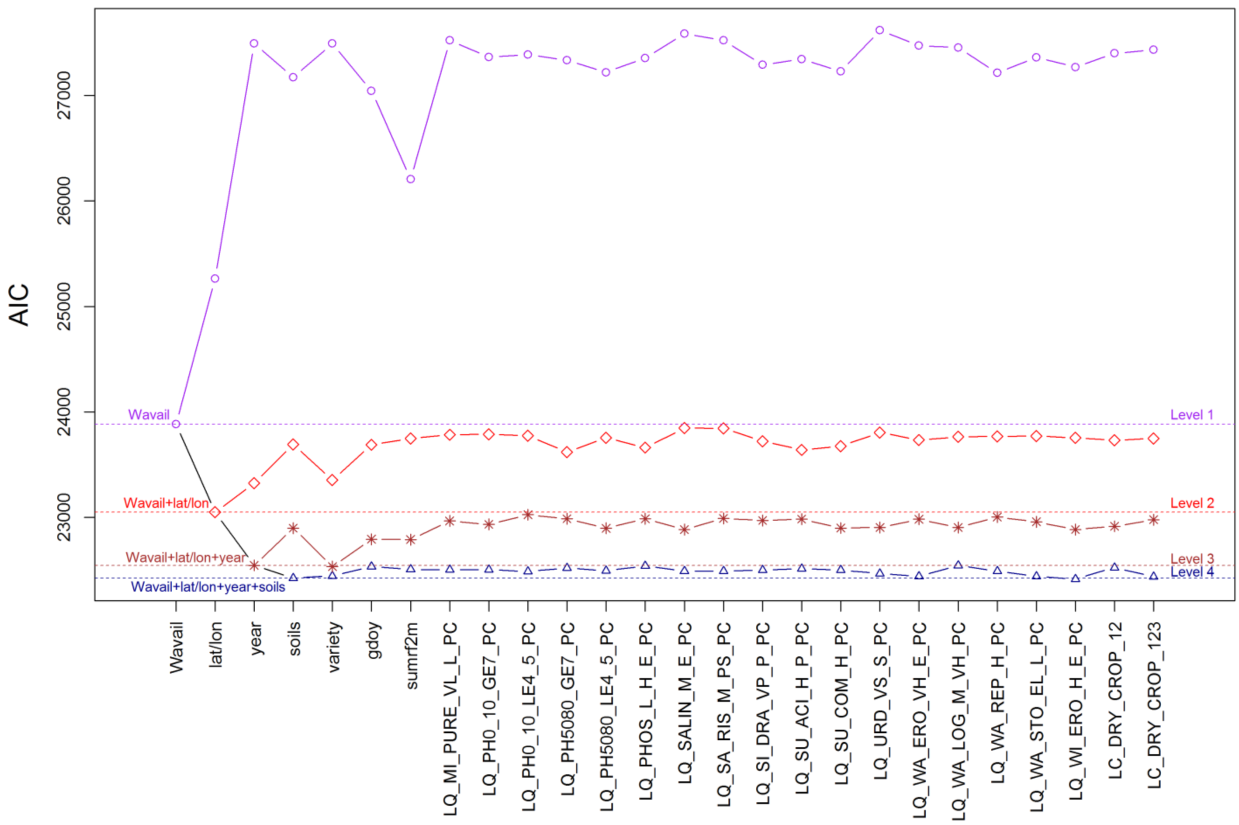


Figure 4: Forward stepwise model selection using AIC. Each variable that was not already in the model was tested for inclusion in the model. The most significant variable of these variables was added to the model at each level. The model started with the null-model, and followed by adding variables to the model one at a time, and continued adding variables until none of remaining variables are significant when added to the model.

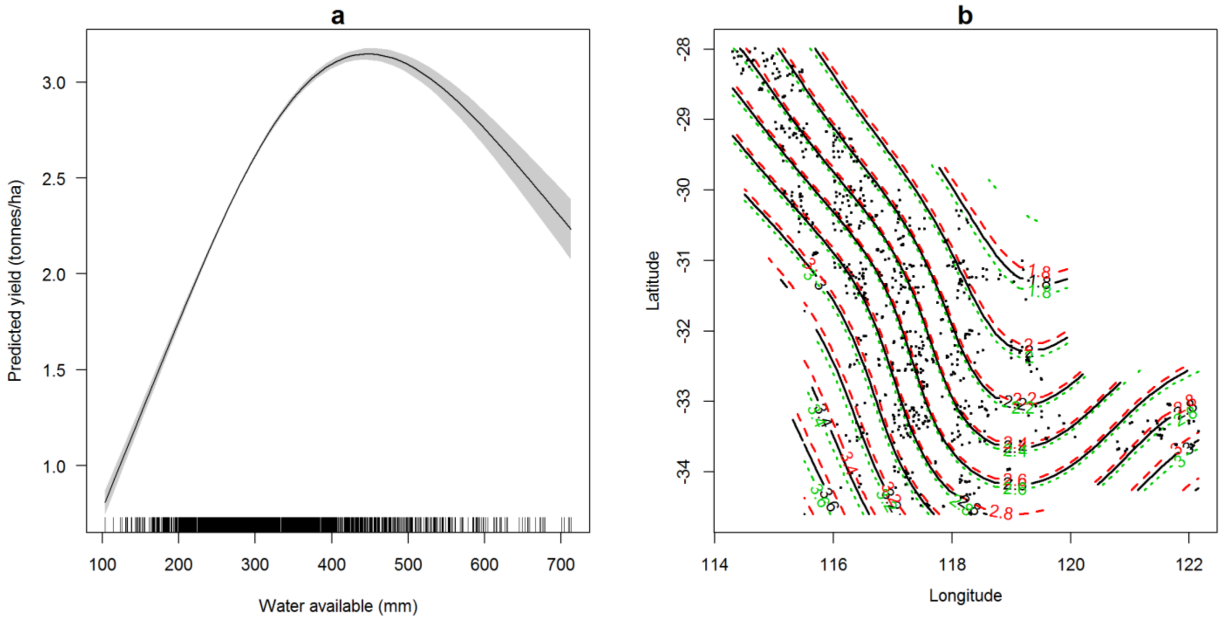
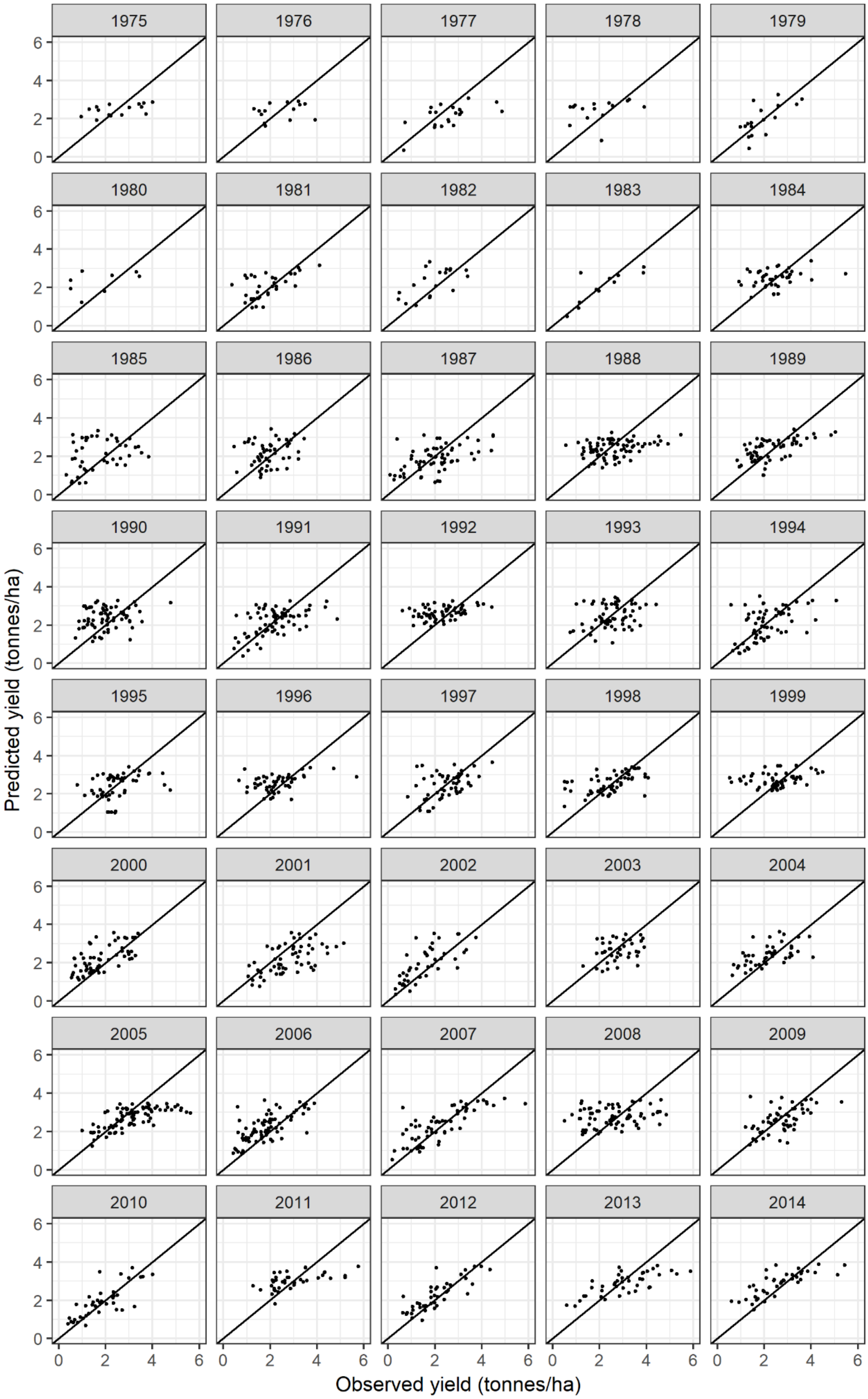


Figure 5: The smooth plots **a)** for *s(Wavail, k=3)* and **b)** for *s(lat, longi, k=6)* terms in the "best" GAM model. Note: The shaded region represents approximate pointwise 95% confidence intervals, the vertical bars at the base of the plots represent a frequency plot of the predictor variable.



678 Figure 6: Plot of the observed versus predicted yields by year (aggregated over variety and soils). The
679 lines show the one-to-one relationship between observed and predicted yields.

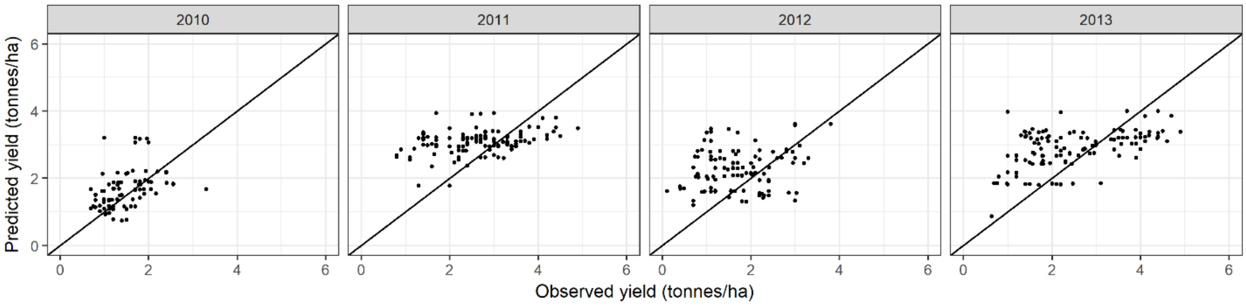


Figure 7: Plot of observed versus predicted yields by year, using the Focus Paddock data. The lines show the one-to-one relationship between observed and predicted yields.

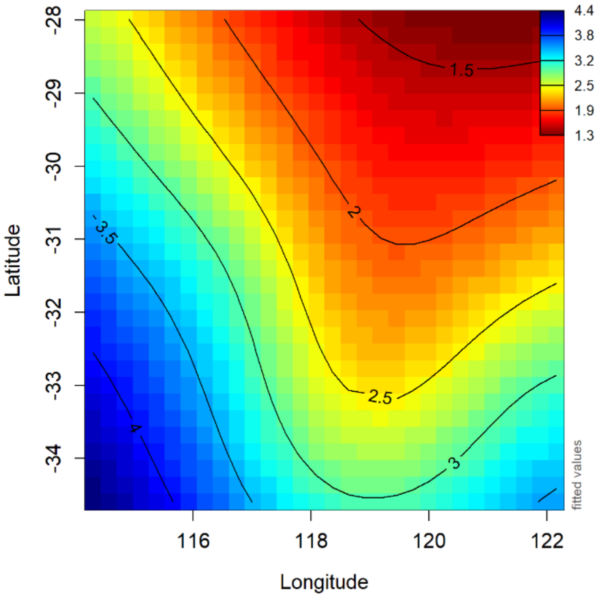


Figure 8: The visualised contour plot of the effect of latitude and longitude on the GAM yield prediction. The contour plot was interpolated from the predicted yield in tonnes/ha, in which year is set at median of the year in the variety trial data and uses the mean of *Wavail*.

12. Online Supporting Information

12.1. Supplementary tables

Table S1: Classification of the agricultural soil groups (Ag_soil_desc) to the functionally simplified soil classes (Soil_6class).

Table S2: Description of the variables for land capability and qualities from the Natural resource information.

Table S3: Mean yield and available water by year.

Table S4: List of models for water-limited yield tested with cross validation in this study.

Table S5: Results of convergence diagnostics for Gibbs simulations.

12.2. Supplementary figures

Figure S1: Map of the 164 location coordinates of Focus Paddocks with interpolation of *avgrf30*.

Figure S2: Plot of observed yield by year at the location of Newdegate Research Station (N=796) in the variety trial data set. Note: The scatter plots were labelled by variety, where red hallow-circle=Calingiri, green triangle=Carnamah, blue plus=Spear, gold cross=Tincurrin, purple diamond=Westonia, pink inverted triangle=Wyalkatchem.

Figure S3: Model checking of the best GAM model using the residual plots including normal Q-Q plot, residuals vs. fitted values, histogram of residuals and response vs. fitted values.

Figure S4: Plots showing how the correlation of annual observed (obs.yield) versus predicted yields (pred.yield) varies with (a) the correlation of observed yield with *Wavail*, and (b) mean annual *Wavail*. Note: the number in the circle represents the last two digits of the year.

Figure S5: The smooth plots for *s(Wavail)* and *s(lat, longi)* terms in the GAM model using Focus Paddocks data. Note: The shaded region represents approximate pointwise 95% confidence intervals, the vertical bars at the base of the plots represent a frequency plot of the predictor variable.

Figure S6: Trace and density plot for all of the parameters for the Gibbs simulations.

Figure S7: Gelman-Rubin-Brooks plot for all of the parameters for the Gibbs simulations.

Figure S8: Geweke-Brooks plot for all of the parameters for the Gibbs simulations.

Figure S9: Using the variety trial data, comparison of the predicted yields from the GAM model and the potential yields from the French & Schultz approach. Where **a)** observed yields versus the French & Schultz's potential yields; **b)** observed yields versus the GAM predicted yields; **c)** GAM predicted yields versus French & Schultz's potential yields; **d)** water available versus French & Schultz's potential yields minus the GAM predicted yields. Note: The scatter plots were labelled by soil types (red hallow-circle=duplex, green triangle=clays, blue plus=gravel, gold cross=loamy, purple diamond=sandy, pink inverted triangle=wet).

Figure S10: Using the Focus Paddocks data, comparison of the predicted yields from the GAM model and the potential yield from the French & Schultz approach. Where **a)** observed yields versus the French & Schultz's potential yields; **b)** observed yields versus the GAM predicted yields; **c)** GAM predicted yields versus French & Schultz's potential yields; **d)** water available versus French & Schultz's potential yields minus the GAM predicted yields. Note: The scatter plots were labelled by soil types (red hallow-circle=duplex, green triangle=clays, blue plus=gravel, gold cross=loamy, purple diamond=sandy, pink inverted triangle=wet).