Forecasting crop yield using multi-layered, whole-farm data sets and machine learning

Filippi P[1], Fajardo M[1], Jones EJ[1], Whelan BM[1], Bishop TFA[1]

[1]The University of Sydney, School of Life and Environmental Sciences, Sydney Institute of Agriculture, Sydney, New South Wales, Australia

A paper from the Proceedings of the

14th International Conference on Precision Agriculture

June 24 – June 27, 2018

Montreal, Quebec, Canada

Abstract. The ultimate goal of Precision Agriculture is to improve decision making in the business of farming. Many broadacre farmers now have a number of years of crop yield data for their fields which are often augmented with additional spatial data, such as apparent soil electrical conductivity (ECa), soil gamma radiometrics, terrain attributes and soil sample information. In addition there are now freely available public datasets, such as rainfall, digital soil maps and archives of satellite remote sensing which can be used to interpret the crop-growing environment. However, rather than analysing one field at a time as is typical in precision agriculture research, there is an opportunity to explore the value of combining all this data for multiple fields/farms and years into one dataset. Using these datasets in conjunction with machine learning approaches offers the possibility of building predictive models of crop yield. In this study, several large farms in Western Australia were used as a case study, and yield monitor data from wheat, barley and canola crops from three sequential that covered approximately 11,000 to 17,000 hectares in each year were used. The yield data was processed to a 10 m grid, and a space-time cube of predictor variables was built at this scalle. This consisted of grower-collected data such as ECa and gamma radiometrics surveys, and the freely-available public data. The data was aggregated to a 100 m spatial resolution for modelling yield. Random Forest models were used to predict crop yield of wheat, barley and canola using this dataset. Three separate models were created based on pre-sowing, mid-season and late-season conditions to explore the changes in the predictive ability of the model as more within-season information became available. These time points also coincide with points in the season when a management decision is made, such as the application of fertiliser. The models were evaluated with cross-validation using both fields and years for data splitting, and this was assessed at the field spatial resolution. Cross-validated results showed the models predicted yield accurately, with a root mean square error (RMSE) of 0.36 to 0.42 t ha-1, and a Lin's concordance correlation coefficient (LCCC) of 0.89 to 0.92 at the field resolution. The models performed better as the season progressed, largely because more information about within-season data became available (e.g. rainfall, remote sensing). The yield forecasts were used to formulate basic nitrogen application scenarios. The more years of yield data that were available for a field, the better the predictions were, and future work should use a longer time-series of yield data. The generic nature of this method makes it possible to apply to other agricultural systems where yield monitor data is available.

Keywords. Yield forecast, empirical yield prediction, remote sensing, big data, machine learning, Random Forest, feature extraction, site-specific crop management, precision agriculture

Introduction

The ability to forecast final crop yields before, and during a growing season is invaluable in guiding management decisions, such as the application of fertiliser or irrigation. The spatial resolution of these yield predictions is a crucial component, as this allows management to be tailored to different fields within a farm, or at the sub-field level. Traditionally, farmers haphazardly estimated their 'yield goals', which was generally based upon previous experience and seasonal conditions, and then used this as a guide to alter management (Dahnke *et al.* 1988; Raun *et al.* 2001). Given that yield is controlled by the interaction between management, pests, soil and weather, this yield goal should vary from season to season, but also vary from location to location.

More systematic, quantitative approaches to predicting crop yield typically consist of using mechanistic/simulation models, such as APSIM (Agricultural Production Systems sIMulator) (Keating *et al.* 2003) or DSSAT (Decision Support System for Agrotechnology Transfer) (Jones *et al.* 2003). These mechanistic models are useful, but there are several limitations. In these approaches, many underlying assumptions are made, a large selection of inputs is required, and the model is often not well-suited to the study area of interest. An alternative approach is to forecast yield using empirical, data-driven approaches. This has typically involved using crop reflectance data from remote or proximal sensing platforms during the growing season to make a prediction of final crop yield (e.g. Raun *et al.* 2001; Boydell and McBratney 2002). Many of these empirical approaches have traditionally ignored other climatic and geo-physical variables, however, more recent research has included a larger suite of data sources with this remotely and proximally sensed data (e.g. Balaghi *et al.* 2008; Walsh *et al.* 2013). Including these additional data sources can be invaluable, and the wealth of agricultural and environmental data available today is an exciting and promising opportunity to further improve yield forecasting through empirical approaches.

In broadacre agricultural systems, such as Australia, Canada and the United States of America, farmers typically have an abundance of spatial agricultural data. This often includes a time-series of crop yield monitor data, which is frequently augmented with auxiliary spatio-temporal data, such as soil test results, and apparent electrical conductivity (EC_a). This data is highly valuable, but is often underutilised due to various limitations, such as being in different formats, located in a variety of repositories, and consisting of different spatial and temporal resolutions. There is often disconnect between these different data sources, and they are rarely combined. Furthermore, publicly-available spatial and temporal environmental data is becoming more available at finer spatial and temporal resolutions, and at declining costs. The nature of these publicly-available datasets is diverse, and includes remotely sensed imagery, geophysical data, and climate data.

In theory, every agricultural crop is an experiment, where the yield is a function of the interaction between a suite of variables that vary in space and time. This abundance of data collected on-farm and publicly-available data describe the conditions under which crops are grown. Machine learning techniques are well-equipped to deal with large datasets with many variables, and they provide the opportunity to create predictive models of crop yield using this mass of data. Traditional agronomy and precision agriculture (PA) was typically concerned with examining single fields in isolation, but there is now the opportunity to explore the value of combining this data over multiple fields and years into one dataset and model. An approach such as this utilises historical information from neighbouring fields to guide yield forecasts, and has the potential to stimulate a paradigm shift in precision agriculture.

Rather than focusing on single fields in isolation, in this study we collate large amounts of spatial and temporal on-farm data and publicly-available data sources in the southern agricultural region of Western Australia (WA). These datasets are then combined with Random Forest models to create predictive yield models of wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.) and canola (*Brassica napus* L.) at three time points in the growing season. This study had a particular focus on the forecasting ability of the models based on pre-, mid-, and late-season information from predictor variables.

2. Methods

2.1. Study area and period

The study was conducted on three large aggregations (A, B and C) of several farms owned by a single corporation (Lawson Grains) that are located in the southern agricultural region of Western Australia (Fig. 1). The soils of the area are typically sandy with notable amounts of gravel. Dryland winter cropping is the sole enterprise in the study regions, with wheat, barley and canola being grown. The study area is characterised by a Mediterranean climate, with cool, wet winters, and hot, dry summers and average annual rainfall of the aggregations ranges from 420 to 533 mm. This study uses data from the 2013, 2014 and 2015 growing seasons. During this period, total annual rainfall values ranged from 389.0 mm to 687.4 mm for the different aggregations (BOM 2017_a; BOM 2017_b).

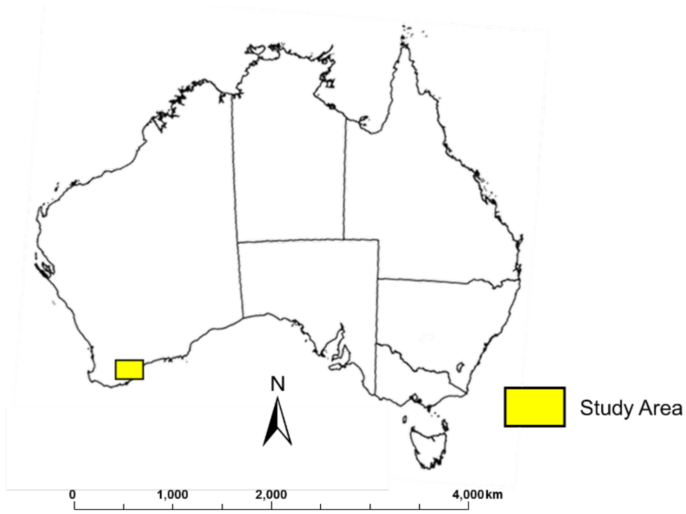

Fig. 1 – Location of the study area within Australia

2.2. Yield forecasting approach

2.2.1. Datasets (space-time cube) and processing

A variety of spatial and temporal data collected on-farm, and publicly-available environmental and agricultural data for the whole study area was collated into a space-time cube (STC). The data was of varying spatial and temporal resolutions, and consisted of yield monitor data, soil information, EC_a and gamma radiometrics surveys (collected on-farm), and remotely sensed information and climate data (publicly-available). Yield from wheat, barley and canola crops from three different seasons that covered 10,587, 16,001 and 16,755 ha in 2013, 2014, and 2015, respectively, were used (Table 1). The amounts of yield monitor data for each crop varied by each season and aggregation, with the most yield data for wheat, followed by barley, and then canola (Table 1). A STC, for the purpose of this study, can essentially be described as a large dataset where each row in the dataset possesses; spatial coordinates (latitude and longitude), year (season), yield, crop type, and a large suite of associated covariates that relate to yield (predictor variables). Each row represents a spatial entity for a particular time point. Some spatial locations in the dataset possessed multiple years of yield data, while others only had one. Despite the varying spatial resolutions of the variables, they were all resampled to a common 10 m grid without changing their native spatial resolution (Table 2).

Table 1 – Number of hectares (ha) of yield monitor data for each crop type and season within each aggregation

Aggregation	Crop	Season 2013	Season 2014	Season 2015	Total (ha)
A	Wheat	1,508	5,296	2,868	9,672
	Barley	1,551	1,398	3,979	6,928
	Canola	0	0	0	0
B	Wheat	1,934	1,723	2,326	5,983
	Barley	1,177	969	986	3,132
	Canola	1,050	2,169	2,219	5,438
C	Wheat	809	1,976	1,886	4,671
	Barley	949	1,570	2,491	5,010
	Canola	1,609	900	0	2,509
Total		10,587	16,001	16,755	43,343

Table 2 – Data sources used in the space-time cube to create the predictive yield models

Data type		Data description	Resolution	Data source
Response	Yield	Yield monitor data	10 m	On-farm
Stationary variables	Soil EC_a (from dual EM)	Shallow (0.5 m)	10 m	On-farm
		Deep (1.5 m)	10 m	On-farm
	Gamma radiometrics	Potassium (K)	10 m	On-farm
		Uranium (U)	10 m	On-farm
		Thorium (Th)	10 m	On-farm
		Total count (K, U, Th)	10 m	On-farm
	Soil maps	Clay content (%)	10 m	On-farm & Project-created
		Sand content (%)	10 m	On-farm & Project-created
Within-season variables	MODIS - EVI	July (middle)	250 m	Global (NASA LPDAAC)
		September (middle)	250 m	Global (NASA LPDAAC)
	Total received rainfall	Jan 1st to Mar 31st	~5,000 m	National (BOM)
		Apr 1st to Jun 30th	~5,000 m	National (BOM)
		Jul 1st to Aug 31st	~5,000 m	National (BOM)
	Forecasted rainfall	Apr 1st to Jun 30th	Regional	National (BOM)
		Jul 1st to Aug 31st	Regional	National (BOM)
		Sep 1st to Dec 31st	Regional	National (BOM)

Before the response and predictor variables were collated into a STC, the varying data sources were processed and feature extraction was performed. Feature extraction and processing of variables is a crucial task, as this transforms the initial set of raw data into useful, usable, and informative data. The yield monitor data at 10 m resolution was corrected and standardised using field-average yields measured after harvest at the silo, as it is known that different yield monitors vary in their measurement accuracy. The whole of the study area had been surveyed with EM and gamma radiometric surveys to 10 m resolution by a single consulting company when the soil profile was dry in summer, ensuring some consistency. An abundance of soil test results were available for the study area, but as this data was in point-form, it was difficult to utilise directly. To transform this soil information into an easily-usable spatial layer, sand and clay content maps were created for the whole study area using Random Forest models and the EC_a and gamma radiometric survey data as covariates. MODIS 16-day Enhanced Vegetation Index (EVI, MOD13Q1) composites at 250 m resolution were acquired from the NASA Land Processes Distributed Active Archive Centre (LPDAAC) portal for the whole study area (https://lpdaac.usgs.gov/, last accessed 21 November 2017). Unlike the Normalised Difference Vegetation Index (NDVI), EVI does not saturate at high canopy density (Huete *et al.* 2002), as such it is better suited as a surrogate of vegetation vigour in high input cropping systems. We selected composites closest to mid-July and mid-September to reflect vegetation conditions at mid- and late growing season. Published research has found that remotely sensed images closest to the middle of September are the most accurate for final yield predictions of winter wheat in the Southern Hemisphere, as this is when flowering and grain-filling occurs (Lyle *et al.* 2013). Total daily rainfall (mm) maps for the study area were obtained from the Bureau of Meteorology (BOM), and this was then aggregated from Jan 1st–Mar 31st, Apr1st–June 30th, and Jul 1st–Aug 31st inclusive, and used as predictor variables in the models (BOM 2017$_{c}$). In addition, the forecasted rainfall from BOM was used as an input, which is the probability of exceeding the median rainfall for the ensuing three months (BOM 2017$_{d}$). The dates for EVI, aggregation of rainfall, and the seasonal forecasts were chosen to coincide with different important points in the winter cropping season, e.g. sowing (April), mid-season N-fertiliser top-dressing allocation (July), and anthesis (September).

2.2.2. Predictive yield modelling

Random Forests (Breiman 2001) were used in conjunction with this STC to create predictive models of crop yield. Rather than creating individual models for wheat, barley and canola, one model was created and crop type was included as a predictor variable. Three models were created based on pre-sowing (April), mid-season (July), and late-season (September) conditions to explore the changes in the predictive ability of the model as more within-season information became available. The models were built at a 100 m resolution, and predicted at the same 100 m resolution. This was then aggregated up to the field resolution, and the prediction quality was then assessed at the field spatial resolution.

The quality of the model predictions were assessed using cross-validation techniques. The first (1) cross-validation method involved creating a model with all seasons of yield data for all fields in the study area, but without all seasons of yield data for a particular field, and then using that model to predict the yield for that field for the missing years. This is identified as leave-one-field-out cross-validation (LOFOCV). The second (2) cross-validation method was similar, and involved creating a model with all seasons of yield data for all fields in the study area, but removing only one year of yield data for a particular field (prior/other yield data retained in the model), and this model was then used to predict the yield for that field in the missing year. This is identified as leave-one-field-year-out cross-validation (LOFYOCV). In both instances, this was repeated for all fields and years, and the average statistics were calculated. It should also be noted that Random Forest models were re-fitted for each data splitting iteration. The aim of performing these different cross-validation techniques was to determine the prediction quality for predicting at a new field with no prior data, as opposed to the predicting at a field with prior yield information included in the model. The root mean square error (RMSE) and Lin's concordance correlation coefficient (LCCC) was used as an assessment of model quality. The LCCC is the fit of the observed and predicted values to the 1:1 line, and is unit-less, making it useful for comparing between models where the magnitude of the predictions may vary (Lin 1989).

3. Results

3.1. Yield modelling predictions

Predictions at the field resolution had a LCCC ranging from 0.19 to 0.27 for the LOFOCV technique, and ranging from 0.89 to 0.92 for the LOFYOCV technique (Table 3). As the season progressed, the models performed slightly better, with the September models possessing the lowest RMSE, and the highest LCCC (Table 3). The significantly improved predictions of the LOFYOCV technique show the important benefit of including prior yield information for a particular field. As an example, cross-validated predictions for the July model improved from an LCCC of 0.20 when no prior yield information was included for a particular field, to 0.91 when prior yield information for the field being predicted was included (Table 3; Fig. 2).

Table 3 – Quality of crop yield predictions at the field resolution with the LOFOCV and LOFYOCV approaches

Cross-validation type	April (pre-season)		July (mid-season)		September (late-season)	
	RMSE (t ha^{-1})	LCCC	RMSE (t ha^{-1})	LCCC	RMSE (t ha^{-1})	LCCC
LOFOCV	0.64	0.19	0.63	0.20	0.62	0.27
LOFYOCV	0.42	0.89	0.39	0.91	0.36	0.92

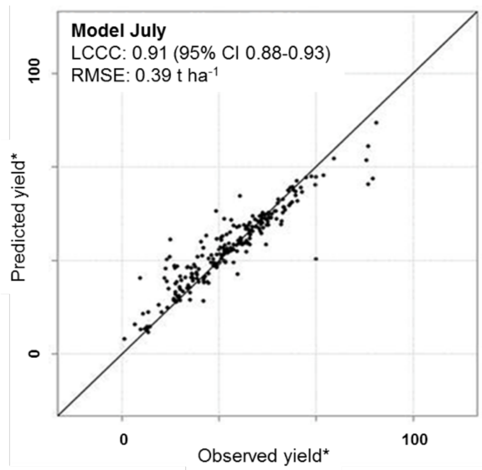

Fig. 2 – Observed and predicted yield for the July (mid-season) model for all fields and years using the LOFYOCV approach at the field resolution (* the scale was altered to range from 0 to 100 for privacy reasons)

The value of including prior yield information for a paddock is also supported by Fig. 3, which shows that as more seasons of prior data were available for an individual field, the predictions substantially improved. This figure was created by using cross-validation, where all data available was used to create a model and this was then used to predict yield for a particular field in a particular year (with that field and year removed from the model). For fields that contained only one year of data (a) (no years of prior yield data in the model), the LCCC was 0.46, and this improved to 0.89 for fields with two years of yield data (b) (one year of prior yield data in the model), and 0.94 for fields with three years of yield data (c) (two years of prior yield data in the model) (Fig. 3).

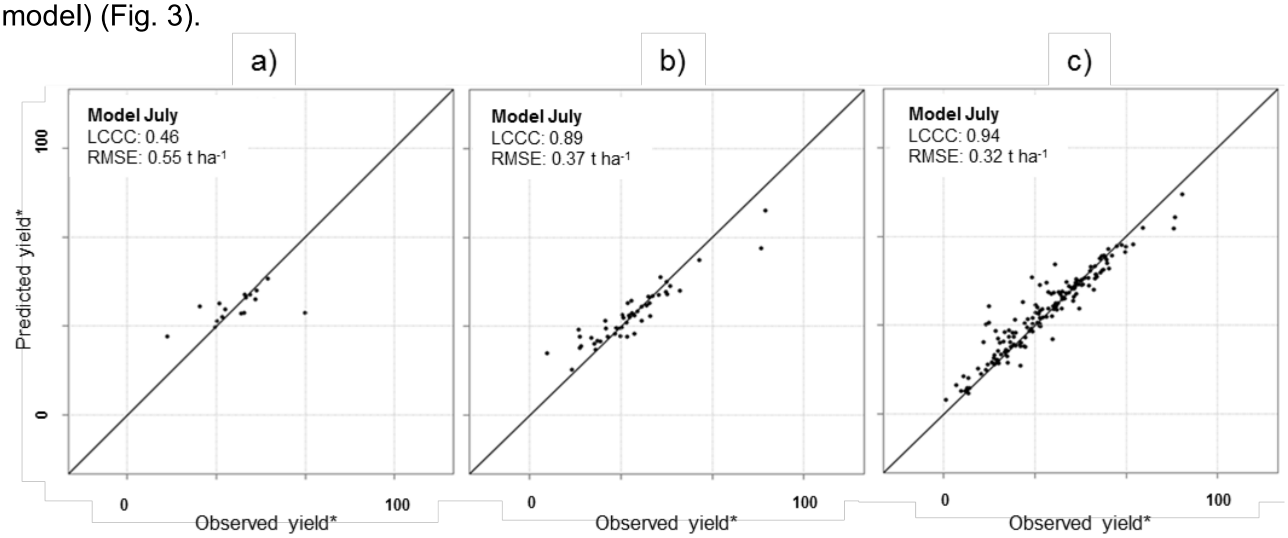

Fig. 3 – Cross validated observed and predicted yield for fields that had a) zero, b) one, and c) two years of prior yield data in the model (* the scale was altered to range from 0 to 100 for privacy reasons)

3.2. Predictor variables and importance

The importance of different predictor variables in the model was tested using the mean decrease in accuracy. The larger the mean decrease in accuracy for a predictor variable, the more important that variable is deemed. Using the July model as an example, crop type was the most important predictor variable (Fig. 4), and this is expected due to the inherent differences in typical yield and yield potential between wheat, barley, and canola. Within-season variables proved to be vital in the models, with received rainfall, forecasted rainfall, and within-season EVI images amongst the most important covariates. The soil maps and geo-physical data (EM and gamma radiometrics) were less important predictors (Fig. 4).

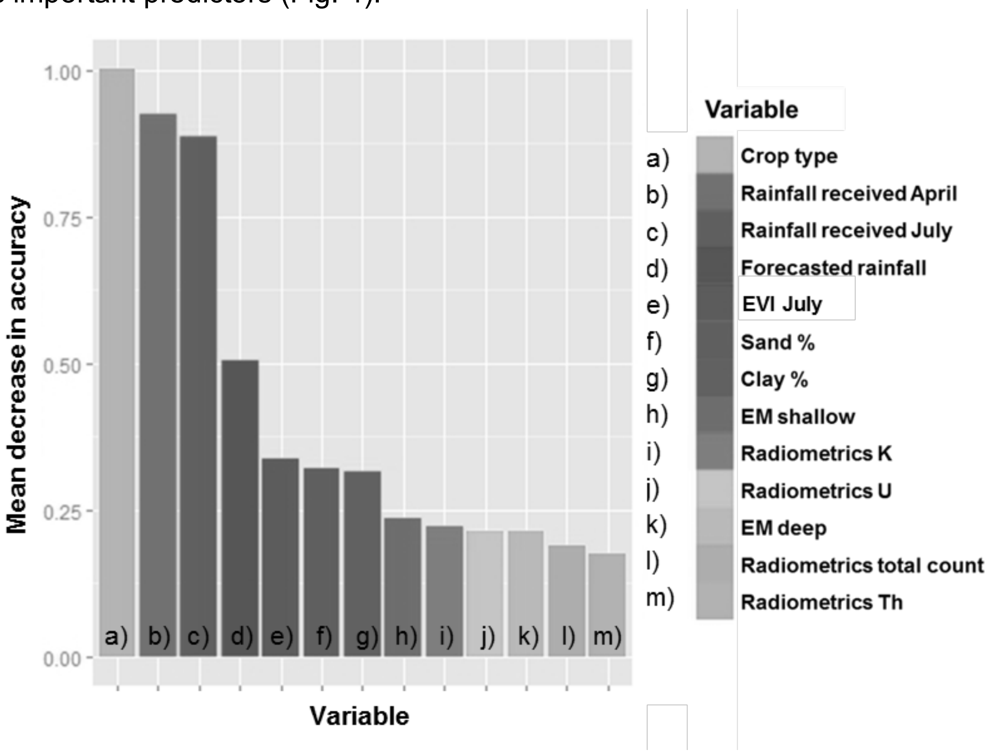

Fig. 4 – Predictor variable importance from the July (mid-season) yield model

4. Discussion

4.1. Model testing and cross-validation approach

Overall, our approach to predict wheat, barley and canola yield showed promising results. An obvious down-fall of both cross-validation approaches used is that information from other fields for the same year is included in the model. Removing all yield data from the same season was restrictive (a leave-one-year-out cross-validation approach), as only three seasons of yield monitor data were available. Furthermore, the expanse of crops grown in each season varied considerably within the different aggregations. For example, we only have canola yield data for the B aggregation in the 2015 season. Despite this limitation, the contrasting results from the different cross-validation techniques suggest that the accurate predictions from our models are not due to the inclusion of yield data from other fields for the prediction year. In the LOFYOCV approach the predictions were very accurate (LCCC of 0.89 to 0.92), however, in the LOFOCV approach the predictions were very poor (LCCC of 0.19 to 0.27) despite the model including data from other fields for the year of prediction (Table 3). While this is not completely robust, it is an indication that prior yield data for the prediction field is the driver for the high quality predictions achieved by the second cross-validation approach, rather than data from other fields from the same year being included in the model.

4.2. Dataset size and resolution (spatial and temporal)

It was clear that including prior yield data for a field resulted in more highly accurate yield predictions, and this is logical, as the model would have a better understanding of expected yield for that field. This is likely due to the consistency of yield patterns between seasons, for example high yielding areas are likely in the same location in each of the three seasons. Consequently, the model then simply needs to scale the yield values from year to year based on seasonal conditions (weather) and observations (EVI). These findings suggest that a larger time-series of yield monitor data, such as 10 or 15 years would greatly improve the prediction accuracy of yield. While there was some climatic variability between the 2013, 2014 and 2015 growing seasons that were used to develop our yield modelling approach, a greater time series would allow the diversity of growing conditions to be better represented in the model, as climatic variables fluctuate significantly from season to season. This would also permit whole-years to be removed from the training model during cross-validation, which would give a more accurate depiction of actual model performance for yield forecasting.

The impact of the spatial extent of the dataset on prediction quality needs to be further explored. In this study we have predicted crop yield for a collection of large farms that covers a large area, but future work should focus on whether this approach can be used on smaller spatial domains – e.g. for a single farm. It may be ideal to have one model for a region, or it may be better for individual farms to have a specific model, and this ideal area that the model covers should be evaluated. For example, it would be interesting to explore whether 10 years of yield data for a 2,000 ha farm results in better predictions than three years of yield data for a 15,000 ha area.

The models in this study were built at 100 m resolution, predicted at 100 m, and then aggregated to the field resolution, but there are opportunities to refine this. The finest spatial resolution of the variables used were 10 m (yield, EM and radiometrics), and it would be possible to build the STC at this resolution, however, 100 m was chosen to make the cross-validation manageable on a desktop computer. While yield predictions at the field resolution are valuable, predictions at finer resolutions within-fields, such as 10 m or management units/zones would be much more useful to guide management decisions and implement spatial precision agriculture (Bishop and Lark 2007; Taylor et al. 2007; Bishop *et al.* 2015), such as the variable rate application of fertiliser. This will be the subject of future work.

4.3. Predictor variables and feature extraction

As the season progressed the model predictions slightly improved and this is likely due to an increased amount of within-season predictor variables being used in these models. The variable importance plot showed that within-season variables were very important predictors in the models, and integrating more of these within-season data sources should be considered in further research. The remotely-sensed EVI images used in this study were sourced from MODIS and are at a 250 m spatial resolution, however, there are opportunities to include finer spatial data, such as Landsat at 30 m resolution as this would give more detailed information (Lewis *et al.* 2017). Both received rainfall and forecasted rainfall were highly important variables in the models, and the inclusion of other climatic data variables, such as temperature could also improve the model predictions. The degree-days (cumulative of the average temperature in a day) are a useful way of measuring the physiological development of a crop, and this can also provide useful insight into the expected final crop yield (McMaster and Wilhelm 1997). The management choices and practices growers implement have a strong impact on final crop yields, and this type of information should be included in the modelling approach in the future. This could include variables such as the crop variety, sowing seeding rate, or amount of applied fertiliser.

Furthermore, additional research should consider the quality of the models under data-poor scenarios, such as when only freely-available datasets are available, and data-rich scenarios, such as when there is an abundance of data collected on-farm available, as was the case here. This could identify the value proposition for growers when deciding on the type of data to collect, as well as the optimal spatial and temporal resolution. For example, this could identify whether growers should invest in surveying their property with EM and gamma radiometrics.

4.4. Potential for using yield forecasts for management interventions

Predictive models of the upcoming season's crop yield are extremely useful, particularly when the predictions are at fine spatial resolutions and of high accuracy. There is a wealth of potential uses of the models presented in this study, and this information could be used to identify yield gaps, decide on futures contracts and market speculation, and to inform decisions on precision agricultural management practices. In particular, the incorporation of management inputs with these models is a promising avenue for future research, for example variable rate application of fertiliser, gypsum, lime, seeding rates, or variety selection.

5. Conclusion and future directions

In this study, a data-driven approach to predicting wheat, barley, and canola crop yield as an alternative to approaches that use mechanistic models was presented. The results from this approach are promising, and the generic nature makes it possible to apply it to many other agricultural systems where yield data is available. Particular benefit was found from including prior yield information for fields, and future work should explore the change in model quality predictions with a larger time-series of yield data (e.g. 10 years). Furthermore, the optimal spatial domain covered for this modelling approach should be investigated, for example; are better predictions obtained if data is pooled among farmers within a similar region (as we have done here), or is it better to have a single model for each individual farm? Future work should also explore integration of more data sources and improved feature extraction, particularly within-season measurements into the model, such as temperature and remotely-sensed information at a finer spatial resolution. Future research should consider forecasting yield at finer spatial resolutions within-fields, such as a 30 m grid or management zones, as this would be more valuable in informing management decisions, such as the application of variable-rate fertiliser.

Acknowledgments

The authors would like to thank Lawson Grains, Precision Agronomics, and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) for providing access to the data.

References

Balaghi, R., Tychon, B., Eerens, H., Jlibene, M. (2008) Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. *International Journal of Applied Earth Observation and Geoinformation* **10**, 438-452.

Bishop, T., Lark, R. (2007). A landscape-scale experiment on the changes in available potassium over a winter wheat cropping season. *Geoderma*, **141**, 384-396

Bishop, T., Horta, A., Karunaratne, S. (2015). Validation of digital soil maps at different spatial supports. *Geoderma*, **241-242**, 238-249.

Boydell, B., McBratney, A.B. (2002) Identifying potential management zones from cotton yield estimates. *Precision Agriculture* **3,** 9-23.

Breiman, L. (2001) Random Forests*. Machine Learning* **45**, 5-32.

Bureau of Meteorology - BOM (2017$_{a}$) Monthly rainfall – Jacup. Available from: <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=010905> [21 November 2017]

Bureau of Meteorology - BOM (2017$_{b}$) Monthly rainfall – Munglinup. Available from: <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=139&p_display_type=dataFile&p_startYear=&p_c=&p_stn_num=009868> [21 November 2017]

Bureau of Meteorology - BOM (2017$_{c}$) Monthly rainfall totals for Western Australia. Available from: <http://www.bom.gov.au/jsp/awap/rain/index.jsp?colour=colour&time=latest&step=0&map=totals&period=month&area=wa> [21 November 2017]

Bureau of Meteorology - BOM (2017$_{d}$) Climate outlooks – monthly and seasonal. Available from: <http://www.bom.gov.au/climate/outlooks/#/rainfall/median/seasonal/0> [21 November 2017]

Dahnke, W.C., Swenson, L.J., Goos, R.J., Leholm, A.G. (1988) Choosing a crop yield goal. SF-822. North Dakota State

Ext. Serv., Fargo.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., (2002) Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices. Remote Sensing of *Environment* **83**, 195–213. doi:10.1016/S0034-4257(02)00096-2.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T. (2003) DSSAT Cropping System Model. *European Journal of Agronomy* **18**, 235-265.

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.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., Wu, W., Tan, P., Li, F., Killough, B., Minchin, S., Roberts, D., Ayers, D., Bala, B., Dwyer, J., Dekker, A., Dhu, T., Hicks, A., Ip, A., Purss, M., Richards, C., Sagar, S., Trenham, C., Wang, P., Wang, L.W. (2017) The Australian Geoscience Data Cube — Foundations and lessons learned. *Remote Sensing of Environment.* Manuscript in press.

Lin, L.I.K. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 255-268.

Lyle, G., Lewis, M., Ostendorf, B. (2013) Testing the Temporal Ability of Landsat Imagery and Precision Agriculture Technology to Provide High Resolution Historical Estimates of Wheat Yield at the Farm Scale. *Remote Sensing* **5**, 1549.

McMaster, G.S., Wilhelm, W.W., (1997) Growing degree-days: one equation, two interpretations. *Agricultural and forest meteorology* **87**, 291-300.

NASA Land Processes Distributed Active Archive Centre (LPDAAC) (2017) MOD13Q1: MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V006. NASA EOSDIS Land Processes DAAC. (https://lpdaac.usgs.gov), accessed 21 November 2017 at https://doi.org/10.5067/modis/mod13q1.006

Raun, W.R., Solie, J.B., Johnson, G.V., Stone, M.L., Lukina, E.V., Thomason W.E., Schepers, J.S. (2001) In-season prediction of potential grain yield in winter wheat using canopy reflectance. *Agronomy Journal* **93**, 583-589.

Taylor, J.A., McBratney, A.B., Whelan, B.M. (2007) Establishing Management Classes for Broadacre Agricultural Production. *Agronomy Journal* **99**, 1366-1376.

Walsh, O.S., Klatt, A.R., Solie, J.B., Godsey, C.B., Raun, W.R. (2013) Use of soil moisture data for refined GreenSeeker sensor based nitrogen recommendations in winter wheat *(Triticum aestivum L.). Precision Agriculture* **14**, 343-356.